# SRAM Read-Assist Scheme for Low Power High Performance Applications

Ali Valaee

A Thesis

In the Department of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements for the Degree

of Master of Applied Science in Electrical and Computer Engineering

Concordia University

Montreal, Quebec, Canada

December 2011

**CONCORDIA UNIVERSITY**
**SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Ali Valaee

Entitled: "SRAM Read-Assist Scheme for Low Power High Performance Applications"

and submitted in partial fulfillment of the requirements for the degree of

**Master of Applied Science**

Complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

________________________________________ Chair
Dr. R. Raut

________________________________________ Examiner, External
Dr. C. Assi (CIISE) To the Program

________________________________________ Examiner
Dr. Y. R. Shayan

________________________________________ Supervisor
Dr. A. J. Al-Khalili

________________________________________ Supervisor
Dr. S. Jahinuzzaman

Approved by: ________________________________________
Dr. W. E. Lynch, Chair
Department of Electrical and Computer Engineering

____________20_____ ________________________________

Dr. Robin A. L. Drew
Dean, Faculty of Engineering and
Computer Science

# ABSTRACT

SRAM Read-Assist Scheme for Low Power

High Performance Applications

Ali Valaee

Semiconductor technology scaling resulted in a considerable reduction in the transistor cost and an astonishing enhancement in the performance of VLSI (very large scale integration) systems. These nanoscale technologies have facilitated integration of large SRAMs which are now very popular for both processors and system-on-chip (SOC) designs. The density of SRAM array had a quadratic increase with each generation of CMOS technology. However, these nanoscale technologies unveiled few significant challenges to the design of high performance and low power embedded memories. First, process variation has become more significant in these technologies which threaten reliability of sensing circuitry. In order to alleviate this problem, we need to have larger signal swings on the bitlines (BLs) which degrade speed as well as power dissipation. The second challenge is due to the variation in the cell current which will reduce the worst case cell current. Since this cell current is responsible for discharging BLs, this problem will translate to longer activation time for the wordlines (WLs). The longer the WL pulse width is, the more likely is the cell to be unstable. A long WL pulse width can also degrade noise margin. Furthermore, as a result of continuous increase in the size of

SRAMs, the BL capacitance has increased significantly which will deteriorate speed as well as power dissipation. The aforementioned problems require additional techniques and treatment such as read-assist techniques to insure fast, low power and reliable read operation in nanoscaled SRAMs. In this research we address these concerns and propose a read-assist sense amplifier (SA) in 65nm CMOS technology that expedites the process of developing differential voltage to be sensed by sense amplifier while reducing voltage swing on the BLs which will result in increased sensing speed, lower power and shorter WL activation time. A complete comparison is made between the proposed scheme, conventional SA and a state of the art design which shows speed improvement and power reduction of 56.1% and 25.9%, respectively over the conventional scheme at the expense of negligible area overhead. Also, the proposed scheme enables us to reduce cell $V_{DD}$ for having the same sensing speed which results in considerable reduction in leakage power dissipation.

# Acknowledgements

I would like to express my gratitude to my advisor Prof. Al-Khalili for his insightful guidance and support throughout the course of this research. His wisdom was not only a source of inspiration for this research, but also for my future life. At the same time, he provided me with encouragement and motivation which helped me to proceed toward my goals. I am grateful to Prof. Jahinuzzaman for his helpful suggestions and financial support during this research.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

VLSI: Very Large Scale Integration

SOC: System-on-Chip

BL: Bitline

WL: Wordline

SA: Sense Amplifier

$V_{TH}$: Threshold Voltage

$T_{OX}$: Oxide Thickness

RBB: Reverse Body Bias

SNM: Static Noise Margin

PVT: Process, Voltage and Temperature

SAE: Sense Amplifier Enable

FSM: Finite State Machine

DNM: Dynamic Noise Margin

$C_{BL:}$ BL Capacitance

VSA: Latch Type Voltage Mode Sense Amplifier

SFT: Simple Four Transistor Sense Amplifier

CBL: Clamped Bit-Line Sense Amplifier

PBT: PMOS Bias Type Sense Amplifier

PDP: Power Delay Product

$T_{ev}$: Evaluation Time

LVS: Layout versus Schematic

# Chapter 1

# Introduction

## 1.1 Benefits of CMOS Scaling

The conclusion of Moore's law has been the main reason for the achievements and evolution in semiconductor industry for more than 45 years. With each technology generation, which spanned from two to three years, the number of transistor per chip has been doubled. Figure 1.1 shows the increase in number of transistors employed in each generation of microprocessors as predicted by Moore's law [1]. Also in each technology node, the operating frequency has been increased around 43% and the switching energy consumption has been reduced around 65% [2]. This is due to the fact that MOSFET is a majority carrier device and its current is almost pure drift which causes the current to be proportional to electric field. Therefore, if the transistor dimensions and voltage are scaled by the same factor ($\beta$), the current density and electric field remain constant and the DC characteristics are unaffected. As a result of this phenomenon, chip area decreases by $\beta^2$, speed increase by $\beta$ and power density remains constant allowing lower power per function or more circuits with the same power dissipation. These scaling benefits and other improvements in photolithography caused CMOS to become the pervasive choice for implementation of digital circuits. This has enabled an astonishing improvement in the performance of very large scale integration (VLSI) systems. However, nanoscale

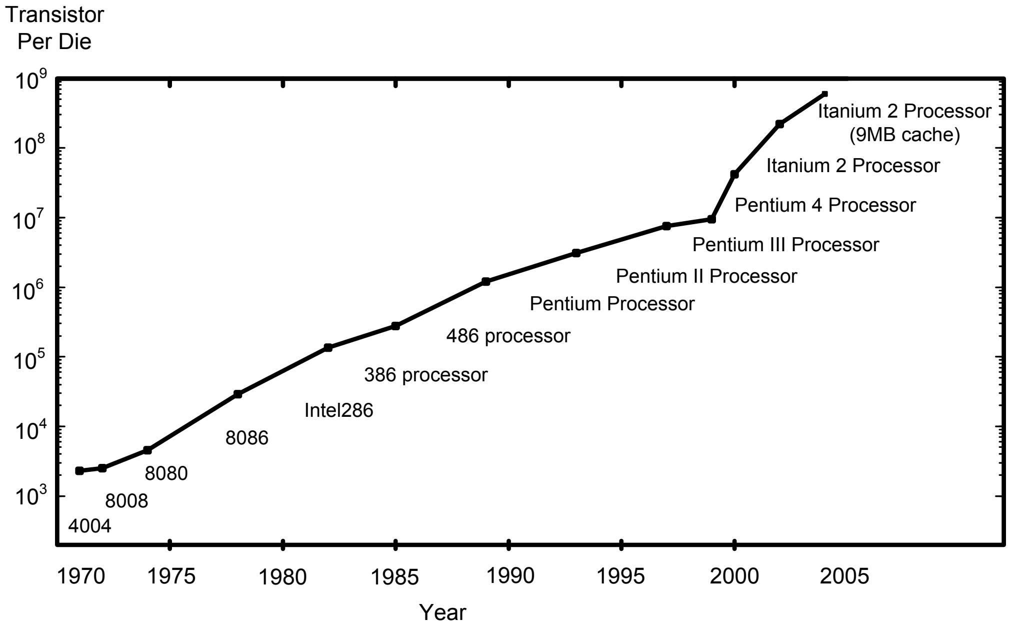


Figure 1.1 Moore's Law of exponential increase in the number of fabricated transistors

CMOS technologies unveiled few significant challenges to the design of high performance integrated circuits and as we continue to move toward exploiting nanometer scale transistors, we have to deal with negative aspects of this reduction in feature size.

## 1.2 Issues in Nanoscale CMOS Technologies

As a result of exponential increase in the device integration density and performance, we are facing some major design challenges due to intrinsic physical limitation of devices. Constant scaling of CMOS technology relied on overlooking several characteristic of semiconductors such as mobility degradation and velocity saturation, thermal energy of carriers, fringing capacitances, random dopant and trap distribution, step-height oxide thickness variation, line edge roughness and tunneling through the gate oxide. After 40 years of continuous scaling of CMOS technology, these simplifications are no longer applicable.

Today's integrated circuits suffer from experiencing two different type of variation [3]:

- Environmental variations such as supply voltage variation, noise coupling between interconnects and temperature variation.
- Physical variation due to imperfection in fabrication process and reliability-related degradation. This source of variation is composed of inter-die and intra-die variation.

One of the major challenges in nanoscale CMOS technology is the increased effect of physical variations i.e. the increased process parameter variation which comes as a result of limitation in the fabrication process such as mask imperfection, sub-wavelength lithography and etching and variation of the number of dopants in the channel of short channel transistors. This will lead to variation in device parameters such as width (W), length (L), threshold voltage ($V_{TH}$) and oxide thickness ($T_{ox}$). Figure 1.2 shows the trend in technology parameter variation [3]. The variation in process parameters can result in degradation of speed or leakage power consumption of different transistors in the same circuit. For example, as explained in [4], process parameter variation in 0.13µm can result in 30% fluctuation in the maximum operating frequency and 5X increase in leakage power. The process parameter variation can be divided into two major categories:

- Inter-die variations
- Intra-die variations

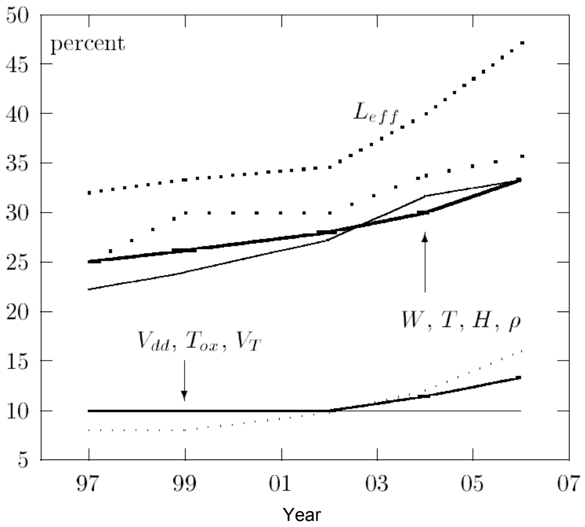


Figure 1.2 Technology parameter variation [3]

Inter-die variation is due to fluctuation in die-to-die, wafer-to-wafer and lot-to-lot process parameters. Inter-die variations affect all the transistors in the chip in the same direction. Hence, it can be interpreted as the deviation of the parameter mean value in the circuit. Inter-die variation can result in performance degradation or even failing hardware when not properly treated. Intra-die variation is characterized as the variations that happen within die. In the applications where the functionality of circuit depends on the matching of the devices, intra-die variation can result in serious deviation from the expected results. For example, in a current mirror configuration, inter-die variation has negligible effect on the performance of the circuit while intra-die variation can cause serious deviation in the value of mirrored current. It has been observed that intra-die variation strongly depends on circuit layout and transistor dimensions. This variability can be divided to random or uncorrelated variation and a systematic or correlated variation. An example of inter-die systematic variation is the variation between different metal layers used for routing. Example of intra-die systematic variation is the variation of

$L_{eff}$ as a function of local layout which is due to optical proximity effects. Threshold voltage mismatch due to quantization effect of doping atoms in increasingly smaller silicon structures is an example of random intra-die variation.

In above nanometer CMOS technology, the inter-die process variation used to be the dominant source of process variation. However, in nanoscale CMOS technology the intra-die process variation has become as important as inter-die process variation [3]. This is due to the large size of today's integrated circuit which makes the intra-die variations comparable to inter-die variation. While inter-die variation is only dependent on fabrication process, intra-die variation is affected by both process and design.

Furthermore, transistors in sub-100nm technologies exhibit higher sub threshold and gate leakage current due to reduction in channel length and gate dielectric thickness, respectively. Hence, there is a continuous increase of 3-5 times in the leakage power consumption with technology scaling as shown in Figure 1.3 [5]. Also, the continuous increase in the subthreshold and gate leakage threatens the noise tolerance of dynamic circuits.

## 1.3 SRAM Design Issues in Nanoscale CMOS Technologies

The use of Moore's law has directed us to integrate large embedded memories. The density of SRAM had a quadratic increase with each generation. More than half of the total transistor count in today's high speed microprocessors is occupied by cache memories and this ratio is expected to increase more [2]. In system level, we are observing a continuous increase in the size of exploited semiconductor memories in

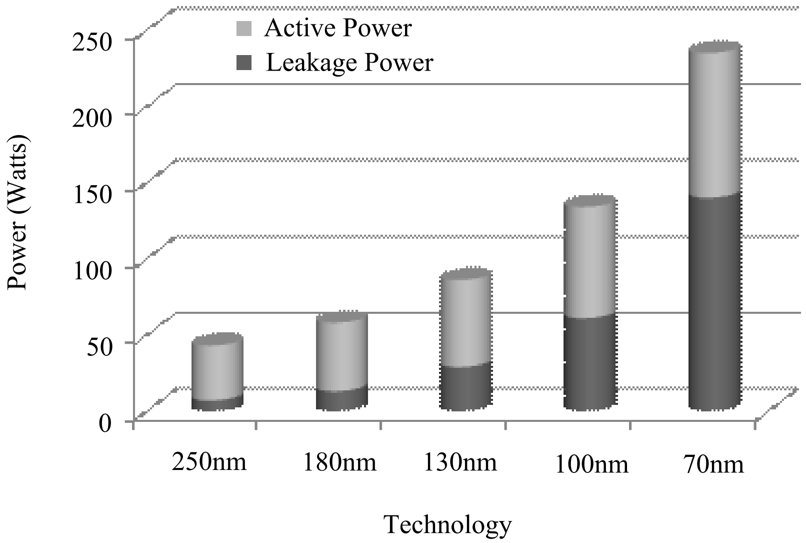


Figure 1.3 Increase of leakage power contribution to the total power consumption of microprocessors [5]

workstations and computers. The size of memories has reached to several Giga bytes in today's systems. As predicted by [6], over 90% of total SOCs area is expected to be occupied by memories in the following decades. Therefore, the speed and power consumption of memories will greatly affect the overall performance and power consumption of digital system. On the other hand, mobile applications constitute huge portion of today's digital systems, where low-power consumption is the main concern in order to prolong battery life time and reduce heat dissipation. The simplest way to reach this goal is to reduce $V_{DD}$ which will result in quadratic reduction in dynamic power consumption and linear reduction in leakage power.

However, as a result of increase in the size of SRAMs, BL capacitance has increased significantly which will result in a higher power consumption and lower speed. Furthermore, the problem is exacerbated by the presence of process variation in nanoscale CMOS technology. This is a major issue in the design of sensing circuitry and can lead to false decision if not properly treated. In order to alleviate this concern, larger signal swings must be developed on the BLs before enabling the sensing circuitry to overcome the effect of process variation and insure correct data at the outputs. But this in turn increases the power dissipation and requires more activation time for WL signals which will decrease the speed. Another challenge is due to variation of cell current which will reduce the worst case cell current. Since this current is responsible for discharging BLs, this problem will translate into reduced speed and longer activation time for WLs in order to develop certain amount of differential voltage on the BLs. The longer the WL pulse width is, the more likely is that the cell to be unstable. A longer WL pulse width can also degrade noise margin [7], [8].

In nanometer CMOS technology, leakage power dissipation constitute considerable portion of overall SRAM power consumption. Since SRAM is the largest block and employs the most number of transistors in microprocessors and SOCs, its leakage power should be kept low in order to reduce overall power consumption and expand battery life time in mobile application. That is why recently several research works address this issue in order to reduce SRAM leakage power [9]-[12]. The leakage power consists of two components i.e. subthreshold leakage and gate-to-channel leakage. Figure 1.4 shows the various leakage power components in an SRAM cell. During the standby mode (WL = ‘0’), there are four OFF transistors and two ON transistors. It is worth mentioning that the gate leakage current of ON transistors (M1 and M4) is higher compared to gate leakage of OFF transistors. To reduce leakage power, several techniques have been proposed. For example, cell $V_{DD}$ can be reduced in order to suppress both subthreshold and gate leakage. Although this method reduces cell leakage, but the BL leakage current (the

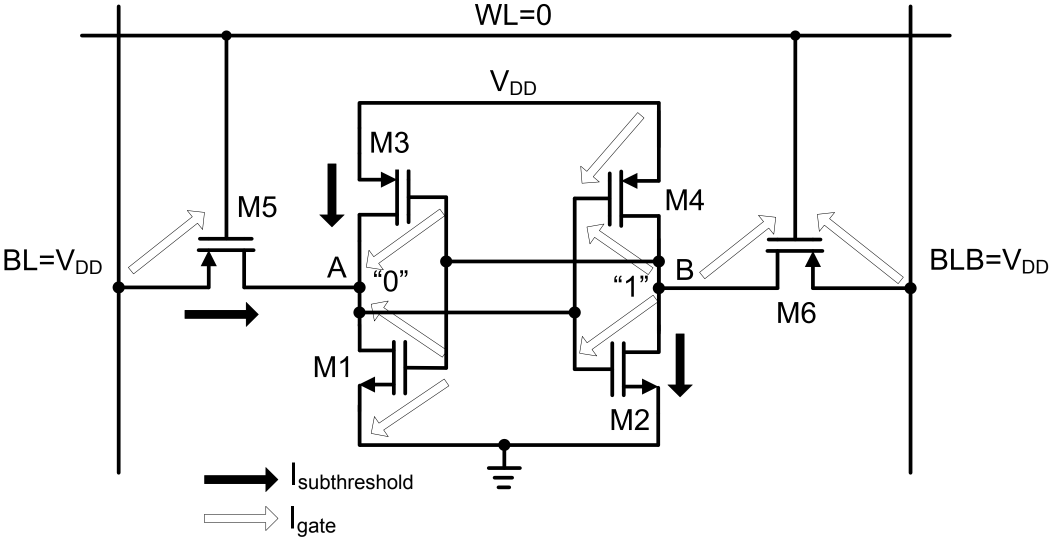


Figure 1.4 Leakage power components

current that flow through access transistors form BL to the internal nodes of cell) will not be suppressed. Hence, another technique to suppress cell and BL leakage at the same time, is to increase the virtual ground voltage in order to apply reverse body bias (RBB) to the access and driver transistors. In either of these two methods, the rail-to-rail supply voltage is reduced. However, in order to ensure adequate read/write margins, SRAM cell is required to operate above a certain supply voltage called $V_{min}$. Due to the increased process variation in nanosclae CMOS technologies and the continuous increase in the size of SRAMs, $V_{min}$ scaling has not kept pace with technology scaling [13]. Also, the low-power operation can be achieved by adoption of high $V_{TH}$ transistors which compromises SRAM performance.

During fabrication process, all the circuits on the same wafer are tested after the processing sequence and before breaking up the wafer into individual dice. The percentage of the circuits that satisfy the expected specifications at this point is called the wafer-sort yield and is usually in the range of 10% to 90% [14]. The malfunction can occur due to number of reasons particularly the point defects of various types that occur during photoresist and diffusion operation. This can be due to mask defects, pinholes in the photoresist, crystalline defects in the epitaxial layer, and so on. If these type of defects happen in the active area of transistor, usually a nonfunctional circuits results. Also, the increased effect of inter-die and intra-die process variations in nanoscale CMOS technologies may lead to several functional failures in SRAM, leading to deterioration of yield i.e. the number of functional chips. This is mainly due to the threshold voltage mismatch between neighboring transistors caused by random dopant fluctuation [15], [16]. The functional failure can occur during read (flipping the data while reading the cell), write (unsuccessful write), access (reduced BL differential voltage while accessing the cell) and hold (data flipping as a result of reduced rail-to-rail supply voltage) [16]. The situation is even exacerbated by reduction in $V_{DD}$ which is used in order to reduce energy consumption and realize low-power SRAMs. This is due to the fact that at lower $V_{DD}$, the static noise margin (SNM) of cell will decrease [17] (Figure 1.5(a)) which makes the SRAM more vulnerable to functional failure. Furthermore, at reduced $V_{DD}$ cells are more prone to the $V_{TH}$ mismatch caused by random dopant fluctuation [18]. In fact, random dopant fluctuation makes independent shift in $V_{TH}$ of transistors within cell, leading to deterioration of cell SNM (Figure 1.5(b)). A column of SRAM is considered faulty if any cell within that column is not functional.

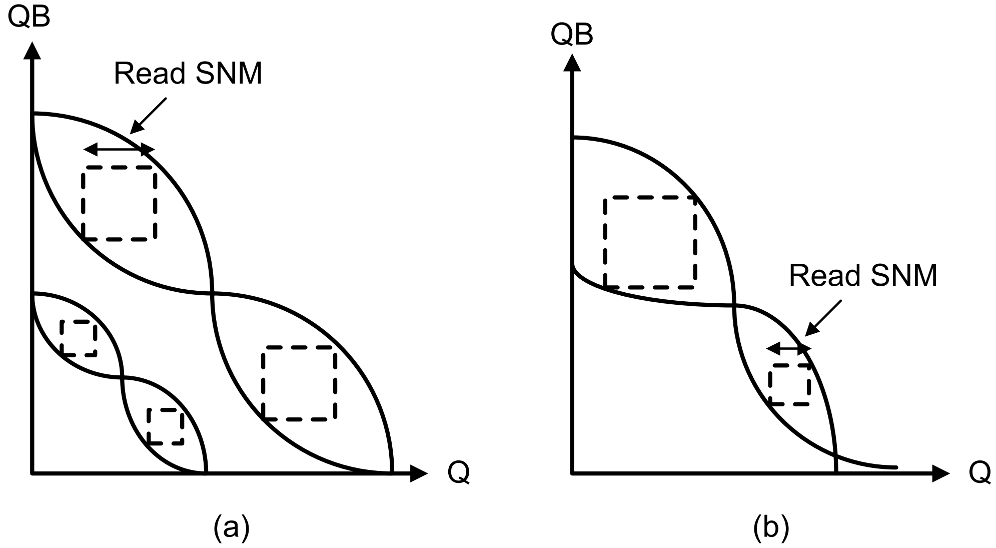


Figure 1.5 Static noise margin degradation as a result of a) reduced $V_{DD}$ and b) $V_{TH}$ mismatch.

## 1.4 Motivation and Thesis Outline

As mentioned above, the speed and power consumption of SRAM greatly affect the total speed and power dissipation of VLSI system. However, embedded memories in nanometer scale CMOS technology suffer from plethora of design challenges such as process variation, increased BL capacitance and leakage power which implies additional techniques and treatments such as read-assist techniques to insure fast, low power and reliable read operation. In this research we try to address these issues in design of sensing circuitry. In particular, we present a read-assist sensing scheme that expedites the process of developing differential BLs voltage that must be sensed by SA while reducing the signal swing on the BLs. This will lead to considerable reduction in power dissipation and increased speed with respect to the conventional scheme. Another advantage of the proposed scheme is that it reduces WL pulse width which will result in an enhancement in SRAM cell stability and noise margin.

The thesis is organized as follows. Chapter 2 describes the general architecture of SRAM and its building blocks as well as SRAM figures of merit. Chapter 3 reviews the existing sensing techniques and circuits along with advantages and disadvantages of each one. Also, it reviews the existing read-assist schemes as well as the advantages and disadvantages of each one. Chapter 4 proposes our read assist sensing scheme along with a complete comparison with conventional circuit and another state of the art design. Chapter 5 summarizes this research and draws the conclusion.

# Chapter 2

# SRAM Architecture and Operation

This chapter discusses the basic architecture of SRAM and its read/write operation. Also, the individual blocks used in a typical column of SRAM will be presented. At the end of this chapter, some of most important SRAM's figure of merits will be discussed.

## 2.1 SRAM Architecture

An SRAM consists of array of bit cells which store data and the peripheral circuitry that enable reading from or writing to the bit cells. Figure 2.1 shows typical architecture of SRAM which consists of $2^n$ rows and $2^m$ columns. A row is selected by inserting one of the $2^n$ WLs which is the output of an n-bit row decoder. On the other hand, a column (BL/BLB) is selected by the output of m-bit column decoder for read or write operation. Typically, a data word which consists of 32 or 64 bits is selected. Hence, for a 64 bit word, each row has $2^m/64$ words and the required number of column decoder inputs will be reduced to $\log_2(2^m/64)$. If the SRAM is large, it may employ several blocks as shown in Figure 2.1. The timing of the peripheral circuits such as sense amplifier, row/column decoders, drivers, etc. is controlled by timing and the control block. The read/write (R/W)

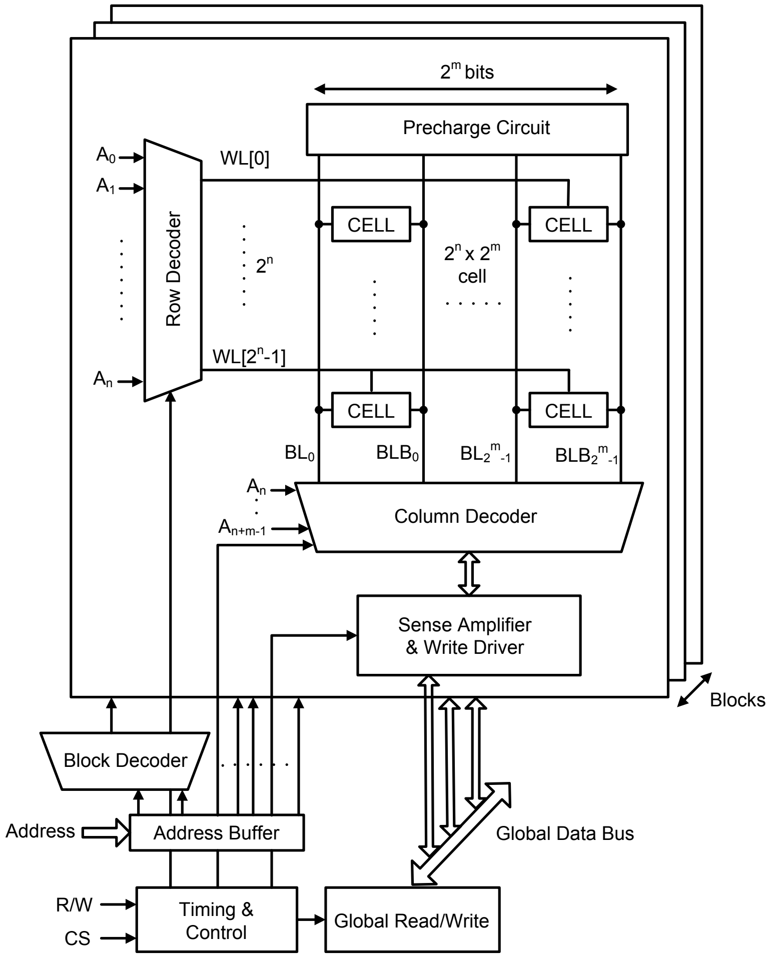


Figure 2.1 SRAM architecture

signals is used for the determination of read or write operation and the chip set (CS) signal is usually employed in multi-chip designs.

During the read operation the integrated SA on each column (sometimes shared between more columns) will be employed to read the data. In write operation, the write drivers will force the BL and BLB of selected column to ‘0’ or ‘1’ and the input data will be written into the internal nodes of the selected cell.

Hence, a typical column of SRAM consists of the following blocks:

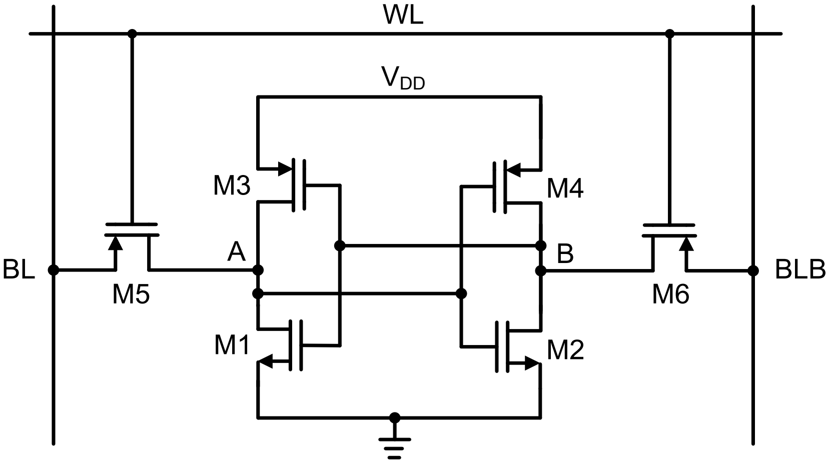


Figure 2.2 SRAM 6-transistor bit-cell

- SRAM cell
- row decoder
- precharge circuit
- sense amplifier
- timing and control

The following subsections will discuss these blocks and present a typical circuit for implementing them.

### 2.1.1 SRAM CELL

The cell is the key element responsible for holding the binary data. Figure 2.2 shows the most common architecture of bit-cell in today's microprocessors and SOCs i.e. the 6T CMOS cell. It consists of two back to back inverters which form a latch and two access transistors. The latch holds the binary data while the access transistors enable the cell to be connected or disconnected to the BLs. In general, a cell should be able to satisfy the following requirements:

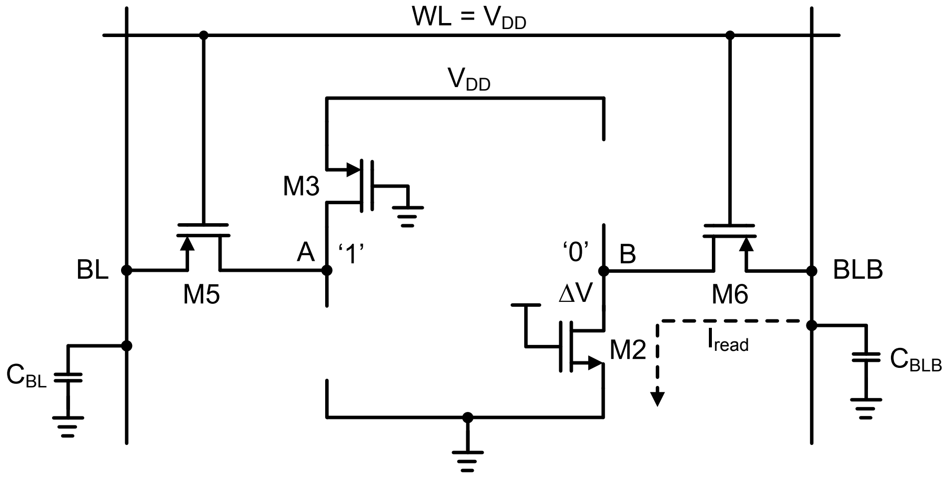


Figure 2.3 SRAM bit-cell during read operation

- Holding the stored data as long as the power is supplied to the cell.
- Providing non-destructive read operation.
- Providing reliable write operation.
- Consuming the minimum area for high storage capacity.

### Read Operation

The read operation initiates by inserting the WL signal which connects the internal nodes of the cell to the precharged BLs. Then, depending on the internal nodes voltage, one of the BLs will develop some differential voltage with respect to the other one. Figure 2.3 shows the cell read operation. In this figure, BL will remain at its precharged value while BLB will be discharged to a lower level through transistor M2 and M6. Due to the voltage divider formed by the M2 and M6, there is degradation (ΔV) in the value of node B which should be kept below the switching threshold of M1-M3 inverter to prevent non destructive read operation. The value of ΔV is dependent on the ON resistance and consequently the sizing of M2 and M6.

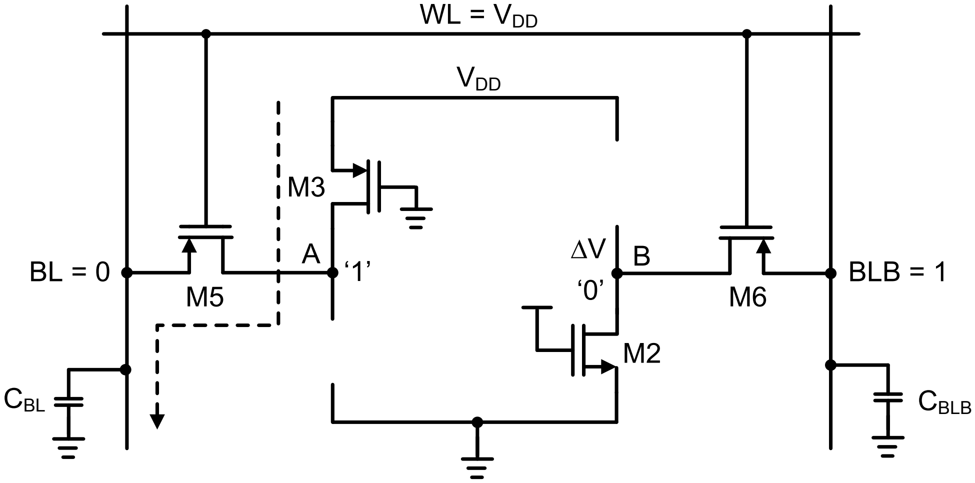


Figure 2.4 SRAM bit-cell during write operation

The developed differential voltage between BLs should be adequate to overcome the effect of offset and other non idealities in the SA. After reaching this point, the SA is enabled which will amplify the small voltage droop on the BLs to full swing signals that can be used by digital logic.

### Write Operation

Figure 2.4 shows the schematic of cell during write operation. Before beginning of write cycle BL is predischrged to 0V while BLB is prechaged to $V_{DD}$. The write cycle starts by asserting the WL signal. The new logic value can be only written by pulling down the value of the internal node that is initially '1' i.e. node A. This is due to the fact that node B voltage cannot be higher than ΔV for read data stability. As a result of current that passes through transistors M3 and M5 to the BL, the voltage at node A will decrease until it reaches the switching threshold voltage of inverter M2-M4. In order to ensure reliable write operation the voltage at node B should be pulled down to a level lower than $V_{THn}$.

From the above mentioned implications of non destructive read operation and reliable write operation, it is obviuos that there is a tradeoff in sizing of access transistors in order to satisfy these requirements. They need to have high ON resistance for read operation while they require low ON resistance for write operation. This tradeoff can be alleviated by choosing minimum width for access transistors (M5 and M6), minimum width for pull-up transistors (M3 and M4) and larger than minimum width (1.5 ~ 1.7) for driver transistors (M1 and M2).

### 2.1.2 Row Decoder

The row decoder activates the WL signal which is necessary for selecting one of the rows. It enables one of the $2^n$ WL signals based on the n-bit address. The number of inputs of row decoder is set based on the required number of bits to access a specific cell in a column. For instance, a column of SRAM which consists of 256 ($2^8$) cells, will require an 8-input row decoder.

Row decoder can be implemented using single or multiple stage architecture. In a single stage architecture, the decoder is realized using a combinational gate such as an n-input NOR gate. Figure 2.5 shows a static PMOS load implementation of row decoder which is employed in order to reduce complexity and area overhead. However, this method suffers from few drawbacks. First, large fan-in of the gates severely degrades the access time of SRAM. Second, fitting the layout of decoder to the WL pitch is not an easy task. Furthermore, it consumes considerable amount of power. Therefore, due to these limitations multiple stage architecture is usually preferred in today’s SRAMs.

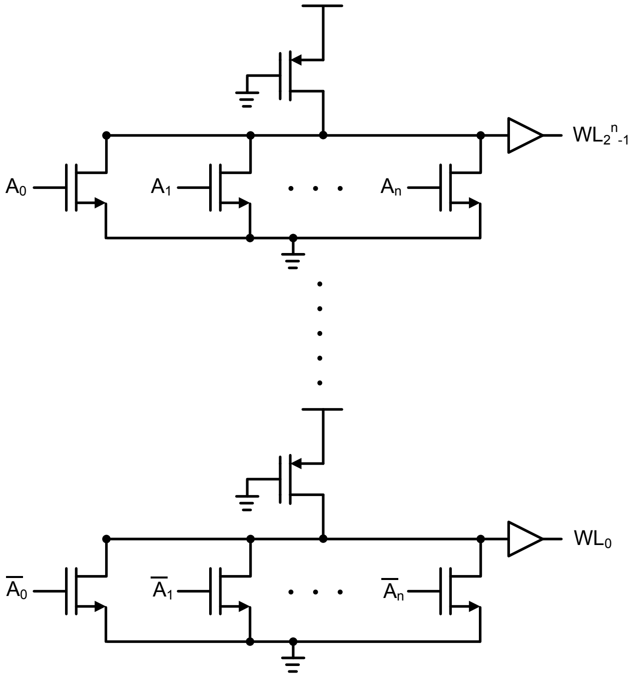


Figure 2.5 Single stage NOR static row decoder

The decoding in multiple stage architecture is done in two stages of pre-decoding and post-decoding. First, the address bits are grouped and applied to the pre-decoder which will prepare the inputs for post-decoder. Then, post-decoder will generate the final WL signals. Figure 2.6 shows a 4-input decoder which is based on two stage of 2-input AND gate. Since in this method the fan-in of the gates is halved, the delay of gates will reduce 4X. Hence, the overall propagation delay will be halved with respect to single stage realization. Furthermore, this method needs fewer numbers of transistors which will result in reduction of area. This characteristic is of particular interest for larger memory size.

### 2.1.3 Sense Amplifier and Precharge Circuit

One of the most important blocks in the periphery of SRAM is sense amplifier (SA) which is employed in order to provide non destructive and fast read operation. Due to the

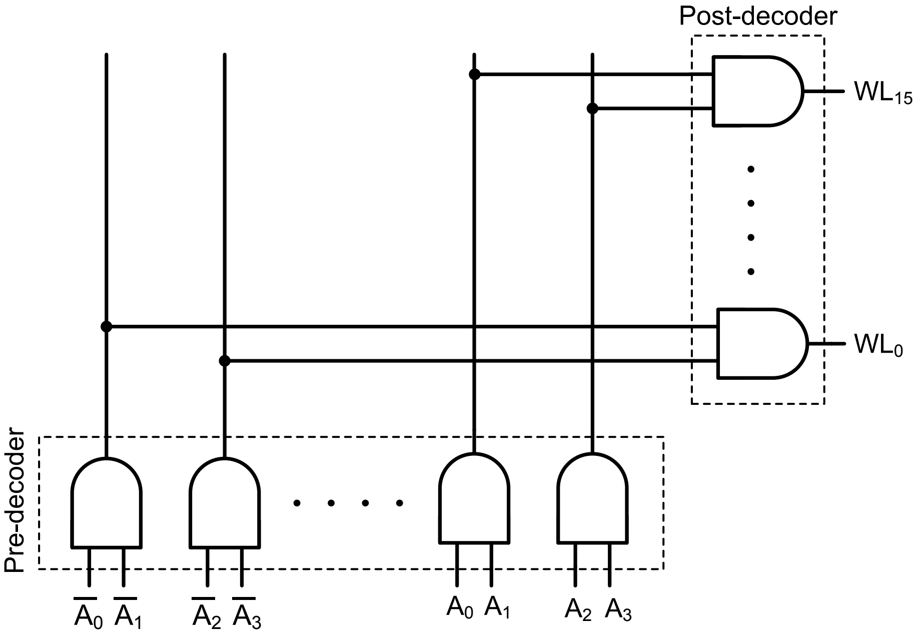


Figure 2.6 Two stage 4-to-16 AND row decoder

small size of memory cell, cell current is not adequate to provide sufficient slew rate during read operation. Hence, SAs are exploited to amplify differential signal on BLs so voltages as small as 200mV can be detected. Since SAs are used to retrieve the stored memory data, their performance and power dissipation strongly affect the overall speed and power consumption of memory. Also, precharge circuits are employed in order to precharge BLs to $V_{DD}$ before the start of read cycle which is necessary for the proper sensing. Figure 2.7 shows the position of SA and precharge circuit in a typical column of SRAM. The employed SA in this figure is a voltage mode latch type SA which is the pervasive choice in today's SRAMs due to its simple structure and low power consumption. In this configuration the positive feedback formed by inverters M1-M2 and M3-M4 will amplify the small voltage droop on BLs to full swing rail-to rail signals at the output.

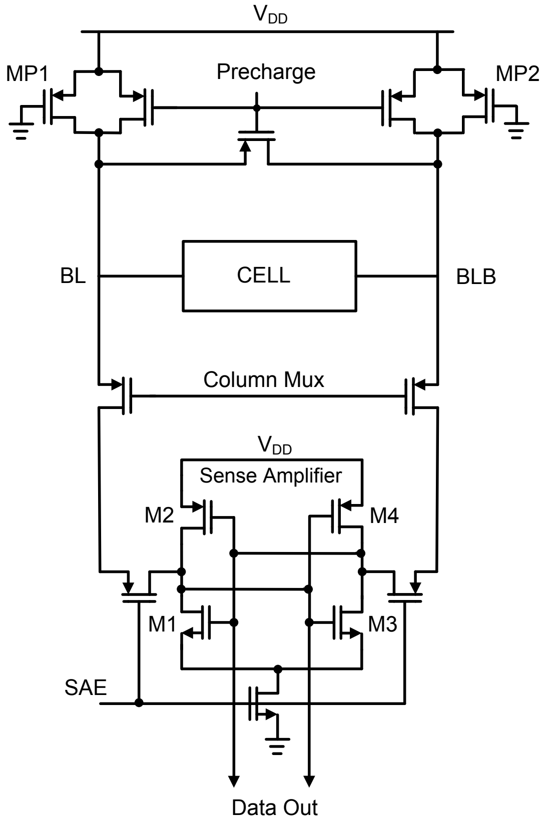


Figure 2.7 Sense amplifier and precharge circuit in a typical column of SRAM

Employing the SA has the following advantages. First, it enables us to reduce voltage swing on BLs, which will result in considerable reduction in power dissipation due to the large capacitive load of BLs. Second, it reduces the access time of SRAM by providing fast read operation. However, in order to function properly they have to meet the following requirements:

- Minimum sensing delay.
- Minimum power consumption.
- High sensitivity.

- Ability to overcome the effect of PVT (process, voltage and temperature) variations.
- Minimum layout area limited to the columns pitch.

Satisfying the above mentioned specifications, require careful understanding and treatment of the specific design limitations that dominate this realm. In fact, the continuous demand for SRAMs with higher speed, lower power and increased storage capacity has opened a new environment for design of SAs. The various type of SA and their design issues, advantages and disadvantages will be discussed in more details in next chapter.

### 2.1.4 Timing and Control Circuit

The timing and control circuit function is to initiate the WL, sense amplifier enable (SAE) and precharge signal for proper read operation. The timing and control block has to meet tough timing requirements during read operation. For example, if the address changes during reading the data, two cells will discharge the BLs at the same time which may lead to reading incorrect value. Also, if the WL signal is activated before precharge signal, the accessed cell may flip its state. In fact, the required delay for activating different signals should be set for the worst case condition which is determined according to process variation and other non idealities. Today’s SRAM employ aggressively small cells which makes them extremely vulnerable to process variation. Therefore, the BL delay dominates the variability in delay. The situation is even exacerbated by the increased effect of process variation in nanoscale CMOS technologies.

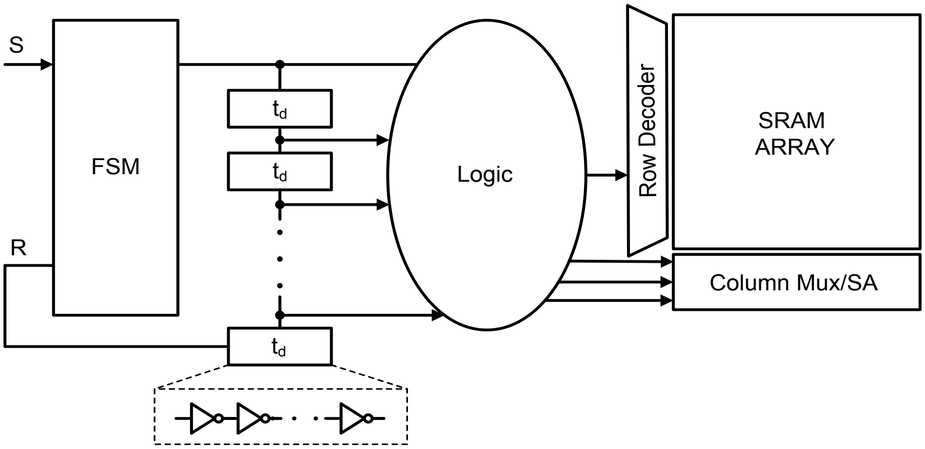


Figure 2.8 Timing and control unit

Figure 2.8 shows a typical implementation of control unit called delay-line timing method [19]. It exploits multiple inverters to define the timing intervals. The control signal S triggers the finite state machine (FSM). The timing is set based on the total delay through the delay elements $t_{d1}$ – $t_{dn}$ in the FSM reset path. Serially connected inverters will be employed to make the delay elements. Then, the timing intervals formed by delay elements along with some logic circuits will generate the required signals for read/write operations.

## 2.2 SRAM Figures of Merit

In this section we explain some of SRAM 's figures of merit which are used in order to characterize SRAM's performance.

### 2.2.1 Data Stability: SNM and DNM

The most common metric for data stability in SRAM is SNM which is defined as the minimum dc noise voltage to flip the state of a cell [20]. In fact, the side of maximum square nested between inverter characteristics that compose SRAM cell, is considered as SNM (Figure 1.5). This is a graphical technique for estimating SNM. During read operation there is degradation in value of the side of accessed cell that stores ‘0’. This is due to the voltage divider formed by the access transistor and the NMOS pull down transistor of accessed cell. The ratio of W/L of NMOS pull down transistor to access transistor which is commonly referred to as β should be adjusted to avoid flipping the cell data. Traditionally SNM is a DC metric which relies on the assumption that the transient time of noise is much lesser than the required time for read and write operation. In modern SRAM this assumption may not be valid and several research works have recently been done in order to take the dynamic behavior of transistors [7], [8] into account which leads to the definition of DNM. In fact, the real noise margin of cell can be higher than what is predicted by SNM once the cell access time is sufficiently shorter than cell time constant [7]. This will result in enhancement of SRAM yield and reduction in cell operating voltage.

### 2.2.2 Read Current Margin

The cell current should be large enough in order to discharge $C_{BL}$ during the activation time of WL signal (Δt) to the required ΔV before enabling the SA. Also, as it was explained in Chapter 1, leakage current is a serious issue in nanoscale CMOS technology. Therefore, the cell current should be sufficient in order to overcome the effect of $I_{leakage}$ of other cells (N cells) connected to the BL. The worst case scenario happens when the accessed cell stores ‘0’ and all other cells store ‘1’. Equation 2.1 gives the read current margin for reliable read operation [36].

$$I_{cell} \geq (N \times I_{leakage}) + (C_{BL} \times \frac{\Delta V}{\Delta t}) \qquad (2.1)$$

### 2.2.3 Access Time

One of the performance metrics of SRAM is access time which is defined as the minimum amount of time required to read the stored data from memory. It is usually measured with respect to the initial rising edge of clock in SRAM read operation. However, access time may be defined in other ways such as the required time for the data to appear at the output pins, after the transition of input data.

### 2.2.4 Power Consumption

As it was explained in Chapter 1, more than half of the total transistor count in today's high speed microprocessors is occupied by cache memories and this ratio is expected to increase more [2]. Furthermore, as predicted by [6], over 90% of total SOCs area is expected to be occupied by memories in the following decade. Hence, SRAM power consumption will greatly affect the overall power dissipation of digital system. That's why several research works have been done to reduce power consumption of memories [9]-[10]. The power consumption of SRAM consists of active power and leakage power. In above micrometer CMOS technologies, the majority of power consumption is due to the active power. However, in sub-micrometer technologies the leakage power becomes as important as active power and techniques for reducing leakage power is highly demanded.

# Chapter 3

# Existing Sense Amplifiers and Read-Assist Techniques

One of the most important building blocks in the design of high performance and low power embedded SRAMs is the sense amplifier. Due to the small size of memory cell, cell current is not adequate to provide sufficient slew rate during read operation. Hence, SAs are exploited to amplify differential signal on BLs so voltages as small as 200mV can be detected. Since they are used to retrieve the stored memory data by amplifying small differential voltage on the BLs to full swing voltages, they are strongly related to the memory access time. Their performance and power dissipation strongly affect the overall speed and power consumption of memory. Due to its critical role in the design of high speed and low power memories, it has become a wide class of circuits by itself and has been the focus of many researchers [21]-[32]. Figure 3.1 shows a typical simplified column of SRAM and the position of the SA.

## 3.1 Sense Amplifier Design Challenges in Nanoscale CMOS Technology

Nanoscale CMOS technology has unveiled few significant challenges to the design of SAs:

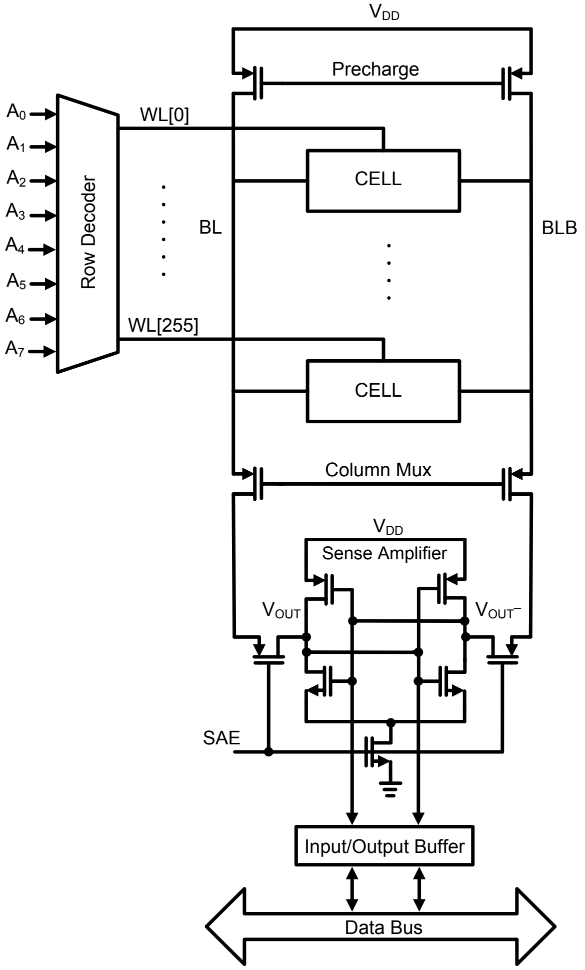


Figure 3.1 Sense amplifier position in a typical column of SRAM

1) Process variation has increased significantly in nanoscale CMOS technology which threatens the reliability of sensing circuitry. In order to alleviate this concern, we need to have larger signal swing on the BLs to overcome the effect of mismatch and process variation in the sensing circuitry, which will in turn degrades speed as well as power dissipation.
2) The second challenge is due to variation in the cell current which will reduce the worst case cell current. Since this cell current is responsible for discharging BLs, this problem will translate into longer activation time for the WLs in order to develop certain amount of differential voltage on the BLs before activating SA which will result in increased

access time. Furthermore, the longer the WL pulse width is, the more likely is the cell to be unstable. A longer WL pulse width can also degrade the noise margin [7], [8].

3) As a result of increase in the size of SRAMs, the BL capacitance has increased significantly which will deteriorate speed as well as power dissipation.

In modern microprocessors and SOCs, there is a continuous demand for high capacity, low power and high speed SRAMs. However, as explained above, there is obvious trade off between these requirements when it comes to the design of SAs. Hence, additional techniques and treatments such as read-assist techniques are highly demanded in order to satisfy the requirements of modern VLSI systems. We present our circuit that tries to target these issues in next chapter.

There are two categories of SAs i.e. voltage mode and current mode. In the next sections we explain these two types and discuss the advantages and disadvantages of each. Also, we present the more frequently used schemes of each category.

## 3.2 Voltage Mode Sense Amplifiers

Latch type voltage mode SA (VSA) is typically used in today's SRAMs (Figure 3.2). Its simple structure and low power dissipation has made it the pervasive choice for modern SRAM's. This type of SA presents high input impedance to the BLs which enables the SA to provide high voltage gain. Ideally, VSA is able to amplify an infinitively small differential voltage. However, due to the increased process variation in nanoscale CMOS technology, SA is experiencing considerable amount of offset. Therefore, in order to alleviate this concern, enough differential voltage should be developed on BLs before enabling SA to overcome the effect of offset.

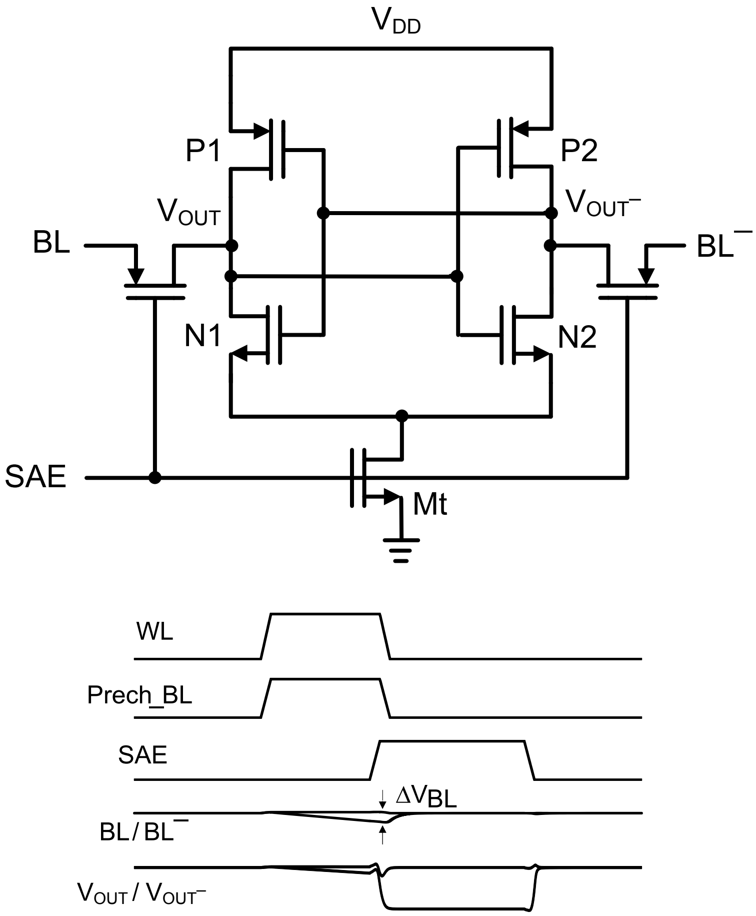


Figure 3.2 Latch type voltage mode SA and associated read waveforms

VSA operates in two phase: precharge and evaluation.

**Precharge Phase:** In the precharge phase both the BLs and internal nodes of SA are percharged to $V_{DD}$. During this phase SAE signal is low.

**Evaluation Phase:** The evaluation phase starts by turning high the WL signal (Figure 3.2). Upon the activation of the WL signal, one of the BLs which is connected to side of accessed cell that stores ‘0’, will start discharging and the other side will stay at the precharged value of $V_{DD}$. Hence, a differential voltage of $\Delta V$ will develop between

BLs. The value of $\Delta V$ depends on BL capacitance ($C_{BL}$), cell current ($I_{CELL}$) and pulse width of WL signal. The amount of $\Delta V$ should be large enough to overcome the effect of offset in SA in worst case scenario. After reaching a certain value of $\Delta V$, we turn on the SA by turning high the SAE signal. As a result of positive feedback formed by cross coupled inverters (transistor N1-P1 and N2-P2), this small $\Delta V$ will be amplified and one of the output nodes goes to $V_{DD}$ and the other one goes to gnd.

The need to develop sufficient differential voltage in VSA internal nodes to overcome the effect of offset will translate into higher power dissipation and lower speed. So, offset is a key limiting factor in the design of VSA. The situation is exacerbated by the increase in the $C_{BL}$ which comes as a result of increase in the size of SRAMs with each generation of CMOS technology. This will make the voltage swing on the BLs even more power hungry. Furthermore, the performance of this type of SA strongly depends on the $C_{BL}$ and will severely degrade with the increase in the number of cells that are connected to the BLs i.e. the size of memory.

Another type of voltage mode SA is shown in Figure 3.3. This type of SA consists of a tail transistor that provides the DC current to bias the amplifier in its high gain region, two input differential pair transistors which are connected to the BLs and two current mirror load transistors. The gain of this SA at node X is equal to:

$$G = -g_{m1}(r_{o1} \| r_{o2}) \qquad (3.1)$$

where $g_{m1}$ is the transconductance of input transistors and $r_{o1}$, $r_{o2}$ are the small signal output resistances of load transistors. This scheme has two advantages. First, since BLs are connected to the gate of transistors, it presents high input impedance to the BLs.

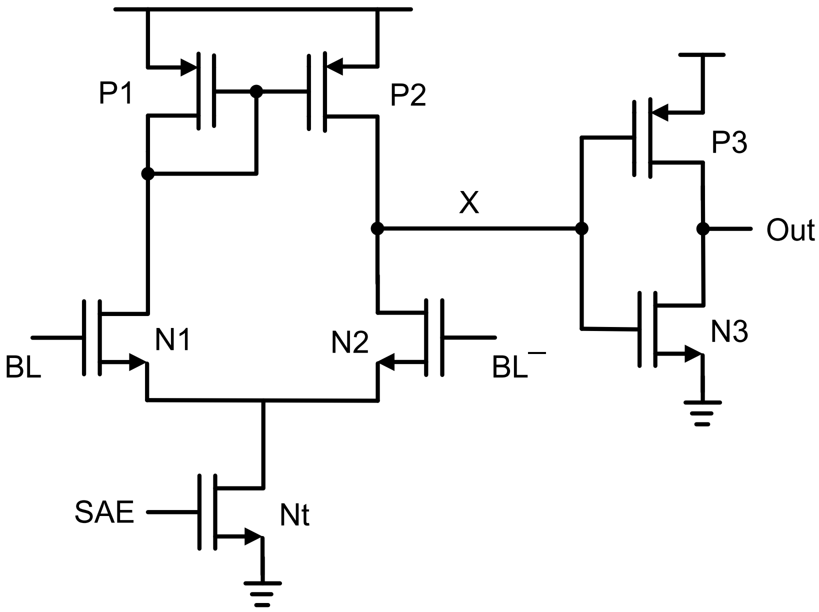


Figure 3.3 Differential sense amplifier with current mirror loads

Second, unlike VSA, it separates inputs and outputs nodes of sense amplifier. In order to function properly the two input differential pair transistor should be completely matched. So, it is highly sensitive to process variation and also, they suffer from the same offset problem as in the case of VSA. In order to have sufficient gain, bias current should be large enough which will increase power consumption. Hence, it is not the preferred choice for low power applications. Furthermore, since the amplifier is turned on after having sufficient voltage swing on the BLs, its performance will degrade with increased BL capacitance.

## 3.2 Current Mode Sense Amplifiers

Current mode SAs have been proposed as good candidates for high speed application. Current mode technique for application in SRAMs was first introduced in

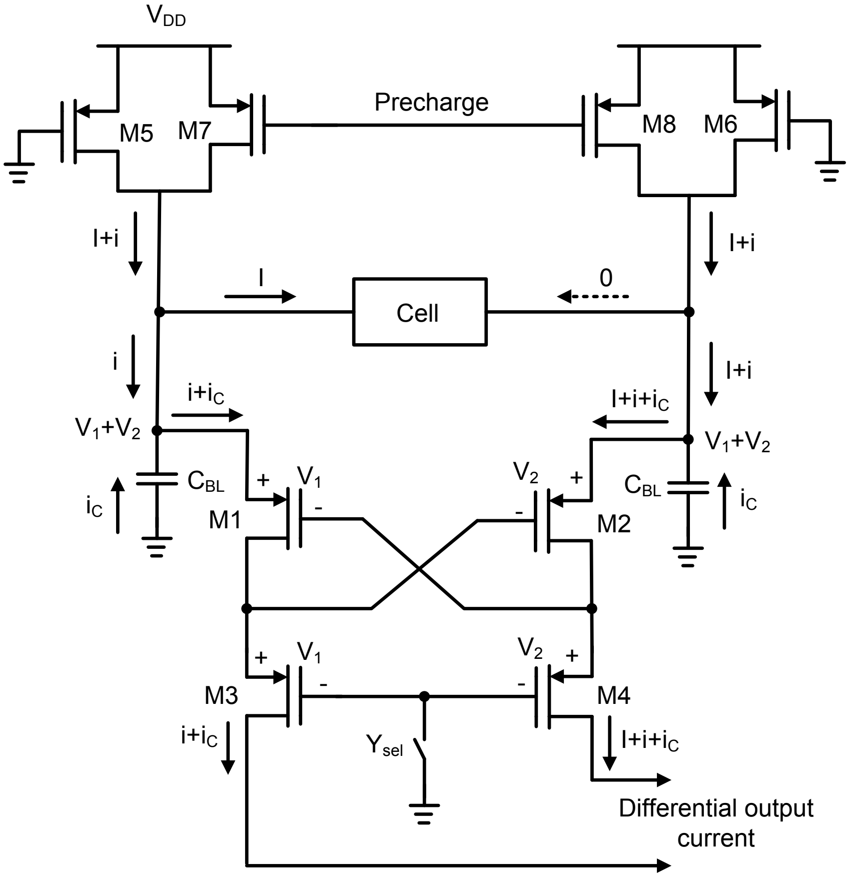


Figure 3.4 Current conveyor

[28]. Since, this type of SA detects differential current rather than differential voltage, they do not require large voltage swing on the BLs. They work as a current buffer for cell current by presenting small input resistance to the BLs. Figure 3.4 shows a current sensing scheme based on current conveyor [28]. The current conveyor consists of 4 equal size PMOS transistors M1-M4. The read operation starts by grounding the $Y_{sel}$ node. As shown in Figure 3.4, the side of accessed cell that stores 0 will draw the current I. Since M1 (M2) and M3 (M4) carry the same current and they are of the same size, their gate-source voltage will be equal which is shown by $V_1$ ($V_2$) in the figure. Since, the gate of M3 and M4 are grounded, the BLs will be both at the same potential $V_1+V_2$ regardless of

BL currents. So, there exists a virtual short circuit between the BLs. Since, the BLs are at the same potential, the load current as well as BL capacitor currents will be equal. As a result of the current drawn by the left side of cell, the right hand side of the circuit will carry more current than the left side. Ideally the difference between these two side's current should be equal to I, the cell current. Hence, a current difference equal to the cell current is implemented at the drain of M3 and M4. Then, this differential current will be utilized to produce full swing voltage at the outputs of SA. Another advantage of this method is that, since no capacitor discharge is required to sense the cell data, the sensing delay is unaffected by the value of BL capacitance.

However as it is shown in [29], the implemented differential current at the output of current conveyor is much smaller than cell current due to the imperfection of the current conveyor. This is due to the assumption that during read operation, there exists a virtual short circuit between BLs. Hence, the load transistors M5 and M6 are expected to source the same current. However, this assumption is valid in an ideal case where process variation and short channel length effects are not present. As a result of increased process variation and short channel length effects in deep sub-micrometer technologies, there exists slight difference between BL voltages. Since the load transistors M5, M6 are operating in the triode region, their drain current is strongly dependent on their drain-source voltage. Therefore, the slight difference between BL voltages will translate into a large difference in drain current of load transistors. Consequently, the implemented differential current at the output of current conveyor is much less than cell current.

Another disadvantage of this technique is that it consumes more area and imposes DC power dissipation. So, another circuit should be employed to turn off the SA when it

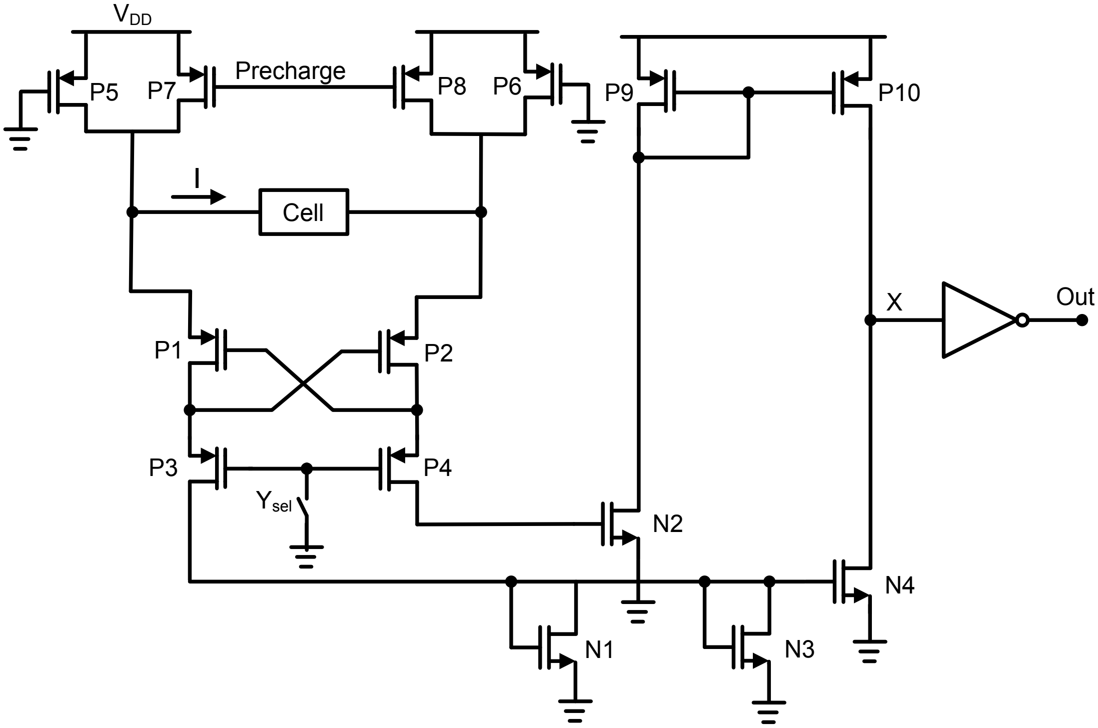


Figure 3.5 Simple four transistor (SFT) SA

becomes idle [30] which will add to the complexity and area overhead of the scheme. Also, since they operate on differential current rather than differential voltage, they are more vulnerable to the variation of cell current. In the next sessions commonly used schemes of current mode SA are presented.

### 3.2.1 Simple Four Transistor Sense Amplifier

The simple four transistor (SFT) SA [28] is shown in Figure 3.5. This type of SA exploits a current conveyor which consists of transistors P1-4. Assuming that the left side of the cell stores 0 and draws current I, it follows that the right hand side of current conveyor pass more current. The differential current at the output of current conveyor is fed to diode connected transistors N1 and N3 in order to convert it to a differential

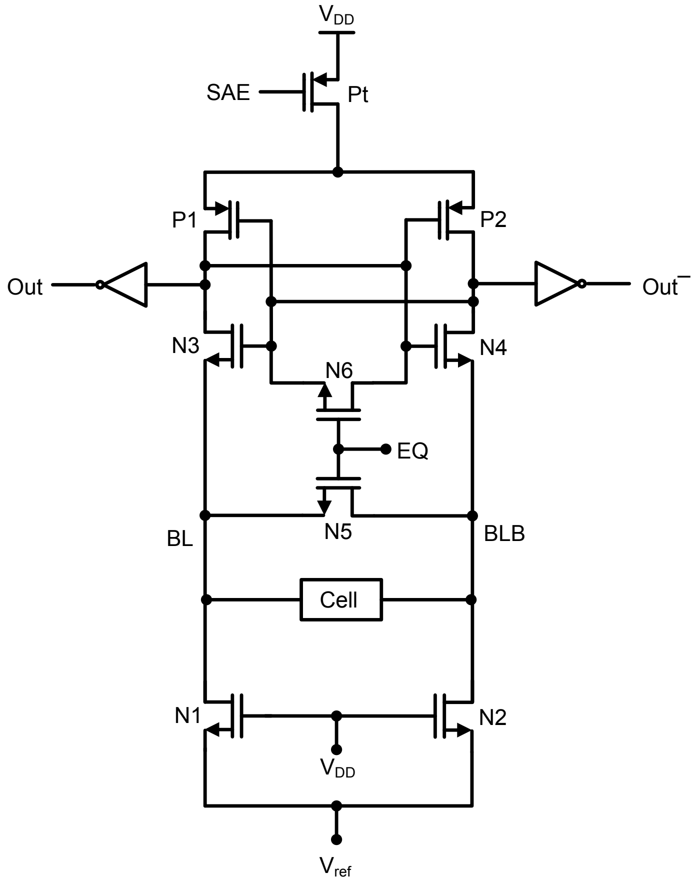


Figure 3.6 Clamped bit-line (CBL) SA

voltage. This differential voltage is then applied to N2 and N4. Hence, N2 will carry more current than N4. Since P9 and P10 form a current mirror, current of P10 will be equal to current through N2 and the difference between the currents of P10 and N4 will charge up node X.

### 3.2.2 Clamped Bit-Line Sense Amplifier

Clamped bit-line (CBL) SA [31] is shown in Figure 3.6. This type of SA exhibits fast speed which is achieved by relocating the BLs at nodes that have less effect on the speed. Transistors N1 and N2 are biased in the linear region by connecting their gate to $V_{DD}$ and provide a low impedance clamp between the BL and $V_{ref}$, hence the name clamped bit-line. Owing to the small resistance of N1 and N2, the BL capacitance will discharge only few tenth of a volt during the read operation. It is worth mentioning that $V_{ref}$ is a low potential used to precharge the BLs. Transistors N5 and N6 serve the purpose of equalizing the BLs and the output nodes of SA during the precharge cycle. The read operation starts by turning low the SAE signal which will provide the SA with power. Transistors P1, P2, N3 and N4 form a cross coupled latch that provides a positive feedback. During the read operation, the difference between drain current of N3 and N4 will be amplified by the positive feedback mechanism of cross coupled inverters and cause one of the output nodes to go to $V_{DD}$ and the other one to go to gnd.

### 3.2.3 PMOS Bias Type Sense Amplifier

The PMOS bias type (PBT) SA [32] is shown in Figure 3.7. Assuming that the BL loads source the current of $I_0$ and the left side of cell stores 0 and draws current I, it follows that the current flowing from right BL to the amplifier is $I_0$ while the left BL passes a current equal to $I_0$-I to the amplifier. Therefore, P1 and P3 carry a current equal to $(I_0-I)/2$ while P2 and P4 have a current of $I_0/2$. Since N1 and N2 form a current mirror, the current that flows through N1 will be the same as P2 and is equal to $I_0/2$. As a result of the difference between drain current of P1 and N1 the capacitance at node X will discharge and the output voltage will rise to $V_{DD}$. In the same manner, since more current flows through P4 than N4, the capacitance at node Y will charge up and the complementary output voltage will fall to gnd.

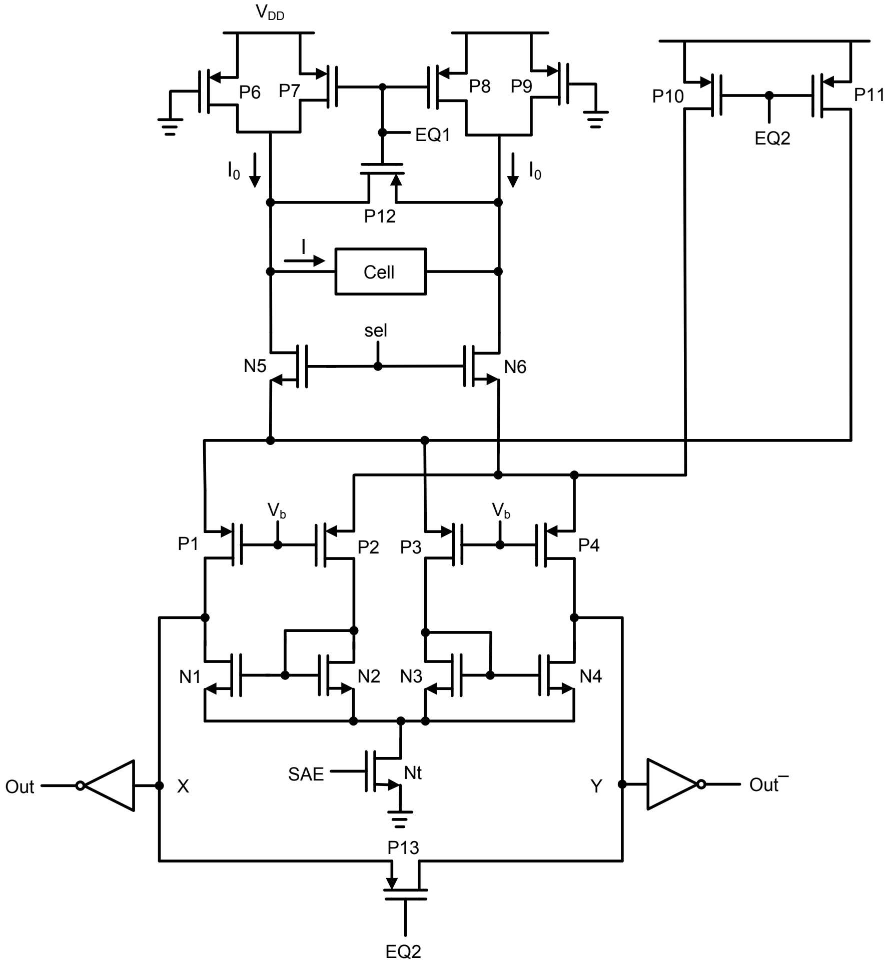


Figure 3.7 PMOS bias type (PBT) SA

### 3.3 Read Assist Techniques

As explained before, nanoscale CMOS technology along with continuous increase in the size of SRAMs has unveiled serious challenges in the design of high performance and low power memories such as increased process variation, reduced cell current and increased BL capacitance. In order to cope with these problems, additional techniques and treatments such as read-assist techniques are highly demanded. In this section, four state of the art read-assist schemes along with their advantages and disadvantages will be discussed.

### 3.3.1 Read-Assist Scheme Using Local Sense Amplifier

During the read operation there is degradation in the value of data in the side of accessed cell that stores ‘0’. This is due to the voltage divider that is formed by the access transistor and the NMOS driver transistor of accessed cell. Read access disturbs can be reduced by decreasing the flow of charge from the precharged BL to the side of accessed cell that stores ‘0’. The quicker the BL can be discharged, the less likely is an unstable cell to lose its data. This is specifically important for half-selected unstable cells i.e. cells on the same row where WL signal is inserted but are not meant for read or write operation. Pilo et al have proposed a read-assist scheme in [33] in which a SA is integrated on each column that will provide full BL amplification on both active and half-selected columns. This full BL amplification enhances the discharge rate of the side of accessed cell which stores ‘0’ and improves the weak cell stability by writing back the original data of cell from the sensed data on the BLs.

Figure 3.8 shows the proposed read-assist scheme in [33]. The integrated SA on each column will expedite discharge of BL capacitance by providing additional discharge current to the cell current. However this method suffers from three major problems:

1- Since BLs undergo full amplification, the power consumption increase drastically.

2- Since BL capacitance in this scheme should be kept low, the number of cells that are connected to each column should be reduced. Hence, in order to have a

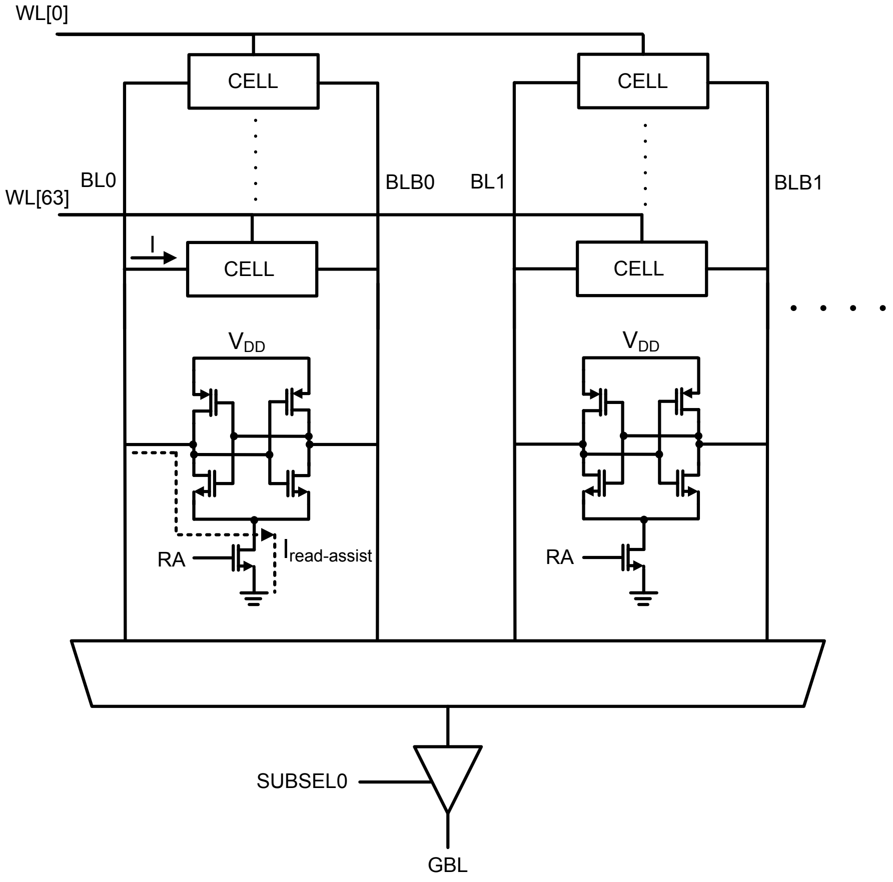


Figure 3.8 Read-assist scheme using local sense amplifier

certain memory capacity, the number of column must be increased which will result in large area overhead.

3- Small sized SA integrated on each column is more prone to process variation as opposed to the conventional case in which a SA is shared between more columns. This is due to the fact that the variance of device parameters is inversely proportional to transistors area [34].

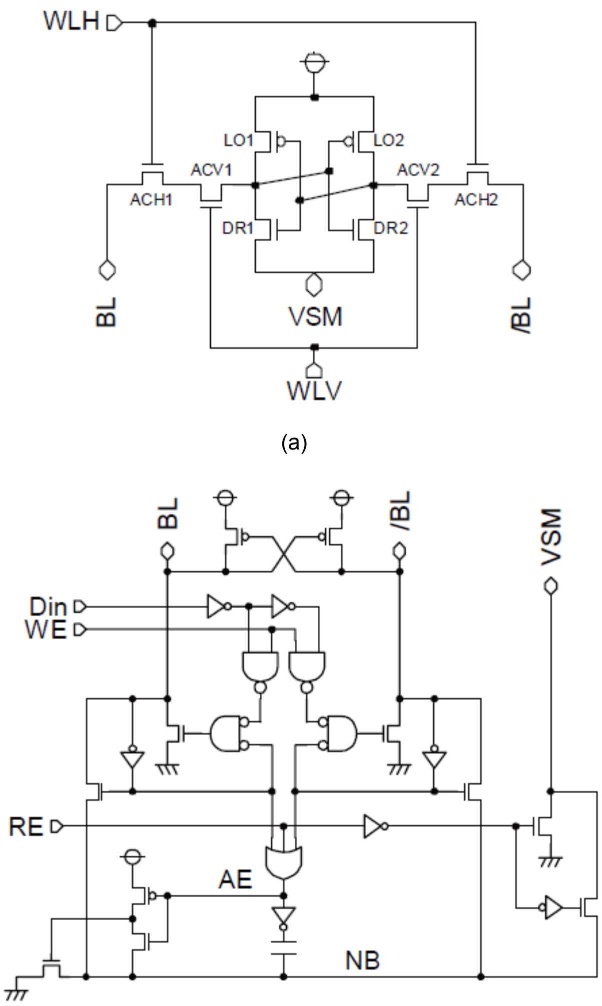
### 3.3.2 Negative Biased Read-Assist

One of the solutions to implement low power applications is to reduce $V_{DD}$. However, as discussed in Chapter 1, reducing supply voltage ($V_{DDmin}$) seriously threatens the cell stability caused by the variation in $V_{TH}$ of transistors. In [35] an 8T-SRAM cell and a new negative bias assist circuit is presented that enables us to reduce $V_{DDmin}$ while enhancing cell stability and reduce access time and power dissipation. Figure 3.9 shows the proposed cell and read-assist scheme. During the read operation, the assist circuit makes VSM signal negatively biased after activating the read enable (RE) signal. This VSM signal actually serves as column $V_{SS}$ during the read operation. Therefore, the enlarged cell bias enhances the static noise margin (SNM). Furthermore, this negative biasing expands the overdrive voltage of access and driver transistor of the cell. Hence, leading to increased access speed. If this method is applied to 6T-cell, it will result in increased power dissipation. This is due to the increase in cell current of half selected columns. Contrary to 6T-cell, in the presented 8T-cell, only a unit $V_{SS}$ in the selected column is forced to negative bias which does not bring any extra power consumption. Furthermore, in this scheme there are no half selected cells, which contribute to the reduction of power consumption.

However, this structure suffers from two major drawbacks:

1- The employed cell in this scheme consists of 8 transistors in contrary to the typical 6 transistors cell. This will lead to large area overhead, especially for higher memory capacities.

2- It exploits complicated negative voltage generator that adds to the complexity of the designed SRAM. From the assist schematic (Figure 3.9 (b)), it seems



(a)

(b)

Figure 3.9 Negative biased read-assist scheme [35]

unlikely that it can fit to the column pitch or even share it between multiple columns. Also, the negative bias generator will add to the power consumption of the SRAM.

### 3.3.3 Integrated Read Assist-Sense Amplifier Scheme

The proposed scheme in [36] speed up the sensing operation by providing a controllable additional BL discharge path while reducing read power consumption by

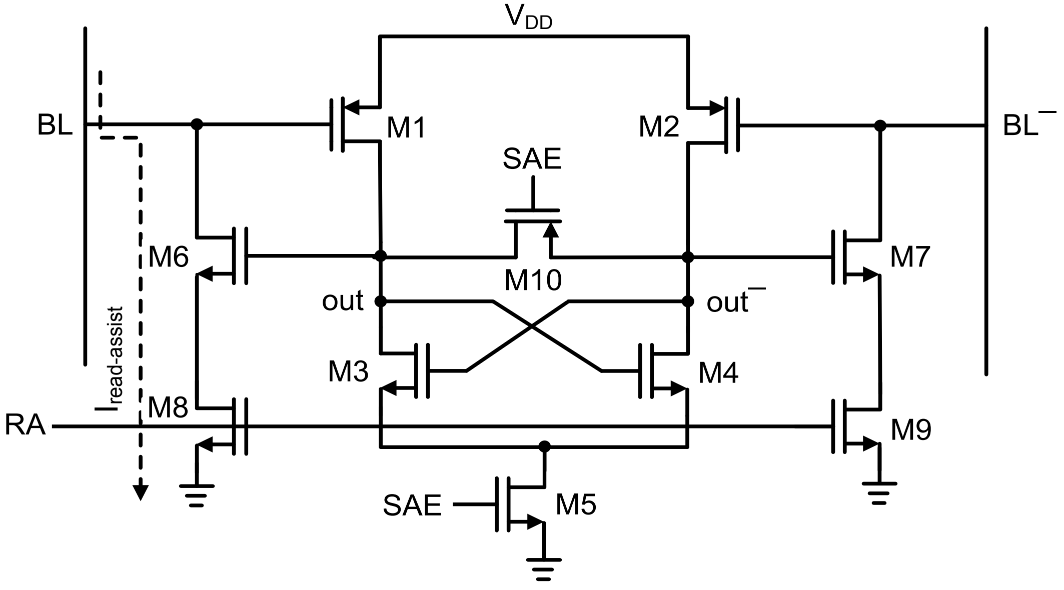


Figure 3.10 Integrated read assist-sense amplifier scheme

preventing the BLs from having full swing (Figure 3.10). This scheme operates as follow. At the beginning of read cycle, the BLs are precharged to $V_{DD}$-$V_{TN}$ by NMOS prechrage transistors. Hence, transistors M1 and M2 will operate in the linear region with $V_{GS}$=$V_{TN}$ and the internal sensing nodes out/out¯will be initially high. After inserting the WL signal, one of the BLs will be discharged to a level lower than precharged value and the other BL will be charged up toward $V_{DD}$ which results in leakage compensation on the opposite BL. Since, the gate voltage of M1 has decreased, it will be shifted towards saturation region while as a result of increase in the opposite BL voltage M2 will move into cut-off region. Hence, out¯ which is connected to the weak PMOS will no longer be able to hold its high value compared to out which is connected to the strong PMOS. By inserting SAE signal, the positive feedback mechanism formed by M3 and M4 will be activated. The difference between the strength of M1 and M2 will cause out¯ to fall and out to rise. The read-assist operation starts by the insertion of RA signal which is delayed with respect to SAE signal.  Transistors M6 and M7 present another positive feedback mechanism to the sensing scheme by providing additional discharge path for the BLs. It is worth mentioning that the duration of the RA signal can be changed in order to control the amount of BL discharge.

However, this scheme suffers from some major drawbacks:

1- The reduced precharge voltage level on the BLs from $V_{DD}$ to $V_{DD}$-$V_{TN}$ will reduce the strength of access transistor and the NMOS driver transistor of the accessed cell. This will translate into the reduced cell current and longer access time.

2- This scheme requires having sufficient differential voltage at the internal nodes of SA (out/out¯) before enabling read-assist mechanism. This will result in reduced speed and in fact, according to my simulation results it will not be effective to improve the sensing speed with respect to conventional VSA. According to the data reported in the paper, the speed improvement has been measured for BLs differential voltage of 200mVand 500mV which is not helpful for real cases. In fact, a BL differential voltage between 100mV and 150mV is adequate for proper sensing. In other words, the scheme is not fast enough to respond from the beginning of read cycle and to reach to BL differential voltage lesser than 200mV e.g. 150mV. Furthermore, the increased voltage swing on the BLs (more than necessary amount i.e. 150mV), will result in increased power consumption.

### 3.3.4 Selective Precharge Read-Assist Scheme

Reducing the precharged value of BLs before inserting the WL signal will improve the stability and SNM of accessed cell which comes as a result of reduction in the strength of access transistor. In the proposed scheme in [37], the upper and lower part of BL is precharged to $V_{DD}$ or predischarged to GND (Figure 3.11). Then, after closing the switches, as a result of charge sharing between $C_{BL}$ and $C_{SL}$, the precharged value of BLs can be precisely controlled. Hence, the SNM of accessed cell will improve. It is worth mentioning that the final value of BLs depends on the capacitance ratios and since the ratio of capacitances shows very weak dependence on PVT corners, this scheme shows high immunity against process variation. Also, in order to generate the required BL voltage, no additional reference voltage is used which reduces the complexity. In the last step, contrary to the selected column, the MUX devices of all unselected columns will be disabled. Therefore, even half-selected cells will have improvement in their SNM since their BL voltage has been reduced.

The proposed scheme requires another time interval for the charge sharing between capacitances before the WL signal can be inserted which will result in increased access time of memory. So, another technique is required to accommodate this problem. As shown in [38], reducing the BLs voltage which is in fact the common mode voltage for the input transistors of SA, will improve robustness and decrease offset of current latch type SA. Since, this scheme reduces BLs voltage, it experience less offset in the SA. Hence, it requires less differential voltage on the BLs before enabling the SA which will translate into reduction in the pulse width of WL signal. So, the increase in the access

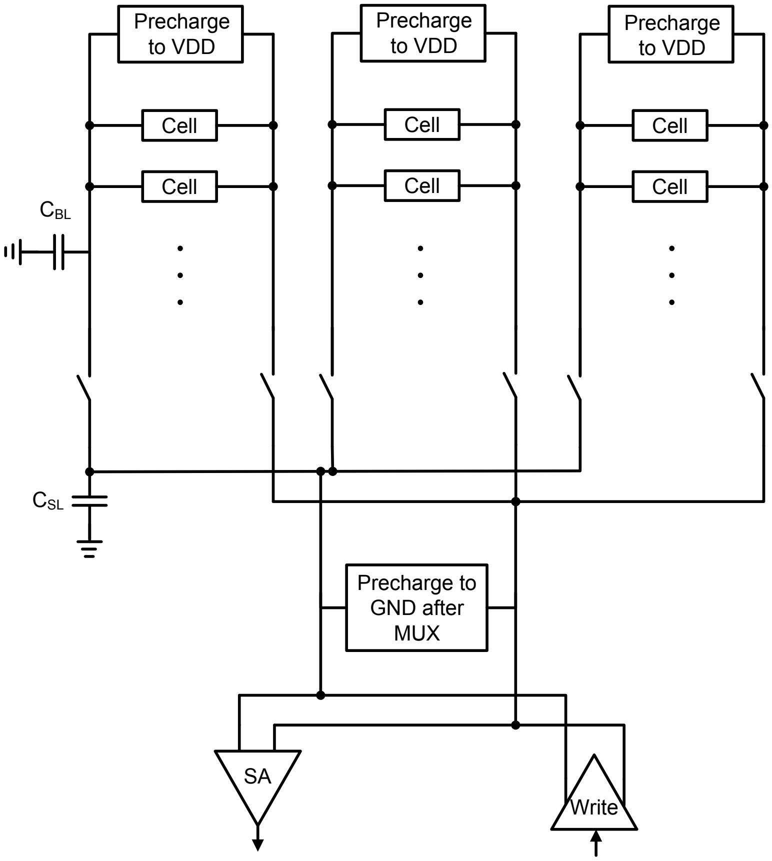


Figure 3.11 Selective precharge read-assist scheme

time which comes as result of additional charge sharing interval before inserting WL signal is compensated by decreasing the WL pulse width.

However, this scheme suffers from the following drawbacks:

1- The reduction in the BL voltage will reduce the strength of access transistor and NMOS driver transistor of accessed cell. This will reduce the cell current which is responsible for discharging the BLs. Hence, it will translate into longer access time and sensing delay.

2- In this scheme, the BL voltage is reduced to a level lower than $V_{DD}$ during the charge sharing phase and raised to $V_{DD}$ again during the precharge phase. Since BLs contain heavy capacitive loads, this charge and discharge of BLs will increase power consumption significantly.

# Chapter 4

# High Speed and Low Power Read-Assist Technique

## 4.1 Proposed Read-Assist Technique

In this chapter we present a new read-assist scheme [39] that attempts to remove the drawbacks of aforementioned schemes. The proposed scheme improves the performance of sensing circuitry by providing amplification for BL voltage droop while reducing the BL voltage swing which results in significant power reduction. Furthermore, the proposed circuit enhances the SRAM cell stability by reducing the WL pulse width. Also, for having the same operating frequency, it enables us to reduce cell $V_{DD}$ with respect to the conventional and referenced schemes which results in considerable reduction in leakage power dissipation. It is worth mentioning that we reached to this scheme by the modification that we applied to our primitive designed circuit which is given in Appendix A.

Figure 4.1 shows the proposed read-assist scheme [39]. It consists of two pre amplifier circuits that are integrated on each column and a conventional SA. The operation of this scheme is as follows. Before the start of a read cycle the BLs are precharged to $V_{DD}$ by precharge transistors. The read cycle starts by activating the WL

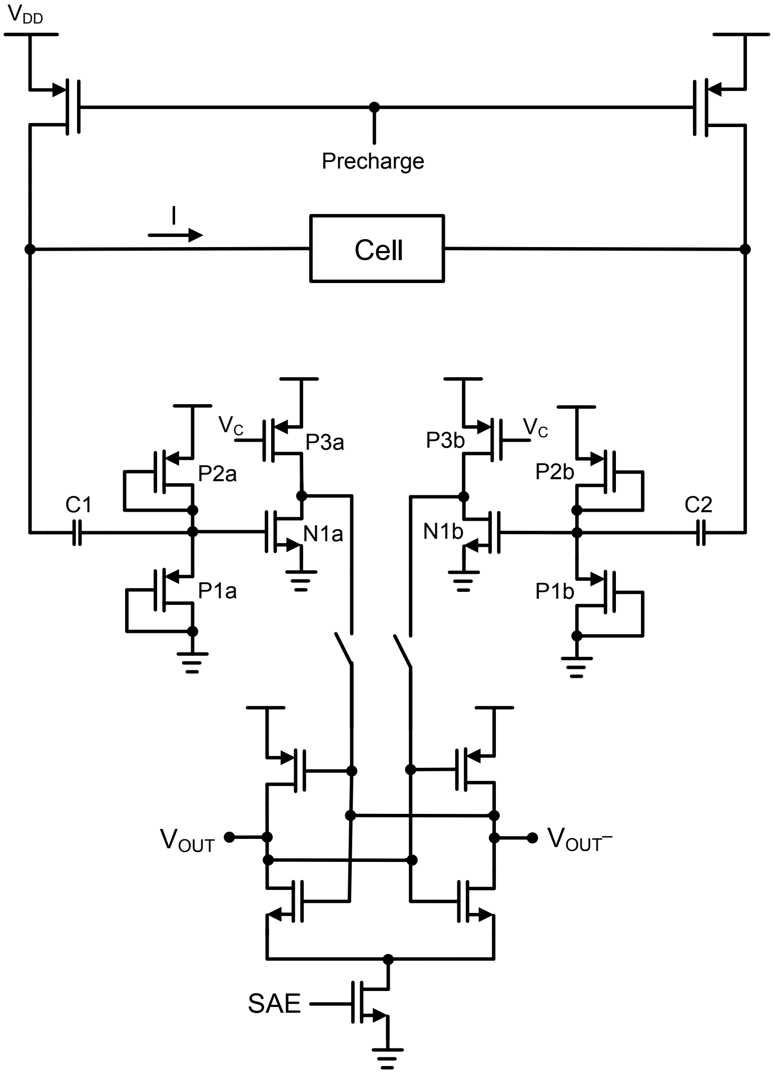


Figure 4.1 Proposed read-assist scheme

signal, which forces the BLs to develop differential voltage. Upon the activation of WL signal, we turn on the read-assist circuit by turning low the $V_C$ signal. Transistors P3a and N1a form a simple common source amplifier which is employed in order to amplify the voltage droop on the BL that is connected to the side of the accessed cell that stores 0. Transistors P1a and P2a serve the purpose of biasing the amplifier by providing appropriate bias voltage for the gate of N1a. A decouple capacitor is employed in order to decouple the DC value of the BLs from the bias voltage at the gate of N1a. It is worth mentioning that this 15fF capacitance is implemented using nmos capacitance. Thanks to the high gain of the preamplifier formed by N1a and P3a, the tiny developed ΔV on the BL will be amplified to a considerably large value at the drain of N1a. On the other BL which is connected to the side of accessed cell that stores 1, we do not have any voltage droop. So, the small signal voltage applied to the preamplifier on that side will be zero and we do not have any amplification on that side. As it was explained in the previous chapter, in a conventional scheme the developed differential voltage on the BLs ΔV should be large enough to overcome the effect of offset and other non idealities in the SA. However, in the proposed scheme, we perform a preamplification on the BL voltage droop before applying it to the SA, in order to reduce the BL voltage swing. In our design we reduced BL voltage swing by 3X i.e. from 150 mV to 50 mV.

Figure 4.2 shows the read waveforms of the proposed scheme in comparison with the conventional one (Figure 3.2). On the other hand, as it is shown in Figure 3.2, the differential amplified voltage at the internal nodes of SA (ΔV') is about 500 mV which is much more than the case of the conventional scheme i.e. 150 mV. Then by turning the SA ON, it will amplify this increased voltage to the supply rails and latch to its stable point quickly.

Using the above mentioned preamplification read-assist technique has the following advantages:

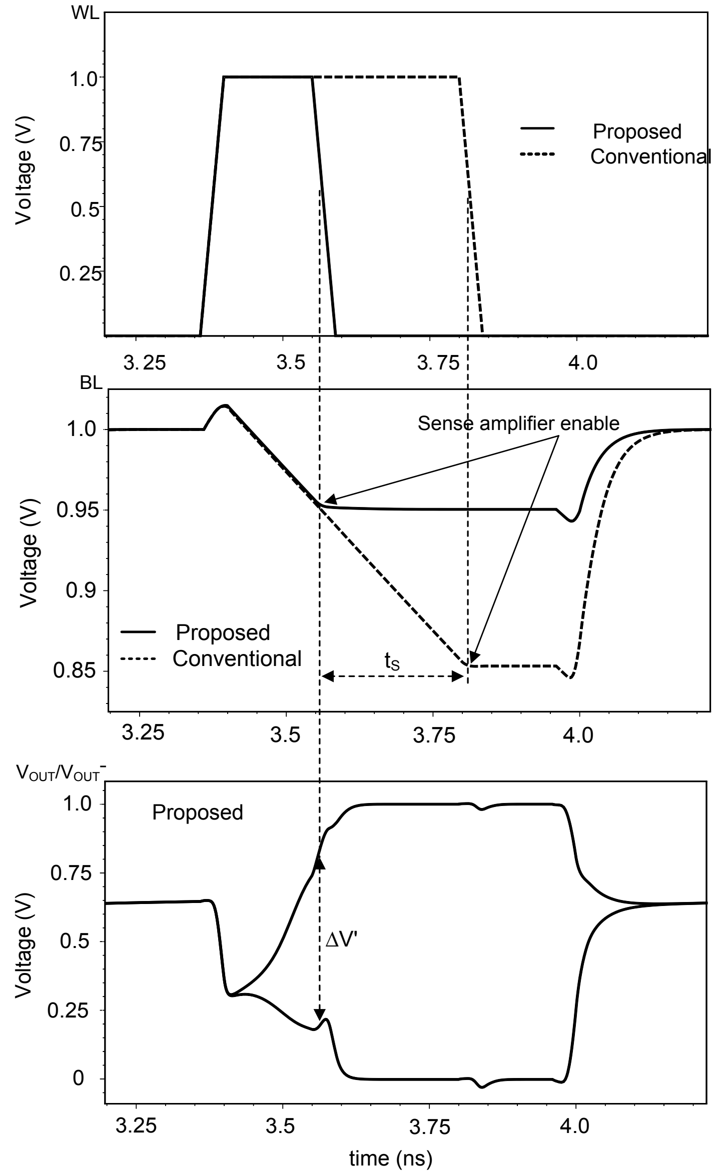


Figure 4.2 Comparison between read waveforms and timing of proposed and conventional scheme

1. Providing amplification for BL voltage droop enables us to reduce BL voltage swing which is necessary before enabling the SA. Due to heavy capacitive load on the BLs, this reduction in voltage swing will lead to considerable reduction in power consumption.

2. This scheme enables us to reach the required differential voltage at the input of the SA much sooner than conventional scheme, which is necessary before enabling it. This will result in significant improvement in the sensing speed. This time saving is shown by $t_S$ in Figure 4.2.
3. The increased differential voltage at the inputs of the SA will expedite the positive feedback mechanism in the SA and cause much faster response than the conventional case.

4. The proposed scheme reduces the WL pulse width i.e. the access time of cell which will enhance cell stability and DNM [7].

## 4.2 Simulation and Comparison Results

The proposed read-assist circuit is implemented in low power (LP) 65nm TSMC technology. In order to check the effectiveness of the proposed scheme, a complete comparison is made between our proposed scheme, the conventional SA and reference scheme [33] using Spectre. In order to have a fair comparison, the sizes of all the transistors in SA are kept the same in proposed scheme and the conventional one.

Figure 4.3 shows a comparison between the conventional SA and the proposed read-assist scheme for BL capacitance of 60 fF and clock frequency of 1GHz at supply voltage of 1 V. The speed improvement and power reduction are 56.1% and 25.9%, respectively.

Figure 4.4 compares the sensing delay of the proposed scheme with conventional and reference [33] circuit as a function of BL capacitance for SAE clock frequency of 500MHZ which shows significant improvement over them. Also, as we expected, the

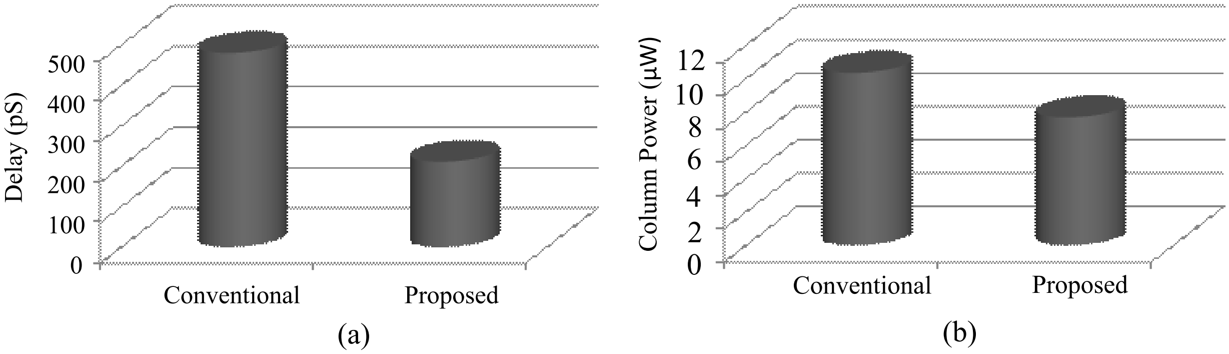


Figure 4.3 a) Sensing delay and b) column power comparison between proposed scheme and conventional circuit at $V_{DD}$=1V, Freq=1GHz and $C_{BL}$=60fF

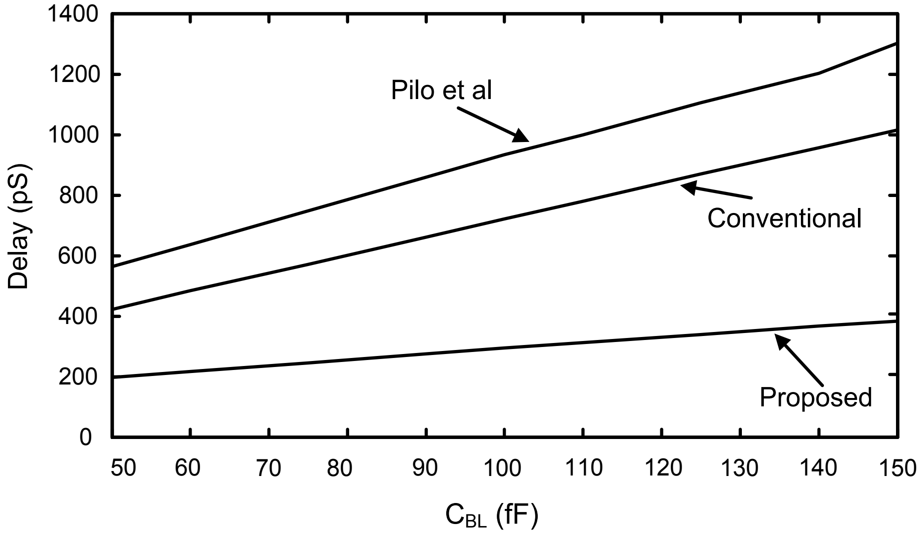


Figure 4.4 Delay comparison

speed improvement is more significant for larger BL capacitance which makes the proposed scheme more attractive for higher memory capacities. This is due to the fact that BL swing is reduced in proposed scheme which makes it less sensible to the increase in BL capacitance. The power comparison is shown in Figure 4.5. As it is obvious from the figure, the proposed scheme consumes the least power among all and shows a drastic power reduction with respect to referenced circuit which is due to the reduced BL voltage

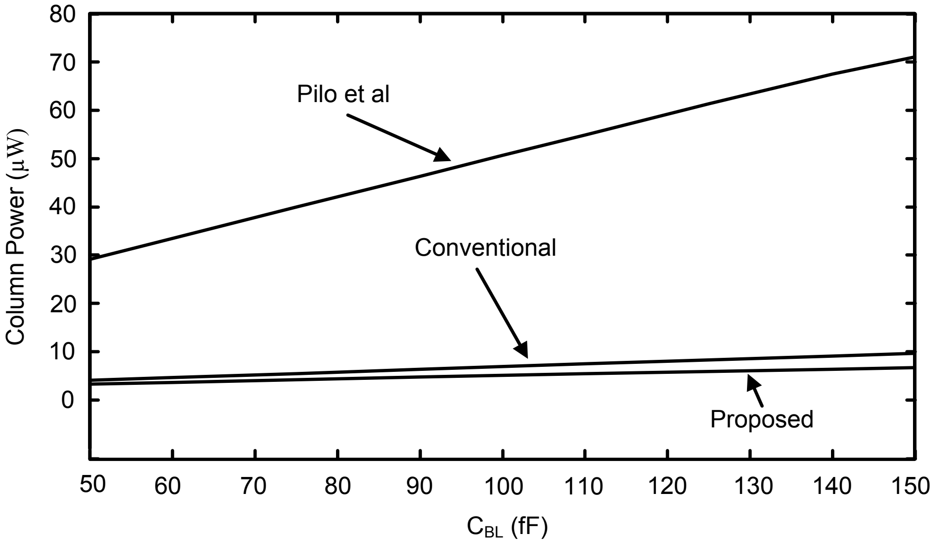


Figure 4.5 Power comparison

swing. Again the power reduction increase for larger BL capacitances which indicate the fact that the proposed scheme is much more effective than others for higher BL capacitance. Figure 4.6 shows the column PDP as function of BL capacitance which shows drastic reduction with respect to the conventional and referenced schemes e.g. for 60fF BL capacitance the PDP of proposed circuit is 0.79fJ while the conventional and referenced circuits PDPs are 2.24fJ and 21.3fJ respectively.

It is worth mentioning that the proposed scheme requires 8 and 10 additional transistors with respect to conventional and referenced circuit which is negligible when considering per column transistor count. Also, as it was mentioned in Chapter 3, BL capacitance in [33] should be kept low. Hence, in order to have the same memory capacity, we need to increase the number of columns in [33], resulting in larger area overhead than the proposed circuit. Furthermore, designing for the same speed and power

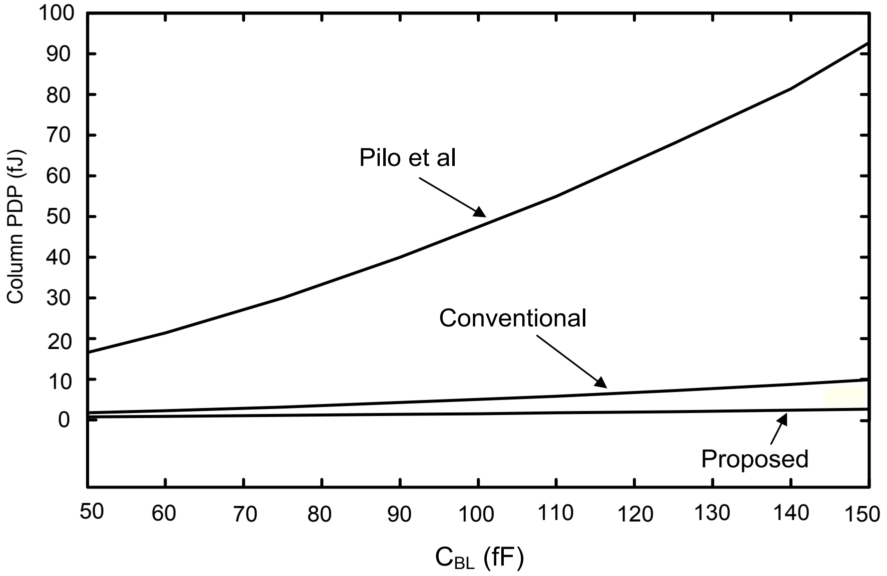


Figure 4.6 PDP comparison

consumption will significantly reduce larger area overhead. Table 4.1 summarizes the comparison.

Table 4.1 Comparisons of various sensing schemes at $V_{DD}$=1V, Freq=500MHz and $C_{BL}$=60fF

| Scheme | Delay (pS) | Column Power (μW) | Column PDP (fJ) | Transistor Count |
|---|---|---|---|---|
| Proposed | 216.5 | 3.65 | 0.79 | 15 |
| Conventional | 484.8 | 4.64 | 2.24 | 7 |
| Pilo et al [33] | 637.0 | 33.48 | 21.30 | 5 |

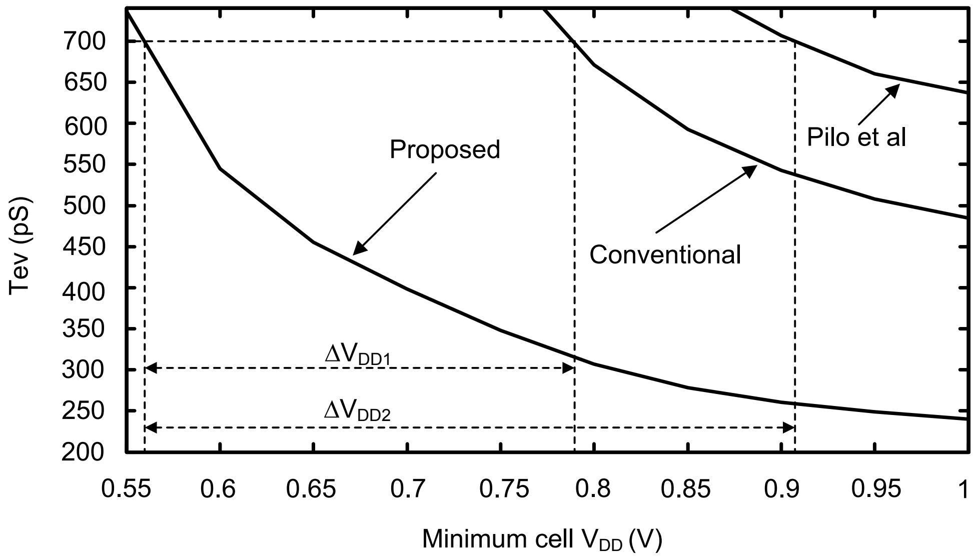


Figure 4.7 Evaluation time as a function of minimum cell $V_{DD}$

## 4.3 Leakage Power Reduction

In order to investigate the functionality of the proposed circuit for reduced cell current, we lowered the cell $V_{DD}$ and measured the evaluation time ($t_{ev}$) to get the final result at the output of SA. Figure 4.7 shows $t_{ev}$ as a function of minimum cell $V_{DD}$. For example, to have $t_{ev}$ of 700 pS the cell $V_{DD}$ of the proposed circuit can be reduced to 560 mV, while conventional and reference circuits require 787 mV and 905 mV, respectively. This shows 28.8% and 38.1% reduction in the minimum required cell $V_{DD}$ to obtain the desired frequency of operation with respect to conventional and referenced circuits. This capability is of particular interest because by reducing cell $V_{DD}$, leakage power which is a considerable amount of the total power dissipation in nanoscale SRAMs, can be reduced. Table 4.2 shows cell leakage power at above mentioned different cell supply voltages. Hence, employing the proposed scheme will result in leakage power reduction of 6.49pW and 11.29pW i.e. 19.7% and 30% with respect to conventional and referenced circuit, respectively. Multiplying these numbers by the total number of cells employed in embedded SRAMs results in considerable reduction in the overall power dissipation.

Table 4.2 Cell Leakage Power at Various Supply Voltage

| Cell $V_{DD}$ (mV) | 905 | 787 | 560 |
|---|---|---|---|
| Leakage Power (pW) | 37.6 | 32.8 | 26.31 |

## 4.4 Yield Improvement

As explained in Chapter 1, all the circuits on the same wafer are tested after the processing sequence and before breaking up the wafer into individual dices. The percentage of the circuits that satisfy the expected specification at this point is called the wafer-sort yield and is usually in the range of 10% to 90% [14].

In order to investigate the effect of proposed circuit on some aspects of the yield, we inserted a 35mV offset to SA of both the proposed scheme and the conventional one and monitored the SA output waveforms. Figure 4.8 shows the simulated waveforms. In order to have the operating frequency of 1.25GHz, the same WL and SAE signal is applied to both schemes. As it can be seen from the figures, the final SA outputs in the proposed scheme toggle in the right direction whereas the conventional circuit outputs toggle in the wrong direction. This observation points to the fact that the proposed scheme is more effective in overcoming some of the non idealities that may arise during the fabrication process and thus results in increased yield. It is important to note that this analysis is primitive and it is only qualitative.

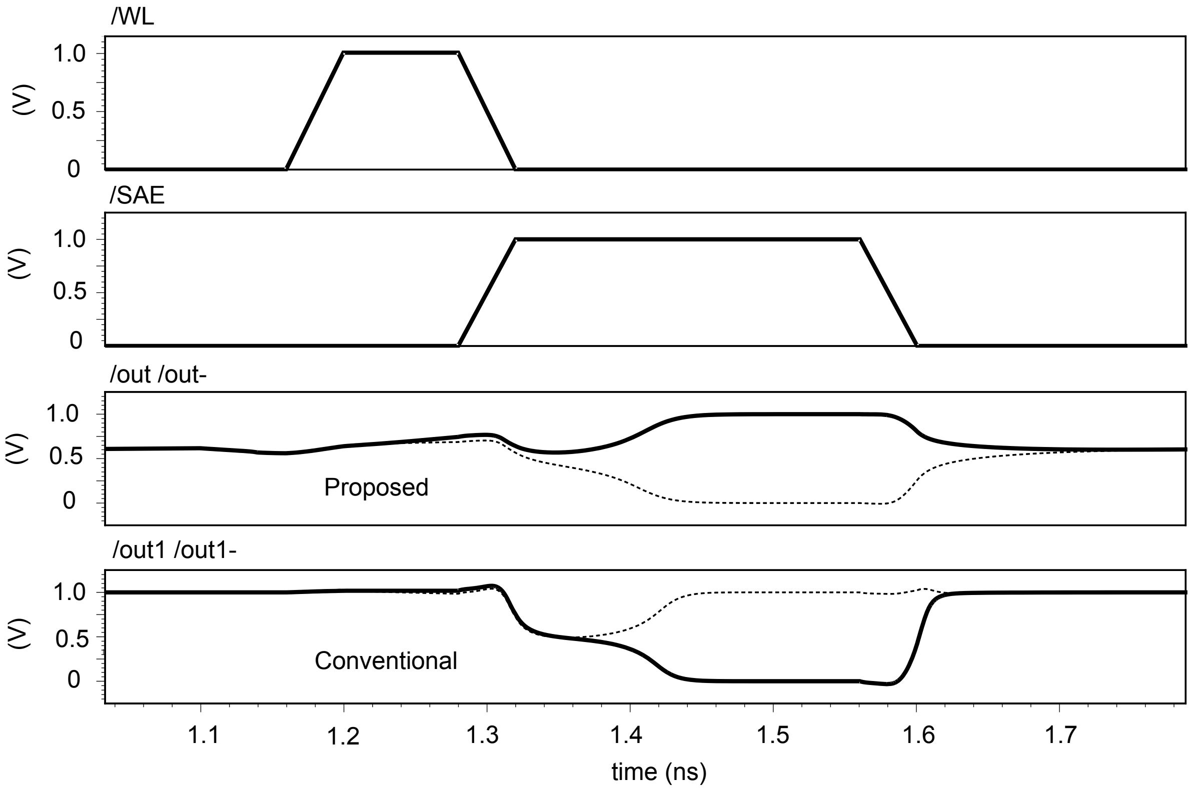


Figure 4.8 Qualitative yield comparison

## 4.5 PVT Simulation Results

In order to make sure that the proposed scheme is able to function properly in the presence of process, supply voltage and temperature (PVT) variations, a complete PVT simulation has been performed taking these effects into consideration.

Table 4.3 shows the sensing delay and power consumption at different process corners. The proposed scheme is able to operate properly at different process corners. As it was expected, at FF corner the scheme has the least sensing delay and the most power consumption while at SS corner the circuit exhibits the most sensing delay and the least power consumption.

Table 4.3 Simulation results at process corners

| Corner | Fast_best (FF) | Typical (TT) | Worst_slow (SS) |
|---|---|---|---|
| Sensing Delay (pS) | 163 | 216 | 225 |
| Power (μW) | 5.69 | 3.65 | 2.61 |

Table 4.4 shows the simulation results with variation in the supply voltage. In order to check the functionality of circuit with variation in supply voltage, the simulations are performed at ±%10$V_{DD}$ i.e. at 900mV and 1.1V. As it is obvious from the table, the scheme has less sensing delay and more power consumption at higher supply voltages.

Table 4.4 Simulation results of voltage variation

| Supply Voltage (V) | 0.9 | 1 | 1.1 |
|---|---|---|---|
| $T_{ev}$ (pS) | 333.1 | 216.5 | 185.4 |
| Power (μW) | 3.37 | 3.65 | 4.5 |

Figure 4.9 shows the simulation results with temperature variation. As it can be seen from the figure, the proposed scheme is able to operate properly from -25ºC to 125ºC. The circuit exhibits more power consumption and less delay at higher temperatures which is due to the increase in the drain current of transistors.

## 4.6 Post-Layout Simulation Result

In order to obtain more realistic results and see the effects of parasitic on the performance of proposed circuit, we laid out the proposed circuit in TSMC 65nm

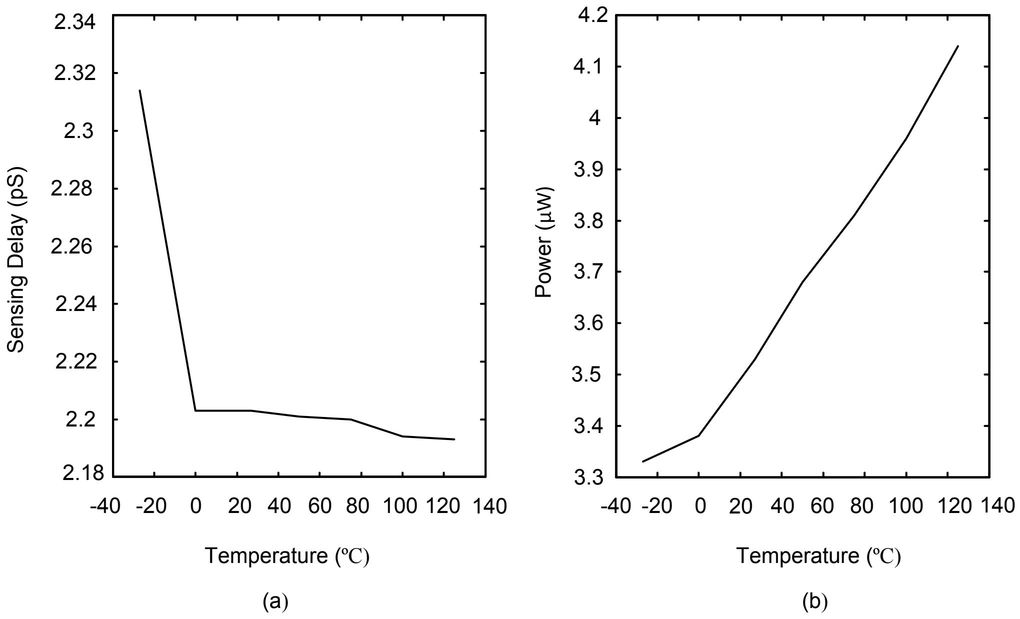


Figure 4.9 Simulation results of temperature variation

technology. Figure 4.10 shows the layout of proposed circuit. Next, we performed the LVS (layout versus schematic) in order to make sure that the laid out circuit is completely matched with schematic of designed circuit. Then, the laid out circuit was extracted in order to do post-layout simulation. Table 4.5 compares schematic and post-layout simulation results. As it can be seen from the table, the sensing delay of post-layout simulation has increased which is due to the extracted parasitic capacitances, resistances and the reduction in the gain of pre-amplifiers and SA. Also, the power consumption of post-layout circuit has increased. In my opinion, this is due to increase dynamic power dissipation i.e. the power for charging and discharging different nodes, which comes as a result of extracted parasitic capacitances.

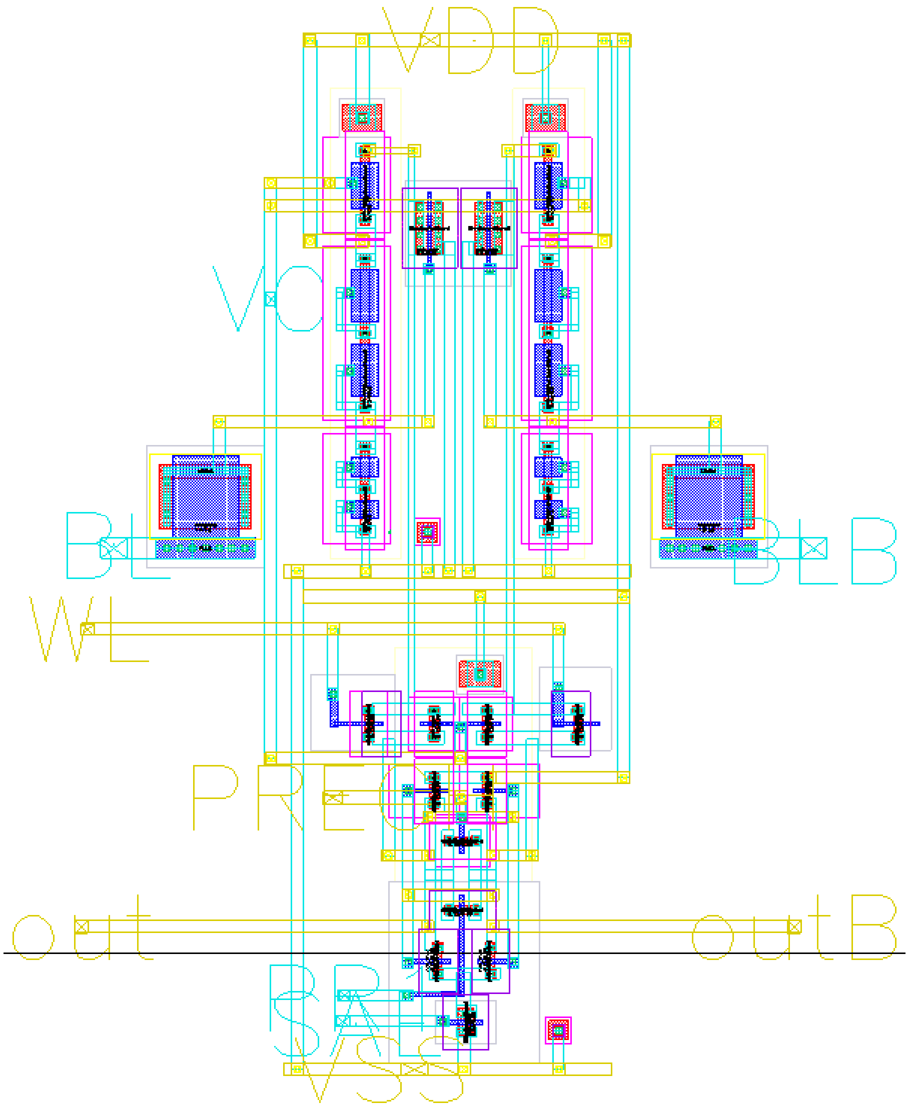


Figure 4.10 Layout of proposed circuit

Table 4.5 Schematic and post-layout simulation comparison at $C_{BL}$=60fF, $V_{DD}$=1V and Freq=500MHz

| | Sensing Delay (pS) | Power Consumption (µW) | RA gain (V/V) |
|---|---|---|---|
| Schematic | 216.5 | 3.65 | 11.6 |
| Post-Layout | 410 | 4.887 | 5.8 |

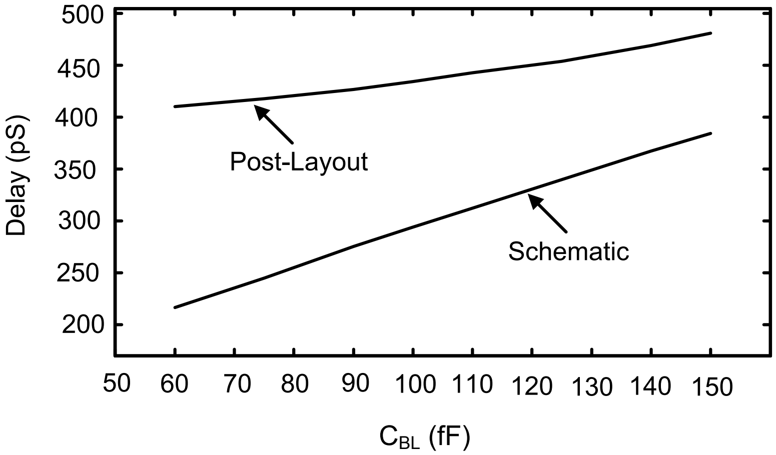


Figure 4.11 Delay comparison of post-layout and schematic simulation

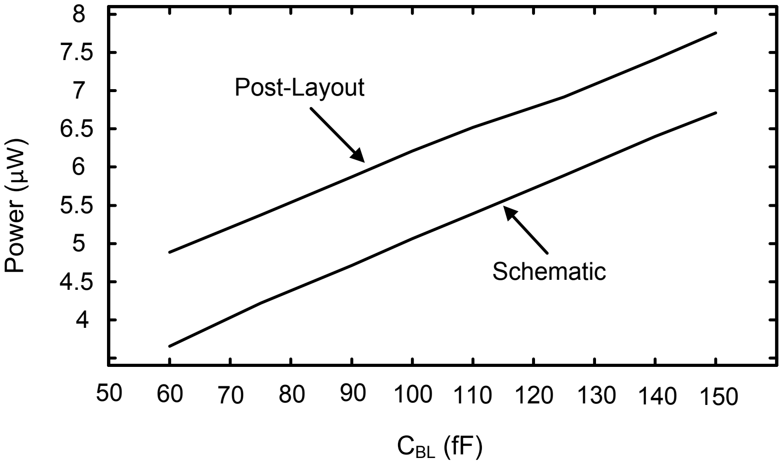


Figure 4.12 Power comparison of post-layout and schematic simulation

Figure 4.11 and 4.12 shows the delay and power comparison of schematic and post-layout simulation as a function of Capacitance of the Bit Line,$C_{BL}$.

## 4.7 Simulation Results of a Column of SRAM

A column of SRAM is designed and simulated using the proposed sensing scheme and the conventional SA and comparison is made between power consumption and access time of these two architectures. The simulated column of SRAM is composed of the following blocks:

- Precharge circuit
- 7-bit row-decoder
- 128 SRAM cell
- Sensing circuit
- Input/Output buffers

Table 4.6 shows the power consumption of individual blocks in these two architectures as well as the total power dissipation. Also, the table shows the access time for having the final data at the outputs. It is worth mentioning that the access time is measured from the transition of signal at the inputs to the time that the final results are available at the outputs.

As it can be seen from the table, by employing the proposed sensing scheme, the access time has been decreased by 44.4% and the total power consumption of a column reduced by 6.7% with respect to the conventional scheme.

Table 4.6 Simulation results of a complete column of SRAM

| Scheme | Proposed | Conventional | % Improvement |
|---|---|---|---|
| Precharge Circuit Power (μW) | 1.882 | 4.658 | + 59.5 |
| Decoder Power (μW) | 4.591 | 4.606 | ≈ 0 |
| Cells Power (nW) | 2.574 | 2.575 | ≈ 0 |
| Sensing Circuit Power (μW) | 2.017 | 0.171 | - 93.6 |
| Buffers Power (μW) | 2.654 | 2.51 | - 5.4 |
| Total Power (μW) | 11.146 | 11.947 | + 6.7 |
| Access Time (pS) | 345.2 | 621.7 | + 44.4 |

# Chapter 5

# Conclusion and Future Work

Semiconductor technology scaling resulted in a considerable reduction in the transistor cost and an astonishing enhancement in the performance of VLSI systems. However, continuous decrease in the size of transistors has unveiled several challenges to the design of low power and high performance SRAMs. Since SRAM employs the smallest transistors and occupies the majority of the die area, it demands special treatments and techniques such as read-assist techniques in order to alleviate the design issues in nanoscale CMOS technology. In this research, we explained SRAM design challenges in nanoscale CMOS technology such as increased process variation, variation in cell current, increased BL capacitance and increased leakage current. Particularly, the effect of these challenges in the design of sensing circuitry has been studied and existing read-assist techniques along with their advantages and disadvantages are discussed. Finally, a novel read-assist technique is presented that address these challenges in the design of nanoscale CMOS SRAMs. The proposed scheme provides amplification for BL voltage droop leading to higher speed while reducing BL voltage swing which results in considerable reduction in power dissipation. Furthermore, it enhances SRAM cell stability by reducing activation time of WL signal.

A complete comparison between proposed scheme, conventional circuit and another state of the art design [33] in TSMC 65nm CMOS technology shows the superiority of the proposed scheme with respect to these schemes in terms of sensing speed and power dissipation. At clock frequency of 500MHz, $V_{DD}$ = 1V and $C_{BL}$ = 60fF the proposed scheme exhibit speed improvement of 55.34% , 66.01% and power reduction of 21.33%, 89.09% with respect to conventional circuit and referenced scheme, respectively. Also, the proposed scheme is capable of working with reduced cell $V_{DD}$ for having the same sensing speed. For instance, in order to have evaluation time of 700pS, the cell $V_{DD}$ of proposed circuit can be reduced to 560mV, while conventional and referenced circuit require 787 mV and 905 mV, respectively. This capability is of particular interest because by reducing cell $V_{DD}$, leakage power which consist a considerable amount of total power dissipation in nanoscale SRAMs, can be reduced by 19.7% and 30% with respect to conventional circuit and referenced scheme, respectively. Furthermore, as it was shown in Chapter 4, the proposed scheme increases some aspects of yield with respect to conventional scheme. It is worth mentioning that the above mentioned achievements comes at the expense of more number of transistors which is negligible when considering the total number of transistors per column. Furthermore, designing for the same speed and power consumption will significantly counteract the larger area overhead.

As we continue to move toward exploiting nanometer scale transistors, we have to deal with negative aspects of this reduction in feature size in SRAMs. Hence, the future research on proposing more novel read-assist circuits and techniques is very crucial in order to insure fast, low power and reliable read operation. There are two possible way to achieve this goal. One way of achieving this goal is to try to modify SA circuit i.e. merging read-assist and SA together, similar to what have been done in [36]. The other way is to separate the read-assist circuit and SA i.e. inserting another stage like what has been proposed in this research. In my opinion, the second method has a large capacity and potential to be the subject of future research works.

# Appendix A

# Primitive Version of Designed Circuit

The proposed circuit in Section 4.1, comes as a result of modification that we applied to primitive designed circuit (Figure A.1). The main idea behind using this circuit was to provide additional discharge path for BL capacitance during read operation in order to reach sooner to the required differential voltage which is necessary before enabling SA and to help weak cells that suffer from having reduced read current. It consists of two pre-amplifiers N1a, P3a and N1b, P3b that amplify the voltage droop on the BLs and two pairs of transistors N2a, N3a and N2b, N3b that provide additional discharge path from BLs to ground. The amplified BL droop at the drain of N1a (N1b) is applied to the gate of N3a (N3b) in order to turn on the read-assist path. Transistors N2a and N2b are employed in order to enable or disable read-assist discharge path. Figure A.2 shows the read waveforms of designed circuit in comparison with conventional circuit (Figure 3.2). As it is obvious from the figure, the designed circuit is successful in achieving the goal i.e. expediting the process of developing differential voltage on the BLs (in this case 150mV). However, later on we reached to this conclusion that the circuit can be modified by removing the extra discharge path (transistors N2a, N2b, N3a, N3b) and applying the amplified voltage at the drain of N1a and N1b directly to the

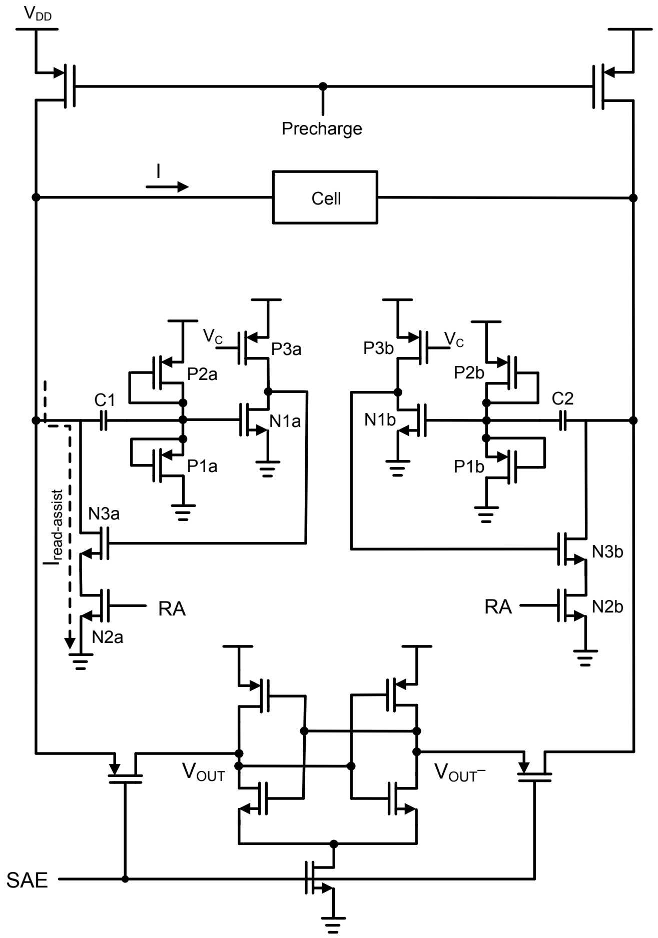


Figure A.1 Primitive version of proposed scheme

inputs of SA. Applying this modification to the circuit has the following advantages. First, since we perform amplification on BL voltage droop, we reach much sooner to the required differential voltage which is necessary before enabling SA. Second, since we are applying the amplified voltage at the output of preamplifier directly to the SA, it enables us to reduce the voltage swing on the BLs which results in considerable reduction in power dissipation due to the large capacitive load of BLs. Also, it enables us to reduce the pulse width of WL signal which results in increased DNM [7].

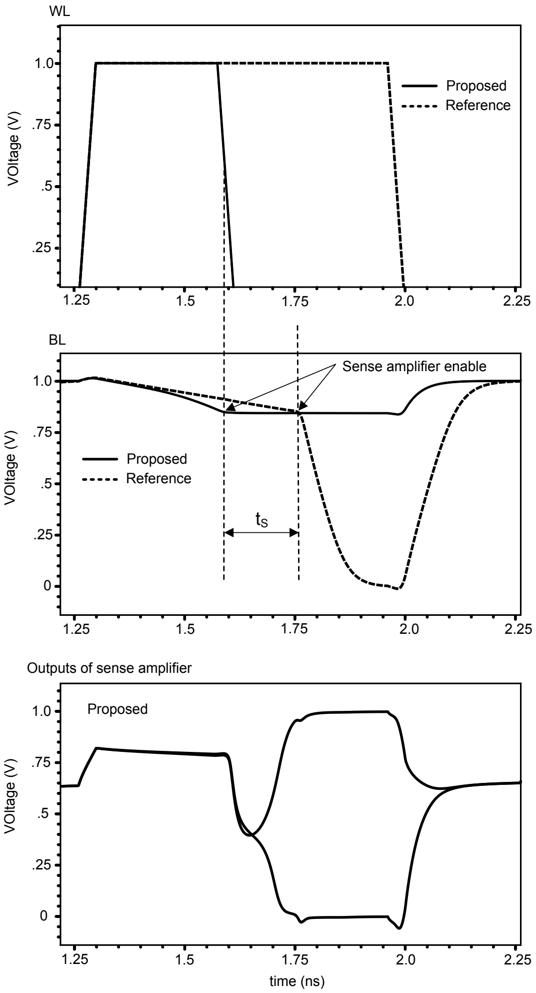


Figure A.2 Power comparison of post-layout and schematic simulation

# References

[1] G. E. Moore, "Cramming more components onto integrated circuits". in *Proceedings of the IEEE ,* pp. 82-85, Jan. 1998.

[2] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, Digital Integrated Circuits – A Design Perspective. Upper Saddle River, New Jersey: Prentice Hall, 2002.

[3] Nassif, Sani R. "Design for Variability in DSM Technologies," *ISQED* , pp. 451-454, 2000.

[4] A. Datta *et al.,* "Speed binning aware design methodology to improve profit under process variations". *ASPDAC*, pp. 712-717, 2006.

[5] R. K. Krishnarnurthy, A. Alvandpour, V. De, and S. Borkar, "High-performance and low-power challenges for sub-70 nm microprocessor circuits," in *Proc. IEEE Custom Integrated Circuit Conf.*, pp. 125–128, 2002.

[6] T. Sakurai, "Perspectives on Power-Aware Electronics," in *International Solid-State Circuits Conference Dig. Tech. Papers,* pp. 26-29, Feb. 2003.

[7] M. Sharifkhani, and M. Sachdev, "SRAM cell data stability: A dynamic perspective," *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 609–619, Feb. 2009.

[8] M. Khellah, Y. Ye, S. K. Nam, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C.Webb, and V. De, "Wordline and bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65 nm CMOS designs," in *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, pp. 9–10, 2006.

[9] Y. Wang et al., "A 1.1 GHz 12 μA/Mb-Leakage SRAM Design in 65 nm Ultra-Low-Power CMOS Technology With Integrated Leakage Reduction for Mobile Applications", *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 172-179, Jan. 2008.

[10] Y.-C. Lai, S.-Y. Huang, and H.-J. Hsu, "Resilient Self-VDD-Tuning Scheme with Speed Margining for Low-Power SRAM", *IEEE Journal of Solid-State Circuits*, Vol. 44, No. 10, pp. 2817-2823, Oct. 2009.

[11] Y.-C. Lai, and S.-Y. Huang, "X-Calibration: A Technique for Combating Excessive Bitline Leakage Current in Nanometer SRAM Designs", *IEEE Journal of Solid-State Circuits,* Vol. 43, No. 9, pp. 1964-1971, Sept. 2008.

[12] M. Goudarzi, and T. Ishihara, "SRAM Leakage Reduction by Row/Column Redundancy Under Random Within-Die Delay Variation", *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, Vol. 18, No. 12, pp. 1660–1671, Dec. 2010.

[13] T. Kawahara, “Low-voltage embedded RAMs in the nanometer era,” in *2005 IEEE Conf. Electron Devices and Solid-State Circuits*, pp. 333–338, Dec. 2005.

[14] P. R. Gray, P. Hurst, S. Lewis, and R. G. Meyer, Analysis and Design of Analog Integrated Circuits, 4th ed., New York, NY: Wiley, 2001.

[15] A. Bhavnagarwala *et al.*, “The impact of intrinsic device fluctuations on CMOS SRAM cell stability,” *IEEE J. Solid-State Circuits*, Vol. 36, No. 4, pp. 658–665, Apr. 2001.

[16] S. Mukhopadhyay *et al.*, “Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS,” *IEEE Trans. Comput.-Aided Des.*, Vol. 24, No. 12, pp. 1859–1880, Dec. 2005.

[17] E. Seevinck, F. J. List, and J. Lohstroh, “Static-noise margin analysis of MOS SRAM cells,” *IEEE J. Solid-State Circuits*, Vol. 22, No. 5, pp. 748–754, Oct. 1987.

[18] G. Chen, D. Sylvester, D. Blaauw, and T. Mudge “Yield-Driven Near-Threshold SRAM Design,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, Vol. 18, No. 11, pp. 1590–1598, Nov. 2010.

[19] S. Schuster et al., “A 15-ns CMOS 64k RAM,” *IEEE J. Solid-State Circuits*, Vol. SC-21, No. 5, pp. 704–711, Oct. 1986.

[20] J. Lohstroh, E. Seevinck, and J. Groot, “Worst-case noise margin criteria for logic circuits and their mathematical equivalence,” *IEEE J. Solid State Circuits,* Vol. SC-18, No. 6, pp. 803–806, Dec. 1983.

[21] M. Sharifkhani, E. Rahiminejad, S. Jahinuzzaman, and M. Sachdev, "A Compact Hybrid Current/Voltage Sense Amplifier With Offset Cancellation for High-speed SRAMs", *IEEE Trans. Very Large Scale Integr. (VLSI) Syst*, Vol. 19, No. 5, pp.883-894, May 2011.

[22] M. Bhargava, MP. McCartney, A. Hoefler, and K. Mai, "Low-Overhead, Digital Offset Compensated, SRAM Sense Amplifiers," in *IEEE Custom Integrated Circuits Conference.*, pp. 705-708, 2009.

[23] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, “A novel high-performance and robust sense amplifier using independent gate control in sub- 50-nm double-gate MOSFET,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst*, Vol. 14, No. 2, pp. 183–192, Feb. 2006.

[24] A. Choudhary, and S. Kundu, “A process variation tolerant self-compensating sense amplifier design,” in *Proc. IEEE Comput. Soc. Ann. Symp. VLSI,* pp. 263–267, 2009.

[25] B. Wincht, J. Y. Larguier, and D. S. Landsiedel, “A 1.5 v 1.7 ns 4 k x 32 SRAM with a fully-differential auto-power-down current sense amplifier,” in *Proc. ISSCC*, pp. 462–508, Feb. 2003.

[26] K. S. Yeo, W. L. Goh, Z. H. Kong, Q. X. Zhang, and W. G. Yeo, “High performance low-power current sense amplifier using a cross-coupled current-mirror configuration,” *IEE Proc. Circuits, Devices, and Syst.*, Vol.149, No.56, pp.308–314, Oct. 2002.

[27] A. Christianthopoulos, Y. Moisiadis, Y. Tsiatouhas, and A. Arapoyanni “Comparative study of differential current mode sense amplifier in submicron CMOS technology,” *IEE Proc. Circuits Device Syst*, Vol. 149, No. 3, pp. 154–158, June 2002.

[28] 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 J. Solid-State Circuits*, Vol. 26, No. 4, pp. 525–536, Apr. 1991.

[29] D. Anh-Tuan, K. Zhi-Hui, and Y. Kiat-Seng, “Hybrid-Mode SRAM Sense Amplifiers: New Approach on Transistor Sizing,” *IEEE Trans. Circuits Syst. II, Exp. Briefs*, Vol. 55, No. 10, pp. 986–990, Oct. 2008.

[30] B. Wincht, J. Y. Larguier, and D. S. Landsiedel, “A 1.5 v 1.7 ns 4 k x 32 SRAM with a fully-differential auto-power-down current sense amplifier,” in *Proc. ISSCC*, pp. 462–508, Feb. 2003.

[31] T. N. Blalock, and R. C. Jaeger, “A high-speed clamped bit-line current-mode sense amplifier”, *IEEE **J.** Solid-State Circuits,* Vol. 26**,** No. 4, pp**.** 542-548, Apr. 1991.

[32] K. Sasaki, K. Ishibashi, K. Ueda, K. Komiyaji, T. Yamanaka, N. Hashimoto, H. Toyoshima, F. Kojima, and A. Shimizu, ‘A 7-11s 140-mW 1-Mb CMOS SRAM with current sense amplifier’. ***IEEE** J. Solid-State Circuits,* Vol. 27, No. 11, pp. 1511-1518, Nov. 1992.

[33] H. Pilo, C. Barwin, G. Braceras, and F. Towler, “An SRAM Design in 65nm Technology Node Featuring Read and Write Assist Circuits to Expand Operating Voltage,” *IEEE Journal of Solid State Circuits*, Vol. 42, No. 4, Apr. 2007.

[34] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, “Matching properties of MOS transistors,” *IEEE Journal of Solid State Circuits*, pp. 1433-1439, Vol. 24, No. 5, 1989.

[35] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, Y. Nakase, and H. Shinohara, “A 45nm 0.6V Cross-Point 8T SRAM with Negative Biased Read/Write Assist,” *IEEE Symposium on VLSI Circuits,* pp. 158–159, 2009.

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

[37] M. Abu-Rahma, M. Anis, and S. Yoon, "A Robust Single Supply Voltage SRAM Read Assist Technique Using Selective Precharge," *Proc. of IEEE European Solid-State Circuits Conference (ESSCIRC),* pp. 234-237, 2008.

[38] B. Wicht et al., “Yield and speed optimization of a latch-type voltage sense amplifier,” *IEEE Journal of Solid State Circuits*, Vol. 39, No. 7, pp. 1148–1158, July 2004.

[39] Ali Valaee, and Asim J. Al-Khalili, “SRAM Read-Assist Scheme for High Performance Low Power Applications”, *International SoC Design Conference (ISOCC 2011)*, Jeju, Korea, November, 2011.