Design and Analysis of Sense Amplifier Topologies for Volatile and Non-Volatile Memories

Student Name: Disha Arora

M.Tech-ECE-VLSI AND EMBEDDED SYSTEMS-14-16

June 15 2015

Indraprastha Institute of Information Technology, New Delhi

Advisor

Dr. Mohammad. S. Hashmi

Submitted in Partial fulfilment of the requirements
for the degree of M.Tech in Electronics and Communication Engineering

Student's Declaration

I declare that the dissertation titled "Design and Analysis of Sense Amplifier Topologies for Volatile and Non-Volatile Memories" submitted by Disha Arora for the partial fulfilment of the requirements for the degree of Master of Technology in Electronics and Communication Engineering is carried out by me under the guidance and supervision of Dr. M. S. Hashmi at Indraprastha Institute of Information Technology, Delhi. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... Place and Date: …….....
Disha Arora

CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

……
Dr. Mohammad. S. Hashmi
Indraprastha Institute of Information Technology, Delhi

ABSTRACT

The growing gap between the processor and embedded memory speed is a major setback in the overall performance of electronic systems. Since the sense amplifier (SA) forms an integral part of the read circuitry in both volatile memories, such as SRAM, and non-volatile memories (NVMs), such as FLASH, its performance has a significant effect on the overall performance of memory. Access time, offset, power and area are the four important performance metrics of SA. The memory access time and input-offset of SA greatly affect the speed of the entire memory and therefore to patch up the gap between processor and memory speed, the SA is required to be fast and efficient. As one SA is employed for each bitline in the memory array, it is required to be compact in size and should have low power consumption. Furthermore scaling in technology makes it difficult to control the fabrication process leading to variation in process parameters causing unpredictability in the performance of SAs. Therefore, it is very important to keep this aspect in mind while designing and estimating the performance metrics of the SA.

This thesis includes the study of various conventional SA designs in detail so as to have a better understanding of a basic SA and its operation and thus helping in understanding what problems are faced by designers in implementing the SA designs and how these problems can be tackled. In addition to the conventional SA analysis, new sense amplifier designs have been proposed for both current sensing in FLASH memory and voltage sensing in SRAM. Keeping the variation in process parameters due to scaling in mind, these proposed designs have been optimized in terms of access time, offset, power and area.

ACKNOWLEDGEMENTS

The work for this thesis was carried out at Indraprastha Institute of Information Technology (IIIT), Delhi, India, during the year 2015-2016. I would like to express my deep sense of gratitude towards my adviser Dr. Mohammad. S. Hashmi for providing expert guidance and encouragement throughout the journey of this work without which, this work would never have been successful. I also take this opportunity to thank Anil Kumar Gundu, teaching fellow at IIITD for the technical discussions and guidance he provided which helped me overcome the hurdles I encountered. My deepest regards to all my friends at IIIT-D who made this journey wonderful. I would like to thank my Parents for the support they provided me spiritually and emotionally.

LIST OF TABLES

LIST OF FIGURES

CONTENTS

1. INTRODUCTION

1.1 NVM OVERVIEW

In most electronic systems, some parts of information must be permanently stored such that, the information could be retained even when the power supply is off. For example, programmable systems require a set of instructions to boot and those particular instructions are often called “firmware”. These particular set of instructions cannot be lost when the power supply is switches off. The memory used for the purpose of storing permanent data is called Non-Volatile memory (NVM). Solid state NVM are used in a variety of applications apart from firmware. In electronic systems, there is usually some data which is set by the manufacturer, distributors and by the users, and this data must be stored in the system permanently, even when the power is switched off. Thus NVM came into existence and today it forms an integral part of almost any electronic system available in market such as set top boxes, printers, laptops, mobile phones and other hand held devices. Also, with time NVM has become advanced and its capacity to hold data has increased manifold. In order to fulfil such a variety of application needs, NVMs are available in a variety of capacities, ranging from a few Kilobytes to Terabytes. With evolution in integration technology, now NVMs are even being embedded into processor chips.

NVM can be basically categorized into electrically addressed systems such as read only memory (ROM) and mechanically addressed systems such as hard disks and optical disks. Electrically addressed systems have a high speed but are expensive whereas mechanically addressed systems are slow but are cost effective. There are various types of NVMs available in market namely ROM, PROM, EPROM, EEPROM and Flash. The Read only Memory or ROM is programmed at the time of manufacturing and its contents can be modified very slowly incurring a lot of difficulty or not at all. The Programmable Read only memory or PROM is manufactured as a blank memory and can be programmed once only using a PROM burner. The erasable programmable Read only memory or EPROM can be erased using Ultra Violet (UV) light through a window which is designed in the memory chip and it can be programed or reused. The EPROMs consist of one transistor per cell and thus are capable of offering a high density memory [1]. The Electrically Erasable Programmable Read only memory or the EEPROM can be erased electrically with a fine granularity of up to a byte and can be reused or programmed. The EEPROMs consist of two transistors per cell. The Flash memory is almost similar to an EEPROM but it does not have a fine granularity of erasure and it can erase a block of memory only. But due to the complex structure of EEPROMs they are not as cost effective when compared to Flash memory which consist of only one transistor per cell and EEPROMs offer lower density of memory when compared to EPROMs. Another performance metric to judge the performance of NVMs is the endurance. Endurance signifies the number of cycles of erasure and programming that a memory device can endure. Even though EEPROMs offer a very high endurance of about one million erase/program cycles, their use is limited only to specific applications due to their low density and high cost. Fig. 1 compares the performance

of various NVMs available in the market in terms of the two main performance metrics namely cost and flexibility.

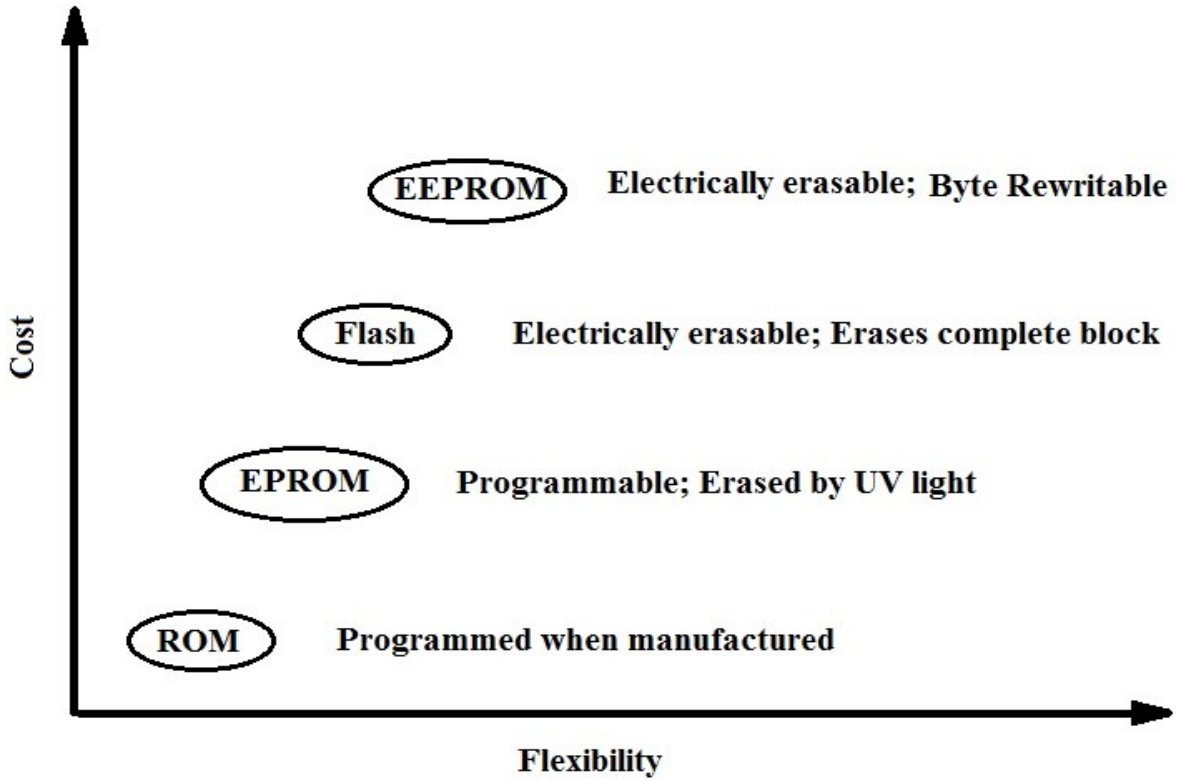

Figure 1. Comparison of various Non-Volatile memories in terms of flexibility and cost

The world has witnessed an impressive growth in the NVM market in the past decade. As the technology has evolved over the years, every electronic system has become advanced and thus requires memory which can perform faster, is more cost effective and consumes less power as compare to its previous counterparts. This impressive growth is mainly due to the development of personal portable electronic appliances and other hand held devices. Electronic systems such as personal digital assistants (PDAs) and mobile phones memories that are smaller in size and consume less power, therefore a demand for mass storage arose. This requirement is mostly fulfilled by Flash memories. Fig. 2. Displays the improvement of the market share of NVM in the semiconductor memory industry. Due to the development of multimedia applications and the movement of personal consumer appliances towards being portable and being able to manage data, images, music and communication has increased the demand for storage. As a result today, memory cards are available in different formats with data capacity ranging from a few KB to GBs. All these factors have led to the growth of NVM market.

Flash Memory holds a significant share of the NVM market. The concept of Flash memory arose due to the requirement of a dense and closely packed memory that can hold a significant amount of data but at the same time the memory is required to be cost effective, should not occupy a large space, should have good endurance and should be flexible enough so as fulfil the constantly evolving demands of the electronics systems we have today. Therefore the last decade has seen a significant increase in the demands of Flash memory.

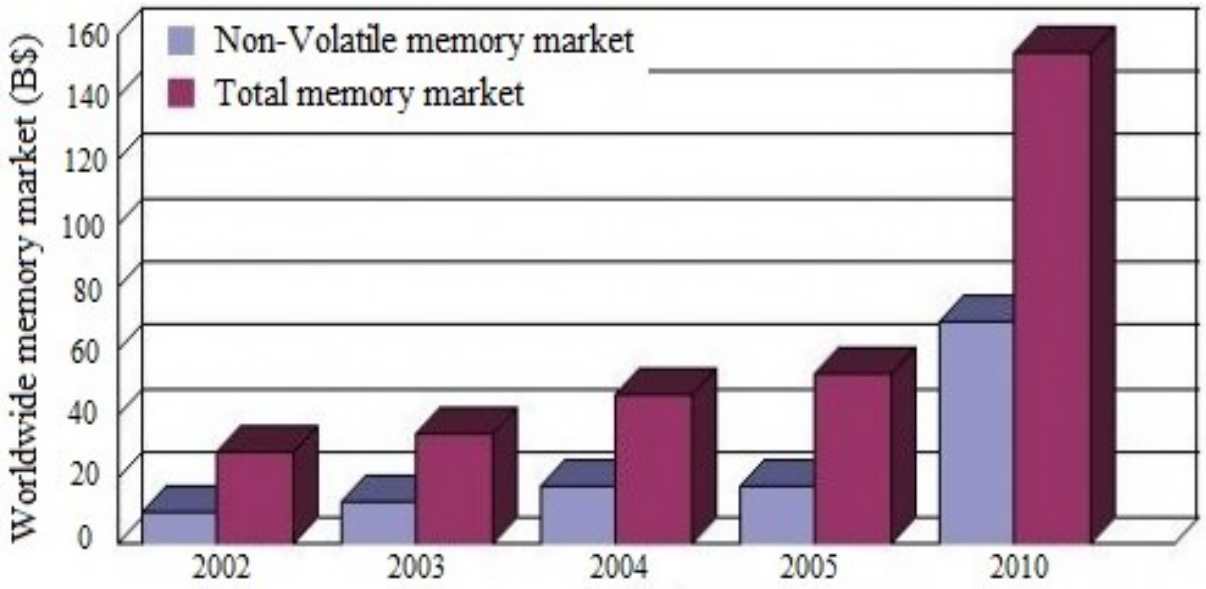

Figure 2. The market share of Non-Volatile memory in the worldwide memory market in Billion $

1.1.1 Types of Flash memory

NOR Flash

The memory cells in a NOR Flash are arranged in a NOR type organization as shown in Fig. 3. The cells are parallel connected with common ground and the bit lines are directly connected with the drains of memory cells. A NOR memory cell consists of one transistor by stacking a double poly floating gate MOSFET. It can be programmed using channel hot electron (CHE) injection and erased using Fowler Nordheim (FN) tunnelling effect [2]. If we do not consider cosmic rays and their effects, charge storage on a floating gate is a very reliable mechanism to store data permanently. Due to the very high (3.2 eV) energy barrier that the electrons need to overcome in order to escape from the floating gate, this mechanism to store data permanently proves to be efficient. The CHE programming technique makes the system immune to program disturbs and scaling tunnel oxide is not required in order to reduce the channel length of the cell, thus allowing a good data retention capability. The NOR Flash proves to be best in case of requirement of high speed and noise immunity, due to the presence of direct access to the memory cell. CHE programming along with the advantages of NOR memory organization make this technology the suitable for multilevel storage of data, which boosts the density making it suitable for cost sensitive applications. NOR memories store two bits in each cell, thus providing high density of data in the overall memory which allows 30-40% saving in cost when compared to memories that store one bit per cell of the same capacity. NOR Flashes have been proven to be compatible with many advanced logic processes making it suitable to be used for embedded memory in system on a chips.

NAND Flash

NAND Flash memory basically has similar memory structure as NOR, but its array is organized in a different way as shown in Fig. 3. The array is organized in NAND arrangement, such that

16 or 32 cells are all connected in series between the ground node and the bit line node. In this way the density of this particular memory array organization is dense when compared with the NOR Flash. Even though the NAND Flash needs a ground line and a bit line contact between every other cell, a dramatic effect is seen in terms of speed, thus slowing it down. Therefore, in order to read a particular cell, 15 or 31 other cells are also read. Thus strongly reducing read current which results in a higher access time (microseconds as compared to nanoseconds in NOR Flash). Furthermore the read through mechanism introduces a lot of noise thus, posing difficulty in usage of NAND Flash in multilevel storage. Even though NOR Flash stores two bits per cell, NAND Flash still stores one bit per cell. The programming mechanism used by NAND Flash is Fowler Northeim tunnelling effect which is less reliable when compared with CHE because it needs a thinner tunnel oxide. But NAND Flash covers up for its reliability with the help of error correction techniques. On the other hand, since FN tunnelling effect is a programming mechanism that requires low current, it allows a very high degree of parallelism for programming and a very high writing throughput. Due to these being the key features for mass storage, NAND Flashes have become very popular. The higher density and the greater programming throughput make NAND the significant technology for memory cards.

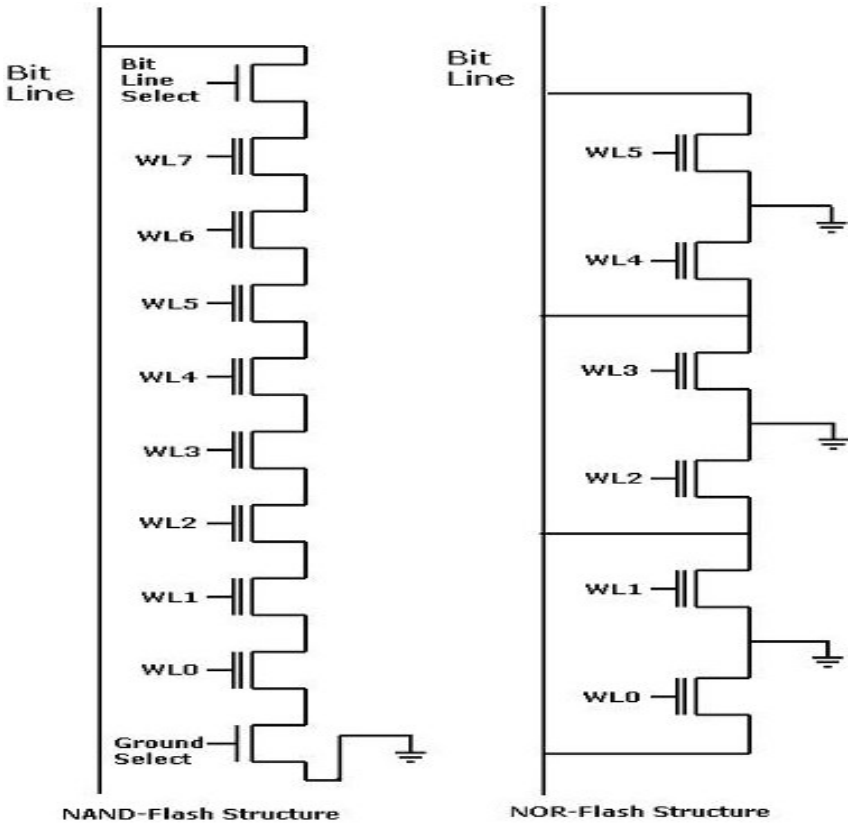

Figure 3. Memory array organization of NOR and NAND Flash memory

1.1.2 Basics of Flash memory cell

The memory cell needs to have a mechanism in order to store data permanently and alter its contents electrically in a non-destructive way. The solution is to alter the threshold voltage of

the cell so that different threshold values may represent different states of the memory. The two basic states of a flash memory cell are called erased and programmed states. An erased cell is signified by a low threshold value whereas a high threshold value signifies a programmed cell. Equation 1.1 expresses the relation between the threshold voltage of MOS with the charge stored on the floating gate:

$$V_t = \mathrm{K} - \frac{\mathrm{Q}}{C_{ox}} \tag{1.1}$$

Where K denotes a constant which depends on gate and substrate material, channel doping and oxide thickness. C_{ox} denotes the gate oxide thickness and Q is the charge trapped into the oxide layer. From the equation (1.1) it is clear that the parameter which can be kept in control to alter threshold of the device is Q which denotes the charge trapped in the oxide layer. There are charge injection techniques available to move charges in and out of the oxide. A normal MOS device cannot be used to retain the charges into its oxide thus the device has been modified. A floating gate (FG) device is used for this purpose. FG transistors have the capability to retain charge in their floating gate for an extended period even after the power supply is turned off. The structure of a floating gate device is explained with the help of a cross section of the device as shown in Fig.4.

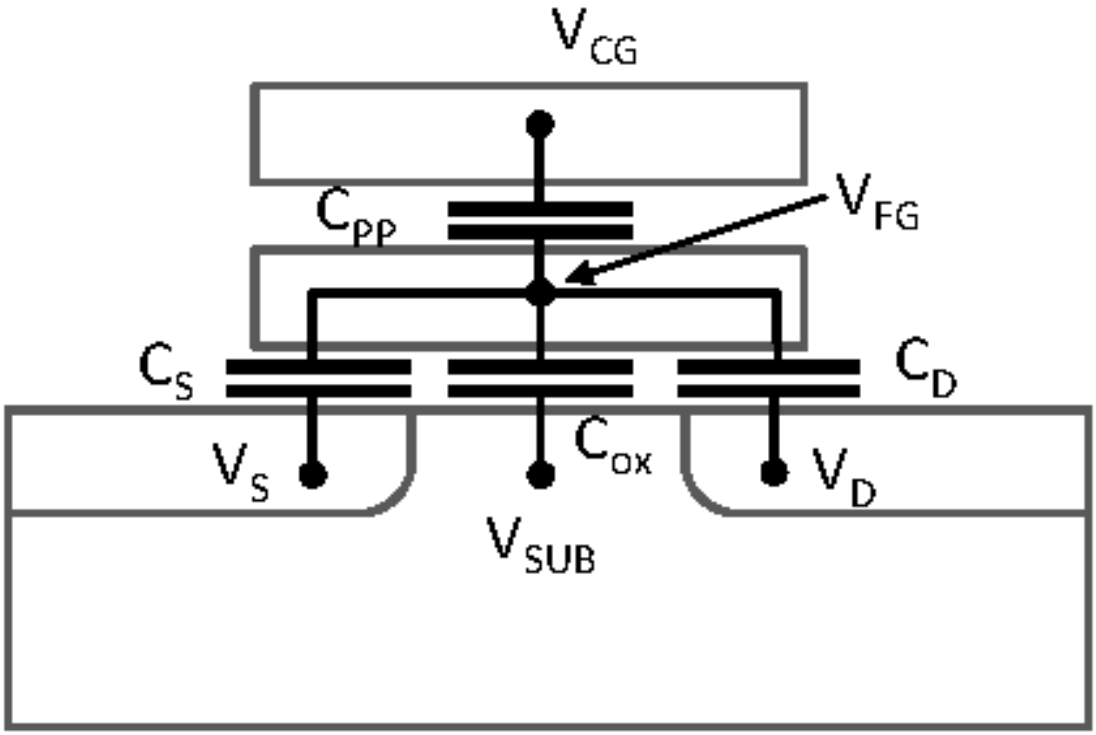

Figure 4. Cross section view of a Floating Gate device

1.1.3 Operations in a Flash memory cell

Program Operation

During the Write/Program operation, the control gate and drain are biased at high voltage of 12V for the gate and 5 V for the drain (the voltages used for biasing are used as convention and may vary for different manufacturers), but the source is kept grounded. Under these circumstances, a very strong electric field develops which lets the electrons pass from the channel to the floating gate. These electrons overcome the potential barrier posed by the oxide layer and this mechanism is called Hot Electrons Injection [2].

Due to the presence of a high voltage on the drain node, the electrons flowing from the source to the drain gain energy due to the orthogonal electric field. Due to the presence of high electric fields, electron energy starts to increase and thus electrons are heated, some electrons gain energy high enough to overcome the barrier between the oxide layer and the silicon conduction band. These hot electrons need to overcome the barrier in the right direction so as to be collected inside the floating gate. The electrons trapped inside the floating gate causing the V_{TH} of the flash memory cell to rise. Thus, when a Read operation occurs, the cell appears to be in the switched off state or is logic programmed '0', since it is unable to conduct current due to its high V_{TH}. Thus writing data in a memory cell brings the cell from an erased state, which is typically called a logic state '1', to a logic state '0' or programmed state. The time required for this process is typically in the range of microseconds. The program operation in a Flash memory cell is explained with the help of Fig. 5 (a).

Erase Operation

In Flash memories, a positive voltage is applied between the source and the control gate by means of grounding the control gate and increasing the source node voltage to 12 V or by lowering the control gate voltage to –8 V and by raising the source node voltage to about 5 V. The first method described above is called the positive source erase operation, whereas the second method described above refers to the negative gate erase operation. The drain terminal, in both these cases, is left floating.

The erase operation utilizes the mechanism known as Fowler-Nordheim (FN) tunneling. The time required for this process is in the range of a few hundred of milliseconds. The erase operation is explained with the help of Fig. 5 (b). For the basic operations of a Flash memory cell described above Table I shows the node voltages for various operations [1]. The shift in VTH values of the memory cell with operations such as program and erase are shown in Fig.6.

TABLE I. Node Voltages of Flash memory cell for Program and Erase operations

Operation	Gate (V)	Drain (V)	Source (V)	Substrate (V)
Program	8	5	0	0
Erase	-8	Floating	8	8

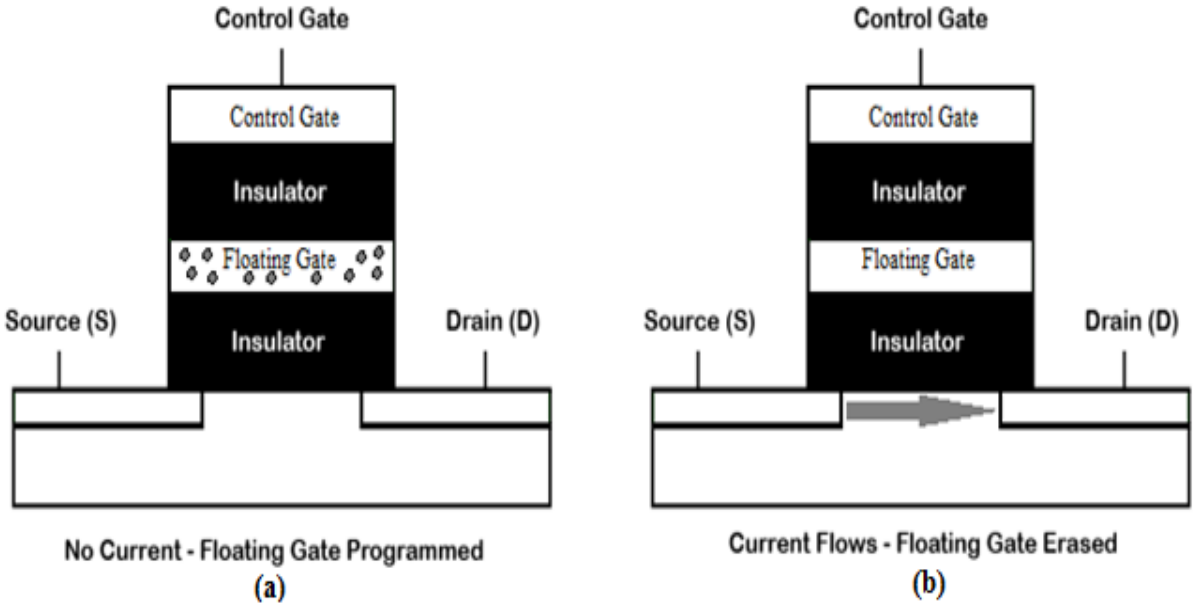

Figure 5 (a) Program operation in Flash memory cell. (b) Erase operation in Flash memory cell

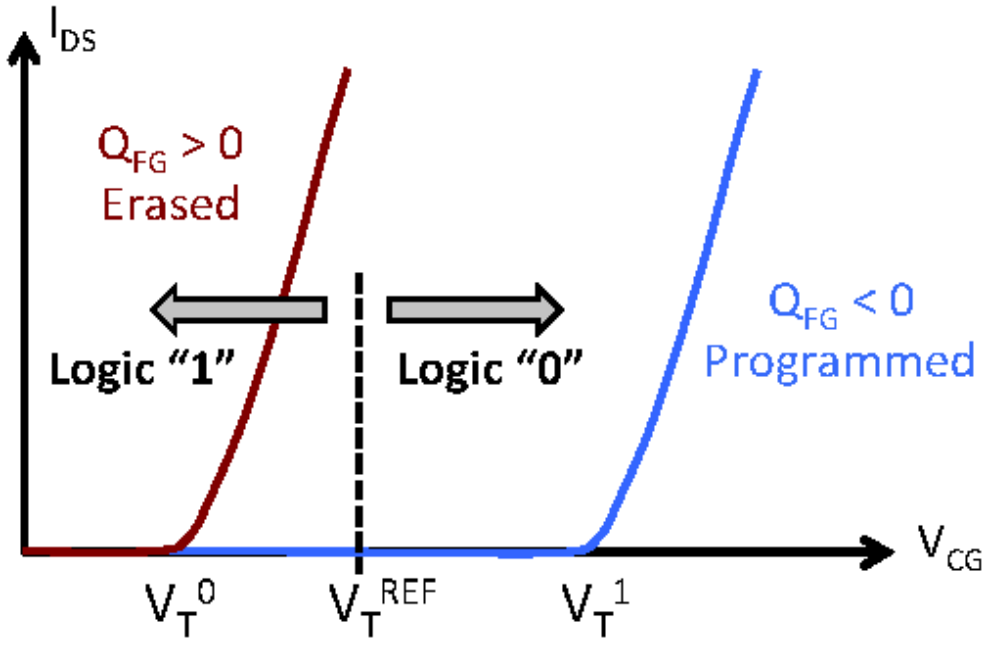

Figure 6. Shifting of Threshold voltage of the memory cell with Program and Erase operations

1.1.4 Test Modes in Flash memory

Apart from the user modes in Flash memory like read, program and erase test modes are included in order to analyse the behaviour of memory. These test modes are not for the use of the end user but are important from the manufacturer point of view. These test modes are used for the purpose of characterization of the memory cell array and reference cell array, to observe their silicon behaviour and to test other blocks of memory. The two main test modes are Direct Memory Access (DMA) and Fast Direct Memory Access (FDMA).

Direct Memory Access (DMA)

DMA test mode is used for the purpose of connecting the cell terminals directly to the external Input/output pads. This helps in characterization of the memory, the matrix and the reference cells in particular. It is a difficult task to filter the interference of the memory array. Many incorrect outputs can be obtained to the faults in circuitry if any. For example, if the voltages are applied in the wrong way or if there is a presence of any voltage spikes or glitches. The possibility of analysing each and every cell is therefore a valuable opportunity. Even a single cell can be analysed with the help of DMA thus proving to be a major test mode. The DMA test mode setup is shown in Fig. 7. It can be observed from Fig. 7. that in DMA the sense amplifier and the output latch are bypassed such that the drain node of the cell is directly connected to the external I/O pad which is further connected to the external supply. The supply voltage of this external supply is equal to that on the drain in case of a read operation. Also, the gate voltage supplied to selected cells in DMA mode is supplied through an external pin. This setup enables the manufacturer to measure cell current, transconductance and V_{TH} of cells under varying conditions.

Fast Direct Memory Access (FDMA)

Operating in DMA mode so as to measure the cell current at different bias voltages is a tedious job. Therefore to increase the speed of this procedure, Fast DMA or FDMA was introduced. FDMA mode is similar to the read mode but in FDMA mode a constant reference current is maintained and the cell current is compared with it with the help of a sense amplifier. The reference current could be generated internally or could be generated externally with the help of the DMA pin. The gate voltage could be controlled by an external I/O pad similar to DMA mode. By varying this current and gate voltage the cell characteristics are plot. FDMA has an advantage over DMA being faster due to the read operation. The advantage in the read operation is that it is very fast because the current to the voltage conversion is carried out by the sense amplifier. For example, if the reference current is 8μA, it is easy to distinguish all the cells absorbing a higher current. By varying the current values it is easily possible to plot a histogram describing the cell distribution. Therefore FDMA test mode helps n analysing the memory cell array in a much faster way as compared to DMA.

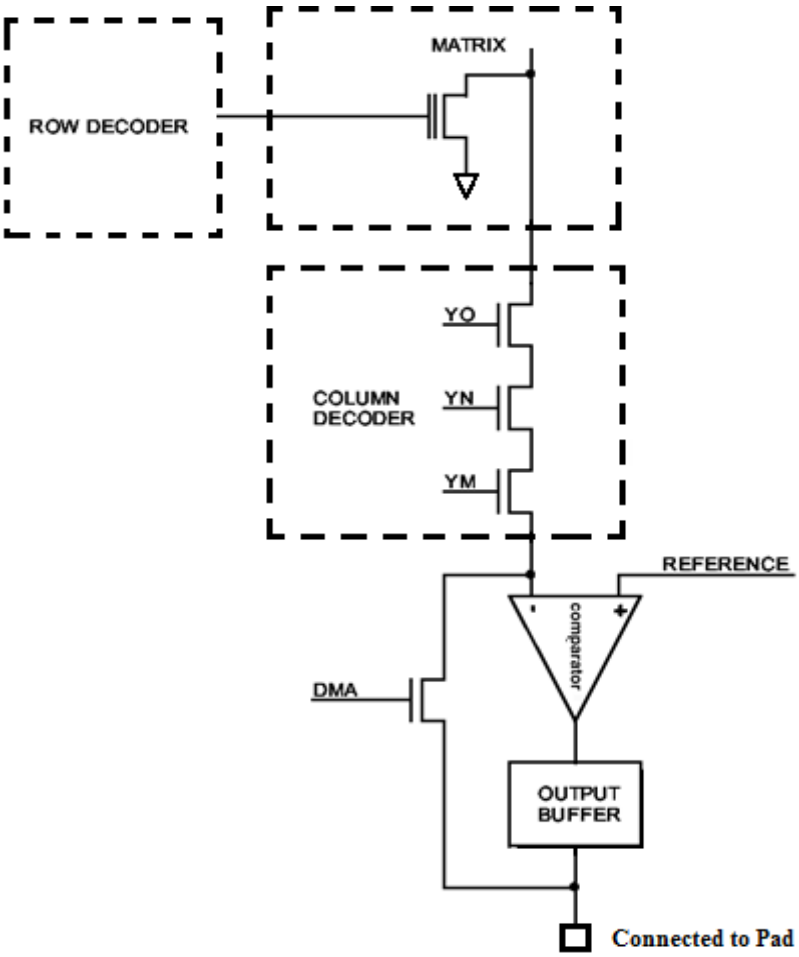

Figure 7. Setup for the DMA test mode

1.2 SRAM OVERVIEW

Static random access memory (SRAM) is a volatile memory which operates as fast as the logic circuits, consuming a very low amount of power in the standby mode. An SRAM is used to provide an interface with the central processing units or the CPU at very high speeds which cannot be attained by DRAM. Thus there is a requirement of a memory which operates consuming very low power and is fast in order to replace the DRAMs. SRAMs are used as caches and to interface between DRAMs and CPU. These features provided by SRAMs are unattainable by any other memories like DRAMs and Flash. The SRAM memory array occupies around 1/4th of the logic circuits in today's time. Therefore the characteristics of any integrated circuit (IC) chip such as operating speed, power consumption, supply voltage, and size is greatly defined by the characteristics of their SRAM memory. A very important driving factor for the development of SRAM technology is its use in low power applications. Also, SRAMs are used in electronic systems that need to be light and portable because the refresh current needed by DRAMs much more than the SRAM standby current thus proving to be heavier and less portable as compared to SRAMs. The access time to access SRAM is almost comparable to DRAM. Therefore, the design of an SRAM cell and array is a major requirement in order to obtain efficient performance, low power operation, low cost, and reliability. A

concept introduced earlier was the high-R cell which was first proposed in the form of low power consuming 4K SRAM. In the high-R cell, there was a high resistivity poly silicon layer which was used as the load of inverter in the SRAM cell. A high-R cell does not require a bulk PMOS thus contributing to the small size of the memory cell, even smaller than the 6T SRAM. The resistivity of the polysilicon layer had a value around 10^{12} thus the standby current for the memory cell was around 10^{-12} per cell which was a very low value. The high-R cell was then extensively used to manufacture high density and SRAM integrated circuits which consumed low power and had a capacity ranging from 4K to 4Mbits. A major drawback of the high-R cell was its low voltage operation, and for supply voltages less than 1.5V the node voltages need to be charged to the supply voltage level required during write operation. As the resistivity of the load polysilicon was high it required a lot of time in order to charge up the high node to the supply voltage. Thus it was difficult to operate the high-R cell at supply voltages less than 1.5V. Then the 6T SRAM cell was introduced but there were a lot of problems in order to obtain low power consuming, reliable and small cell size of the 6T SRAM cell. As the 6T SRAM cell gets scaled with Moore's law, the size of the transistors in 6T SRAM cell is also reduced with Moore's law. Supply voltage of 6T cell is also reduced as the feature size is reduced. Due to scaling in technology, some unpredictability is introduced in the process parameters thus the variation in the threshold voltages of the transistors has increased. Also, the leakage of the transistors has been observed to increase due to the process of scaling. Due to scaling, the supply voltage of the SRAM memory cell has also been reduced. Recently, other low power consuming circuit techniques like Dynamic Voltage Frequency Scaling (DVFS) further require operation at low voltage. The design of the 6T SRAM cell is such that it must be electrically stable at even low supply voltages even though there are large variations in the transistors. Today, due to the need for smaller hand help electronic devices, the memory cell size needs to be as small as possible in order to obtain integrated circuits with a small chip size. Even the leakage in 6T SRAM cells, needs to be very small even though there is large leakage in the transistors of the cell. Also, there should be immunity to soft errors caused by any alpha particles in order to have reliable integrated circuits. In order to reduce operating power of the integrated circuits, the SRAM cell needs to work at a low voltage and should be able to retain data with a very low leakage in the standby mode. Many low power techniques have been proposed over the years to obtain low power SRAM memory. Reliability is another issue which is inevitable for SRAM memory cell and array design. Soft errors which occur due to alpha particles are one of the main reliability issues.

1.2.1 6T SRAM Memory cell

The SRAM cell basically consists of a flip flop. On the internal or storage nodes of this flip flop, binary data "0" or "1" is stored. The most common configuration of the SRAM cell consists of a full CMOS 6-transistor or 6T memory cell configured as a cross coupled latch [3]. It can be seen from Fig. 8, the SRAM cell consists of two inverters connected as shown. From Fig. 9 the SRAM cell consists of two load transistors M1 and M3, two driver transistors M2 and M4, two access transistors M5 and M6 which are connected to a pair of bit lines (BL and BLB). The gates of the two access transistors are connected to a word line (WL). In order to

form a flip flop, the input and the output of one of the inverter are connected to the output and input of the other inverter. A system on chip (SOC) is fabricated by the same process as the CMOS logic which is used to fabricate the CMOS 6T memory cell and thus SRAM memory is the most suitable choice for SOCs.

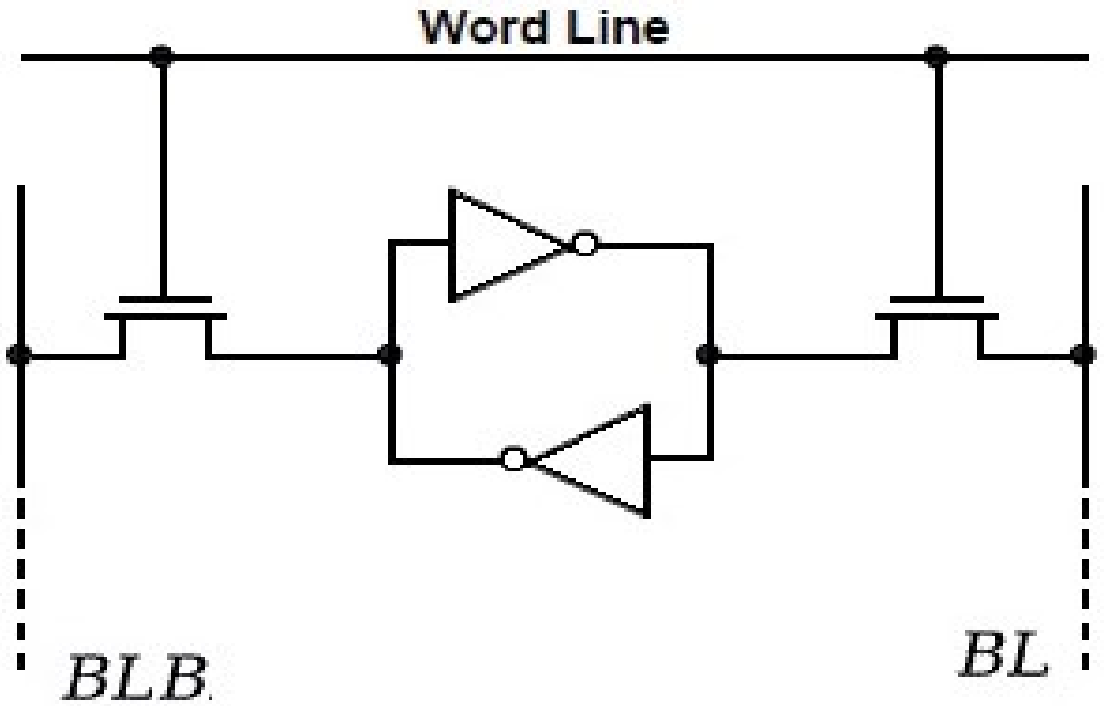

Figure 8. SRAM memory cell as two interconnected inverters

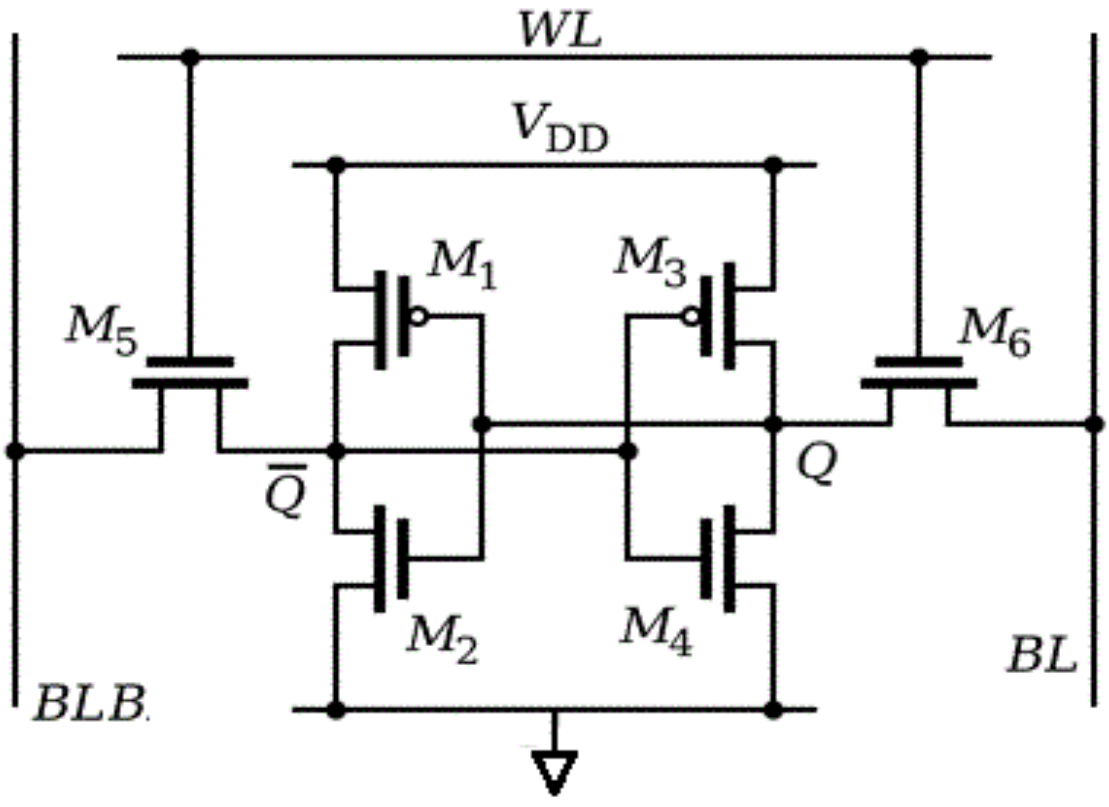

Figure 9. SRAM memory cell

1.2.2 SRAM Cell Operation

Read

For the read operation, first the BL and BLB are precharged to Vdd and then the word line (WL) is enabled. During the read operation, the bit line voltage of the node BLB at that side of the SRAM cell which stores the '1' bit (assuming that it stores the bit '1' just for the sake of explanation), remains at the precharge voltage equal to Vdd. The complementary bit line voltage BL at that side of the SRAM cell which stores the '0' bit is discharged through transistors M4 and M6 connected in series as shown in Fig. 9. Thus, the transistors M4 and M6 form a voltage divider whose output is the connected with the input of the inverter M1−M2 as shown in Fig. 9. The Sizing of M4 and M6 is kept in such a way that the voltage bump which occurs at the node BL should not be more than the trip voltage of the other inverter M3−M4 which will otherwise produce an incorrect output. Fig. 10 shows a block diagram for the read operation.

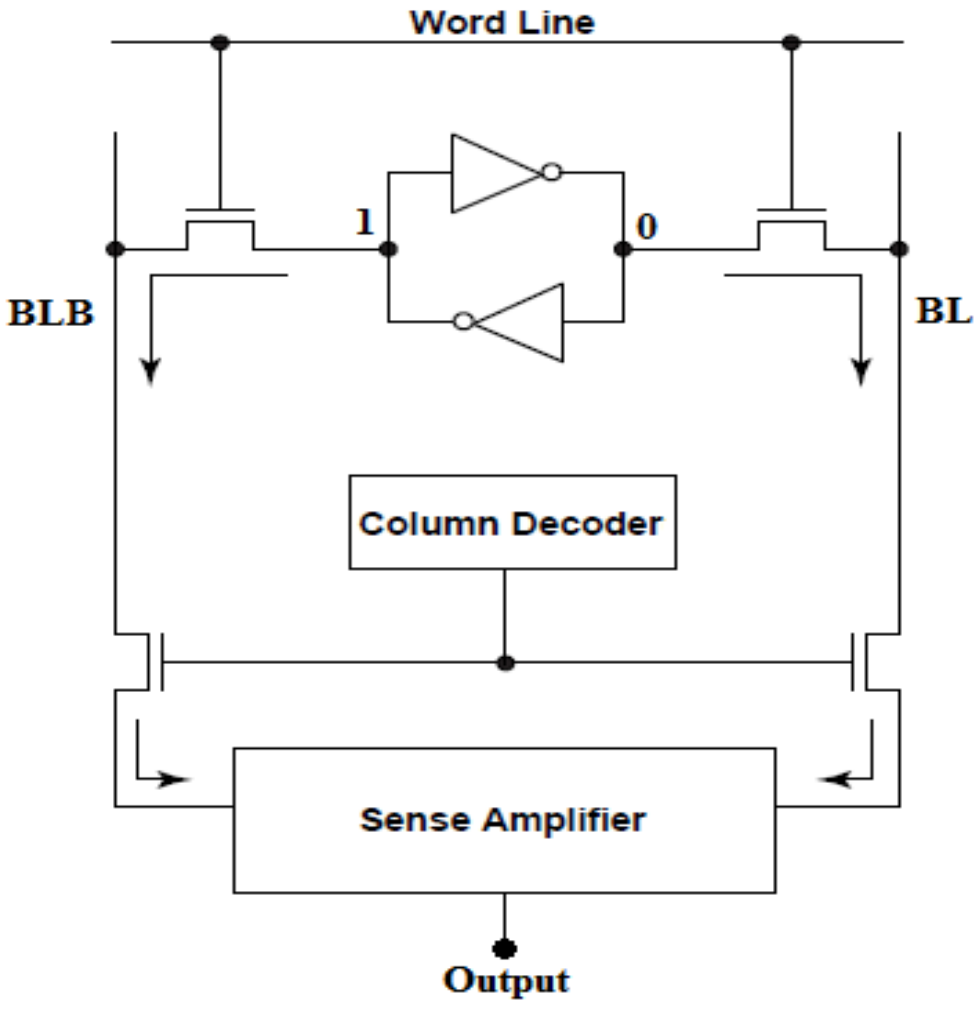

Figure 10. Block diagram for Read operation

Write

Let us consider the write '0' operation assuming that the bit '1' is already stored in the memory cell at the node Q beforehand. M1 and M4 are initially turned off, while M2 and M3 operate in linear mode. For the write operation, both the bit lines BL and BLB need to be precharged and the bit line BL needs to be discharged to zero in such a way that the internal node Q discharges and thus the positive feedback effect causes flipping of data. It is important to note that the strength of the access transistor should be larger than that of the pull down transistor. Therefore, the ratio of W/L of the pull up should be at least 3.5 to 4 times less than that of the W/L of the pull down transistor. Usually, in order to decrease the cell area which will lead to an increase in the density, the sizing of the pull up and access transistors needs to be chosen to minimize the area. But stronger access transistors or weak pull ups may be required to ensure an efficient write when considering various process corners. Fig. 11 shows a block diagram for the write operation.

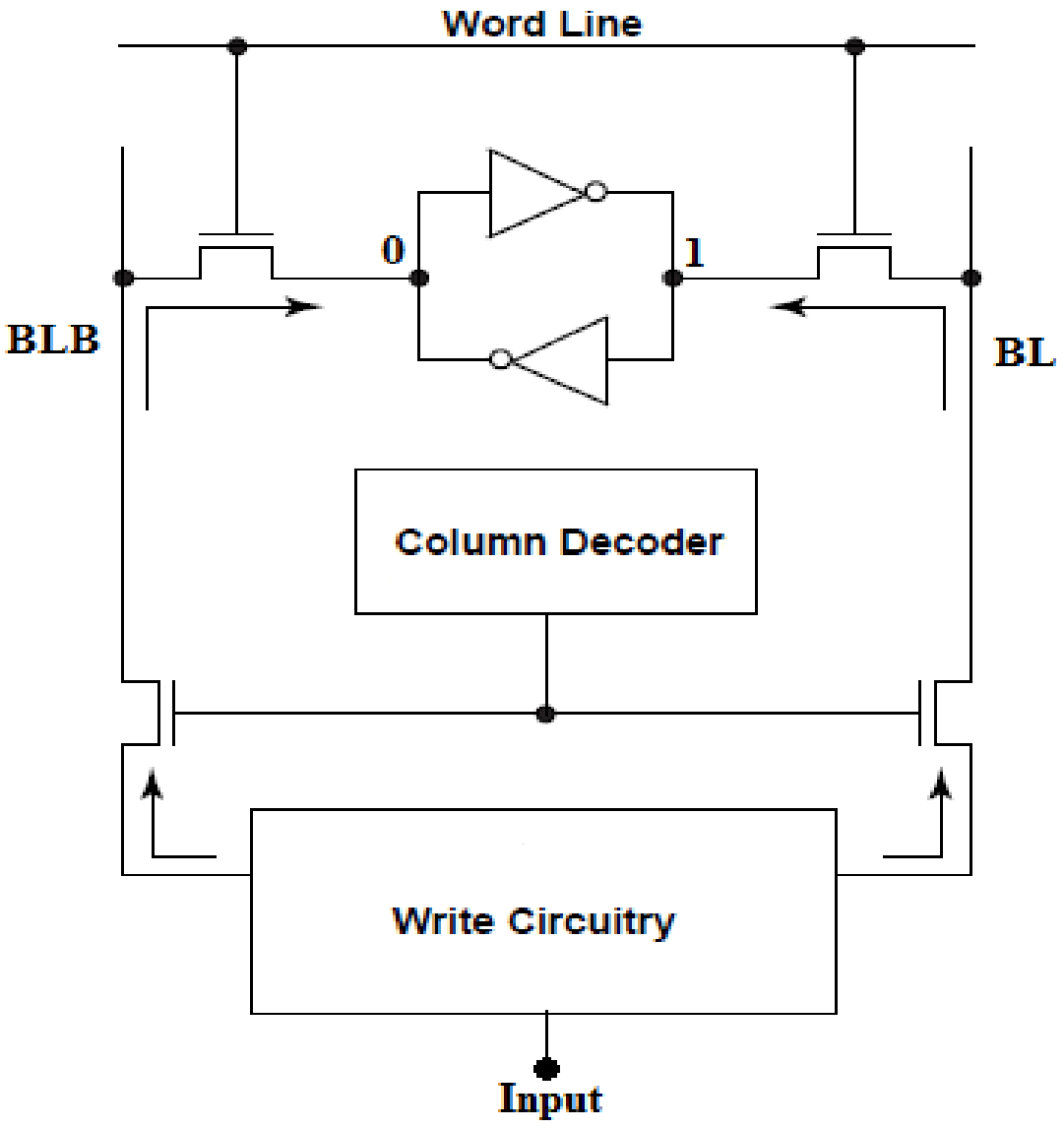

Figure 11. Block diagram for Write operation

1.3 MOTIVATION AND AIM FOR RESEARCH

Nowadays as the demands and requirements of electronic systems is increasing, memory plays a vital role in all these electronics devices. The electronics are becoming smaller, faster and more intelligent thus increasing the expectation from memory as well. The speed of memory mainly depends on the sense amplifier block. The sense amplifier is responsible for accessing the memory, and converting the voltage fluctuations in the bitline into output data. A lot of research has gone into sensing techniques for memory, thus making the memory we use today compact, fast, low power and efficient. Obtaining the correct output when accessing the memory is the most important process and thus a lot of research and effort goes into designing a sense amplifier. Also due to scaling, controlling all the parameters during fabrication is very difficult thus causing unpredictable behaviour due to parameter variations, particularly the threshold voltage V_{TH}. Thus these parameter variations need to be kept in mind when designing a sense amplifier. In this work, two new sense amplifiers designs have been proposed. A cross coupled latch type SA has been proposed for NVM, particularly flash memories which works on a power supply as low as 0.85mV and shows high speed. Another low offset high speed cross coupled latch type SA has been proposed for SRAM in which body biasing has been used in order to reduce the offset. These two designs have been implemented and simulated using ELDO simulator in 65nm technology. Keeping the variation in parameters in mind, while analysing these designs, Monte Carlo (MC) simulations are launched incorporating V_{TH} variation of 10mV in all the devices. Also, this work includes a detailed study and analysis of a few conventional SA designs.

1.4 Thesis Organization

This thesis has been organized into five sections. Section 2 discusses about a few conventional SA designs and includes their detailed study. Section 3 firstly introduces a high speed low offset cross coupled latch type current sense amplifier for NVM applications and the Flash memory in particular. The design and its working has been explained in detail, the design has been simulated an analysed by incorporating V_{TH} variations and the outputs have been presented. Section 4 discusses about the low offset high speed cross coupled latch type SA for SRAM applications, in which offset and sensing delay lowing have been achieved using body biasing techniques. In this section, the design has been explained in detail, it has been simulated and analysed by incorporating MC V_{TH} variations of 10mV in all the devices. The outputs for this sense amplifier design for SRAM have been presented in section 4. Finally the conclusions and scope for future work have been discussed in section 5.

2. Conventional Sense Amplifier designs

2.1 Conventional Sense Amplifier for NVM

A sense amplifier has three sections, the first section coverts the current signals from the reference and cell matrix into voltage signals, the second section compares these voltage signals from cell and reference matrices and the third section latches the output [4]. For current sensing techniques, the idea is to convert the currents which are drawn from the reference side and cell side into voltage signals so that they can be compared by a comparator circuit which produces the output in the form of recognizable logic levels such as '0' and '1'. The current which is drawn from the reference cell (I_{ref}) is slightly more as compared to the current drawn from a programmed cell (I_{cell}) due to the high V_{TH} value of a programmed cell. The current which is drawn from reference cell (I_{ref}) is slightly less as compared to the current drawn from the erased cell (I_{cell}) due to the low V_{TH} value of the erased cell. This difference in currents between I_{ref} and I_{cell} are required to be converted into a differential voltage which is sufficiently recognized by the comparator. The suitable setup to measure the cell current and reference current is shown in Fig. 12. In order to select a suitable comparator circuit for the SA, the gain required by the comparator for minimum input differential voltage needs to be calculated. In order to resolve this minute difference in current, the gain requirements are calculated. The minimum difference in voltage at the cell side and reference side is observed through all PVT corners. This minimum voltage difference thus observed is used to calculate the minimum gain requirement as following:

$$\text{Gain} = \frac{(V_{OH} - V_{OL})}{\Delta V_{in}} \tag{2.1}$$

$$\text{Gain (dB)} = 20\ \log_{10} \frac{(V_{OH} - V_{OL})}{\Delta V_{in}} \tag{2.2}$$

Where V_{OH} denotes the logic '1' voltage, V_{OL} denotes the logic '0' voltage and ΔV_{in} denotes the minimum voltage differential between the cell side and the reference side. Then a single stage differential amplifier as shown in Fig. 13 is employed which acts as the comparator and compares the voltages at the cell side and reference side to give the required outputs [5]. The gain is calculated as equation (2.2). The following steps are used in order to realize the required comparator:

$$A_V = g_m \cdot R_{out} \tag{2.3}$$

$$g_m = \mu_n C_{ox}(W/L)(V_{gs} - V_{TH}) \tag{2.4}$$

$$R_{out} = r_{01} || r_{03} \tag{2.5}$$

Where g_m denotes the transconductance of the input devices to the differential amplifier and R_{out} is the resistance as seen from the output. The power consumed by the circuit is calculated as:

$$P_{dissipation} = V_{dd}(I_{M1} + I_{M2}) \tag{2.6}$$

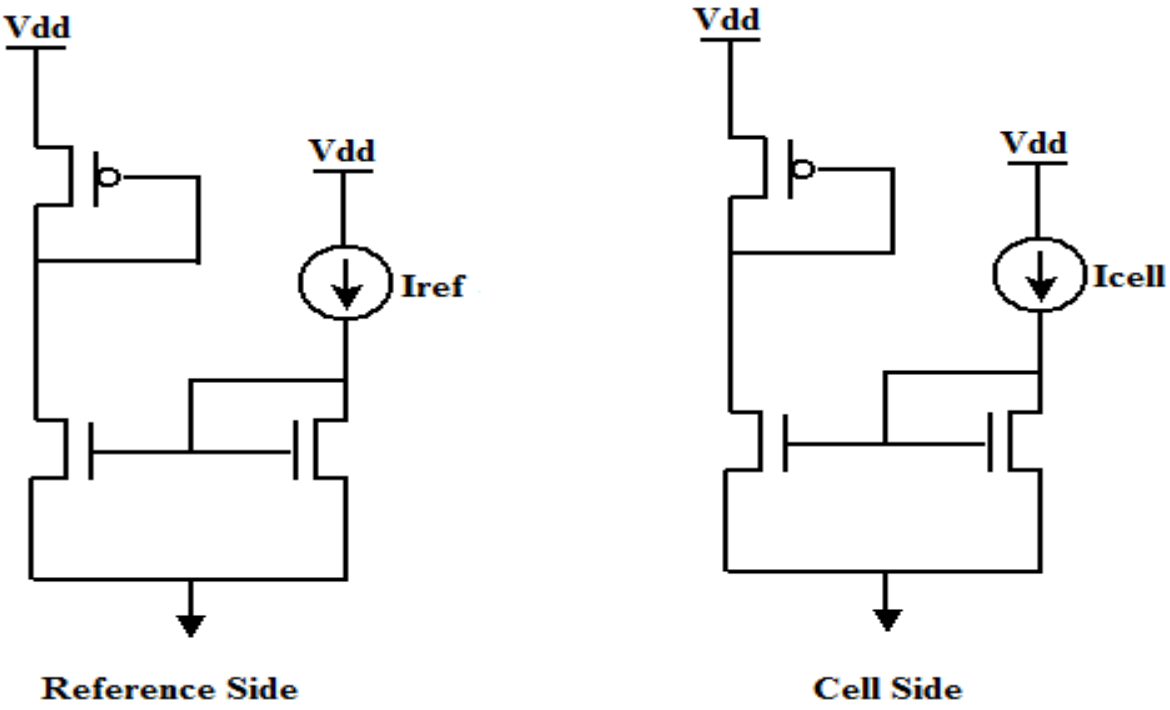

Figure 12. Setup to measure the cell and reference currents

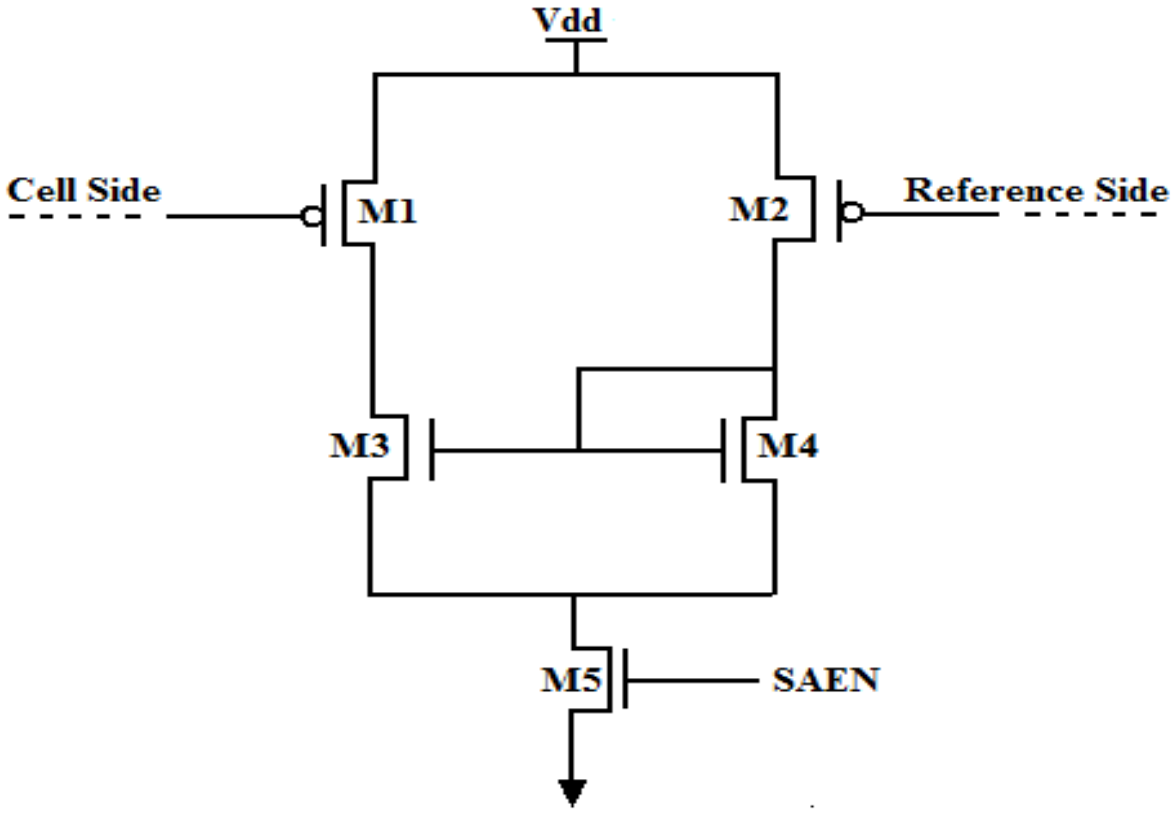

Figure 13. Single stage differential amplifier

2.2 Conventional Sense Amplifier for SRAM

A conventional sense amplifier circuit for an SRAM consists of a cross coupled inverters along with access transistors that couple the bitline voltage with the SA [6]. This whole circuit is also provided with proper pre-charge circuitry. A basic SRAM SA is shown in Fig. 14. Initially the internal nodes and the bit line nodes of the circuit are pre-charged, then the wordline signal is pulled to high voltage. The bitline nodes discharge or stay at their initial voltage in accordance with the data stored on the SRAM cell. As soon as the access transistors to the SRAM SA are switched on, the bot line voltages are coupled to the SA. After a short time, the SAEN signal is switched on thus activating the SA. The circuit is then simulated for different conditions and through different pvt corners and the minimum bit line differential voltage is observed. The probability of success is calculated as:

$$\text{Success} = \frac{\text{No.of correct sensing}}{\text{No.of trials}} \tag{2.7}$$

The SAEN signal is triggered at various instants and the data is observed for various percentages of success until the probability of success is 1. The data points are then used plotted in order to obtain the CDF and PDF plots. The standard deviation and offset voltage are hence determined. The sensing delay is calculated from the transient behaviour of the SA.

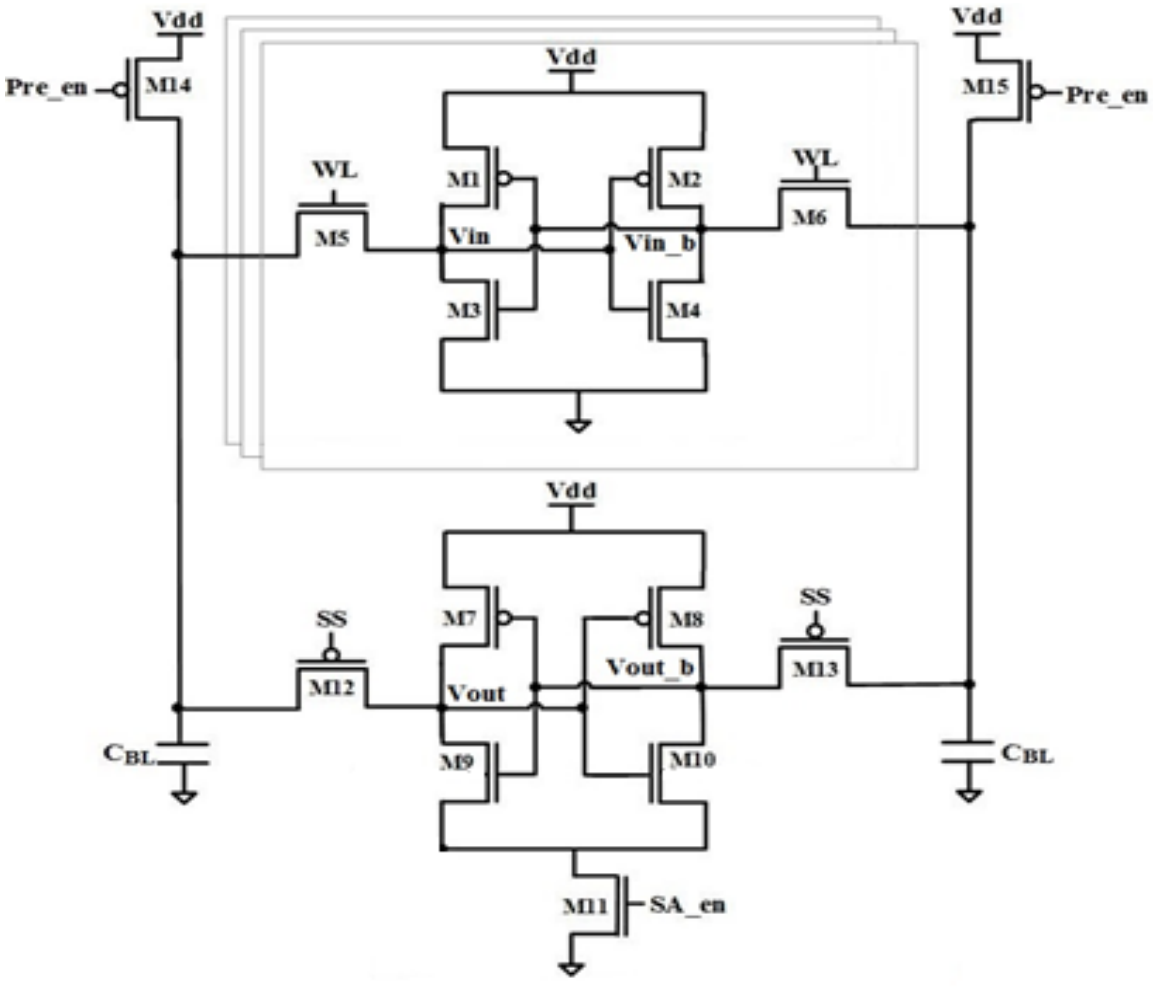

Figure 14. Basic schematic of an SRAM Sense Amplifier

3. High Speed Low Offset Modified Latch Type Sense Amplifier for NVM

Sense Amplifiers (SAs) form an integral part of memories. These circuits are responsible for reading the contents of memory. This is done by amplifying the small signal bit line differential voltage and converting them into recognizable output logic levels such as '0' and '1'. Out of all the SA topologies available today, it has been reported that a cross coupled latch type sense amplifier possesses high sensitivity. It is capable of achieving a low sensing delay due to the presence of strong positive feedback in the cross coupled topology [7]. The use of such a cross coupled latch type topology, with proper modifications for current sensing in non-volatile memory (NVM) particularly the flash memories, could be beneficial in order to achieve faster sensing using a lower supply voltage. Today as the electronic market is evolving, there is a requirement of faster memories even though the load capacitance has increased over the years, and also these memories need to be operated under lower power supplies so as to save power. The CMOS cross coupled inverter pair is used in order to design a fast and reliable sense amplifier. The main advantage of this design is the positive feedback provided by the cross coupling. The positive feedback can thus be exploited in order to achieve faster sensing during read operation. A lot of research and analysis has been done on the cross coupled inverter pair so as to be able to use it as an SA. Such a topology can therefore find applications in the lower CMOS technology nodes in which the SAs are highly constrained in terms of access time and offset due to increase in bitline capacitance [8].

Furthermore, with scaling in technology it is difficult to control the fabrication process. Most of the semiconductor products available today are due to the improvements in recent times which were a direct or indirect result of the shrinking in devices and circuits. Therefore scaling of devices has resulted in performance enhancements at low fabrication costs. At the same time, the process variations and intra-die variations have been observed to increase with each the shrinking technology nodes. Since major high performance analog circuits require matched devices and signal paths, this scaling has led to decreasing yields and reliability of integrated circuit chips. Basically, the main problem is that the parameters of devices on the die exhibit variations which results in different characteristics leading to variations in process parameters which causes unpredictability in the performance of SAs. For example, there may be variation in oxide thickness and the number of dopant atoms in the transistor channel due to various reasons which leads to variation in the threshold voltage (V_{TH}) and other parameters of the device [2]. These variations may result in dramatic changes in performance metrics of the circuits, in both positive and negative directions. Therefore, the design process changes because the design is constrained by particular specifications like speed. To account for these variations, the circuit needs to be designed according to the worst case values for all device parameters. Monte Carlo (MC) analysis simulates the circuit over a huge range of randomly chosen parameters. Thus it is extremely important to keep this aspect in in mind when designing an SA and estimating its performance in terms of access time, offset, power, and area.

This work proposes a modified latch type SA for current sensing. The design has been implemented in 65nm CMOS technology and exhibits a small sensing delay at a supply voltage of 1.2V. The SA topology is also capable of working at lower power supplies albeit with a substantially degraded performance. The reported topology can be easily modified in terms of sizing and sense enable triggering in order to achieve the desired specifications of the supply, sensing delay, and offset for various applications.

3.1 Design of the Modified Cross Coupled Latch Type Sense Amplifier

The SA circuits has been designed as shown in Fig. 15. It consists of a modified cross coupled latch. It can be seen from Fig. 15 that there is resistive access coupling in which the resistors are realized with MOSFET devices. In this topology, the idea is to couple the cross coupled SA latch with the load during sensing phase, capacitively using the devices M5 and M6. The benefit obtained by such a scheme is that during the discharging stage, these devices will be automatically in the OFF state. Therefore this scheme will substantially reduce the power consumption during read operation. Another huge advantage of using this configuration is the reduced coupling effect at nodes V_out and Vout_b thus reducing the voltage bump in the output signals due to the switching of input signals to the SA. This also results in the reduction of the overall power dissipation.

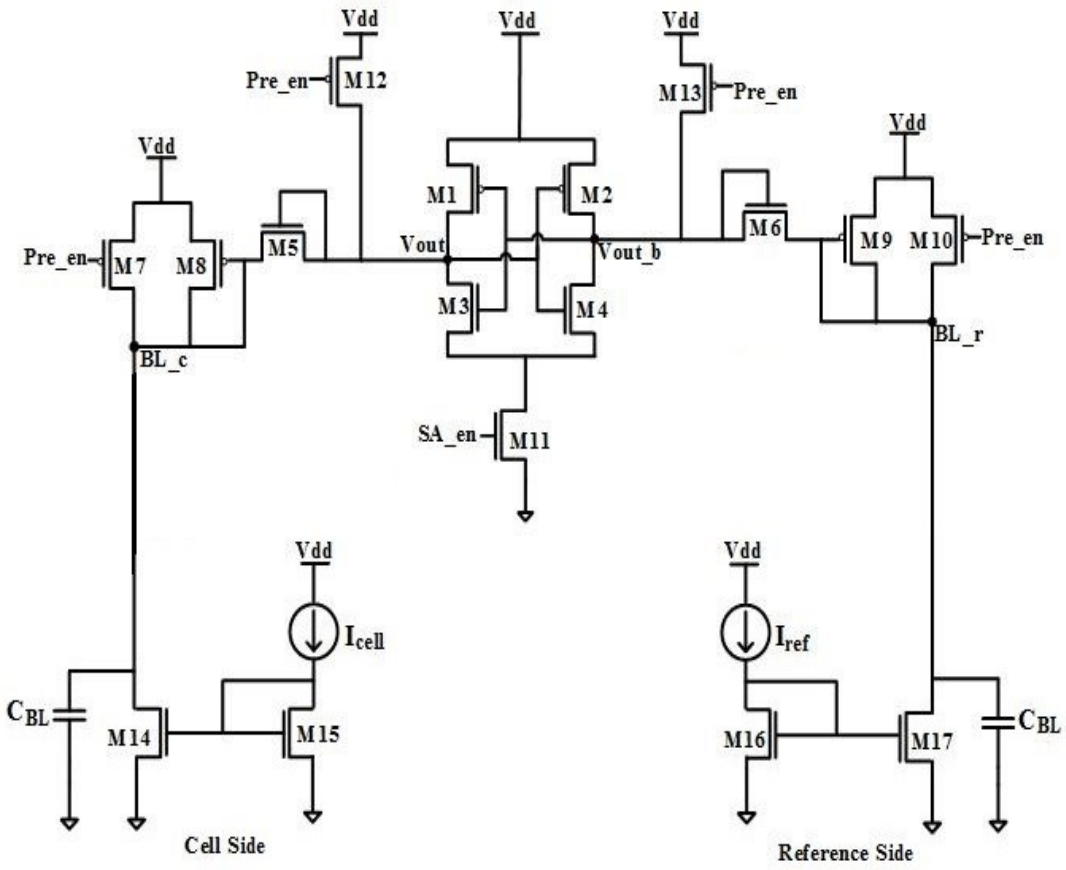

Figure 15. Schematic of the Modified Latch Type SA for current sensing

3.2 Operation

A sense amplifier has three sections, the first section coverts the current signals from the reference and cell matrix into voltage signals, the second section compares these voltage signals from cell and reference matrices and the third section latches the output. For current sensing techniques, the idea is to convert the currents which are drawn from the reference side and cell side into voltage signals so that they can be compared thus producing the output in the form of recognizable logic levels such as '0' and '1'. The current which is drawn from the reference cell is I_{ref} is fixed at 8μA (The value of reference current is fixed buy convention and may vary from manufacturer to manufacturer). At the same bias condition, current drawn from a programmed cell (I_{Cell}) observed to be less than I_{REF} due to the high V_{TH} value of a programmed cell. The current drawn from an erased cell (I_{cell}) is observed to be greater than I_{ref} due to the high V_{TH} value of the erased cell. This difference in currents between I_{ref} and I_{cell} are required to be converted into a differential voltage which can sufficiently trigger a correct sensing operation. During pre-charging phase, the signal Pre_en is triggered causing devices M7, M10, M12 and M13 to switch ON. As soon as the pre-charge is switched off, the voltage at nodes BL_c and BL_r is coupled to the internal nodes Vout and Vout_b through the devices M5 and M6 respectively. The timing diagram for operation of the proposed sense amplifier is shown in Fig. 16. For a programmed cell, the bit line node on the cell side discharges slower as compared to the reference side and for an erased cell, the bit line node on the cell side discharges faster as compared to reference side. As soon as a sufficient bit line differential voltage is formed, the signal SA_en is triggered and the device M11 switches ON so as to obtain the output.

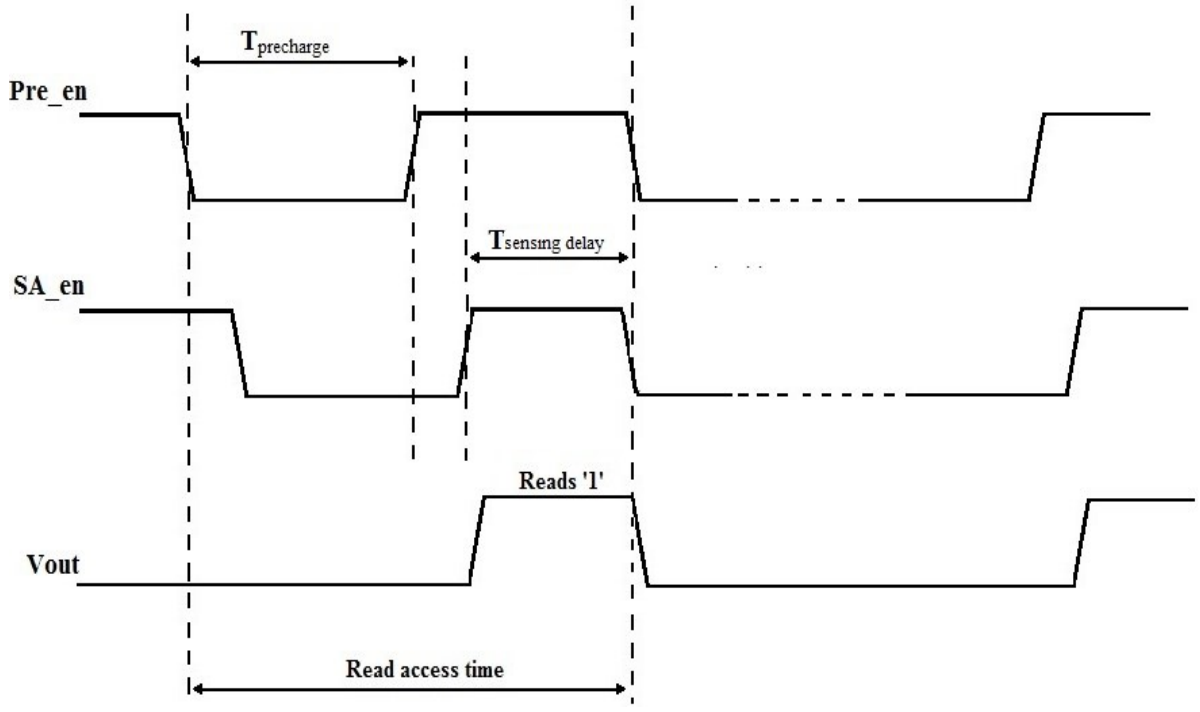

Figure 16. Timing Diagram for signals

3.3 Offset and Delay Estimation

Offset estimation is one of the most important tasks in the design of SAs as it is one of the performance metrics of the SA. The offset of an SA characterizes its performance. Offset for a current sense amplifier is measured in terms of current which is defined as the minimum difference of cell current and reference current when the output voltage crosses Vdd/2. For proper operation, the current difference between cell and reference must always exceed this offset. Hence sense amplifier offset must be as low as possible [9]. In addition to current offset, there exists a voltage offset for the latch type sense amplifier. The minimum differential voltage between the bit line nodes which is able to trigger a correct sensing is referred to as the offset voltage. An ideal SA exhibits infinitely small offset but in all practical designs there will be always a finite offset. This finite offset if not accounted for in the design of the SA, can cause failure in memory. Therefore, it is a critical task to characterize and account for this finite offset in order to ensure the efficient working of memory and avoid any failures. Along with incorporating any mismatch in the values of the V_{TH} of the devices due to parameter variations, the SA requires a minimum voltage difference V_{diff} between the bit lines BL and BLB. This V_{diff} needs to be above a certain threshold for the circuit to operate efficiently. To identify the effect of process variations, 1000 Monte Carlo (MC) simulations were performed by incorporating a total of 10mV variation in VTH of devices. In the current situation, current offset is defined as the minimum current difference between the bit lines of the cell and reference sides for achieving full success rate (i.e. success rate of 100%). Similarly, voltage offset is defined as the minimum differential voltage between the bitline nodes for 100% success rate. The success rate (S), can therefore be formulated as:

$$S = \frac{\text{No. of correct sensing of SA Output}}{\text{Total No.of Trials (N)}} \tag{3.1}$$

Furthermore, the read access time, T_{Access}, which signifies the time utilized by the SA to read the contents of memory, in the case of NVM can be expressed as:

$$T_{Access} = T_{Precharge} + T_{Sensing\ delay} + T_{Latching} \tag{3.2}$$

As seen from equation (3.2), the read access time has been divided into three parts namely $T_{Precharge}$, $T_{Sensing\ delay}$ and $T_{Latching}$. Apparently, appropriate design and topology of SAs can enable the control of first two parts, namely $T_{Precharge}$ and $T_{Sensing\ delay}$, while latching time is fixed depending upon the output load and buffering time. As the load on memory has increased over the years, this latching time $T_{Latching}$ has also been observed to increase. In the present work, sensing delay, $T_{Sensing\ delay}$, is predicted by utilizing the transient behaviour of SA at the estimated current and voltage offset for 100% success incorporating V_{TH} variations in all devices in order to account for any variation in parameters at the fabrication stage [10].

3.4 Simulation Results and Comparison

The proposed SA is implemented in 65nm CMOS bulk technology at a supply voltage (Vdd) of 1.2V. The selected aspect ratios for respective transistors of the core sense amplifier circuit are mentioned in Table II. In order to demonstrate the performance of sense amplifier with respect to Offset Voltage of sense amplifier, Monte Carlo (MC) simulations are launched by incorporating a total threshold voltage variation of 10mV in all the devices and simulations were carried out in ELDO simulator for a reference cell current (Iref) of 8μA, bitline capacitive load of 1pF at temperature 27°C. In order to calculate the current offset, two different simulation sets were carried out for programmed and erased memory cells. The term yield (S), is determined at each differential bitline voltage by projecting SA_en signal for specific differential bit line voltage by running 1000 MC simulations using equation (1). It is important to note that the analysis carried out here in terms of voltage offset assumes that the transistors in the designed circuit are with random mismatches. The Cumulative Distribution Function (CDF) and the Probability Density Function (PDF) of the sensing probability for the proposed SA at supply voltages of 1.2V are shown in Figs. 17 and 18 respectively. The corresponding bit line voltages (VBL_c and VBL_r) were maintained at 1.2V. The triggering of sense enable signal has substantial effect on the access time of the SA and therefore MC simulations were carried out by including VTH variations and applying SA_en signal at various time instants till 100% success is achieved in order to calculate the Cumulative Distribution Function (CDF)**.** Samples from the CDF were fit into the nearest Gaussian distribution function to obtain the Probability Density Function (PDF) which is given in equation (3):

$$f(x) = a1 * e^{-((x-b1)/c1)^2} \tag{3.3}$$

After the analysis, the coefficients in equation (3.3) are obtained as a1=63.77, b1=0.0002097 and c1=0.008847. It can be observed that 100% success in sensing the correct value is achieved when SA_en signal was applied at 7ns at a differential bitline voltage of 26.23mV which is shown in Fig. 17. In addition, the standard deviation in the offset voltage is 6.002 mV as obtained from Fig. 18.

TABLE II. Transistor Aspect Ratios

Transistor	W/L (μm)
M1=M2	0.4/0.09
M3=M4	0.4/0.09
M5=M6	9/0.1

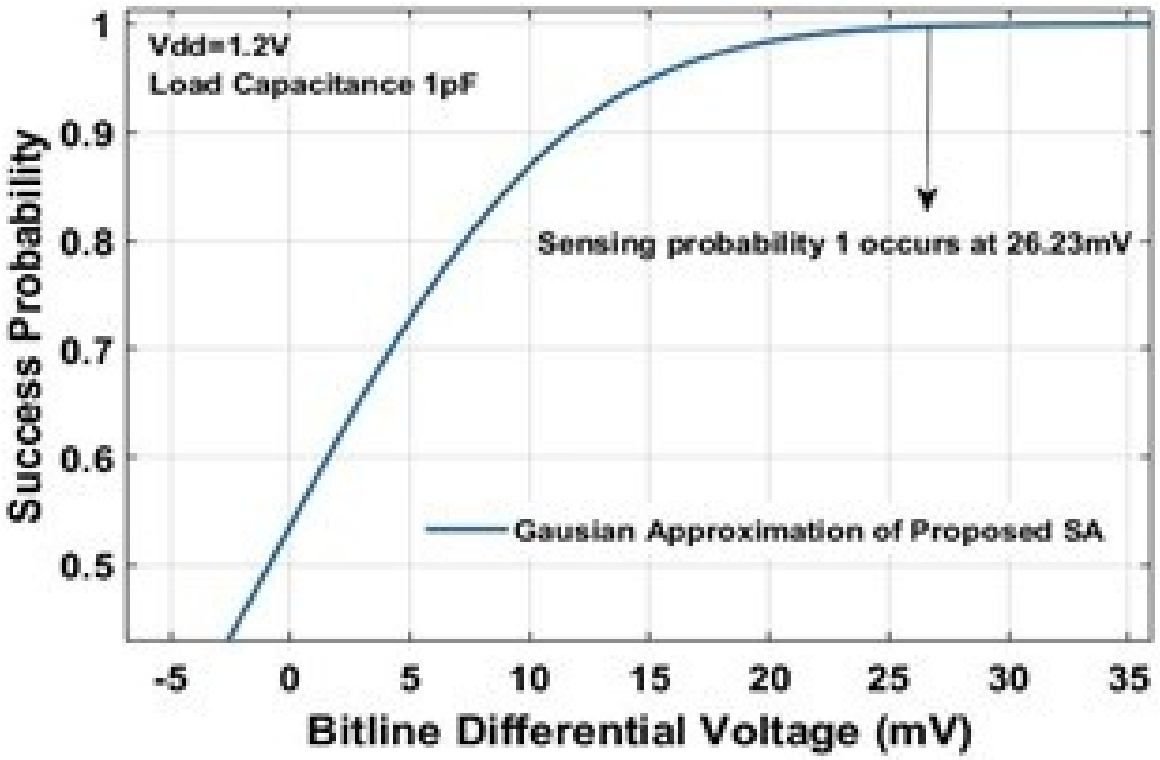

Figure 17. Sensing Probability versus Bit line differential Voltage

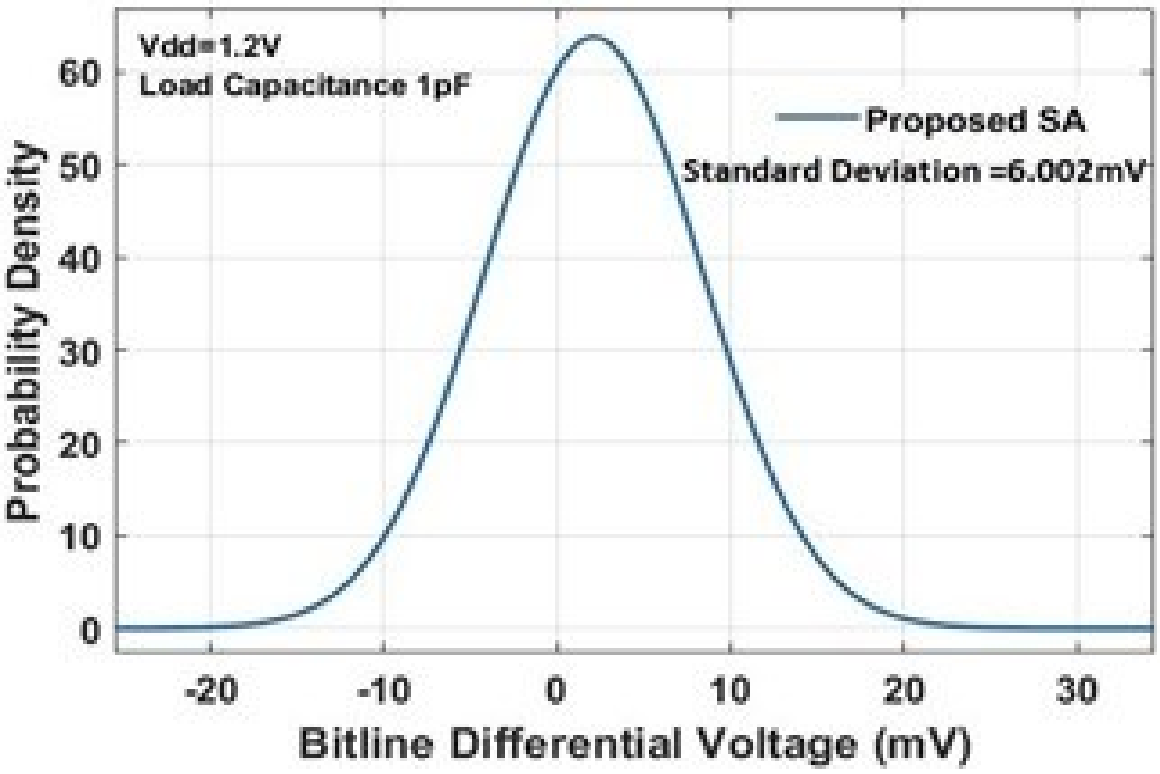

Figure 18. Probability Density function of the Sensing Probability depicted in Fig. 17

The transient behaviour of the modified sense amplifier topology is shown in Fig. 19. It can be observed from Fig. 19 that the sense amplifier achieves sensing delay of 0.946ns. Furthermore, in order to check the effectiveness of the sense amplifier with respect to sensing delay, V_{TH}

variations were introduced in all the devices using Monte Carlo method at a biasing voltage of 1.2V and temperature of 27°C. The distribution of the obtained sensing delay for programmed cell is shown in Fig. 20. In order to verify the effect of supply voltage on offset voltage, the power supply of the sense amplifier is swept from 1.2V to 2V. It is observed from Table III that with increase in supply voltage the bitline differential voltage increases thus decreasing the sensing delay which is an intuitive result [11]. Subsequently, the sense amplifier is assessed in order to check the distribution of offset voltage under Max-Max and Min-Min corners. The outcome is plotted in Fig. 21. It can be observed from the plot that when both the NMOS and PMOS are Max-Max, the bitline differential voltage has lower values as compared to Min-Min corners because of higher leakage.

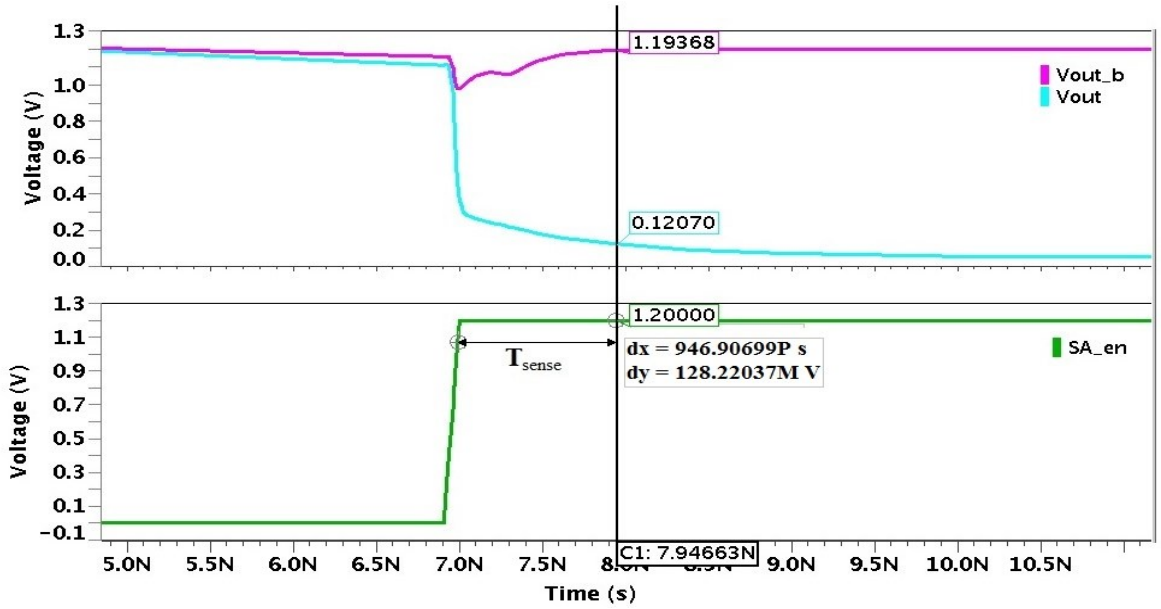

Figure 19. Output Waveform

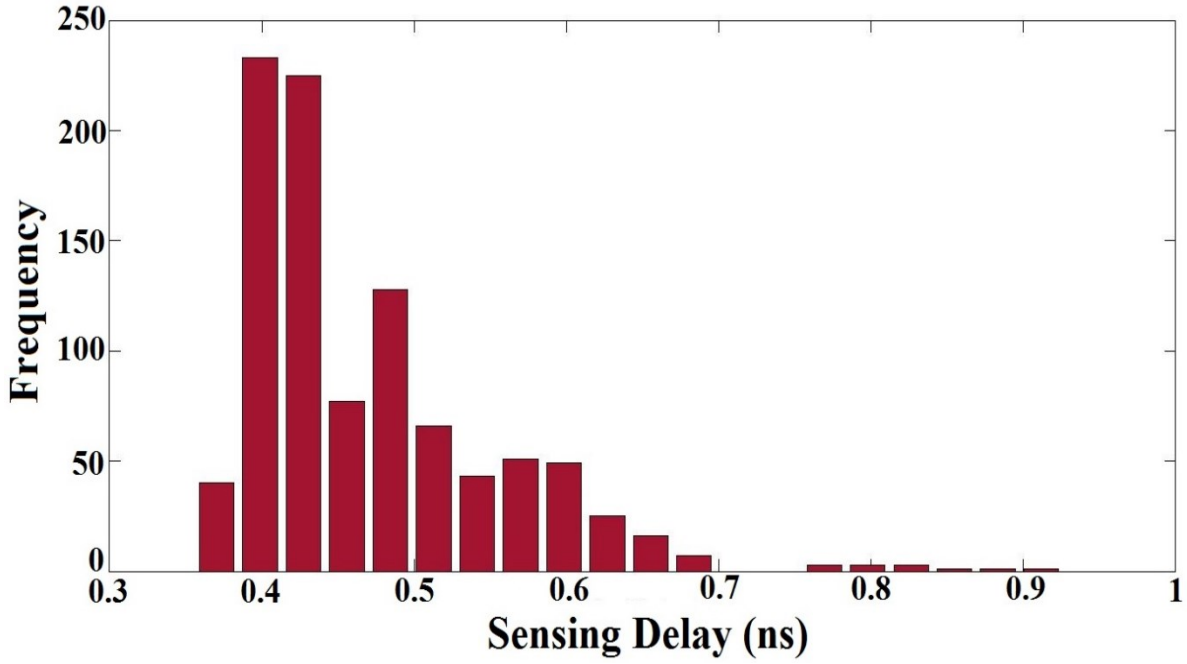

Figure 20. Distribution of Sensing Delay

TABLE III. Variation of Bit line Differential Voltage with Supply Voltage

Supply Voltage (V)	Bit line Differential Voltage (V)
1.2	0.0240
1.4	0.0241
1.6	0.0279
1.8	0.0303
2.0	0.0309

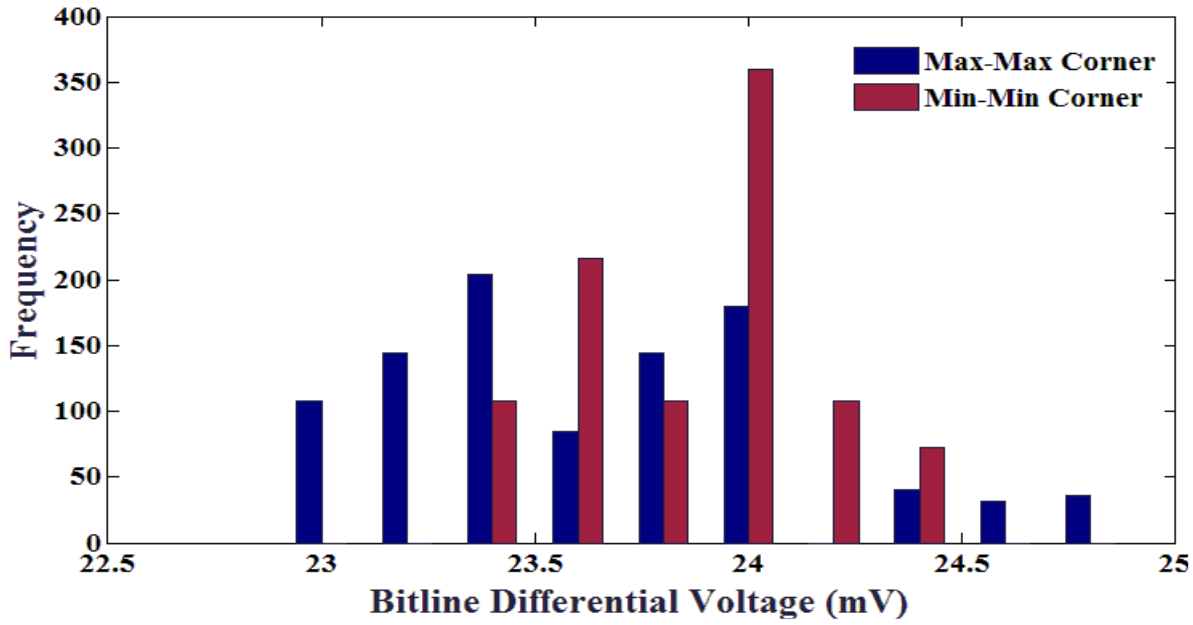

Figure 21. Variation of Bit line Differential Voltage at different PVT corners

In order to compare the performance of the proposed sense amplifier with other conventional designs, an equivalent bitline load of 0.5pF is considered with an equivalent Iref of 16µA according to the setup given in [12]. The supply voltage is swept from 1.3V to 2V for the modelled setup and the results for sensing delay are shown in Fig. 22. To observe the effectiveness of the proposed SA in terms of sensing delay, equivalent bitline load for the SA is varied from 0.2pF to 1pF at a fixed voltage supply of 1.5V according to the setup given in [12]. The corresponding results for sensing delay are shown in Fig. 23. It can be observed from Fig. 22 and 23 that the obtained results for the proposed sense amplifier compare favourably with [12], [13] and [14]. Also, as reported in [12], the power consumed by the design at a power supply of 1.5V and a bitline load of 0.5pF is 85µW whereas the power consumed by the latch type proposed SA under the exact same conditions is 57.27µW, thus exhibiting an improvement of 32.6% in terms of power.

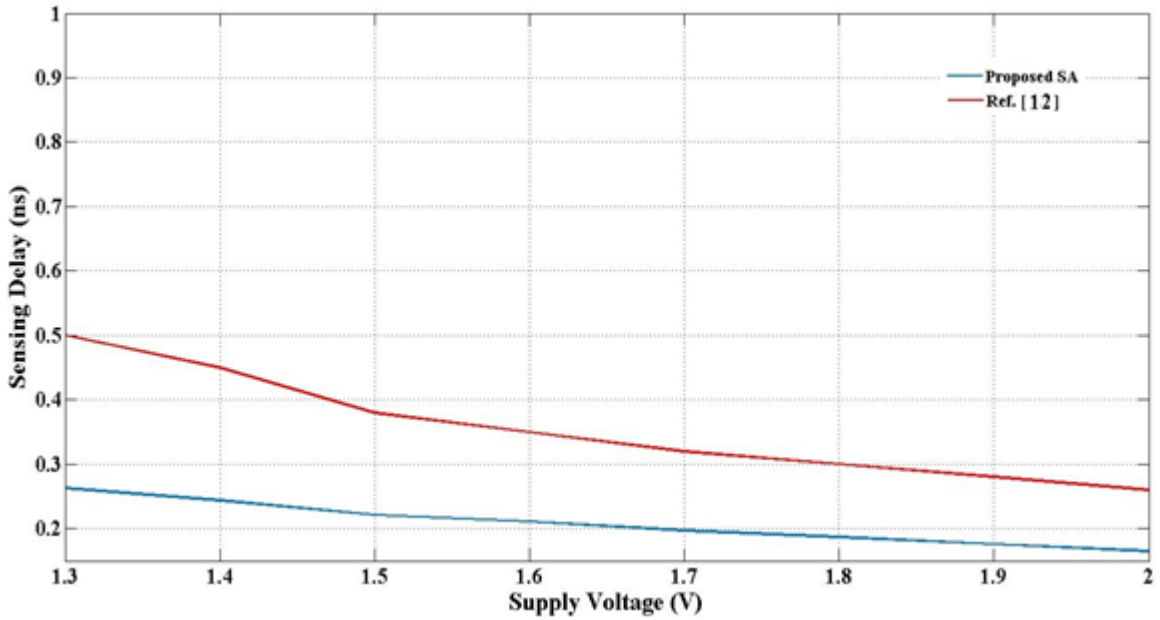

Figure 22. Variation of Sensing Delay with supply Voltage

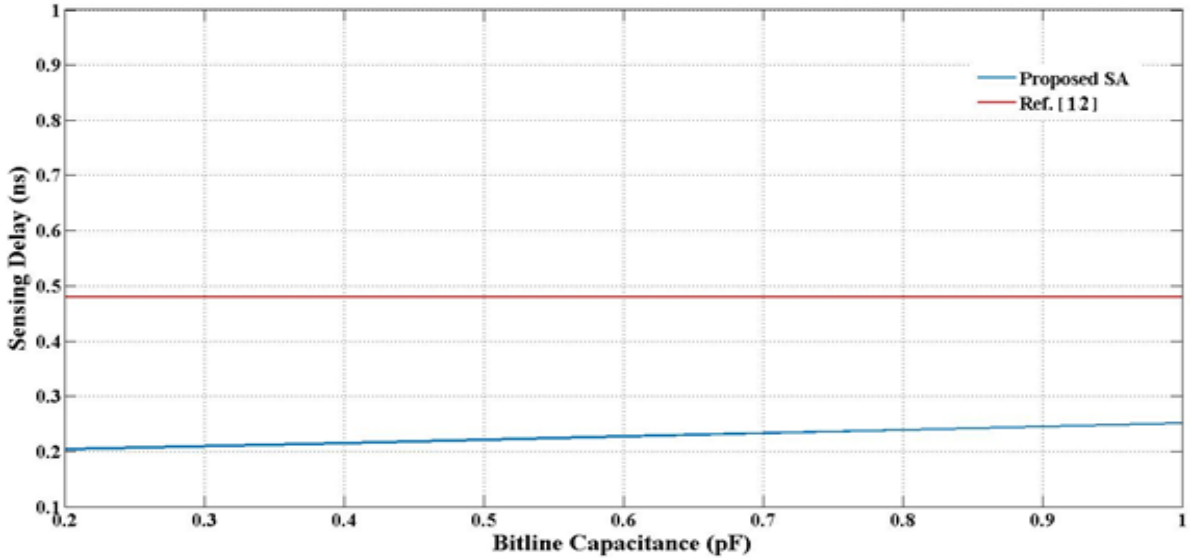

Figure 23. Variation of Sensing Delay with Bit line Capacitance (pF)

Then, the proposed SA is simulated under different supply and temperature conditions, the results for which are mentioned in Table IV. Furthermore, it can be seen from the results in Table V that the success rate decreases with the decrease in offset current. Essentially, it can be observed that the SA exhibits 100% success rate at a current offset equal to 1μA and it degrades as the current offset is decreased. Finally, the proposed topology was analysed for power consumed in the cell. The total power consumed can be expressed as:

$$P = I_{Avg} . Vdd \tag{3.4}$$

The total power consumed can be thought in two parts namely the power consumed during sensing and the power consumed during differential bitline formation gives as:

$$P_{Total} = P_{Differential\ Forrmation} + P_{Sensing} \quad (3.5)$$

For calculation of the above power values, the average current I_{Avg} through the comparator or core of the SA is measured and then the power is calculated. Then the various consumed power at varying supply level was obtained which are given in in Table VI. It can be seen in Table VI that the total power consumed by the SA decreases with decrease in supply voltage.

TABLE IV. Variation in Sensing Delay and Current Offset with variation in supply voltage and Temperature

Vdd (V)	Temperature (^oC)	Sensing Delay (ns)	Current Offset (µA)
1	27	1.298	0.4
1.35	125	1.493	0.7
1.65	27	0.246	0.7
1.95	-125	0.356	0.4

TABLE V. Variation in Percentage Success and Voltage Offset with Cell current

Cell Current (µA)	Percentage Success (%)	Voltage Offset (V)
7.0	100	0.0262
7.1	98.765	0.0198
7.2	95.37	0.0138
7.3	88.88	0.0116
7.4	79.11	0.0102

TABLE VI. Power consumption at different values of power supply

Vdd (V)	Power consumed during Bit line Differential Formation (µW)	Power consumed during Sensing (µW)	Total Power consumed (µW)
1.2	18.458	27.054	45.512
1	15.175	20.082	35.257
0.9	13.642	16.589	30.051
0.85	12.859	14.801	27.66

3.5 Conclusions

A new SA topology for non-volatile memory has been presented in this work. The proposed topology has been analysed in 65nm CMOS technology operating at a supply voltage of 1.2V and allows operation at supply voltages as low as 0.85V. The proposed topology when compared to the existing topology exhibits better performance in terms of sensing delay and offset voltage. The proposed SA exhibits a sensing delay of 0.452ns and 0.946ns in the best and worst case scenarios respectively at typical corners with a voltage offset of 26.23mV and current offset of 1μA. The minimum differential bitline voltage for 100% success was found out to be 23.39mV and 23.85mV at Max-Max and Min-Min corners respectively. The SA consumes a total power of 45.512μW at a power supply of 1.2V at typical corners. Thus it can be inferred from the above discussion that, the proposed topology could be potentially useful for high speed, low voltage, and low power applications.

4. A Cross Coupled Latch Type Sense Amplifier for SRAM with Delay Reduction and Offset Lowering Using Body Bias

A sense amplifier (SA) is one of the most essential and crucial parts of memory as these circuits define the read access time of memory. An SA is used to retrieve the data stored in memory. The whole purpose of a memory is to store data and to be able to retrieve that data whenever required, and thus SA is one of the most crucial parts. SAs amplify any small signal variations in the bit lines so as to obtain a recognizable output in the form of bits '0' and '1'as fast as possible. The minimum sensing voltage at which the SA is able to trigger a successful sensing operation is called the voltage offset. The smaller this sensing voltage is, the faster the SA will produce the output. Also, smaller sensing voltage reduces the overall power dissipation because as soon as the contents of the memory are sensed, the bit lines can be restored to their original states without any further delay. In today's world there is a significant gap between the performance of memory and the central processing units (CPUs). Therefore there is a need for highly efficient, fast and stable memories on chip in order to decrease this performance gap, thus faster SA designs will definitely help in bridging up this gap. For an SA design, it is important to focus on two factors which are high yield and low delay [15]. High yield is crucial because there is one SA for each bit line and hence is responsible for a large segment of memory. To design a fast, high yielding, low power and efficient SA circuit, one may have to face many challenges owing to the fact that in today's designs bit lines exhibit high load because of the requirement of high capacity. Therefore the speed of the SA is limited. Even though with scaling in technology and supply voltage, the logic circuit delay has shown improvement over the years with each technology node, the speed of the memory as a whole has been limited due to the delay caused owing to longer interconnects and bit lines which carry huge loads due to increased capacitance. Also, there are constraints on SRAM design such as the requirement of being compact, thus the need to use minimum sizing in transistors for the memory cell. Thus these small memory cells are required to drive bit lines bearing large capacitance which results in small voltage signal swings thus limiting speed of the memory.

The use of a latch type SA in order to read the contents of memory has been promoted over the years. A latch type SA possesses strong positive feedback which helps to achieve the required output faster. As the dimensions of the devices scale, with scaling in technology, process variations pose a significant threat to the designs. Due to process variations, there is mismatch between the devices on the same die thus leading to decrease in yield [16]. The SA circuits thus suffer due to mismatches in devices, as mismatches lead to changes in operating conditions like offset. Thus resulting in errors and failures, ultimately leading to a significant decrease in the yield. In modern technology nodes, there may be many sources of variations in parameters pf the devices. A major cause of device mismatches is the variation in parameters like effective channel length and V_{TH} within the same die thus affecting the performance of the circuit crucially. An SA essentially requires matching devices and identical device parameters in order to perform efficiently. Random dopant fluctuation (RDF) has been observed as the major cause

for variations in the device parameters such as effective channel length and V_{TH} [17]. Lithography has also been observed to cause of process variations. Variations in the exposure of light and varying resist thickness also cause variations in V_{TH} and effective channel length. Thus with scaling in technology it is becoming more difficult to control these variations in parameters. Nowadays SRAM design are required to be denser thus resulting in the use of minimally sized devices, also for a dense SRAM design, the SA needs devices which have been matched carefully so as to minimize any variations in the device parameters. However, even with devices which have been matched so carefully, the yield and performance are of major concern. Out of all parts of the SRAM circuit, an SA forms the vital component. A latch type SA as shown in Fig. 24 is used because of its advantages such as low power dissipation and low sensing delay due to the presence of strong positive feedback.

4.1 Design of the Modified Cross Coupled Latch Type Sense Amplifier

The design of the cross coupled latch type SA has been shown in Fig. 24. A basic SRAM cell has been shown in the figure along with the SA. It can be seen from Fig. 24 that there is a potential divider circuit represented by M16 and M17 which produces an output of Vdd/2.

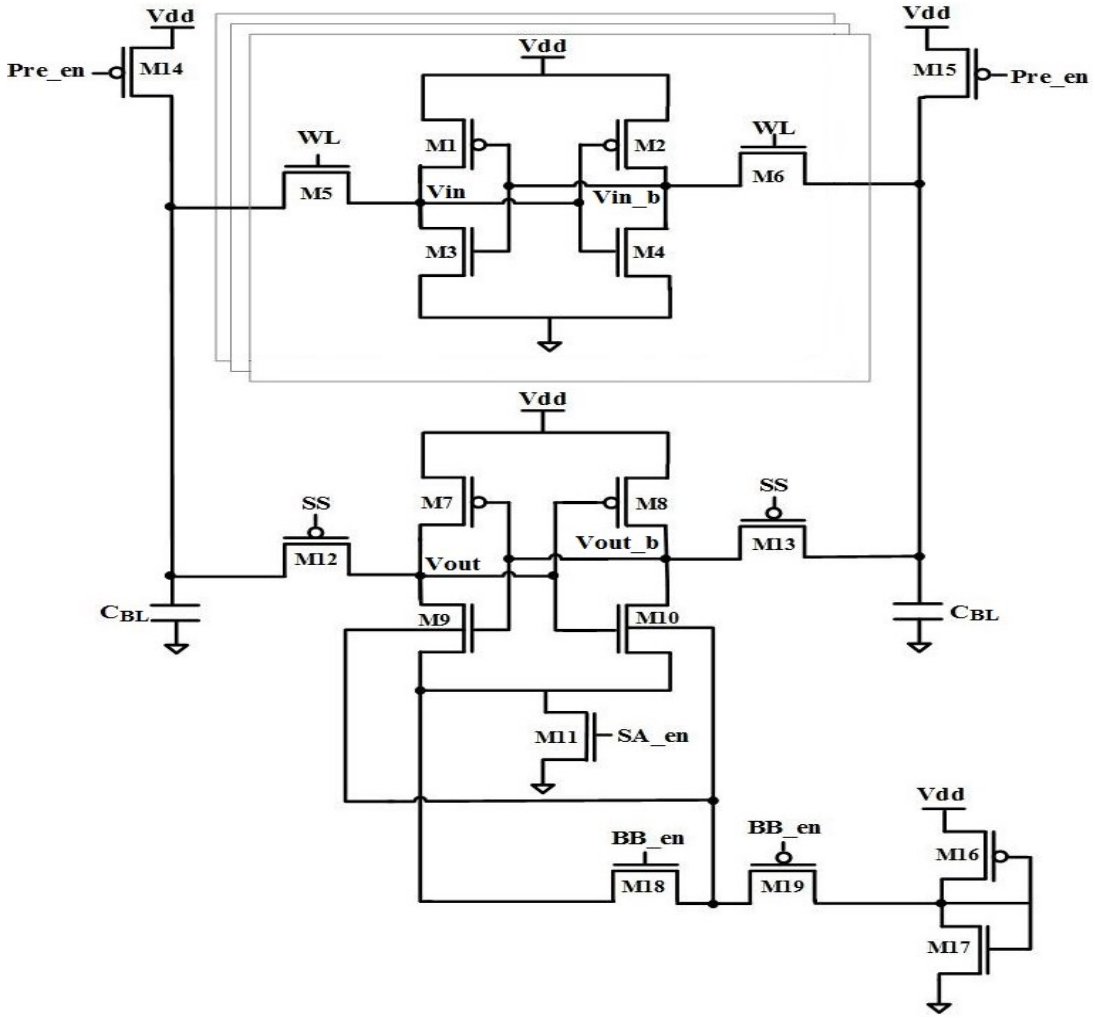

Figure 24. Modified Cross Coupled Latch Type Sense Amplifier along with SRAM cell

4.2 Operation

As seen from Fig. 24, the SA has been modified by providing body biasing to it. The potential divider formed by devices M16 and M17 produces an output of Vdd/2 which is 0.6V as the power supply has been fixed at 1.2V in this case. According to the signals shown in Fig. 25, initially the bit line nodes and the nodes Vout and Vout_b are precharged to Vdd. Assuming that the data has been stored in the SRAM cell already, the wordline WL is switched on. As soon as the WL is switched on, the signal SS is switched on to allow the coupling of bitline voltages with the cross coupled SA. At the same time the body bias signal BB_en is switched on. Initially, before the switching on of the SS signal, BB_en remains at high voltage, switching on M18 thus shorting the source and substrate of the pull downs M3 and M4 ($V_{SB} = 0$). After SS is switched on, BB_en switches to low voltage thus switching on M19 thus providing the substrates of the pull downs with a voltage of 0.6V. As soon as SS is switched off and SAEN is switched on, the source voltage of the pull downs is pulled down to 0. Now for M3 and M4, V_{SB} = -0.6V. This enables the lowering of threshold voltages of the pull downs and thus speeding up the read operation and enhancing the positive feedback and resulting in a lower offset voltage in accordance with the following equation:

$$V_{TH} = V_{t0} + \gamma(\sqrt{|V_{SB} + \Phi_F|} - \sqrt{|\Phi_F|}) \quad (4.1)$$

Where V_{t0} denotes the threshold voltage of the device when $V_{SB} = 0$, γ denotes the body effect parameter and Φ_F denotes the fermi potential of the substrate.

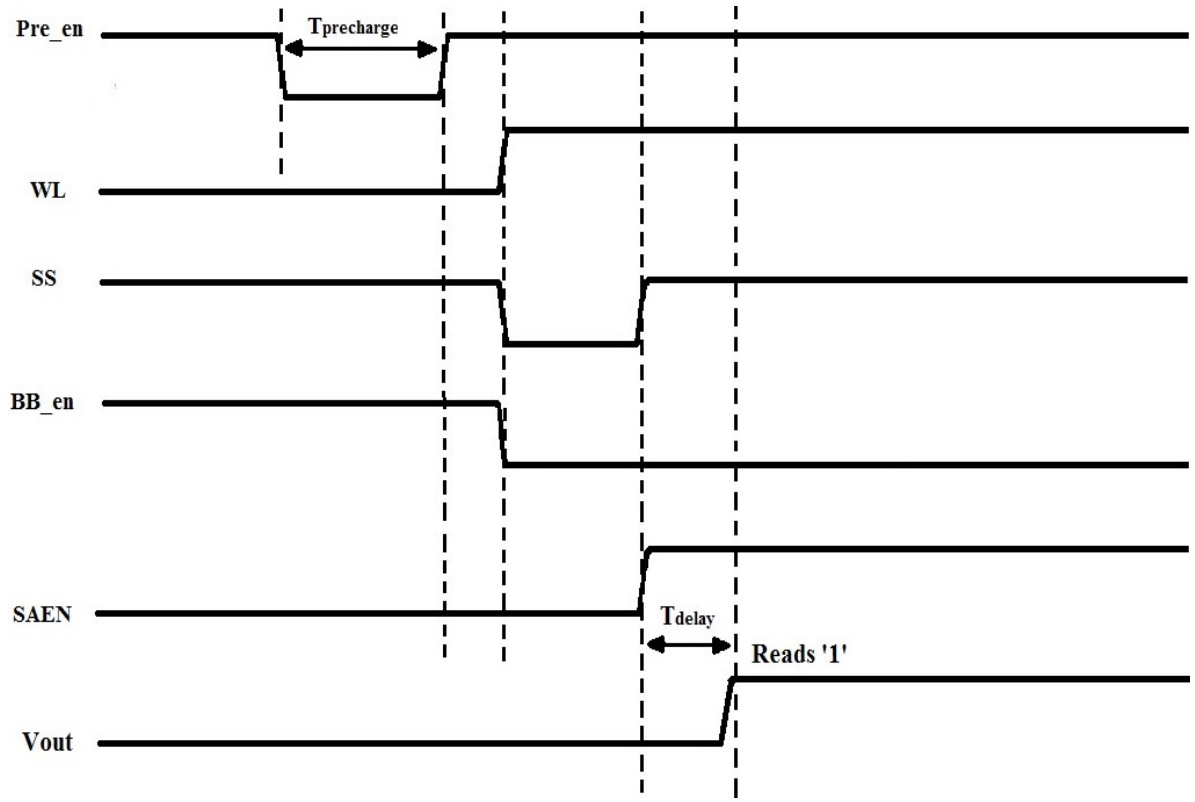

Figure 25. Signal diagram for sense amplifier operation

4.3 Offset and Delay Estimation

One of the major performance metrics in SA design is the offset and hence its estimation is a crucial task. The offset of an SA characterizes its performance. Offset for a voltage sense amplifier is measured in terms of the minimum voltage differential between the bit line nodes which is able to trigger a correct sensing. An ideal SA exhibits infinitely small offset but in all practical designs there is always a finite offset. This finite offset if not accounted for in the design of the SA, can cause failure in memory. Therefore, it is a critical task to characterize and account for this finite offset in order to ensure the efficient working of memory and avoid any failures. Hence sense amplifier offset must be as low as possible. Along with incorporating any mismatch in the values of the V_{TH} of the devices due to parameter variations, the SA requires a minimum voltage difference V_{diff} between the bit lines. This V_{diff} needs to be above a certain threshold for the circuit to operate efficiently. To identify the effect of process variations, 1000 Monte Carlo (MC) simulations were performed by incorporating a total of 10mV variation in VTH of devices. In the current situation, voltage offset is defined as the minimum voltage difference between the bit lines in order to achieve a full success rate (i.e. success rate of 100%). The success rate (S), can therefore be formulated as:

$$S = \frac{\text{No. of correct sensing of SA Output}}{\text{Total No.of Trials (N)}} \tag{4.2}$$

Furthermore, the read access time, T_{Access}, which signifies the time utilized by the SA to read the contents of memory, in the case of NVM can be expressed as:

$$T_{Access} = T_{Precharge} + T_{Delay} + T_{Latching} \tag{4.3}$$

As seen from equation (4.3), the read access time has been divided into three parts namely $T_{Precharge}$, T_{Delay} and $T_{Latching}$. Apparently, appropriate design and topology of SAs can enable the control of first two parts, namely $T_{Precharge}$ and T_{Delay}, while latching time is fixed depending upon the output load and buffering time. As the load on memory has increased over the years, this latching time $T_{Latching}$ has also been observed to increase. In the present work, sensing delay, $T_{Sensing\ delay}$, is predicted by utilizing the transient behaviour of SA at the estimated current and voltage offset for 100% success incorporating V_{TH} variations in all devices in order to account for any variation in parameters at the fabrication stage.

4.4 Simulation Results and Comparison

The proposed SA is implemented in 65nm CMOS bulk technology at a supply voltage (Vdd) of 1.2V. The selected aspect ratios for respective transistors of the core sense amplifier circuit are mentioned in Table VII. In order to demonstrate the performance of sense amplifier with respect to Offset Voltage of sense amplifier, Monte Carlo (MC) simulations are launched by incorporating a total threshold voltage variation of 10mV in all the devices and simulations were carried out in ELDO simulator with a bit line capacitive load of 1pF at temperature 27^{o}C.

The term yield (S), is determined at each differential bit line voltage by projecting SAEN signal for specific differential bit line voltage by running 1000 MC simulations using equation (4.2). It is important to note that the analysis carried out here in terms of voltage offset assumes that the transistors in the designed circuit are with random mismatches. The Probability Density Function (PDF) of the sensing probability for the proposed SA at supply voltages of 1.2V is shown in Figs. 26. The corresponding bit line voltages were maintained at 1.2V. The triggering of sense enable signal has crucial effect on the access time of the SA and therefore MC simulations were carried out by including V_{TH} variations and applying SAEN signal at various time instants till 100% success is achieved in order to calculate the Cumulative Distribution Function (CDF)**.** Samples from the CDF were fit into the nearest Gaussian distribution function to obtain the Probability Density Function (PDF) which is given in equation (4.4):

$$f(x) = a1 * e^{-((x-b1)/c1)^2} \tag{4.4}$$

After the analysis, the coefficients in equation (4.4) are obtained as a1=0.201, b1=14.412 and c1=5.545. It can be observed that 100% success in sensing the correct value is achieved when SAEN signal was applied at 5.12ns, at a differential bitline voltage of 18.014mV which is shown in Fig. 26. In addition, the standard deviation in the offset voltage is 3.454mV.

TABLE VII. Transistor Aspect Ratios

Device	W/L (µm)
M1=M2	0.4/0.09
M3=M4	0.9/0.09
M5=M6	9/0.09

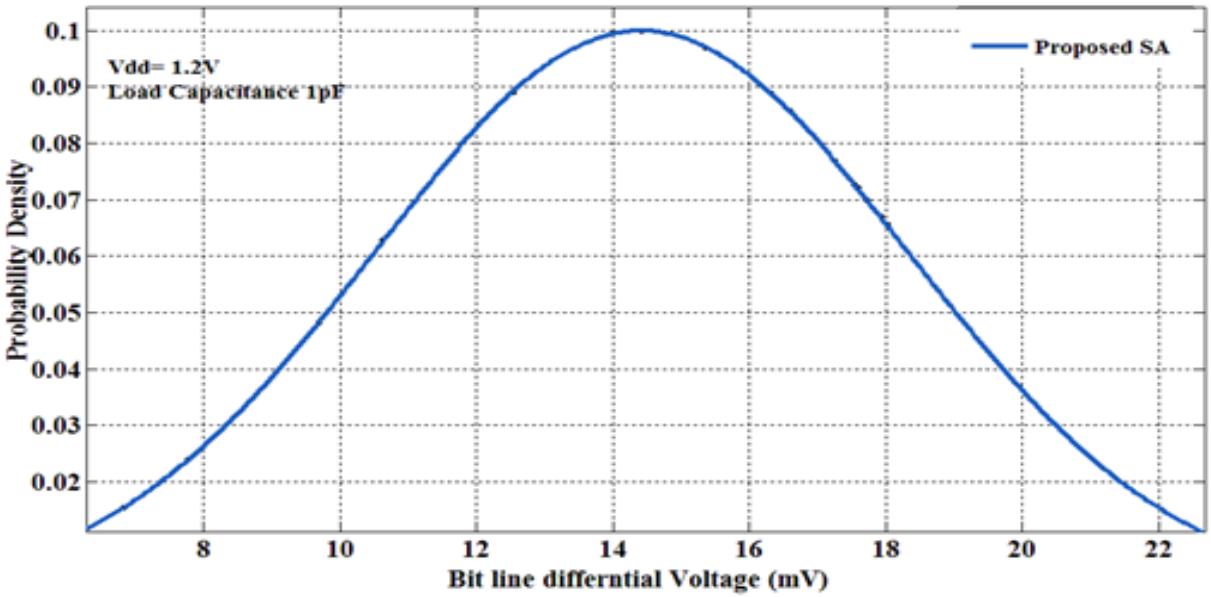

Figure 26. Probability Density function of the Sensing Probability

The transient behaviour of the modified sense amplifier topology is shown in Fig. 27. It can be observed from Fig. 27 that the sense amplifier achieves sensing delay of 42.432ps when a single cell was simulated. Fig. 28 and Fig. 29 compare the performance of the cross coupled latch type SA with body biasing and without body biasing in terms of sensing delay and voltage offset performing MC simulations and incorporating V_{TH} variations. From Fig. 28, an improvement of 10.861% for the circuit with body biasing has been observed over the circuit without body biasing in terms of sensing delay. From Fig. 29, an improvement of 8.098% has been observed for the circuit with body biasing over the circuit with no body biasing in terms of voltage offset. It can be observed from Fig. 28 and Fig. 29 that the worst case sensing delay and voltage offset for the proposed SA are 42.127ps and 18.014mV respectively, in the presence of V_{TH} variations. Furthermore, in order to check the effectiveness of the sense amplifier with respect to sensing delay, V_{TH} variations were introduced in all the devices using Monte Carlo method at a supply voltage of 1.2V and temperature of 27°C. The distribution of the obtained sensing delay and offset voltage is shown in Fig. 30 and Fig 31. In order to verify the effect of supply voltage on offset voltage, the power supply of the sense amplifier is swept from 0.9V to 1.6V. It is observed from Fig. 32 that with increase in supply voltage the bitline differential voltage increases thus decreasing the sensing delay which is an intuitive result [6]. Subsequently, the sense amplifier is assessed in order to check the distribution of offset voltage under Max-Max and Min-Min corners. The outcome is plotted in Fig. 33 and Fig. 34. It can be observed from the plots that when both the NMOS and PMOS are Max-Max, the bitline differential voltage has lower values as compared to Min-Min corners because of higher leakage thus resulting in the corresponding values of sensing delay for both the corners.

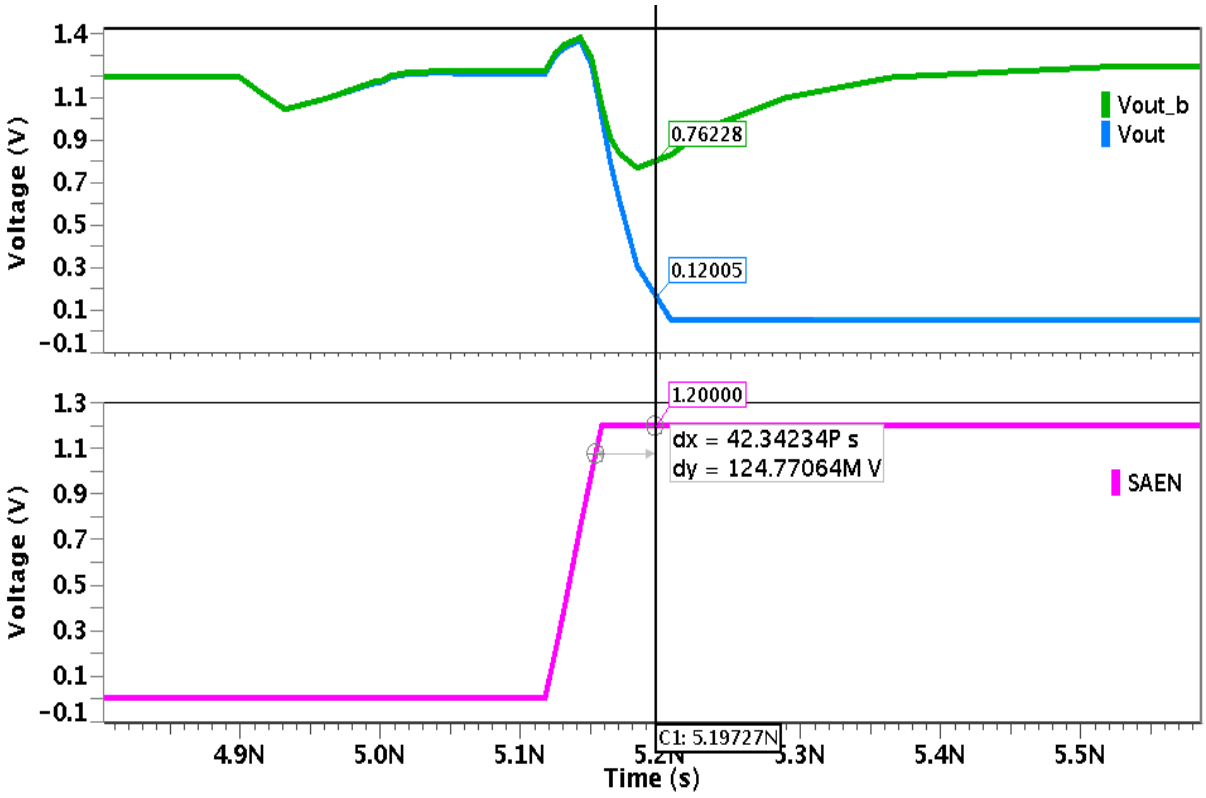

Figure 27. Output Waveform for the proposed Sense Amplifier

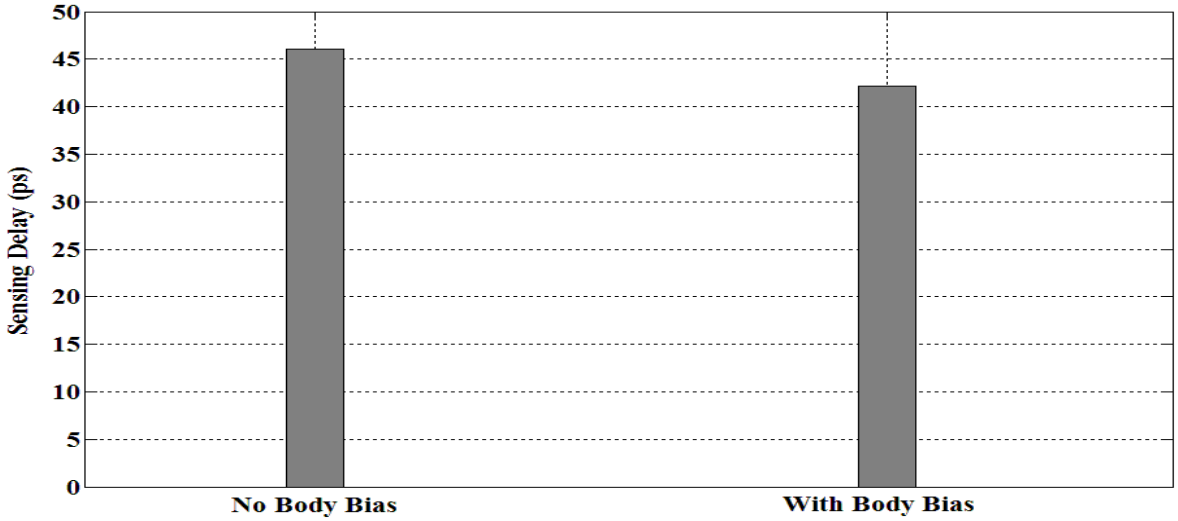

Figure 28. Sensing Delay for the proposed Sense amplifier with and without Body Biasing

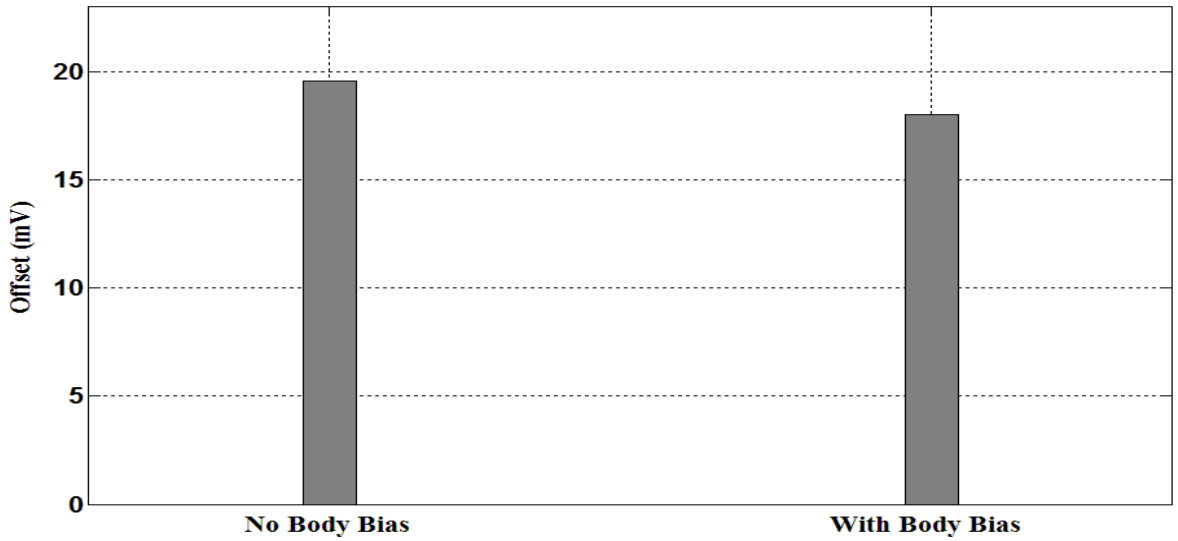

Figure 29. Offset Voltage for the proposed Sense Amplifier with and without Body Biasing

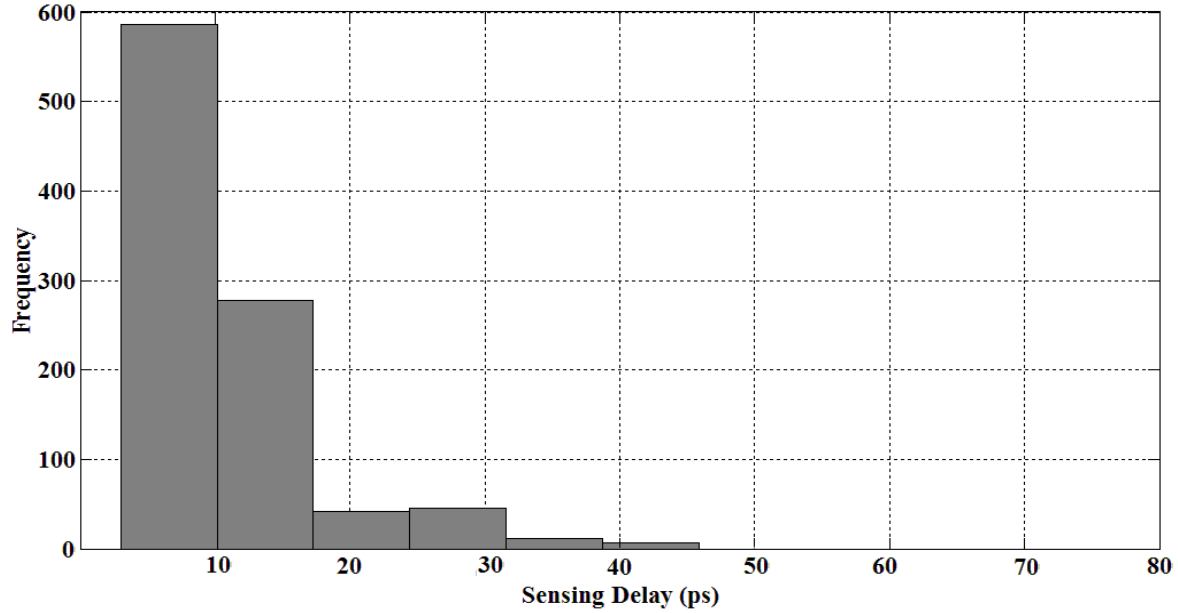

Figure 30. Histogram depicting variation in the values of Sensing Delay

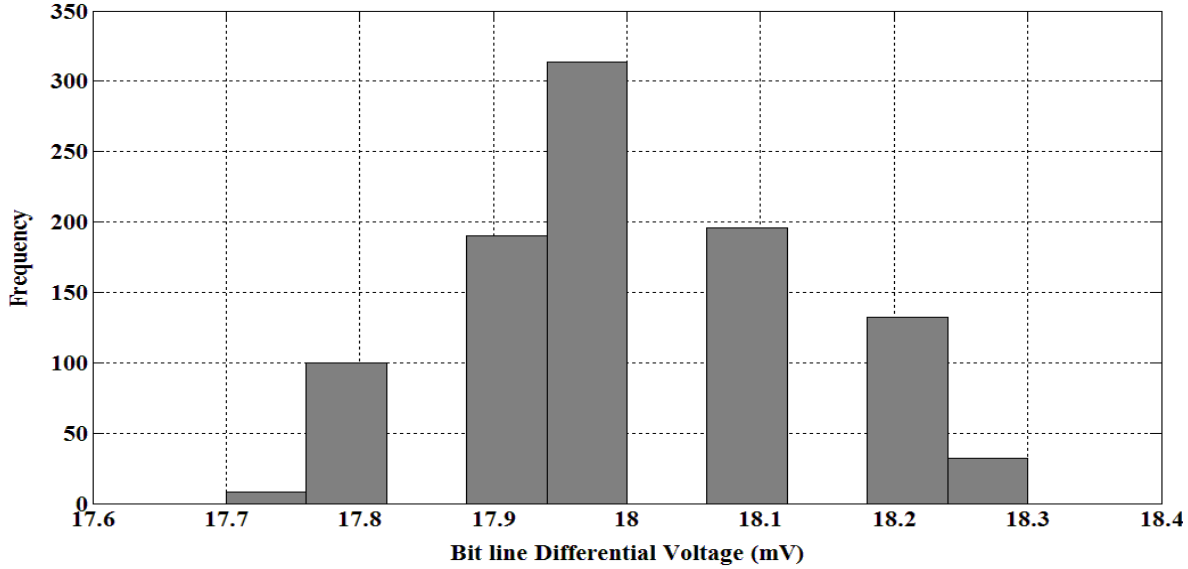

Figure 31. Histogram depicting variation in the values of Offset Voltage

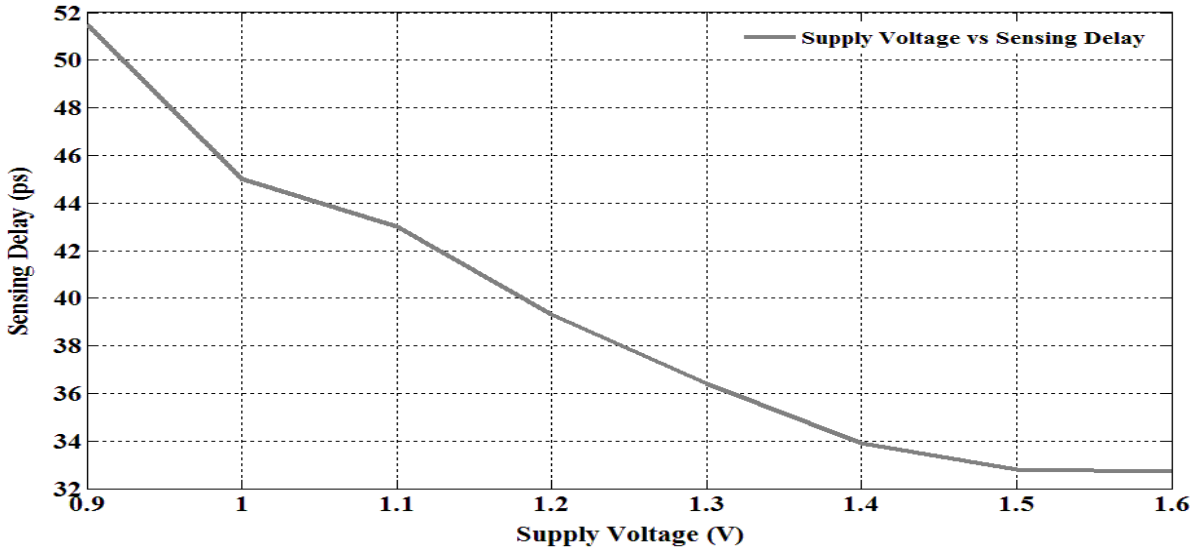

Figure 32. Variation in Sensing Delay with Supply Voltage

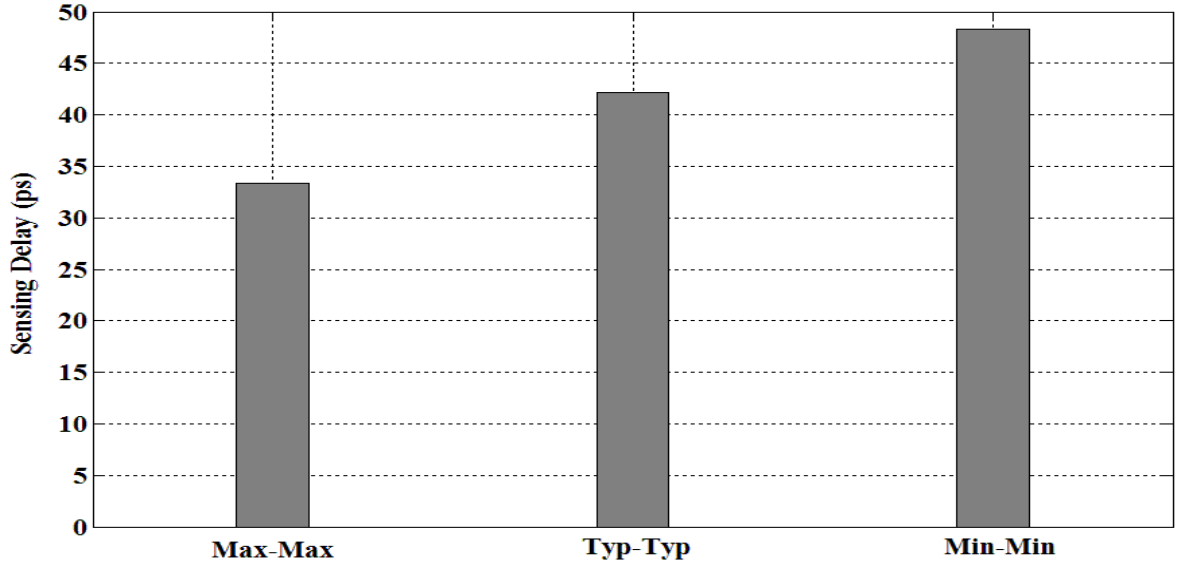

Figure 33. Sensing Delay for various pvt corners

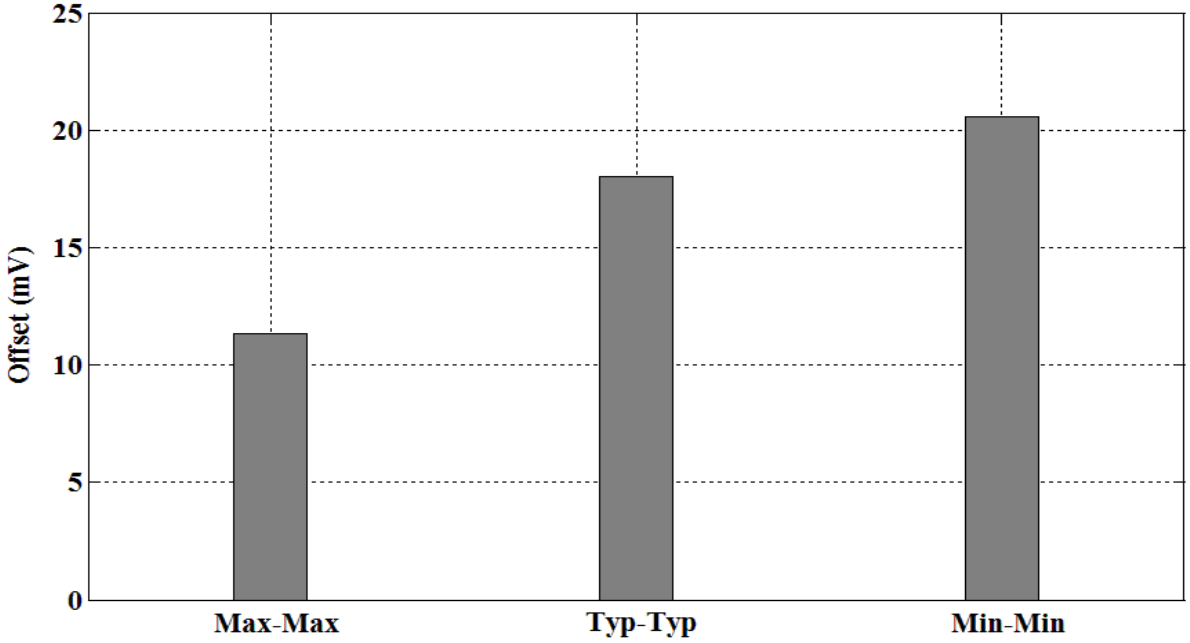

Figure 34. Offset Voltage for various pvt corners

In order to compare the performance of the proposed sense amplifier with other conventional designs, an equivalent supply voltage of 1Vis fixed and the bit line load is varied according to the setup given in [18]. The results for the sensing delay are shown in Fig. 35. The SA proposed in this paper shows favourable results as compared to the Current Mirror Sensing (CMS) scheme proposed in [18]. In order to demonstrate the effectiveness of the proposed SA with respect to sensing delay, an equivalent bitline load and supply voltage is considered according to the setup given in [19]. The sensing delay values were observed for different bit line differential voltages for the modelled setup and the results for sensing delay are shown in Fig. 36. It can be observed from Fig. 36 that the obtained results for the proposed SA compare favourably with the Charge Transfer Sense Amplifier (CTSA) proposed in [19].

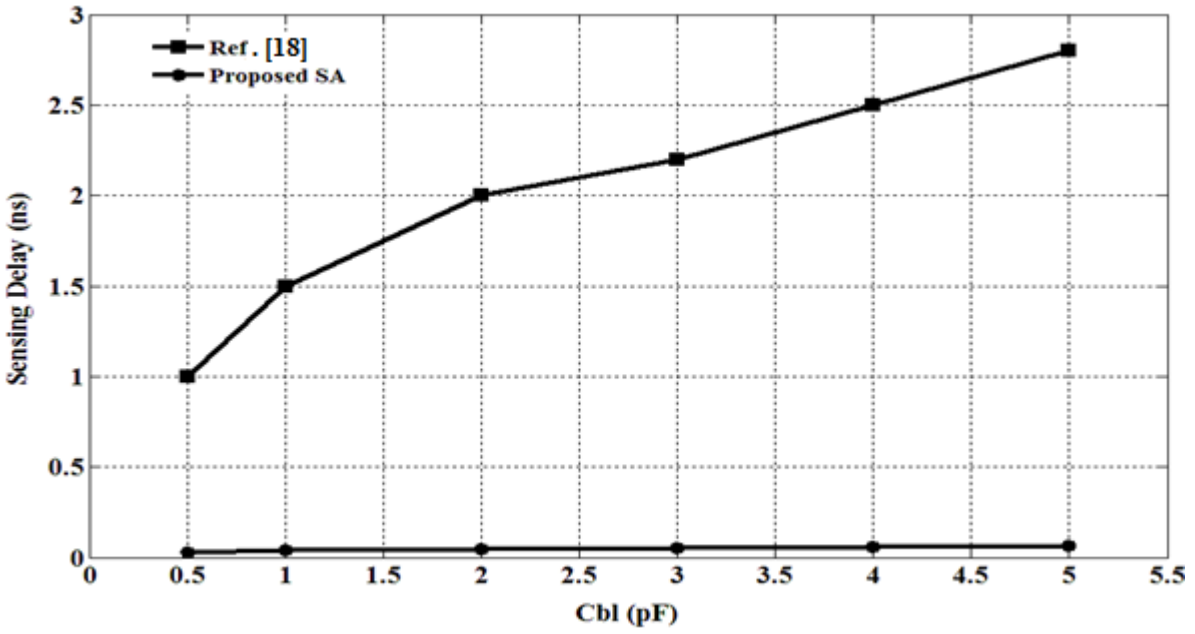

Figure 35. Variation in Sensing Delay with varying Bit line load

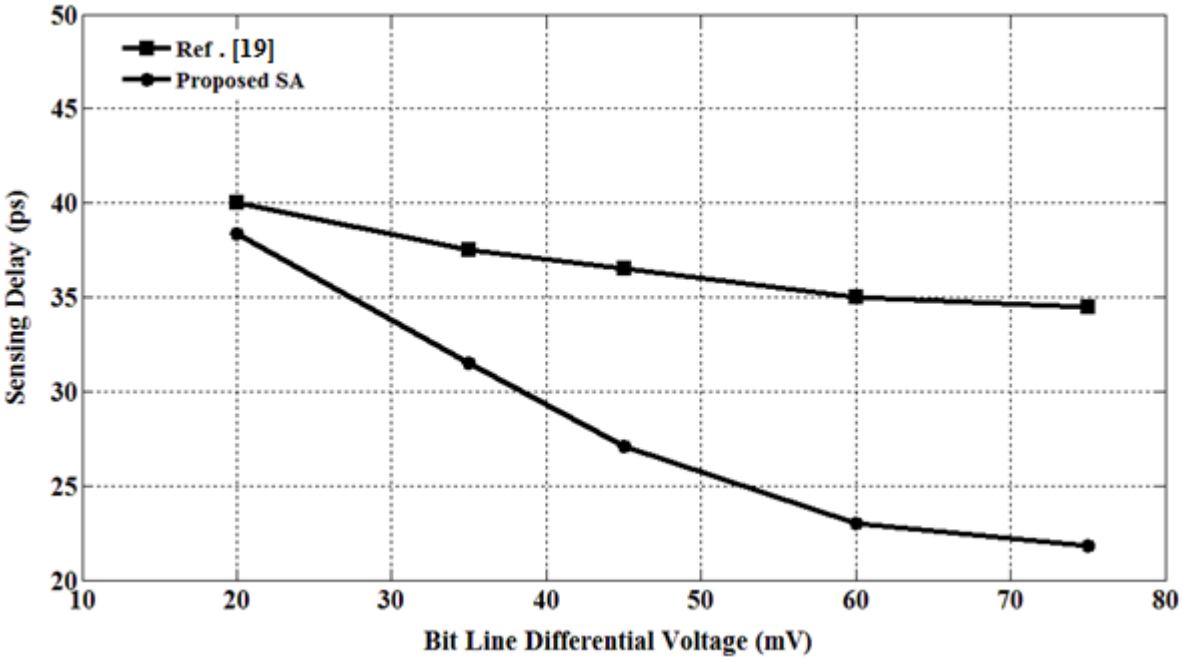

Figure 36. Variation in Sensing Delay with varying Bit line Differential Voltage

Then, the proposed SA is simulated under different supply and temperature conditions, the results for which are mentioned in Table VIII. Finally, the proposed topology was analysed for power consumed by the SA. The total power consumed can be expressed as:

$$P = I_{Avg} . Vdd \tag{4.5}$$

The total power consumed can be thought in two parts namely the power consumed during sensing and the power consumed during differential bitline formation gives as:

$$P_{Total} = P_{Differential\ Forrmation} + P_{Sensing} \tag{4.6}$$

For calculation of the above power values, the average current I_{Avg} through the core of the SA is measured and then the power is calculated. Then the various consumed power at varying supply level was obtained which are given in in Table IX. It can be seen in Table IX that the total power consumed by the SA decreases with decrease in supply voltage. Also it is observed that the proposed SA consumes a total power of 31.996µW at a supply voltage of 1.2V.

TABLE VIII. Sensing Delay and Offset Voltage for varying supply voltage and temperature

Vdd (V)	Temperature (°C)	Offset (mV)	Sensing Delay (ps)
1	27	17	45
1.35	125	17.48	126.55
1.65	27	24.28	33.04
1.95	-40	78.9	90.8

TABLE IX. Power consumed by the proposed Sense Amplifier for varying supply voltage

Vdd (V)	Power consumed during Bit line Differential Formation (µW)	Power consumed during Sensing (µW)	Total Power consumed (µW)
1.2	7.488	24.211	31.996
1.1	3.609	16.846	20.455
1	0.945	10.638	11.583
0.9	0.496	6.248	6.744

4.5 Conclusions

A new SA topology for SRAM has been presented in this work. The proposed topology has been analysed in 65nm CMOS technology operating at a supply voltage of 1.2V and allows operation at supply voltages as low as 0.9V. The proposed topology when compared to the existing topology exhibits better performance in terms of sensing delay and offset voltage. The proposed SA exhibits a sensing delay of 9.921ps and 42.127ps in the best and worst case scenarios respectively at typical corners with a voltage offset of 18.0146mV. The minimum differential bitline voltage for 100% success was found out to be 11.812mV and 20.951mV at Max-Max and Min-Min corners respectively. The SA consumes a total power of 31.996µW at a power supply of 1.2V at typical corners. Thus it can be inferred from the above discussion that, the proposed topology could be potentially useful for high speed, low voltage applications.

5. Conclusions

5.1 Summary

In this work, a detailed study of conventional sense amplifiers for both volatile and non-volatile memories has been done. New sense amplifier topologies have been proposed which user lower power supply, sense outputs faster and consume lesser power when compared to conventional designs. The proposed SA for NVM shows that capacitive coupling with the SA in order to couple the load results in lower power dissipation due to lowering of the coupling effect at nodes, also the proposed SA senses output faster at a lower voltage offset. The proposed SA for SRAM cell shows that when body biasing is used in order to strengthen the positive feedback in the cross coupled SA topology, the SA gives faster results due to lowering of threshold voltages of pull downs.

5.2 Future Work

It has been established that SAs form an integral part of any memory and thus any improvement in the speed, yield and offset of the SA will contribute to significant improvement in the performance of memory circuits and such improvements will help bridge up the gap between processor speed and memory. In the future more such topologies could be explored and the current topologies could be analysed for layout work and chip area.

REFERENCES

[1] R. Bez, E. Camerlenghi, A. Modelli and A. Visconti, "Introduction to Flash Memory," in *Proccedings of the IEEE*, 2003.

[2] G. Campardo, R. Micheloni and D. Novosel, VLSI-Design of Non-Volatile memories, Berlin: Springer-verlag, 2005.

[3] A. Pavlov and M. Sachedev, CMOS SRAM Circuit Design and parametric test in nano-scaled technologies, Berlin: Springer-verlag, 2008.

[4] A. A. A. M, "Design and analysis of a high-speed sense amplifier for single-transistor nonvolatile memory cells," in *IEEE Proccedings G- Circuits, Devices and Systems*, 1993.

[5] B. Emoto and A. A. A. M, "High speed differential sense amplifier for use with single transistor memory cells". 1992.

[6] D. J. Rennie and M. Sachdev, "SRAM Sense Amplifier". USA Patent US 8536898 B2, 17 september 2013.

[7] B. Azeez, T. Xinghai and J. Meindl, "The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability," *IEEE Journal of Solid-State Circuits,* vol. 36, no. 4, pp. 658-665, 2001.

[8] B. Wicht, Current Sense Amplifiers for Embedded SRAM in High-Performance System-on-a-Chip Designs, Berlin: Springer-verlag, 2003.

[9] R. Micheloni, M. Crippa, M. Sangalli and G. Campardo, "The flash memory read path: Building blocks and critical aspects," *IEEE,* April 2003.

[10] A. Conte, G. Giudice, G. Palumbo and A. Signorello, "A High-Performance Very Low-Voltage Current Sense Amplifier for Nonvolatile Memories," *IEEE Jornal of Solid-State Circuits,* vol. 40, no. 2, pp. 507-514, Feb, 2005.

[11] N. N. Wang, "On the Design of MOS Dynamic Sense Amplifiers," *IEEE Transactions on Circuits and Systems,* Vols. CAS-29, no. 7, pp. 467-477, July 1982.

[12] L. Jiang, W. Xueqiang, W. Qin, W. Dong, Z. Zhigang, P. Liyang and L. Ming, "A low-voltage sense amplifier for high-performance embedded Flash memory," *Journal of Semiconductors,* vol. 31, no. 10, 2010.

[13] S. Kajiyama, M. Fujito, H. Kasai, M. Mizuno, T. Yamaguchi and Y. Shinagawa, "A 300 MHz Embedded Flash Memory with Pipeline Architecture and Offset-Free Sense Amplifiers for Dual-Core Automotive Microcontrollers," *IEEE Asian Solid-State Circuits Conference,* pp. 257-260, 2008.

[14] M. K. Seo, S. H. Sim, H. S. Lee, S. W. Kim and I. W. Cho, "A 130-nm 0.9-V 66-MHz 8-Mb (256Kx32) local SONOS Embedded Flash EEPROM," *IEEE Journal of Solid-State Circuits,* vol. 40, no. 4, pp. 866-883, April 2005.

[15] B. Wicht, T. Nirschl and D. Schmitt-Landsiedel, "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.

[16] R. Venkatraman, R. Castagnetti and S. Ramesh, “The statistics of device variations and its impact on SRAM bitcell performance leakage and stability,” in *Proc. 7th Int. Symp. Quality Electronic Design,* 2006.

[17] B. Cheng, S. Roy and A. Asenov, “The impact of random doping effects on CMOS SRAM cell,” in *Proceedings of 30th European Solid-State Circuits*, 2004.

[18] Y. Tsiatouhas, A. Chrisanthopoulos, G. Kamoulakos and T. Haniotakis, “New memory sense amplifier designs in CMOS technology,” in *7th IEEE international conference on electronics, Circuits and systems* , 2000.

[19] S. H. A. A. Manoj Sinha, W. Burleson, R. Krishnamurthy and S. Borhr, “High-Performance and Low-Voltage Sense-Amplifier Techniques for sub-90nm SRAM,” in *SOC conference*, 2003.