Multilevel Sense Amplifier sensing for Phase Change Memory

by

Aakash Tyagi

Under the Supervision of Dr Anuj Grover and Harsh Rawat

Indraprastha Institute of Information Technology Delhi

June, 2020

Multilevel Sense Amplifier sensing for Phase Change Memory

by

Aakash Tyagi

Submitted in partial fulfilment of the requirements for the degree of
Master of Technology

to

Indraprastha Institute of Information Technology Delhi

June, 2020

Certificate

This is to certify that the thesis titled "**Multilevel Sense Amplifier for Phase Change Memory**" being submitted by **Aakash Tyagi** to the Indraprastha Institute of Information Technology Delhi, for the partial fulfilment of the requirements for the degree of Master of Technology, is his original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Dr. Anuj Grover

Department of ECE

Indraprastha Institute of Information Technology Delhi

New Delhi 110 020

Mr. Harsh Rawat

Embedded Memories Group

ST Microelectronics

Greater Noida 201308

Acknowledgement

I would like to express my gratitude to my thesis advisors Dr Anuj Grover of Indraprastha Institute of Information technology and Mr Harsh Rawat of ST Microelectronics for their constant guidance and motivation. Their valuable inputs helped me in various technical aspects of this work. I would also like to thank Mr Promod Kumar for giving me the opportunity to work on this research problem.

I would like to thank Aman Tyagi for helping me in troubleshooting during layout routing. I am also thankful to Prateek Singh who helped me in understating the technical parts of the circuit, particularly the comparator latch. Finally, I would like to thank my family for their constant moral support and prayers which helped me sustained this far.

Abstract

In recent years, there has been a very rapid advancement in the area of phase-changing materials, due to which there has been a substantial improvement in Phase-change memories (PCM) and its technology. Due to scaling limitations in flash memories, PCM seems to be the most promising alternative among the various emerging Non-Volatile Memories (NVM). Their ability to store more than two levels of data can be exploited to store multiple bits within a single PCM cell. In this work, we have defined a sensing scheme for reading the data stored in the PCM cell, which has the capability of storing 2bits/cell.

This sensing architecture has the capability of sensing both the bits in one read cycle, also called parallel sensing. The access time of the proposed scheme is 12.04 nS with a targeted nominal current offset of 1uA. The area of the circuit is $327um^{2,}$ and the average power per reading is 106uW.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Memory Devices

In the modern era where electronics plays a significant role in the life of every individual, it becomes vital that each semiconductor component scale in line with the other components as we move down the technology nodes [1]. However, it was observed that gradually, speed of memory became a bottleneck in the performance of the system as it was not increasing with the same rate as that of the processing or computational speed (Fig.1). The system designers and computer architects are aware of the fact that memory devices play the most crucial role in the overall performance and speed of the system. For example, in the smartphone market, the capacity of memory is used to justify the cost of the product. The power consumption of the product is also highly correlated with the type of memory chips. In personal computing system, there is a recent transition from the magnetic hard drive to the NAND based SSD, which significantly improved the system performance. Along with this, In-memory computing (IMC) [2] has also been explored in the past few years to overcome the bandwidth limitations between processor and memory. Therefore, the selection and type of memory design in the system have become very important in today's age.

Embedded memories can be broadly classified into two subcategories i.e. volatile and Non-volatile memories. Volatile memories which need constant power supply for data retention can be further classified as SRAM (Static Random-Access Memory) and DRAM (Dynamic Random-Access Memory). In contrast the categories of Non-volatile memories include Flash, ROM, EPROM (Erasable Programmable Read-Only Memory), EEPROM (Electrically Erasable and Programmable Read-Only Memory) and other electron and non-electron-based memories as shown in Fig.1.2. RAM is used in system caches and main memory. Its contents can be altered theoretically infinite amount of times while ROMs do not allow to alter their content at all or in some cases a smaller number of times. Therefore, they are used in places where there is the least requirement of the data to be changed, for example, BIOS code of a computer system. The main advantage of using ROM is its data retention capability. In contrast, the main advantage of RAM is its read/ write speed and its endurance, i.e. its ability to alter content many numbers of times.

An ideal memory must have two main properties i.e. high endurance and data retention capability. Therefore, much research has been done in the field of NVM. The major NVM technologies include Flash memory, Phase Change Memory (PCM), MRAMS, FeRAMS.

The comparison of SRAM, DRAM with Flash and PCM is shown in Table1.1 [7]. Flash memory performs poorly when it comes to speed, and its scaling is not possible beyond a specific limit due to its constructional limitations. Further, the scaling of DRAM is very cost-intensive when compared with other memories. Therefore, there is a need for alternate technology which is faster than Flash, maintains its non-volatile nature and is scalable. An overview of emerging non-volatile technologies is discussed in [3].

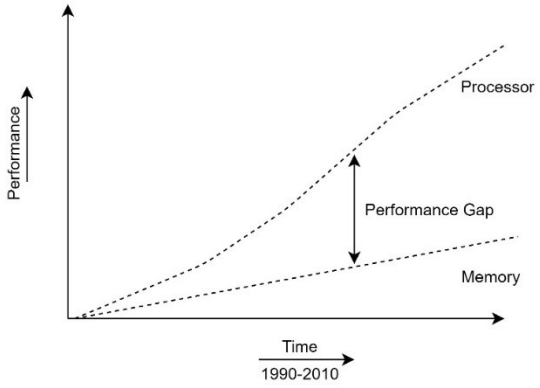


Fig. 1. 1 Memory vs processor performance gap

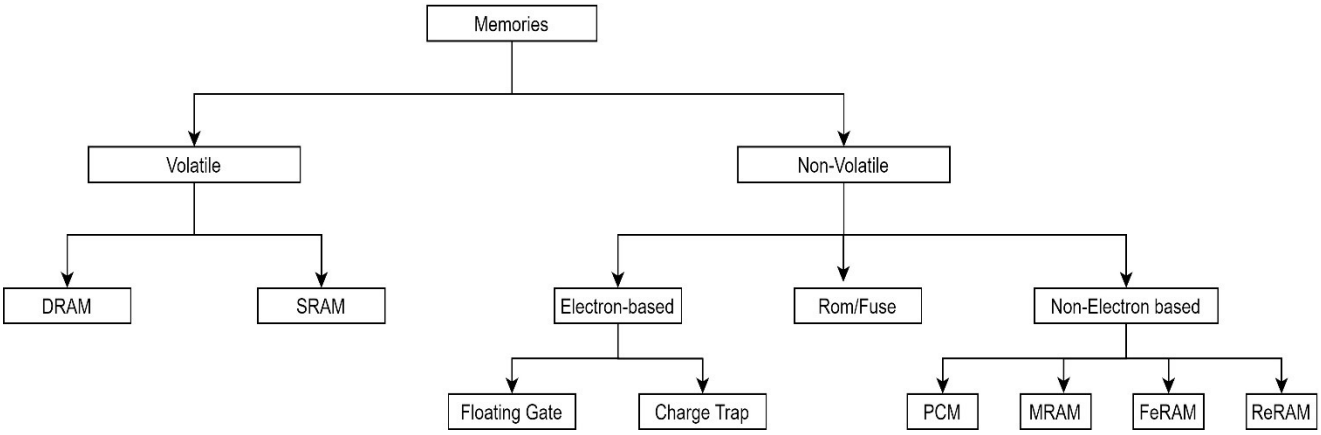


Fig. 1. 2 Classification of memories

Parameters	SRAM	DRAM	FLASH	PCM
Non-Volatile	No	No	Yes	Yes
Read Time(ns)	< 2	30	10^3	2
Write Time(ns)	< 2	50	10^6	10^2
Cell size(F^2)	50-120	6-10	5	4-19
Scalable	Yes	Limited	Limited	Yes
Endurance	Unlimited	Unlimited	10^4	10^6

Table 1. 1 Comparison of different memories [7]

In the following section, we review a mature NVM technology, the flash memory along with its limitations with respect to its scaling challenges.

1.2 Overview of Flash Memory.

The first non-volatile memory was proposed as a Floating Gate (FG) device by Sze SM and Kahng D during 1967 at Bell Labs. The basic circuits of all the three significant memories, i.e. SRAM, DRAM and FLASH, is shown in Fig.1.3. It can be seen that DRAM and FLASH use only 1 transistor to store 1 bit while SRAM uses 6 transistors to store a single bit. Thus, there is a considerable area trade-off when using SRAM. Therefore, it can be concluded that Flash memories are better when area is considered as the deciding parameter.

A flash cell device is shown in Fig. 1.4 (a). The flash cell can be defined as a floating- gate MOS transistor, which is electrically governed by a capacitively coupled control gate [4]. Since the FG is electrically isolated, it stores the charges to modulate the threshold voltage (Vth). In flash memories, the data is stored in terms of the Vth value of the FG device. To program the cell, a high Gate Voltage is applied to attract some of the electrons from the channel to the FG as shown in Table 1.2. These electrons are trapped in the oxide layer of the floating gate even after the gate voltage is removed. This electron trapping increases the threshold voltage of the device when compared to the threshold voltage of an erased cell. Thus, the charge stored in the FG defines the threshold voltage of the flash cell which in turn defines the state of the device.

During the read operation, the drain voltage is reduced to avoid read disturbs. (False writing in the neighbouring erased cells). A reference cell is used to compare the programmed and erased cells. This reference voltage is compared with the voltage of the cell to be read using a sense amplifier. For low Vth (Erased cell) the current is more as compared to high Vth (programmed cell) cell.

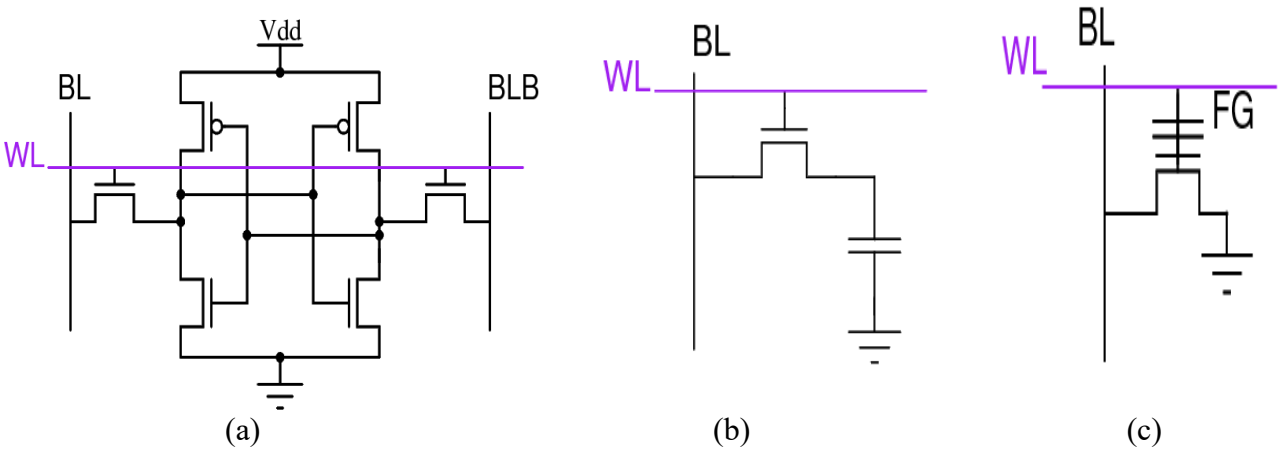
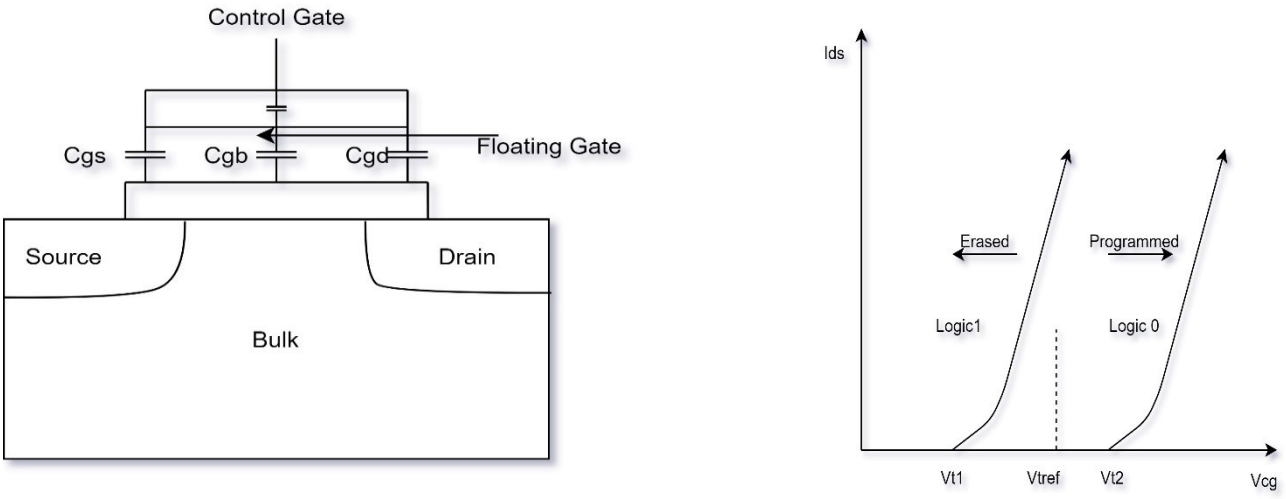


Fig. 1. 3 Schematic of various memory devices

(a) SRAM (b) DRAM (c) Floating gate Flash cell



(a) (b)

Fig. 1. 4 (a) Structure of Flash cell (b) Flash cell characteristics

Operation	Selected Sector				Non-Selected Sector			
	Gate	Drain	Source	Bulk	Gate	Drain	Source	Bulk
Read	4.5	SA	0	0	0	SA	0	0
Program	8.0	5.0	0	0	0	Float	0	0
Erase	-8.0	Float	8.0	8.0	0	Float	0	0

Table 1. 2 Cell voltages for different operation of the flash cell [15]

The next section deals with various challenges associated with the flash memories and the need for new NVM technology.

1.3 Need for new NVM technology.

In recent years, flash memories have seen a constant trade-off between the scaling of its lateral dimensions and maintaining the coupling between the floating gate and the control gate [5]. Due to device scaling, the oxide layer becomes so thin that it becomes under constant electrical stress. There have also been challenges related to Stress-Induced Leakage Current which arises due to large programming voltage stress over ultra-thin oxides. As discussed in [6], there are three key areas in terms of challenges for flash memory scaling.

1)Physical Scaling- Process dependent and defined by lithography

2)Electrical Scaling - Defined by the voltage requirements for read/program and erase.

3)Reliability Scaling - Defined by device physics of the cell.

To overcome these challenges, many alternative designs have been proposed. These designs generally involve replacing the floating gate poly with some other charge trapping layer, such as SONOS (Silicon-Oxide-Nitride-Oxide Semiconductor) structures [14]. The problem with such SONOS devices is that they use very thin tunnel and blocking oxides, and thus suffer from various data retention issues. Therefore, as the scaling of flash memory occurs, the design rules become more constrained. The cell to cell interface and Hot carrier disturbances increases as we scale down and the design rules shrink.

Considering the above-mentioned challenges in flash memories, there has been a need for replacing flash memories with newer NVM technologies. As shown in Fig. 1.2 several possible candidates are under consideration. These include technologies such as FeRAM, MRAMs, ReRAM and PCM.

In this work, we will consider PCM as a viable replacement for flash memories and propose a reading mechanism for Multilevel Cell storage with this technology. This work is divided into 7 chapters. In chapter 2 we will discuss the PCM cell and its characteristics along with its array and read operation. Chapter 3 introduces some conventional sense amplifier schemes along with various figures of merit of sense amplifier. In chapter 4 we propose a delay programmable sense amplifier which can be used to sense the read current of the PCM cell, however this scheme is not feasible for multilevel cell sensing therefore a new modified current mirror-based sense amplifier is proposed in chapter 5. The multilevel sensing scheme using modified current sense amplifier is proposed in chapter 6 and finally the conclusion and future work is given in chapter 7.

2 Phase Change Memory

The advancements in PCM technology in recent years has improved to a higher degree with some very promising results in terms of scalability. It has even surpassed the older technologies like MRAM and FeRAM in demonstrated scaling [5]. PCM has gained a significant share due to its high scalability and better read/write latency than flash memories. There has been an increasing amount of research publications in the field of non-volatile memory [7], particularly the PCM due to the advancements in material physics and device technology. However, we need to verify the reliability of the PCM at the array level to consider it a viable technology for the future

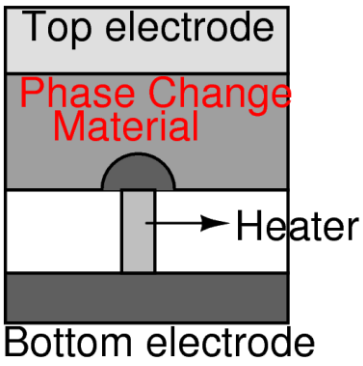
The operation of PCM is based on the large contrast between the resistance of the Phase change material between the SET and RESET states of the cell. The two states, i.e. amorphous and crystalline, has an order of magnitude difference when it comes to resistance. The resistance of the amorphous state is more than that of crystalline state with 3 to 4 order difference [5]. This enormous contrast between the resistance can also be exploited for MLC (Multi-Level Cell) where we can have multiple levels of resistance to store the data between the 2-bit SET and RESET states. These different analog current levels are required for MLC. In the following section, we will discuss the PCM cell and its characteristics.

2.1 Phase Change Memory Cell

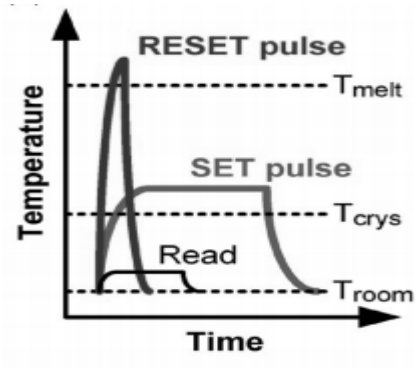
It is observed that the alloys of group VI elements of the periodic table (called chalcogenides) are stable at room temperature in both the amorphous as well as crystalline states. This property of the elements is exploited to make the PCM cell. The most promising of the alloys are GeSbTe, also called GST. They can reverse between the high resistance amorphous state and low resistance crystalline state in a very short duration of time (order of few nanoseconds).

PCM cell is composed of a thin film chalcogenide material resister that changes the magnitude of its resistance depending on the phase of the GST material in the active region. In order to switch between the phases, the local temperature needs to be altered. When the temperature reaches above the crystalline temperature, the crystal nucleation occurs, and chalcogenide goes to the crystalline state (set operation). For RESET operation the material is heated above the melting point and then very quickly quenched not to let the material crystallize. The SET pulse is current pulse just above the crystallization temperature and it sets the PCM cell to low resistance [7]. The

current pulse heats the cell portion. The SET pulse is directly related to the write speed of the device because it needs to be wide enough so that crystallization can occur. The set-reset pulses for PCM programming are shown in fig 2.1(b) [7]. Fig.2.1(a) shows a PCM cell. It contains a top and a bottom electrode with the GST material infused between them. The heater is used to increase the temperature of the GST using electrical pulses. The area directly connected to the heater is called the active region of the cell.



(a)



(b)

Fig. 2. 1 (a) PCM cell cross-section schematic. [7] (b) Set and Reset pulse of PCM [7]

Fig.2.2 shows the voltage vs current characteristics of the PCM cell [7] in both the crystal and amorphous phase. The region below the threshold switching is used for reading, since there is a big difference in resistance between both the states of the device. Here it is seen that the reset state remains in a high resistance state for voltages below Vth. If the voltage above Vth is applied for a duration longer than the crystallization time, the memory switching takes place and cell reaches the low resistance state. In the next section, the architecture of the PCM array is discussed.

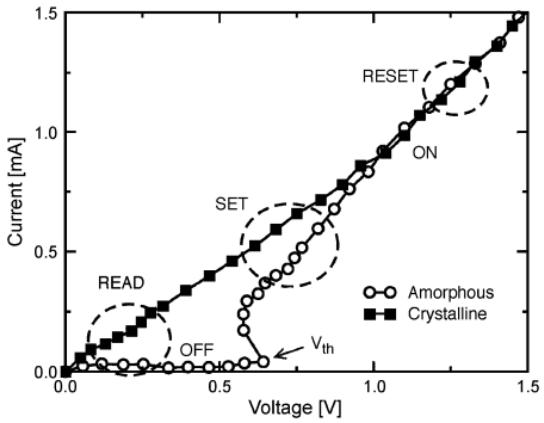


Fig. 2. 2 I-V characteristics of PCM cell. [7]

2.2 PCM Array

There are several parameters in consideration before making the architecture choice. These parameters include power, read and program latency and so forth. The high-density array needs to be fragmented into several sub-arrays to minimize the leakage, power and control the speed. The PCM array architecture is shown in fig.2.3 [8]. The array core is divided into several tiles, each having a column decoder, a row decoder and a local Sense Amplifier. These tiles are grouped horizontally making a partition [8]. Tiles comprise of word lines and bit lines. Bit lines are connected to the sense amplifier for reading while the word lines are used to select one or more cells in parallel during read and write operations. To activate the cell, the access circuitry diode needs to be in forward biased.

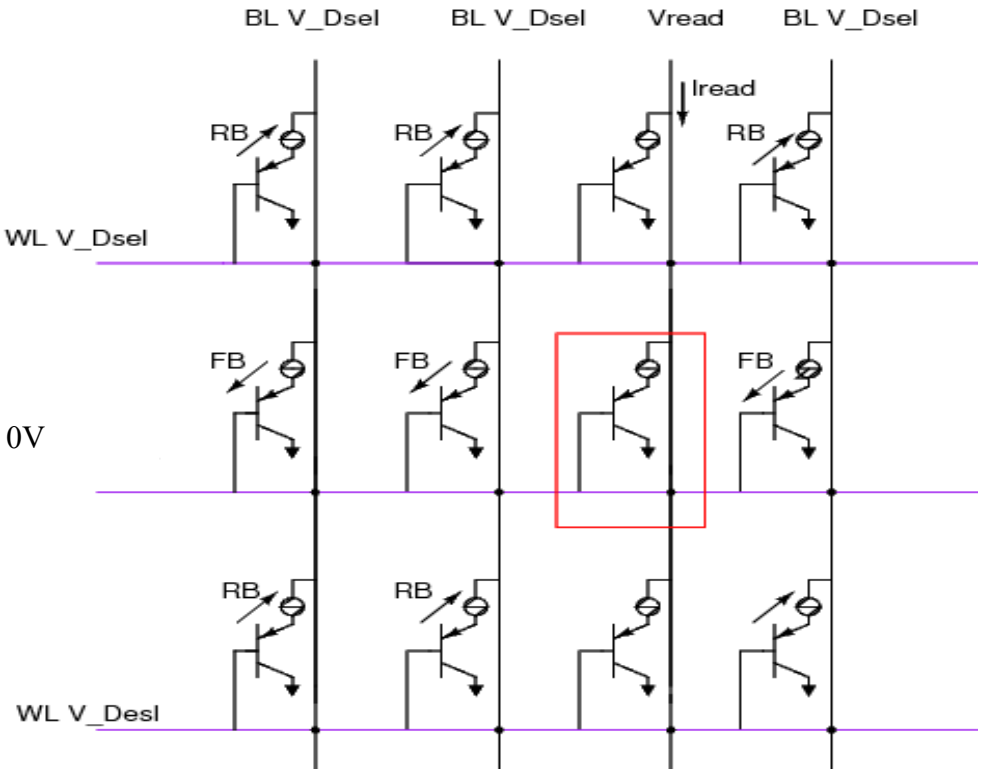


Fig. 2. 3 PCM Cell array

The tile size is limited by the IR drops on the Bit line and word line. Thus, to determine the tile size, these two drops need to be considered. Increasing the physical size of wordline and bitline will, in turn, increase the IR drops, which increases the programming voltage to be delivered to get the required amount of programming current. This will also increase the leakage from the unselected cells.

2.3 Access Circuit

We require an access device that is used to access the PCM cell. This device can be either a diode, a field-effect transistor or a bipolar junction transistor [5].

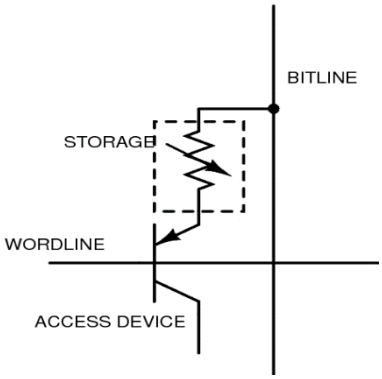


Fig. 2. 4 PCM cell with BJT access device

The advantage of using such a device is to reduce the leakage current of the unselected cells. The primary consideration while designing the access device is if it can provide sufficient current to RESET the cell. Also, the designer needs to keep in mind that due to high programming currents required for PCM, the voltage drop along the metal lines, i.e. the wordline and bitline will account to significant power loss. Thus, the array is generally split to reduce this loss. This splitting of the array can cause the chip real estate to reduce, thus reducing the density. One such access circuitry is shown in fig.2.5. It uses a PNP BJT, with its base connected to the word line and emitter connected to the storage element of the PCM cell.

It must be noted that this BJT selector has a very strong implication on the organization of the PCM array. The main difference between FG device and this structure is the word-line drop due to the base resistance of the BJT. Therefore, the access circuit is a very important consideration while designing the PCM array.

2.4 Read Operation in PCM

For a read operation, each idle tile is biased such that its diode access circuits are reversed biased. The wordline is biased at V_{WL}, and all bitines are left floating or grounded. To select a cell, the particular wordline is driven low, and bitline is driven high such that the GST structure has a voltage equal to the voltage of BL minus the diode drop.

$$V_{gst} = V_{BL} - V_{diode} \quad (2.1)$$

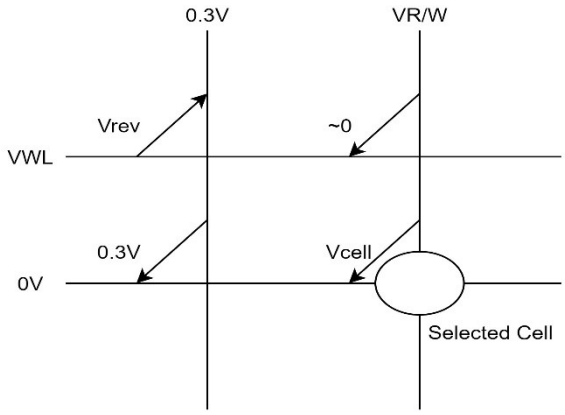


Fig. 2. 5 Neighbouring cells during read.

The read current flows in the bitline of the selected cell. From equation 2.1 we can calculate the bitline voltage to be equal to the sum of the voltage across GST and the diode voltage drop. The typical range for bitline voltage is 1.2V-1.6V. The minimum bitline prechagre voltage is a function of the voltage across the diode of the access circuit. The read current is compared with a reference current using a sense amplifier, and the output is generated.

It must be noted that in section 2.1, we discussed that PCM stores data in terms of two analog levels of resistances. Each resistance level gives a different value of cell current which is used to distinguish between the set and reset state during the read operation. However, it is possible that more than two analog levels used to distinguish between the two cells which are called multilevel cells (MLC). Therefore, to increase the array density, we use MLC. In this work, we have discussed the challenges in the current sensing schemes for MLC read and proposed a new circuit which overcomes these challenges. In the next chapter, we will discuss some basic sensing schemes and figures of merit of Sense amplifiers.

3 Sense Amplifier Design

Sense amplifier is one of the most crucial circuits in the memory design. During the read operation, sensing a particular bit (the state of PCM) is done using a sense amplifier. The content stored in the memory is computed by comparing the current drawn from the memory on the matrix side with the current drawn by the reference side under constant bias conditions. The sense amplifier outputs are fed to the output buffers. Thus, if there is an error in the sensing, it leads to an erroneous bit at the output. It becomes essential to design a sense amplifier which is highly robust towards variations in temperature, process and voltages.

The sense amplifier is broadly classified as static sense amplifiers and dynamic sense amplifiers. The static design is a latch-based design that samples the difference between the two bitlines once, while dynamic design continuously checks the difference between the two-bit lines.

3.1 Differential Amplifier

Fig.3.1 shows a differential sense amplifier which is of dynamic voltage mode type. The bit lines are connected to the M1 and M2 devices, whose drains are connected to active PMOS current mirror load (M3 and M4). A bias current source is provided as footer NMOS. The main advantage of this type of sense amplifier is its easy timing requirements, simple design and reliability; thus, it is used in many systems. It can be turned on at the same time with wordline. The current I_{ss} flows continuously through the bottom NMOS.

Consider a case when the input voltages are exactly equal, then equal current ($I_{ss}/2$) flows in both halves of the amplifier. P1 and P2 must have equal size and must be matched. When there is a differential voltage due to discharging of either of the bit line, the gate voltage of one of the transistors reduces, thus decreasing drain current of that half. Since the total current needs to be I_{ss}, the current on the other half increases. For example, consider BL is more than $\overline{BL}$ which causes Id1 to increase by ΔI and Id2 to decrease by ΔI. Since Id3= Id1 and Id4=Id3, current Id4 also increases by ΔI. Therefore, the current at Iout is increased by 2 ΔI. However, the main disadvantage of this circuit is that we need to provide a constant bias, thus at all time, current I_{ss} flows, which consumes a high amount of power.

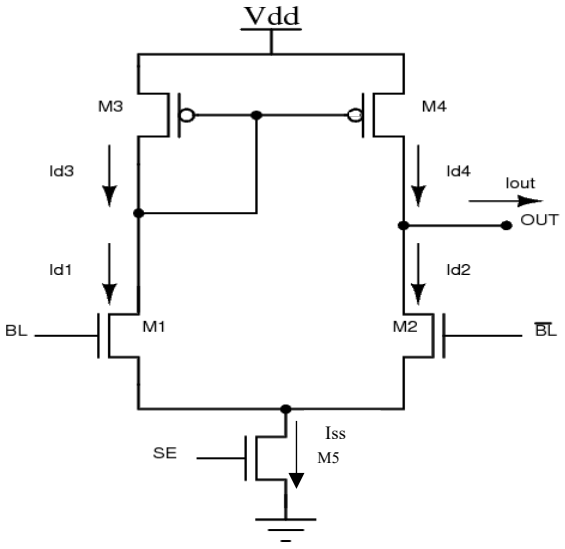


Fig. 3. 1 Differential amplifier

3.2 Latch Type Sense Amplifier

Fig 3.2. shows a basic latch type sense amplifier. It is a static voltage mode sense amplifier. The circuit is made of two cross-coupled back to back inverters. These inverters use positive feedback for latching operation.

The difference between the differential amplifier and latch type amplifier is that the latter is turned on only after there is sufficient differential between the bit lines. This is to avoid the output to get stuck in the metastable state. The pass transistors are used to decouple the amplifier input and outputs from the bit lines. To turn on the sense amplifier SE signal is turned high. When there is sufficient differential (offset) voltage between the BL and BLB, it means that there is sufficient bias for the inverters to latch the output. At this instant, SE is turned on, which enables the sense amplifier and decouples the input/output nodes from the bit lines. The challenge with such a scheme is that one must account the voltage drop of the pass transistors because a slight mismatch between the pass gates devices during fabrication might lead to high errors in reading.

3.3 Figures of Merit

Access Time

This is one of the essential parameters for Sense amplifier as it is directly related to the read time of the memory. Thus, the performance of memory highly depends on the Access time. It is the total time taken from pre-charging stage to the state when the sense amplifier latches the output.

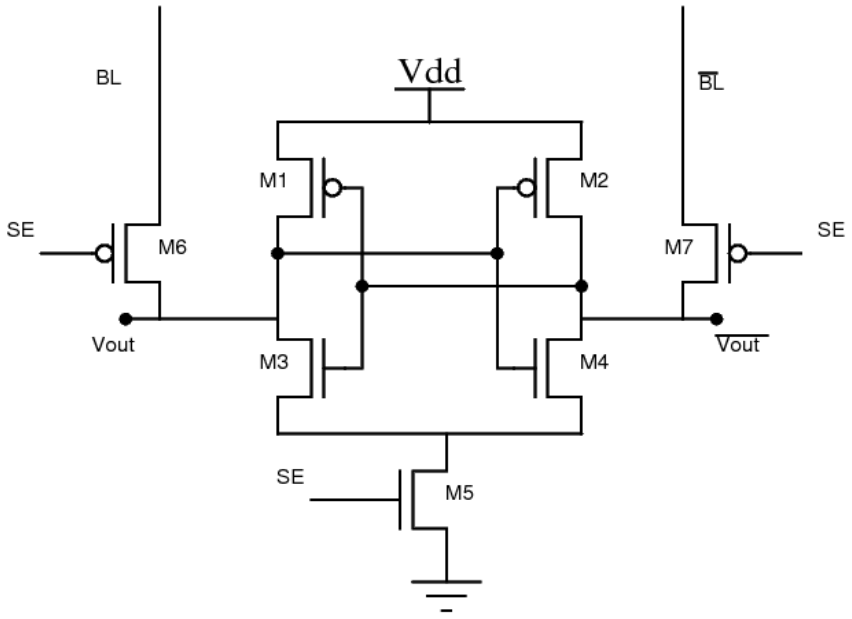


Fig. 3. 2 Latch type sense amplifier

$$T_{Access} = T_{Precharge} + T_{Discharge} + T_{Latch} \quad (3.1)$$

The precharge time depends on the bitline capacitance. In the case of PCM, bitline capacitance is reduced by using tiles. $T_{Dischagre}$ depends on the bit line capacitance and cell current, while T_{Latch} depends on Sense amplifier and latch design.

Power

Power is another critical parameter for sense amplifier design, which becomes significantly crucial in the context of multilevel sensing as the number of reference increases. The sensing power per reading cycle is given by:

$$P_{AVG} = V_{dd}\, I_{AVG} \quad (3.2)$$

where I_{AVG} is the average current flowing in the Sense Amplifier, which is calculated by integrating the total current consumed per read cycle.

Offset

As discussed earlier, there is a minimum differential voltage required between the bit lines so that the sense amplifier can work. Ideally, the offset of the sense amplifier must be zero. However, due to process variation at the sub-micron level, there is a minimum offset that needs to be overcome before turning on the sense. Offset is also directly correlated with the access time of the sense amplifier. Lesser the offset less will be the time taken for the bit line to reach the sufficient differential voltage. For Multi-level sensing, this offset becomes extremely important since the sensing margin between the two different levels is less.

Area

In NVMs most of the area is occupied by the array. The sense amplifiers are placed after the column decoders to reduce the number of sense amplifier requirements (impacts access time). However, in the case of 2bits/cell sensing, the area required by the reference matrix increases three folds.

In this chapter, we have discussed the basic sense amplifier schemes figures of merit of the sense amplifier. In the following chapter, we discuss a delay programmable sense amplifier that is used in PCM for reading the bitcell.

4 Delay Programmable Sense Amplifier

The main disadvantage of using latch type sense amplifier is its high offset. The variations around the PVT corners for such sensing scheme is too significant, and thus this type of sensing scheme cannot be used in multilevel sensing. We propose a sense amplifier with offset cancelling and charge sharing mechanism which can be used for multilevel phase change memories. Fig. 4.1 shows the schematic of the proposed sense amplifier.

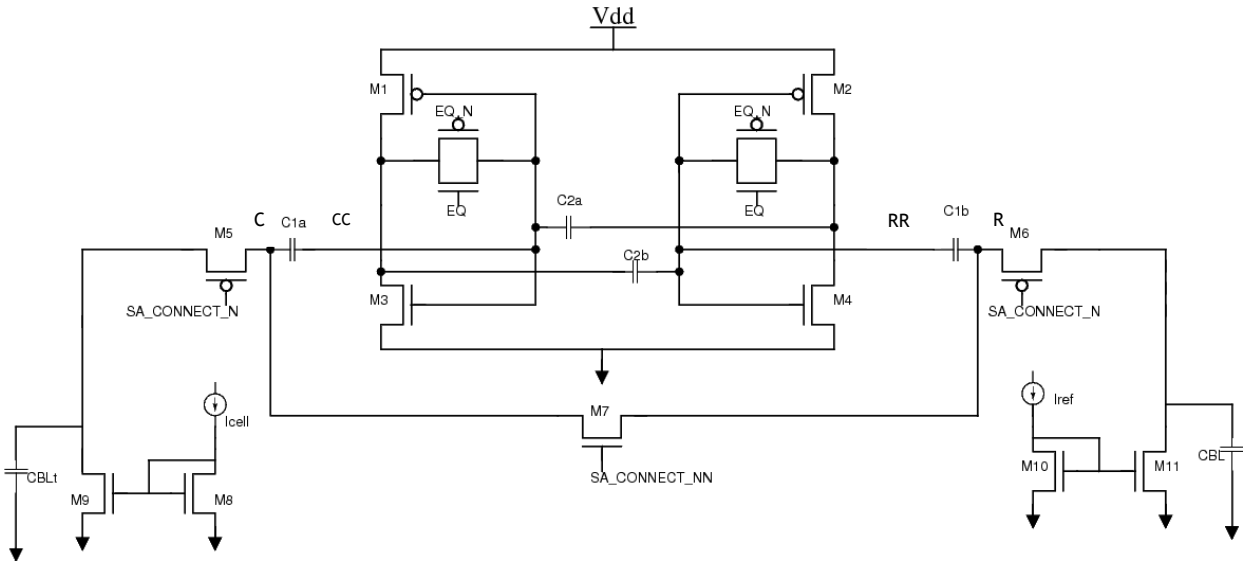


Fig. 4. 1 Schematic of Delay programmable SA

The circuit can be considered as a current sense amplifier. For current sensing scheme, the current difference if first converted into the voltage difference after which this difference is sent to a comparator. The cell current side and the reference current side is modelled using a simple current mirror, as shown in Fig 4.1.

The given Sense amplifier can be divided into three parts

1) Back to back inverter latch
2) Offset cancelling mechanism
3) Pass gates to enable discharge.

4.1 Circuit Operation

The Sense amplifier works in three phases, namely offset cancellation, access phase and Evaluation phase.

Table 4.1 shows the device type used in the sense amplifier. A particular type of extended gate (eg) devices are used in this sense amplifier to prevent the sense amplifier during programming state.

DEVICE	**TYPE**
M5, M6	Egpfet
C1a, C1b	Egncap
C2a, C2b	C_{eql} (MOM)
EQ Switches	Lvt
M3, M4	Lvt
M1, M2	Lvt
M7	Egpfet

Table 4. 1 Types of devices used

During the programming phase, the bit lines are charged up to 5V. To protect the latch circuit, an eg device is used as pass gates. These are special extended gate devices which can bear high voltage stress. Thus, these devices help in protecting the inner sense amplifier circuit. On the other hand, if lvtfet had been used instead of eg devices, there would be very high voltage stress on the device during programming stage since it would be directly connected to bit lines which are at very high potential.

The waveform in Fig 4.2 shows the switching operation for the Sense amplifier device.

The first step is to charge the bit lines to the precharge voltage. The pulse width of this signal depends on the bit line capacitance. During precharge, we also start the offset compensation by turning on the signal EQ. The voltage of the capacitors c1a and c1b are charged to bit line voltage. When the word line is turned on, the capacitor c1a and c1b starts discharging along with the bit line through the cell and reference currents respectively. When there is sufficient differential

created between the matrix side and reference side, the pass gates M5 and M6 are turned off after which both the capacitor c1a and c1b are connected for charge sharing using the SA_CONNECT_NN signal and the EQ switches are turned off to latch the output.

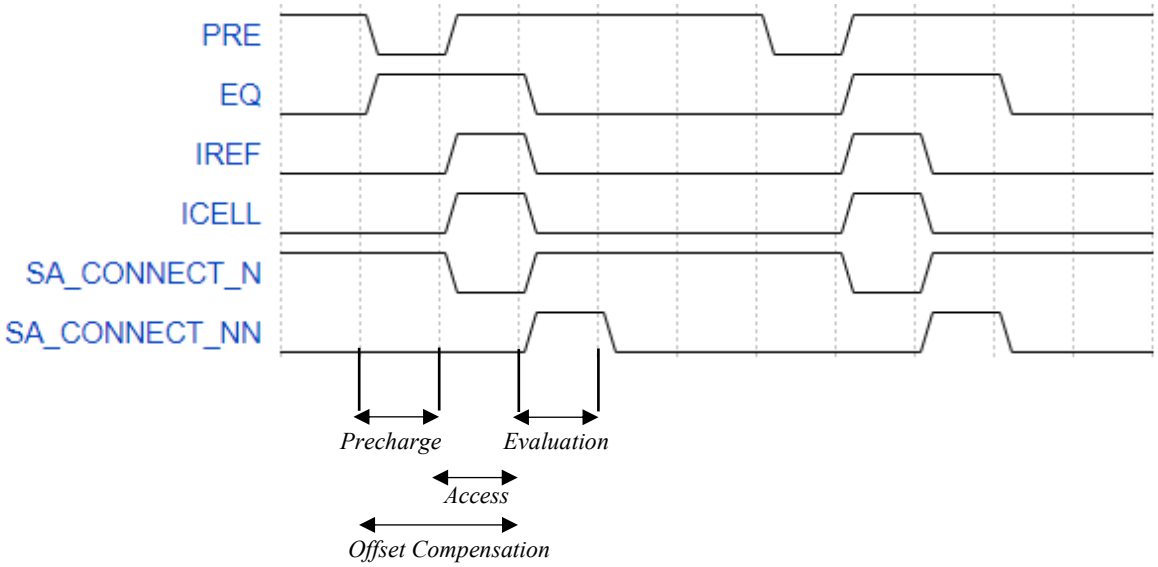


Fig. 4. 2 Input stimulus for Delay programmable SA

4.2 Offset Compensation

As discussed earlier, the main disadvantage of using a latch type sense amplifier for multi-level sensing is its intrinsic offset. This offset arises due to device mismatches during fabrication. This offset is more in lower technology nodes since the channel lengths of the devices is less at lower technology nodes. To counter this problem, we have used an offset compensation technique in our circuit, as shown in Fig.4.3(a). The input and output of both the inverters are shorted using transmission gates, as shown in Fig. 4.3(b).

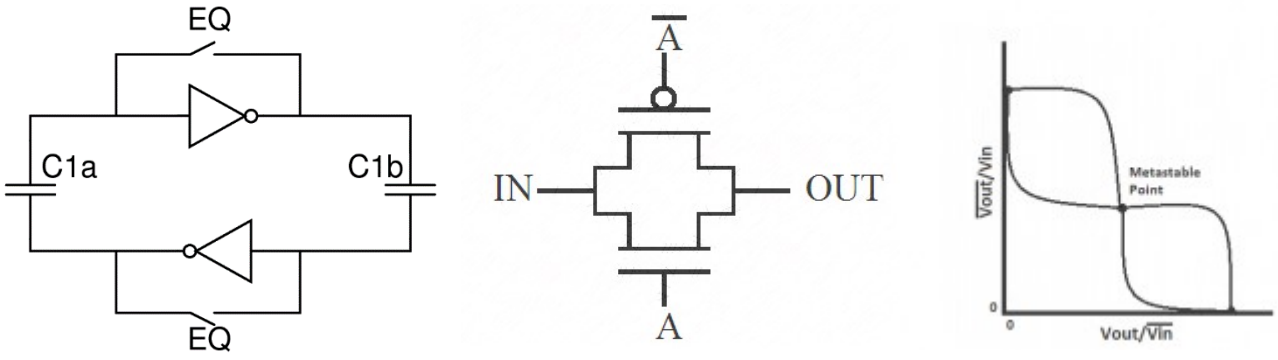


Fig. 4. 3 (a) Offset compensation (b) EQ switch (c) Metastability

This helps in biasing the inverters at the metastable state, as shown in Fig. 4.3(c) [9]. Due to this, the mismatches because of vt variations between the inverters nmos and pmos devices are reduced since the input and output of the inverter is shorted. A slight shift in the input of the inverter may result in disturbing the metastable state of the input, and the output may latch to one of the stable states. Thus, even a small differential voltage of 2mV is sufficient enough to latch the outputs of the sense amplifier.

We must note that both the inverters of the back to back latches are isolated with each other using the capacitor c2a and c2b. This enables them to be compensated independent of each other so that they can be initialized at their metastable points individually.

4.3 Access Phase

After successful completion of offset compensation, the output of the inverters is at a voltage nearly equal to $V_{dd}/2$. During the access phase, the state of the memory cell to be read is selected by decoding the address lines to wordline, and bitline. The cell current flows, and the bitline starts discharging. At this time, we turn on the M5-M6 devices using the SA_CONNECT signal. This enables the discharging of capacitors c1a and c1b. The rate of the discharge depends on the cell current(I_{cell}) and the reference current(I_{ref}). If I_{cell} is higher than I_{ref}, C1a discharges at a rate faster than the discharge rate of C2b. The voltage at time=t is given by the equation:

$$V(t) = V(0) - \frac{1}{C} I * t \tag{4.1}$$

Where V(t) is the voltage at time t, V(0) is the initial voltage, C is the capacitance value, I is the discharging current, and t is the pulse width of SA_CONNECT signal.

However, the current through the devices M5 and M6 is limited by its V_{ds} and given by the equation:

$$Id = \frac{uCoxW}{L}\left[(Vgs - Vth)Vds - \frac{Vds^2}{2}\right] \tag{4.2}$$

If $I_{cell} > I_{ref,}$ the voltage at node R becomes more than the voltage at node C. Thus, the current difference between the matrix cell and the reference cell is converted into voltage difference between nodes R and C in this phase.

4.4 Evaluation Phase

When the differential voltage between the nodes C and R exceeds the intrinsic offset of the latch, the M5 and M6 devices are turned off, and the next phase called the evaluation phase begins. In this phase, charge sharing takes place. The EQ switch is turned off, and the transistor M7 is

switched on using the SA_Connect_NN signal. This connects the nodes R and C. Since Voltage at node R is more than the voltage at node C, after SA_Connect_NN goes high, the voltage at RR becomes more, disturbing its metastable point and turning on the device M4 of the inverter. Similarly, the voltage at CC becomes less, and therefore the cross-coupled inverters go into positive feedback, and the output is latched.

4.5 Results

Fig.4.4 shows the output waveform of the proposed Sense Amplifier. It can be seen in the waveform that before the EQ switch is released, the output nodes are at the same voltage level equalized at $V_{dd}/2$. When the M7 device is turned on, and the charge sharing takes place, the output is latched.

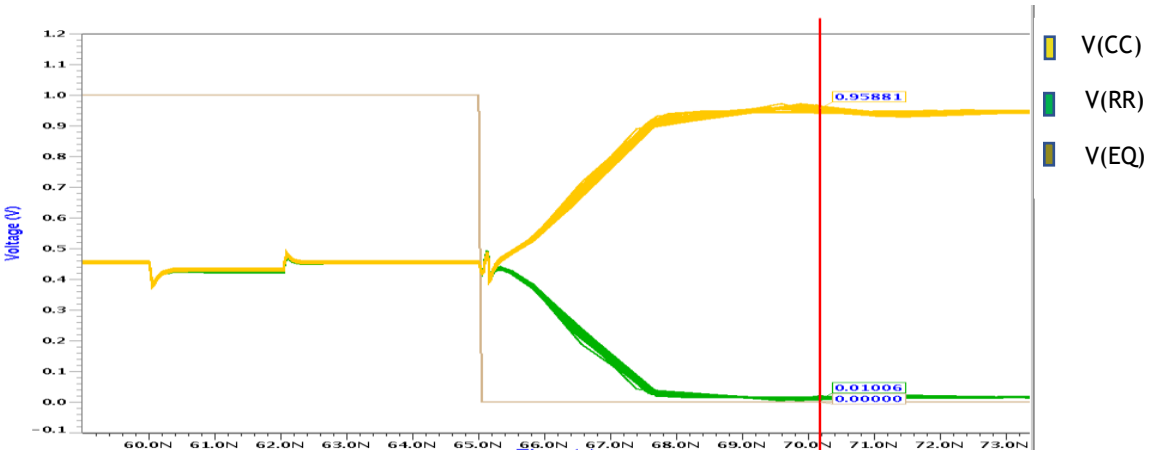


Fig. 4. 4 Output waveform for delay programmable SA

Table 4.2 shows the current discharging through capacitor c1a with respect to different bitline capacitances. It can be observed that the current is very less or higher values of capacitances. Therefore, for a higher value of bitline capacitance, the access time of the sense amplifier is more when compared to lower values of bitline capacitance. Therefore, it can be concluded that the figures of merit of this amplifier depend on bitline capacitance.

The differential voltage generated just before the charge sharing between the nodes C and R is very significant because this voltage will be used to shift the operating point of the latch from its metastable state. The differential voltage generated before turning on the SA_CONNECT_NN signal is shown in fig. 4.5. The mean differential voltage generated is 5.5mV for Bitline capacitance value of 500 fF.

Bitline Capacitance (fF)	Current Through M6 (uA)	Reference current I_{ref} (uA)
200	9.70	25
500	5.20	25
1000	3.00	25
1200	2.60	25
1500	2.13	25

Table 4. 2 Change in discharging current for Node R with respect to different values of bit line capacitance

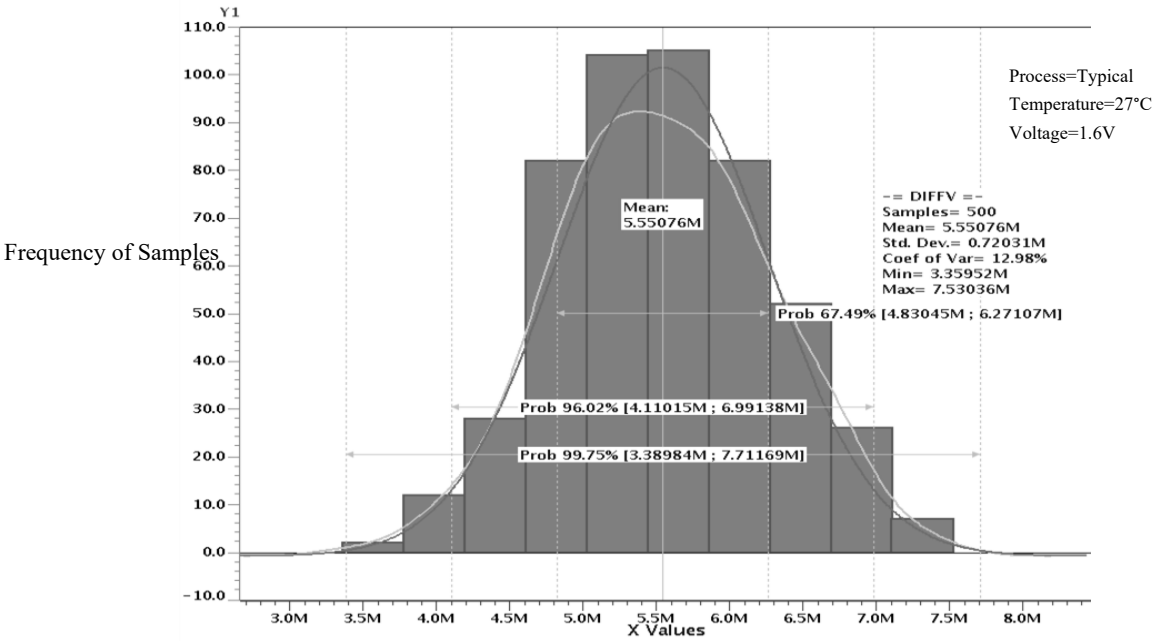


Fig. 4. 5 Monte Carlo for differential voltage generated

The sense amplifier discussed in this chapter cannot be used for MLC sensing because for different values of cell current the pulse width of the SA_CONNECT_N signal needs to modulated. For higher values of current the capacitors, C1a and C1b tend to discharge fast dismissing the differential voltage created between the two nodes, whereas for lower values of cell current they might need time to create sufficient differential for the evaluation phase to begin. To overcome this problem, we have proposed a modified current mirror-based sense amplifier, which is discussed in detail in the following chapter.

5 Modified Current Sense Amplifier for Multilevel Cell

The Sense amplifier discussed in chapter 4 has a few shortcomings. It is difficult to use this sensing scheme for multilevel cells because of the challenges to tune the timing of the circuit for different current values. If a single finite state machine is used for timing the circuit for the SA_CONNECT_N signal, the capacitor C1a and C1b are subjected to discharge for higher current values. In comparison, for lower current values, they might not be able to create sufficient differential voltage for sensing. Thus, a single finite state machine for timing the circuit cannot be used for multiple values of current. Another shortcoming for this circuit is its dependence on bitline capacitance. The discharging rate and, therefore, the sensing time depends on bitline capacitance of the memory array. Therefore, to overcome these problems, a new circuit is proposed, as shown in Fig.5.1.

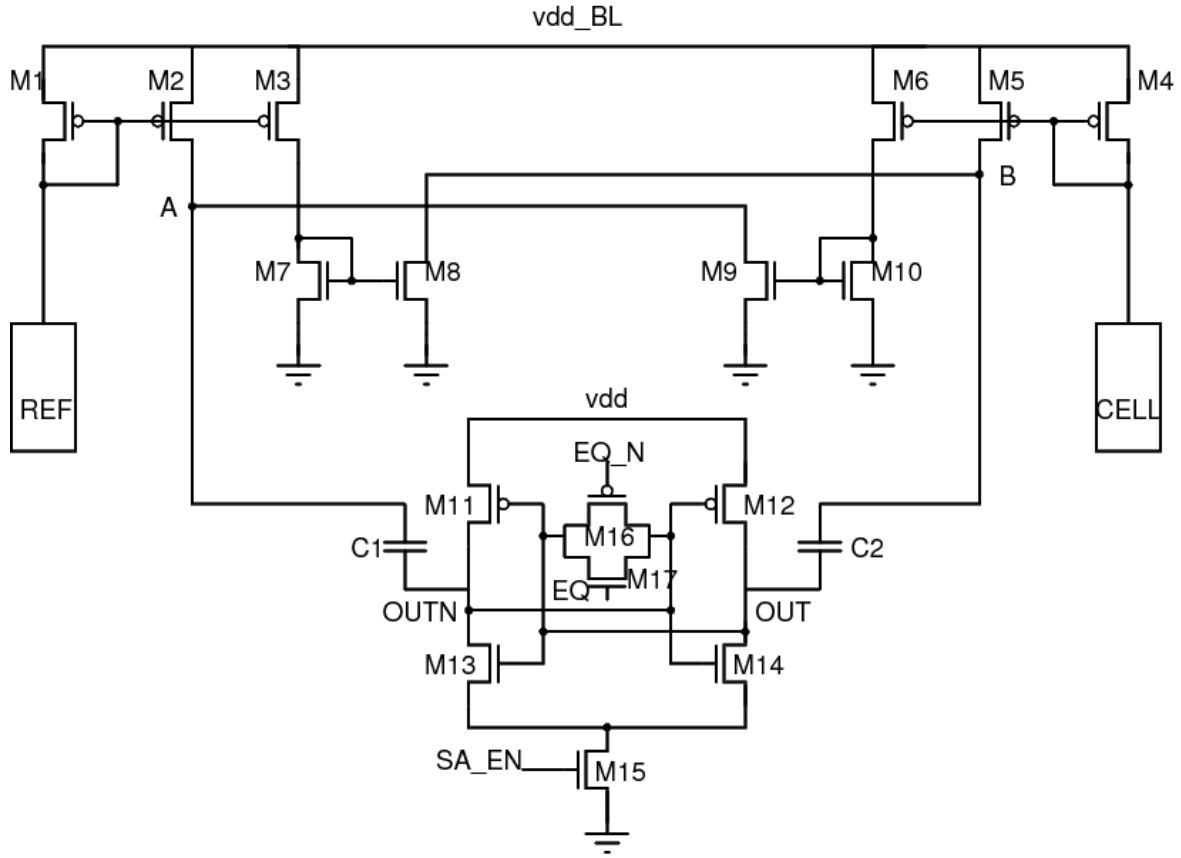


Fig. 5. 1 Schematic for modified SA

The main advantage of this circuit over the previous circuit discussed in chapter 4 is that instead of comparing the absolute value of cell current (I_{cell}) with the reference current (I_{ref}), it compares the difference between Icell and I_{ref}. Therefore, the problem of the capacitor discharging for higher values of current is reduced using this circuit. This circuit is designed for the current range of 5uA to 40uA with four levels of storage of the PCM cell.

5.1 Circuit Components

We can divide a current sense amplifier into three subparts.

1. Current mirror circuit to mirror the bitcell current to the Sense circuit.
2. A current to voltage conversion mechanism.
3. Latch type voltage amplifier.

Apart from these three, a decoder is also used in case of multi-bit sensing schemes.

In this circuit, transistor M1-M10 consists of the current mirror scheme. All these current mirrors are identically sized, to avoid mismatches. The reference side consists of M1-M4 PMOS and M7-M8 NMOS, while the cell side consists of M4-M6 PMOS and M9-M10 NMOS.

The nodes A and B are of significant importance because the differential voltage is generated at these nodes. The two capacitor C1 and C2, serves two purposes:

1. Current to Voltage conversion
2. Power saving

Current to voltage conversion is done by discharging the capacitor nodes A and B. The difference in cell current and reference current is reflected as the difference in voltage at nodes A and B. At the same time, the capacitors help in power saving because the comparator latch can be used at different voltage level than the bitline voltage as shown in Table 5.1.

The transistor M11-M15 comprises of a latch-type voltage comparator. The input nodes of both the transistor are equalized before the sensing operation begins, using the transistor M16 and M17 as a transmission gate.

	Without Capacitor	With Capacitor
Bitline Voltage (vdd_BL)	1.6 V	1.6 V
Sense amplifier (vdd)	1.6 V	1.2 V
Power	38.86uW	37.5uW

Table 5. 1 Power comparison with coupling capacitor vs no coupling capacitor

5.2 Circuit Operation

To understand the working of the circuit, the input stimuli is shown in Fig. 5.2

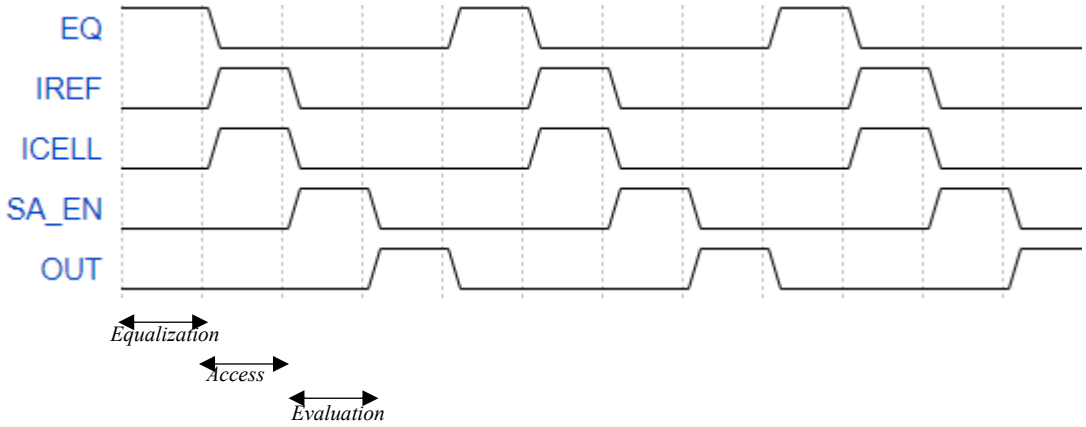


Fig. 5. 2 Input stimulus for modified SA

During the RESET state, the nodes A and B are precharged at a voltage equal to the bit line voltage vdd_BL. Before the read operation, the comparator output nodes OUT and OUT_N are equalized using the transmission gate by turning on the M16 and M17 transistors using the signals EQ and EQ_N. After successful equalization of the comparator nodes, the wordline is turned on, and the current I_{cell} and I_{ref} flow in the cell and reference bitlines of the array respectively.

The current I_{ref} starts to flow from the devices M2 and M3 while I_{cell} starts to flow from the devices M5 and M6. Note that I_{ref} is also mirrored in the device M8 and M9 through M7 and M10 respectively. It can be concluded by applying Kirchhoff current law at node A, and node B that the current discharging the Cap C1 equals to I_{cell}-I_{ref} and for Cap C2 is I_{ref}-I_{cell}. Thus, instead of taking the absolute value of cell and reference current, we are taking the algebraic difference to discharge the capacitors.

At the time when the wordline is selected, we also turn the EQ and EQ_N signal off so that the comparator nodes start to couple with the voltage difference created at nodes A and B through capacitors C1 and C2 respectively. When the differential voltage is sufficient, the footer transistor M15 is turned on using the Sense Enable signal (SA_EN), and the output is latched using the positive feedback of back to back inverter.

5.3 Circuit Design

It is evident from the above discussion that the current mirrors in the circuit need to be identically sized. The Table 5.2 and Table 5.3 shows the parametric analysis of the length of the current mirror transistors M1-M10 for the Monte Carlo analysis of the differential voltage generated at nodes A and B for the minimum and maximum value of reference current in the current range 5uA-40uA. Note that the coupling capacitor and the comparator are not considered in the

following table, because we don't know the value of the coupling capacitor yet. Therefore, only the parasitic capacitance at node A and B are used to create the differential voltage here

I_{cell} (uA)	I_{ref}(uA)	Len(um)	Mean(mV)	Sigma
14	15	0.1	144	322
14	15	0.2	231	292
14	15	0.3	271	235
14	15	0.4	298	192
14	15	0.5	318	159
14	15	0.6	332	131
14	15	0.7	339	107
14	15	0.8	341	86
14	15	0.9	337	67
14	15	1.0	325	51

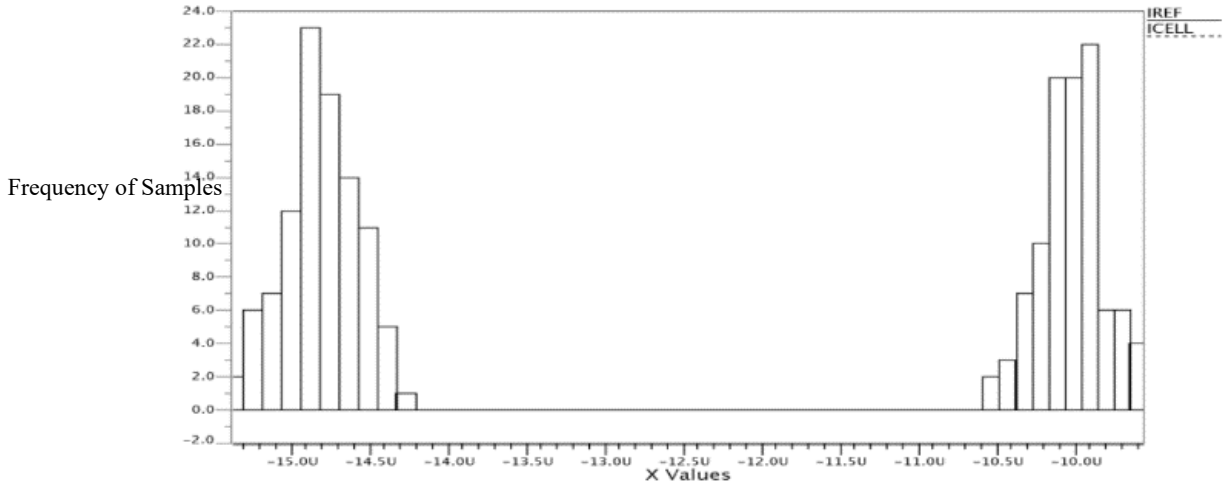
(a)

I_{cell} (uA)	I_{ref}(uA)	Len(um)	Mean(mV)	Sigma
33	35	.1	153	327
33	35	.2	284	335
33	35	.3	359	286
33	35	.4	411	243
33	35	.5	441	205
33	35	.6	451	161
33	35	.7	434	128
33	35	.8	377	98
33	35	.9	289	75
33	35	1.0	198	52

(b)

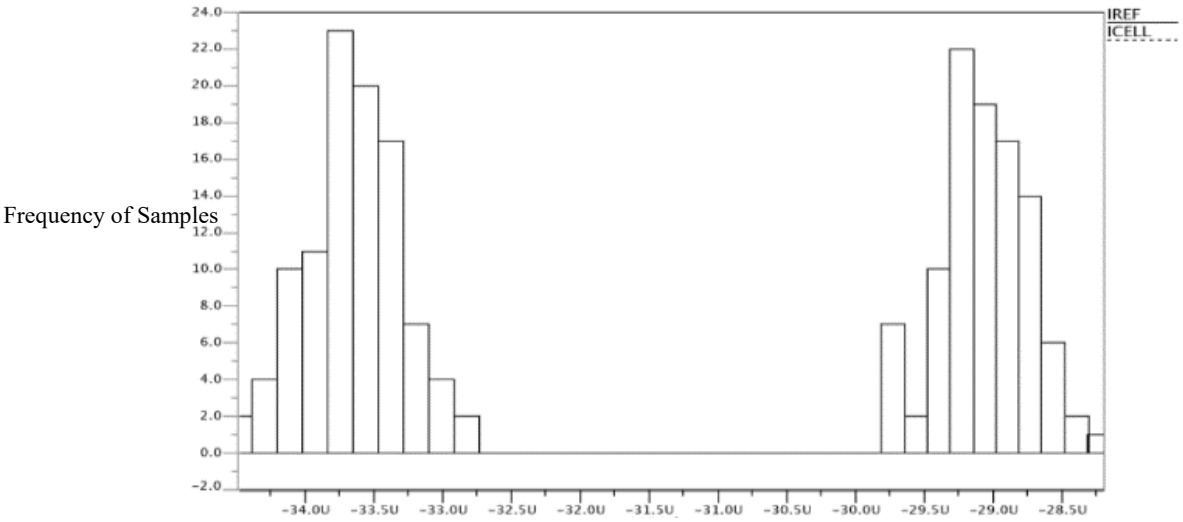
Table 5. 2 Length vs differential voltage for (a) Icell=14uA (b) Icell=33uA

It can be observed that for length=0.7nm, 3 sigma is qualified for both the values of current. For the higher value of currents, it is observed that the sigma becomes worse even if the current difference between I_{cell} and I_{ref} is increased from 1uA (14/15) to 2uA (33/35). This behaviour can be explained using the current equation of MOSFET in saturation. The V_{th} variations or the process variations are more for the higher value of drain current. This is also shown in the distribution function in Fig.5.3 The distribution is more spread in the Fig.5.3(b), which is for I_{cell}=35uA and I_{ref}=33uA, whereas distribution is more compact in the Fig.5.3(a), which is for I_{cell}=15uA and I_{ref}=14uA.



Current I_{cell} & I_{ref}

(a)



(b)

Fig. 5. 3 Distribution of Current values

(a) Icell/Iref=14uA/15uA. (b) Icell/Iref=33uA/35uA

It is observed that for lower values of current the current distribution between Iref and Icell is far as compared to higher values of current.

The capacitor C1 and C2 are metal capacitors with value 10 fF. To obtain the value of the coupling capacitor, we sweep it 1 fF to 50 fF and check where there is the maximum transfer for the voltage difference between the capacitor nodes A and B to comparator nodes OUT and OUT_N. It can be seen in the plot in Fig. 5.4 that 10fF seems the most suitable value for the same. The Capacitors are made in layout from metal layer metal 4 and metal 5. DIFF represents the voltage difference between node A and node B, and DIFF2 is the voltage difference between OUT and OUT_N. We say that coupling is good when the DIFF is equal to DIFF2.

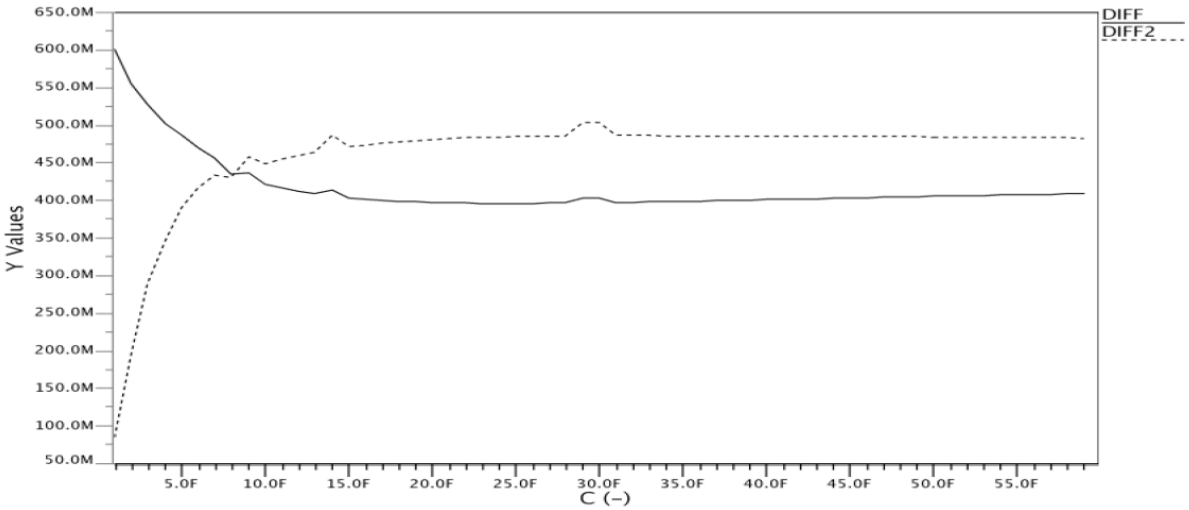


Fig. 5. 4 Capacitor coupling

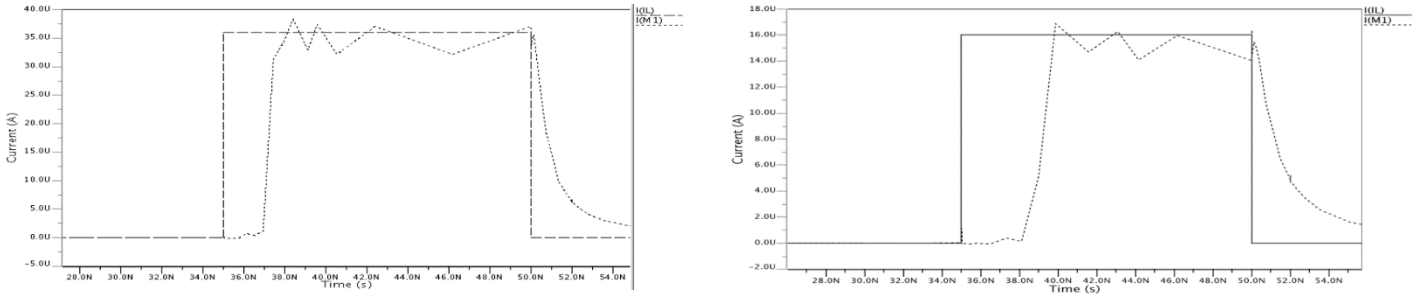
The design of the comparator latch is done for worst-case voltage offset, which can be obtained from Table 5.3. i.e. mean-3*sigma= 50mV.For this offset, the sizes of M13-M15 are taken as 2um, and the sizes of M11-M12 are taken as 1um.

DEVICE	LENGTH (um)	WIDTH (um)	TYPE
M1-M6	0.7	2.0	HVT
M7-M10	0.7	2.0	HVT
C1-C2	-	-	MOM
M11-M12	0.09	1.0	LVT
M13-M14	0.09	2.0	LVT
M15	0.09	1.0	LVT
M16	0.09	1.0	LVT
M17	0.09	0.5	LVT

Table 5. 3 Device sizes and types

It is observed from Table 5.3 that high threshold voltage devices are used for the current mirror design. This is done because the sense amplifier is more sensitive to the process variations in these devices.

From the waveform shown in Fig.5.5(a) and Fig.5.5(b), we can observe that for the lower end of the current range (I_{ref}=15uA) the coupling begins at 3ns after the reference current is given, i.e. the wordline is turned on. However, at the upper end of the current range (I_{ref}=35uA), the coupling begins at 2ns after the wordline is turned on. This is because the device length of the current mirrors is significant to avoid process variations making the device having a slow response.



(a) (b)

Fig. 5. 5 Delay in current flow after wordline is on due to the current mirror response.

(a) I_{ref}=35uA (b)I_{ref}=15uA

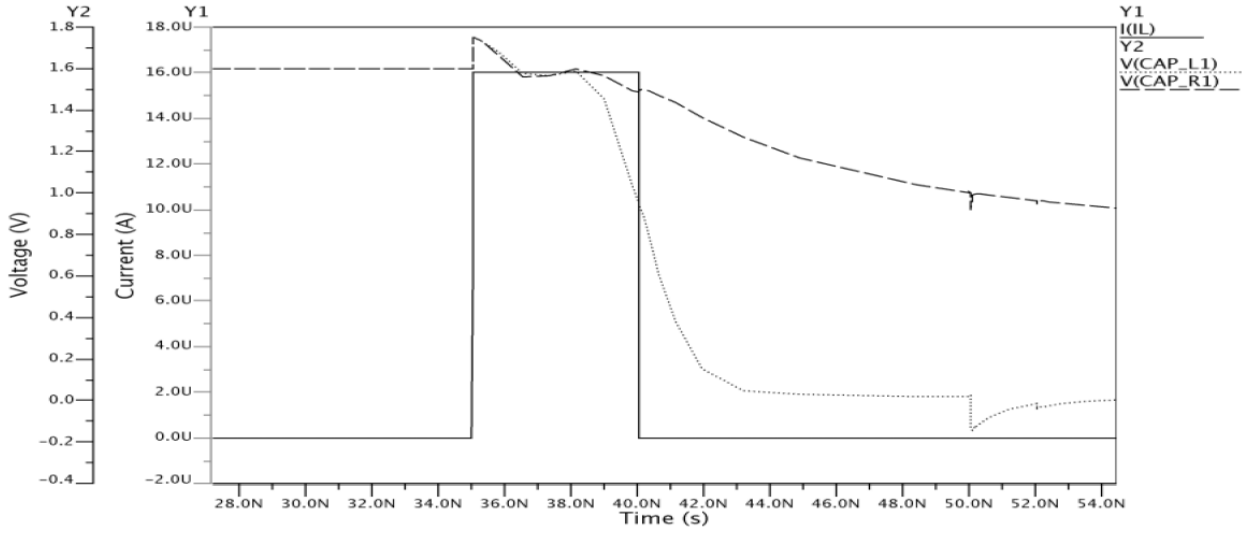


Fig. 5. 6 Delay in coupling after wordline is on.

The output waveform of the circuit is shown in Fig.5.7 During the equalization phase, the EQ is turned on; thus, the OUT and OUT_N ports are shorted. After successful equalization, the worline is turned on, and the nodes start discharging, the coupling begins. When sufficient differential voltage is created at nodes A and B, we turn on the footer transistor using the SA_EN signal.

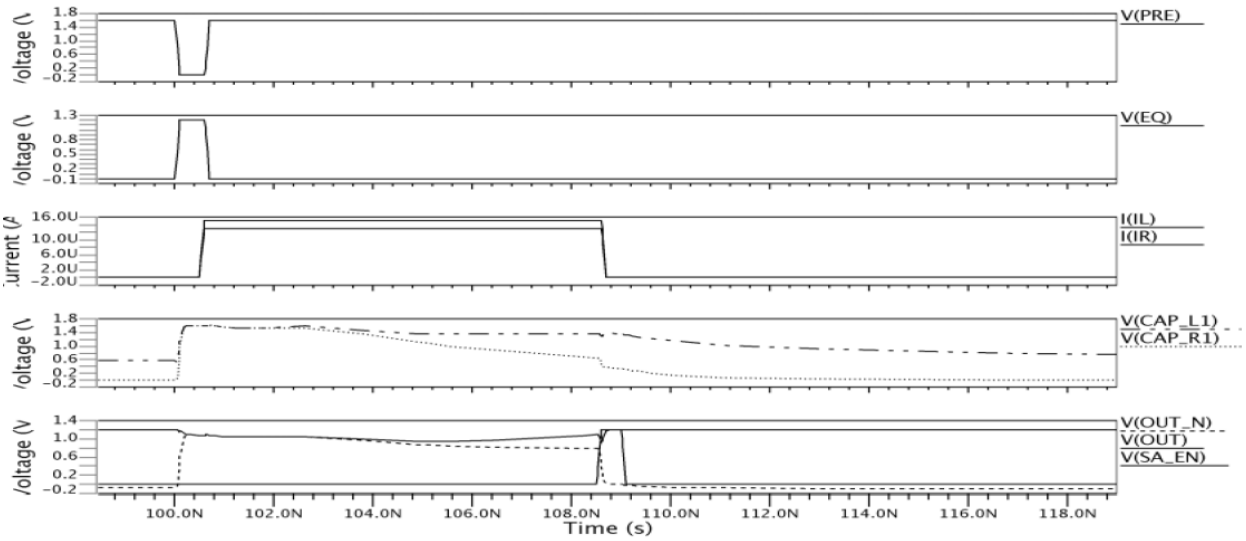


Fig. 5. 7 Output waveform for modified SA

5.4 Sense Amplifier performance

The access time of the sense amplifier is given by the equation:-

$$T_{Access} = T_{Precharge} + T_{Equalize} + T_{Discharge} + T_{Latch.} \quad (5.1)$$

Where

$T_{Precharge}$ = The time required for the node A and node B to get charged at bit line voltages.

$T_{Equalize}$ = The time to equalize the OUT and OUT_N nodes of the latch.

$T_{Discharge}$ = The time required to discharge the nodes A and B.

T_{Latch} = The time required for the output nodes to reach 80% of $V_{dd.}$

It must be noted that the precharge and equalization can be done simultaneously which reduces the access time and the equation eq 5.1 changes to

$$T_{Access} = \max(T_{Precharge}, T_{Equalize}) + T_{Dischage} + T_{Latch} \quad (5.2)$$

The monte Carlo simulations for the access time is shown in Fig.5.8 with the mean access time equal to 8.6ns

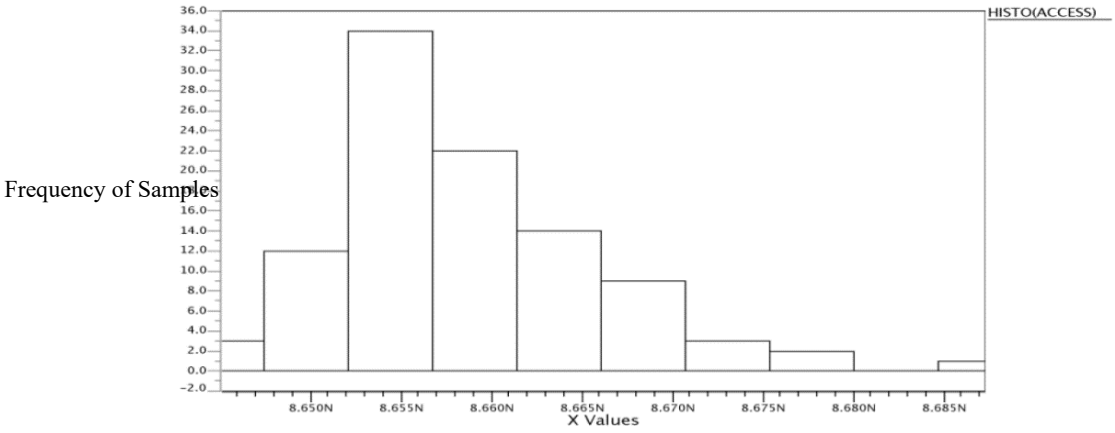


Fig. 5. 8 Access Time for modified SA

Offset

Fig. 5.9 shows the differential voltage distribution function over 1000 monte Carlo runs. The X-axis corresponds to the voltage difference between OUT and OUT_N nodes just before the SA_EN signal is turned on. Fig.5.9(a) shows the distribution for I_{cell} =13uA and I_{ref}=15uA, Fig. 5.9(b) shows the distribution for I_{cell}=23uA and I_{ref}=25uA and Fig5.9(c shows the distribution for I_{cell}=33uA and I_{ref}=35uA.

IREF	ICELL	MEAN	SIGMA	MEAN-3SIGMA
15	13	242	64	50
25	23	283	82.5	35.5
35	33	292	88	28

Table 5. 4 Mean vs Sigma for all three reference currents

AREA

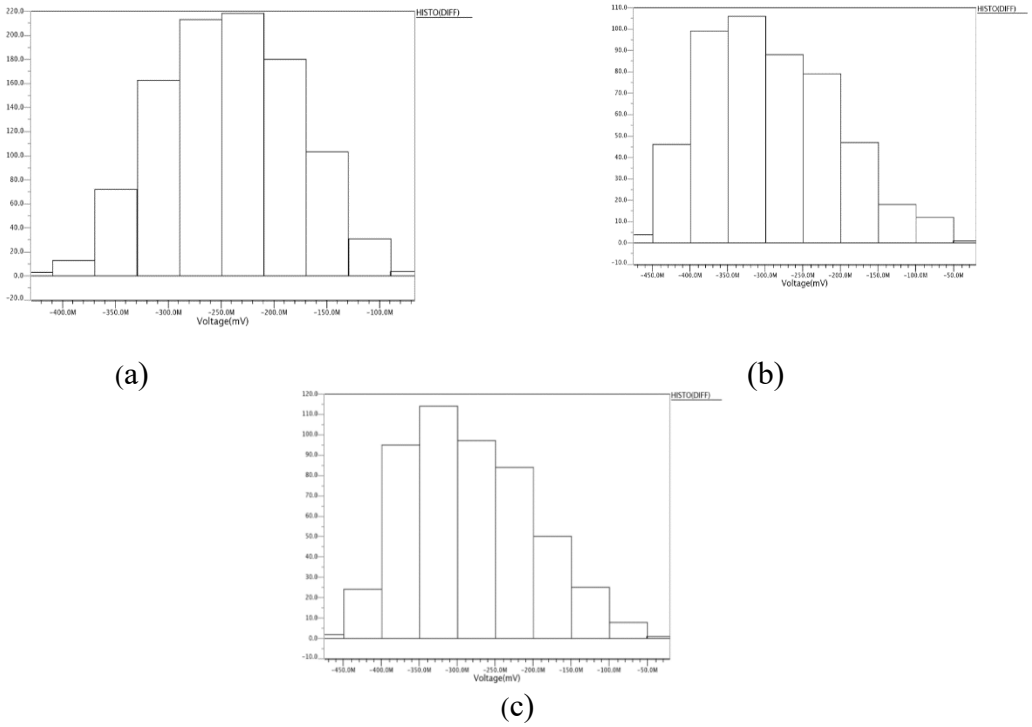
The layout is divided into three parts:

1) Reference Side

2) Cell Side

3) Comparator Latch

The area of the sense amplifier is 110um^2

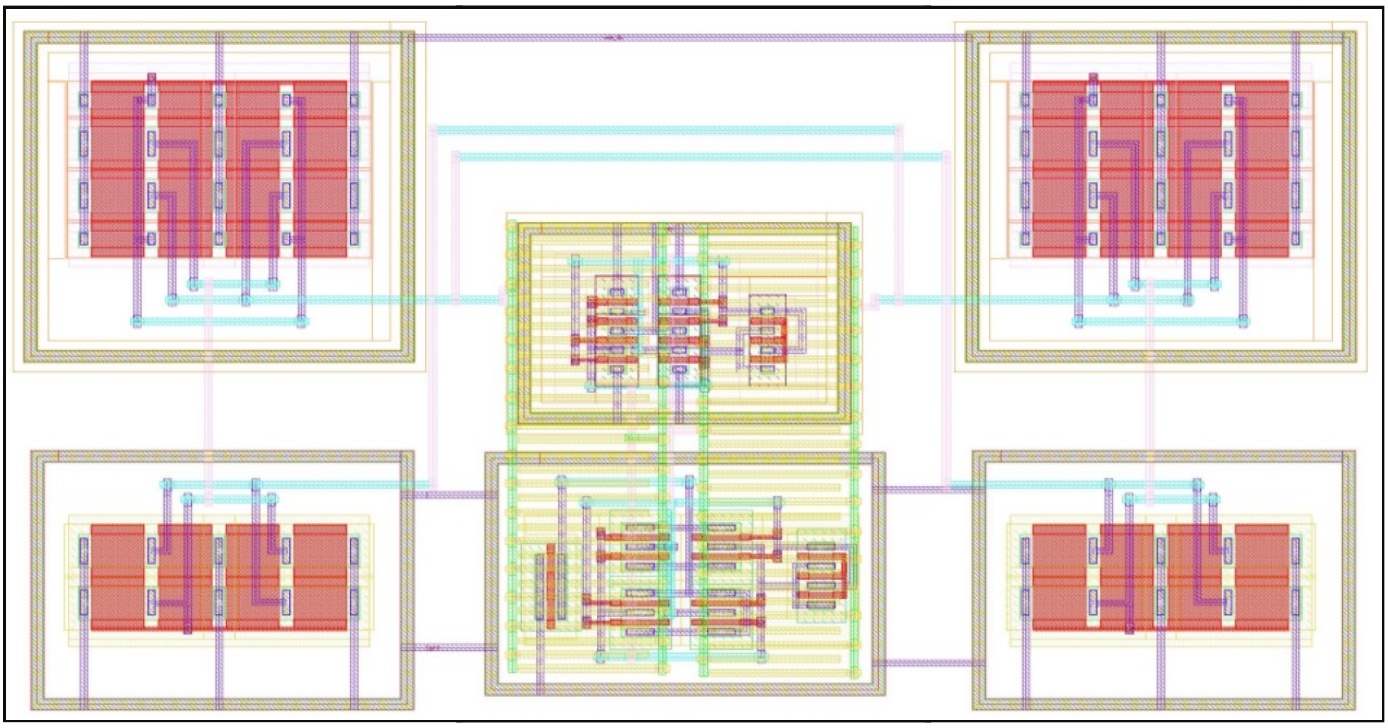


(a) (b)

(c)

Fig. 5. 9 Monte Carlo of differential voltage for all three reference currents.

(a) Iref=15uA (b) Iref=25uA (c) Iref=35uA



REFERENCE SIDE COMPARATOR CELL SIDE

Fig. 5. 10 Layout for single bit sensing.

The detailed layout of all the three subdivisions is shown in Fig.5.12 The reference and cell side are identical and is matched with common centroid with ABBA, as shown in Fig. 5.11.

M1	M1
M2	M3
M3	M2
M1	M1

M7	M8
M8	M7

Fig. 5. 11 Common centroid matching

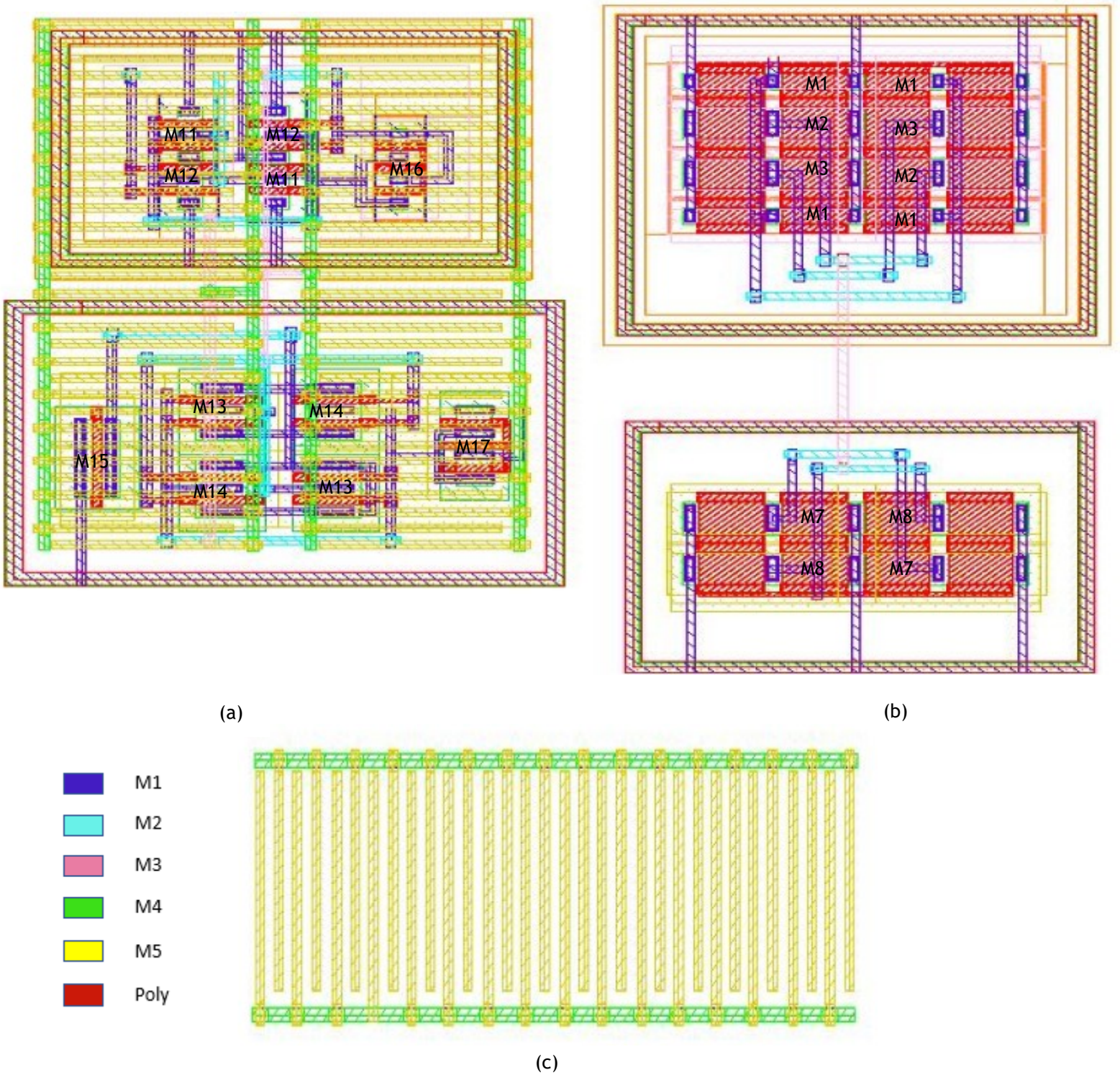


Fig. 5. 12 Layout of individual subparts

(a)Comparator (b)Current Mirror (c)MOM Capacitor

The capacitors C1 and C2 are made as Metal Oxide Metal type capacitor using the M4 and M5 metal layer and thus can be placed over the latch comparator.

M1, M2 and M3 metal layers are used for routing the layout.

Table 5.5 shows the PPA comparison between the pre-layout and post-layout versions of the modified sense amplifier.

Parameter	Pre-layout	Post-layout
Access time	8.58ns	8.65ns
Power	37.2uW	37.5uW
Offset	1.0uA	1.3uA
Area	-	110um2

Table 5. 5 PPA comparison of Pre layout vs Post Layout

In this chapter, we have proposed a modified current mirror-based Sense amplifier that has an offset value of 1.3uA and it can be used for a wide range of current as opposed to delay programmable sense amplifier discussed in chapter 4. In the next chapter, we propose a multilevel sensing architecture using this sense amplifier.

6 Multilevel Sensing Design and Architecture

The idea of multilevel storage is that if we can store more than 2 bits/cell in a single PCM device, the gain in density becomes very high. Since the data is stored in the form of resistance levels in PCM, varying level for resistance can correspond to different value of bits. In this work, we have assumed that there are 2 more states between the set and the reset state of the PCM cell. Thus, multilevel cell (MLC) enables us to exploit the intrinsic capability of storing an analog data to store more than a single bit per cell.

The various challenges in multilevel cell storage [5] are discussed below: -

1) Resistance Drift- Experimentally it has been shown that the resistance in the amorphous phase increases over time. This phenomenon is called resistance drift. The formula of cell resistance at time t is given by:

$$R(t) = R(t0) * \left(\frac{t}{t0}\right)^{v}$$

Here R(t0) is the initial resistance at time t0 and v is the drift exponent. In multilevel storage, this phenomenon of drift resistance reduces the separation between two levels of storage.

2) Variability – Multilevel storage solution is directly correlated with cell variability in terms of reliability of the circuit. The cell variability in PCM is of two types, the inter-cell variability that arises due to the fabrication process and intra cell variability, which is due to repeated write/programming cycles. Due to structural variations of the cell during fabrication, causing various electrical variations, the same programming pulse applied to different cells may result in different resistances. This might cause errors during the read operation. The centre cell variations due to repeated programming cycles may lead to different temperature distributions resulting in different resistance levels. However, this type of variability is small compared to inter-cell variations.

3) Reliability – The three factors affecting the reliability of PCM are data retention, Endurance and data disturb. Endurance refers to the number of repeated programming cycles that can be reliably performed in a memory array. Data retention is the device

capability to store the data over time, and data disturbs refer to various read and write disturbs that result in unintentional programming of the neighbouring cells.

Fig.6.1 shows the distribution of the 4 levels of storage. To detect these 4 levels labelled as “00”, “01”, “10”, “11”, we have used three reference values of current. Table 6.1 shows the relationship between cell current values and reference current.

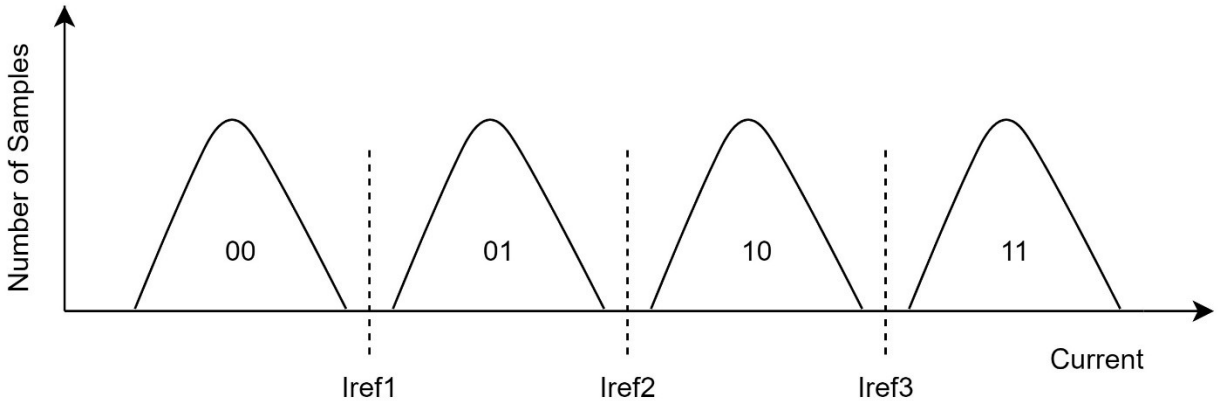


Fig. 6. 1 Current distribution for 4 levels of storage

For reading data, three sense amplifiers are used, following a combinational decoder circuit for parallel sensing. Fig.6.2 shows the block diagram for the multilevel sensing architecture, Three Sense Amplifiers SA1, SA2 and SA3 are used, the output of each Sense amplifier is fed to a decoder which will give the output according to the tree diagram shown in Fig. 6.3.

Cell value	Cell current	Reference current
00	N	$N < I_{ref1} < N+\alpha$
01	N+α	$N+\alpha < I_{ref2} < N+\beta$
10	N+β	$N+\beta < I_{ref3} < N+\gamma$
11	N+γ	

Table 6. 1 Values of Cell current and Reference current for different bits.

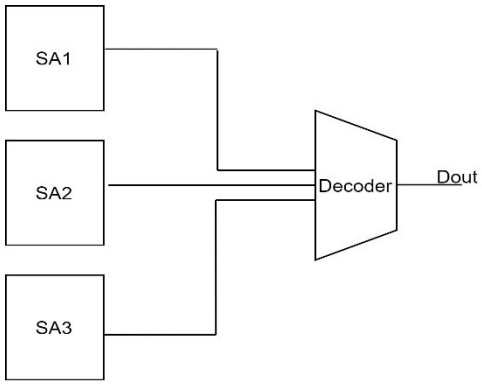


Fig. 6. 2 Block diagram for MLC read

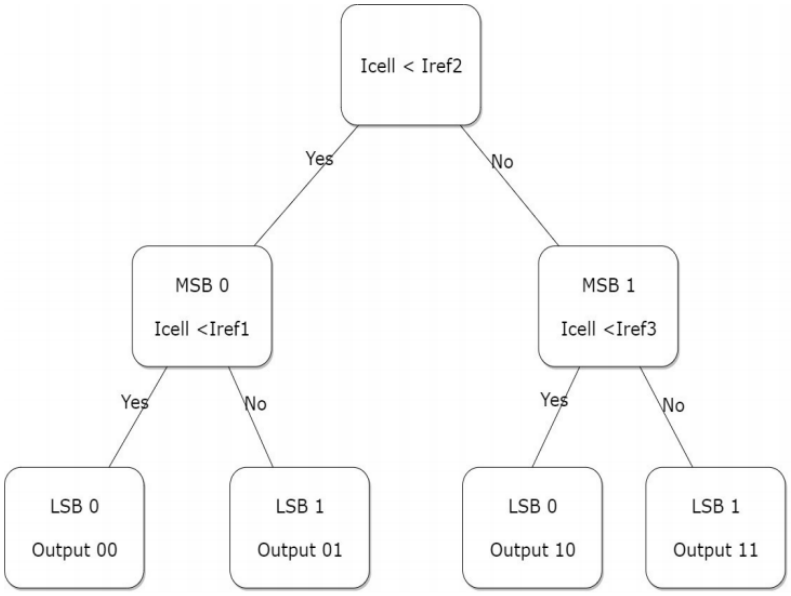


Fig. 6. 3 Flowchart for MLC read

It can be seen that the value of the output can be obtained in just two cycles since the height of the tree is two. First, the value of cell current is compared with reference2. This defines the value of MSB. If the cell current value is less, the MSB is fixed as 0, and if the cell current is more, the MSB is fixed as 1. Thus, the value of MSB is directly obtained using the SA2. To get the value of LSB, the second comparison is made. If the value of cell current is less than reference, the LSB Is set to 0, and if the value of the cell is more than reference LSB is set to 1. However, this type of approach will require two clock cycles to give output and thus increasing the access time. The first clock cycle decides the MSB and the second clock cycle decides the LSB based on the MSB. An alternative approach for computing the same is to compare all the three reference currents with

the cell current in parallel and then use a decoder to generate the output thus reducing the access time from two read clock cycles to a single read clock cycle.

The schematic of the sense amplifier with multibit sensing is shown in the Fig.6.4. We have three Sense amplifier units which include current mirrors and latch type comparators. The output of each latch type comparator is fed to the 3:2 Decoder, which gives B0 and B1 as the output of the device.

From Table.6.1, we see that we have 4 values of cell current corresponding to different resistance levels of the PCM bitcell. To sense these 4 levels, we have to generate 3 reference levels in between as shown in Table 6.2 with I_{ref1}=15uA, I_{ref2}=25uA, I_{ref3}=35uA

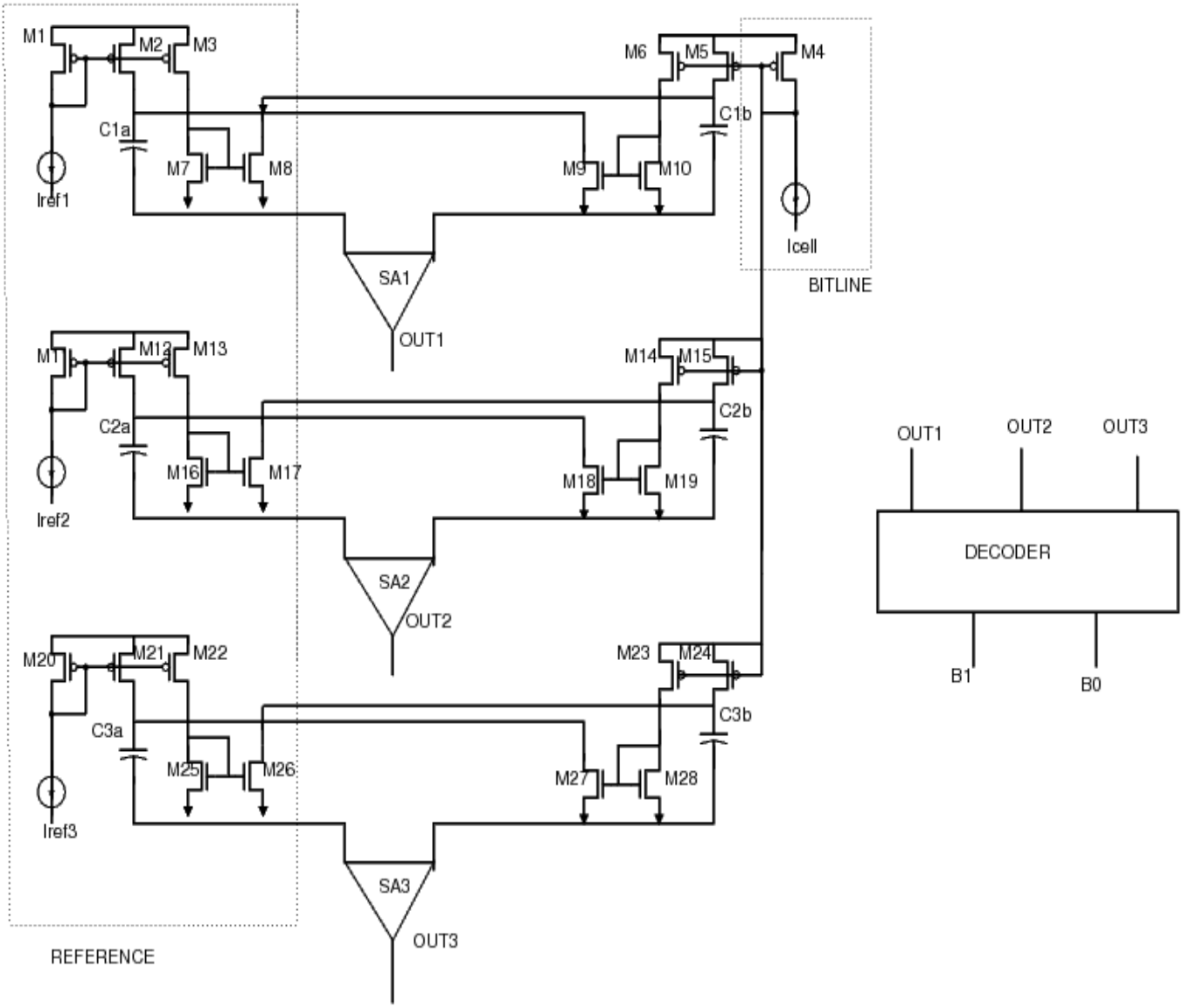


Fig. 6. 4 Schematic for MLC read

BIT0	BIT1	CELL CURRENT (uA)
0	0	10
0	1	20
1	0	30
1	1	40

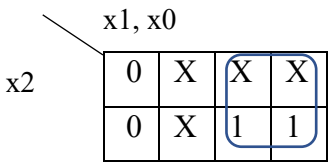
Table 6. 2 Output bits for four levels of cell current

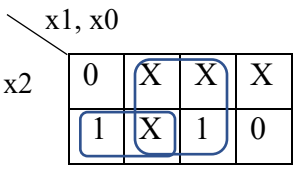
To size the devices, we must note that the device M4 has thrice the loading of that of M1, M11 or M20 transistors on the reference side, therefore to equalize the loading we make the M4 device 3 times than other devices. Thus, all the reference currents are also divided by 3.

The decoder design is done by using the truth table given in Table 6.3

OUT2	OUT1	OUT0	B1	B0
0	0	0	0	0
0	0	1	X	X
0	1	0	X	X
0	1	1	X	X
1	0	0	0	1
1	0	1	X	X
1	1	0	1	0
1	1	1	1	1

Table 6. 3 Truth Table for decoder design





Therefore B1=OUT1

$B0=OUT2.\overline{OUT1}+OUT0$

The schematic of the decoder is shown in the Fig.6.5

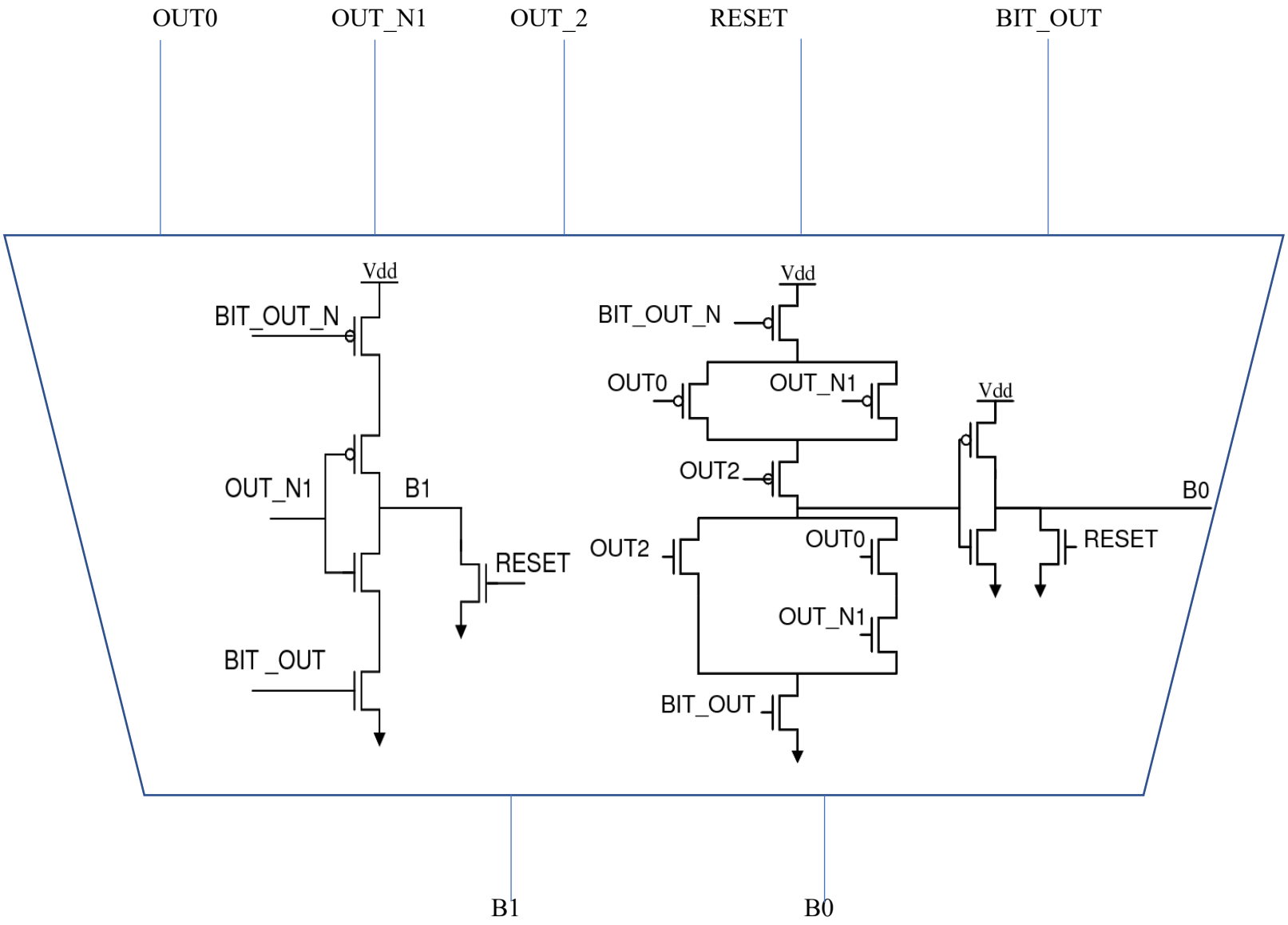


Fig. 6. 5 Decoder Design

The full custom layout for the Multilevel sensing scheme is shown in Fig.6.7. The left side of the layout corresponds to the reference side, while the right side is the bitcell side. Note that a single reference side can be used for the entire macro. The middle part of the layout contains the three comparators with their capacitors made using metal layer 4 and 5. Below the comparators in the middle is the 3:2 decoder which takes the outputs from the comparator and converts it to output bits.

The waveform stimulus is shown in Fig.6.6 All the capacitor plates are precharged to vdd_BL, and the comparator nodes are equalized. The wordline is turned on allowing all the three reference current in the left half of the circuit and cell current in the right half of circuit. The capacitor nodes discharge with I_{cell}-I_{ref1}, I_{cell}-I_{ref2} and I_{cell}-I_{ref3} and coupling begins. When sufficient differential is generated, the footer transistor of all the three comparators is turned on using the SA_EN signal. The reset is turned off, and the BIT_OUT signal of the decoder is turned on to give the output bits.

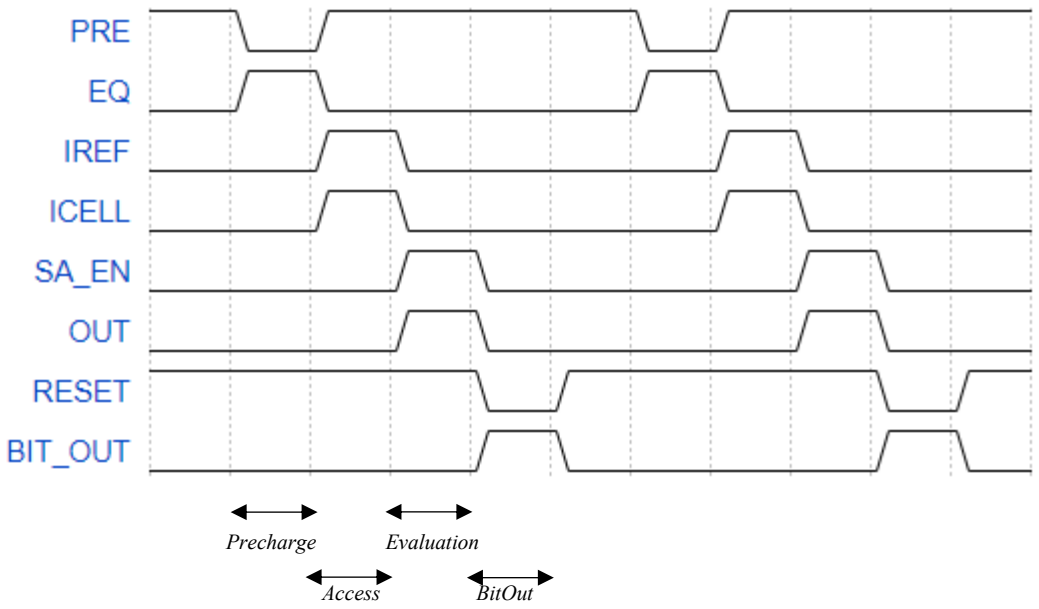


Fig. 6. 6 Input Stimulus

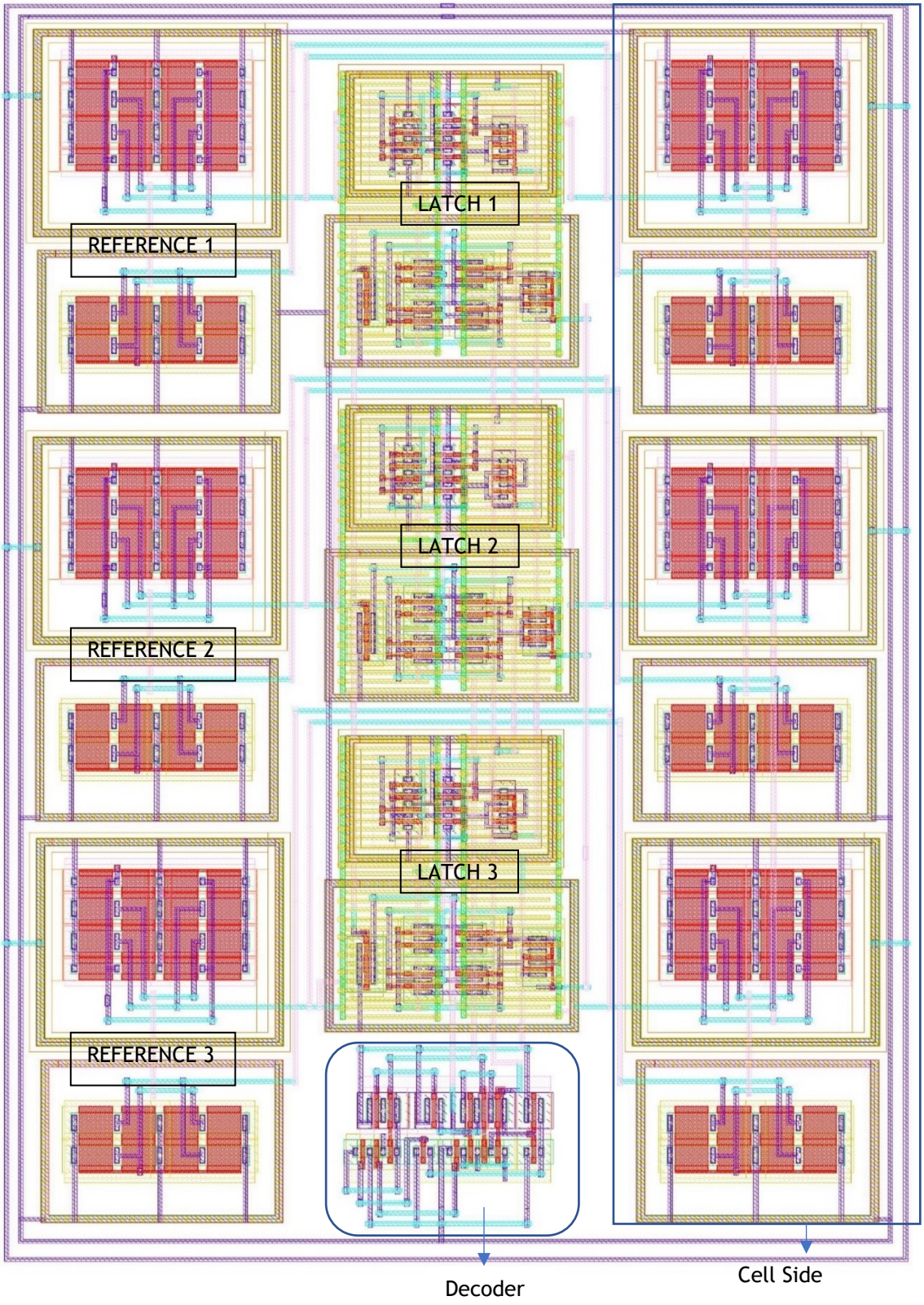


Fig. 6. 7 Full custom layout of the Sense Amplifier

Total Area=327um^2

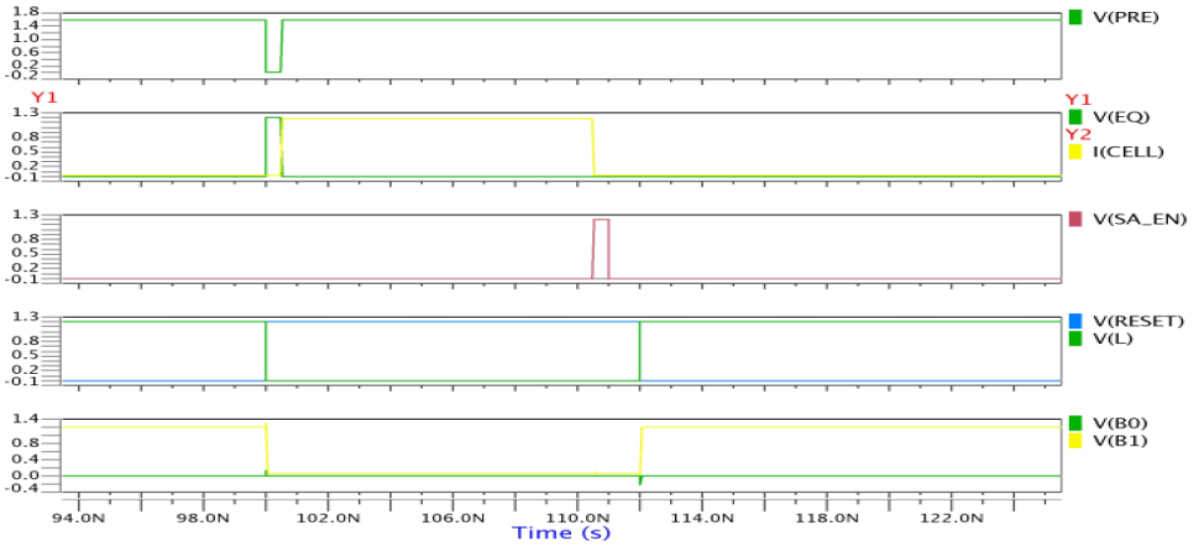
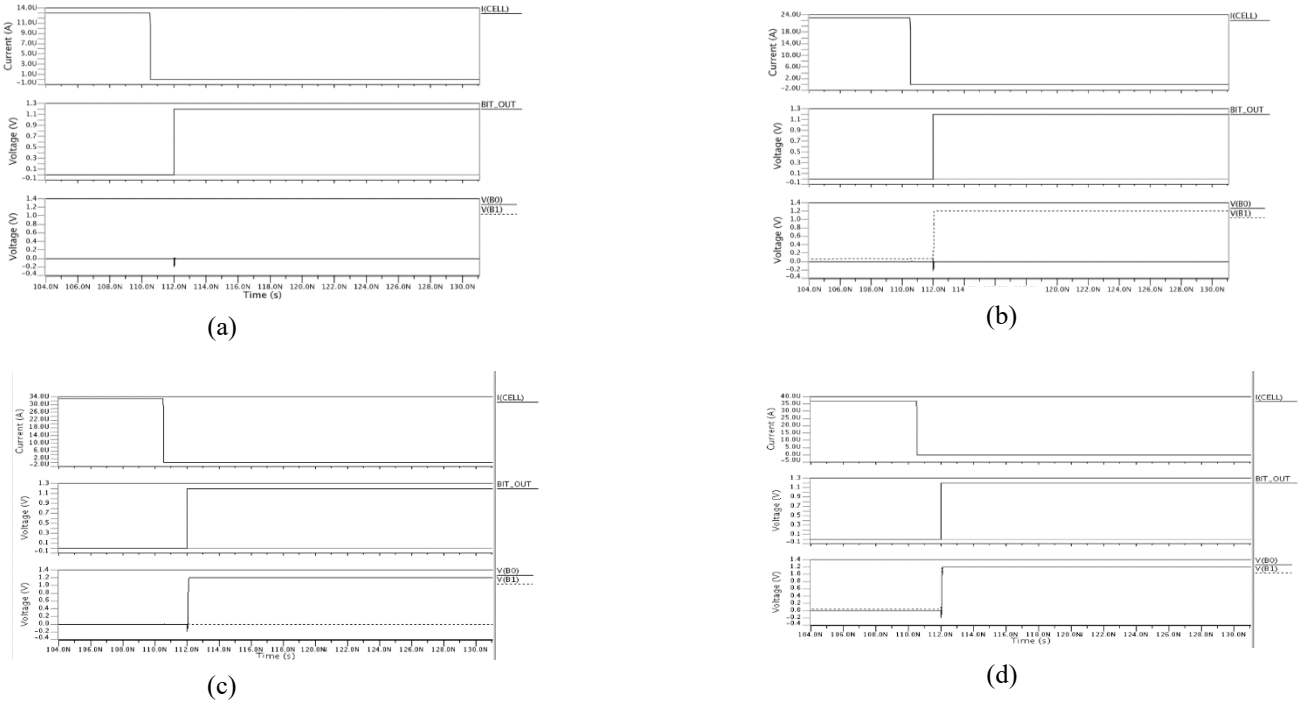


Fig. 6. 8 Output waveforms for MLC

Fig. 6.8 shows the output waveforms along with the input stimulus. The access time for the sense amplifier is 12.04nS, and the average power consumption per reading operation is 106 uW/read.



(a) (b)

(c) (d)

Fig. 6. 9 2 bit/cell sensing

(a) ’00’ stored with Icell=10uA (b) ‘01’ stored with Icell=20uA (c) ‘10’ stored with Icell=30uA

(d)’11’ stored with Icell=40uA

Table 6.4 compares this work with [10] and [12]. It can be seen that this sense amplifier has better access time and area overhead compared to the other schemes and can be used for multilevel sensing.

	[10] Single Bit	[12] Single Bit	This work	
			Single Bit	MLC
Year	2019	2015	2020	2020
Technology	40nm	130nm	65nm	65nm
Supply	1.1V	1.2V	1.2V	1.2V
Access Time	17.5nS	25nS	8.65nS	12.04nS
Offset	1uA	3.5uA	1uA	1uA
Area	-	$300um^2$	$110um^2$	$327um^2$

Table 6. 4 Comparison of Modified Current Sense Amplifier with references.

7 Conclusion and Future Work

7.1 Summary

In this work, we have proposed two different topologies of the sense amplifier, and their performance metrics were evaluated. The delay programmable type sense amplifier has a disadvantage that for multilevel reading the capacitors have a variable response with respect to different values of cell currents. This problem has been eliminated in the modified current mirror type sense amplifier by converting the absolute values of cell and reference currents to their algebraic differences.

It is concluded that for multilevel reading, the margins between two levels of cell currents should not overlap to improve sensing reliability. Since power is major performance metrics, especially in case of a multilevel read, we have applied coupling capacitors so that the comparator latch can work on lower voltages.

7.2 Future Work

This work can we extended by integrating the modified current sense amplifier on a PCM array and testing the circuit on the system level so that it can be implemented on silicon.

References

[1] Mutlu, Onur. "Main memory scaling: Challenges and solution directions." *More than Moore technologies for next generation computer design*. Springer, New York, NY, 2015. 127-153

[2] Sebastian, Abu, et al. "Memory devices and applications for in-memory computing." *Nature Nanotechnology* (2020): 1-16.

[3] Meena, J.S., Sze, S.M., Chand, U. *et al.* Overview of emerging nonvolatile memory technologies. *Nanoscale Res Lett* **9,** 526 (2014). https://doi.org/10.1186/1556-276X-9-526

[4] Bez, Roberto, et al. "Introduction to flash memory." *Proceedings of the IEEE* 91.4 (2003): 489-502.

[5] Burr, Geoffrey W., et al. "Phase change memory technology." *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 28.2 (2010): 223-262.

[6] Lai, Stefan K. "Flash memories: Successes and challenges." *IBM Journal of Research and Development* 52.4.5 (2008): 529-535.

[7] Wong, H-S. Philip, et al. "Phase change memory." *Proceedings of the IEEE* 98.12 (2010): 2201-2227.

[8] Redaelli, Andrea, Redaelli, and Lekhwani. *Phase Change Memory*. Springer, 2017.

[9] Kim, L-S., and Robert W. Dutton. "Metastability of CMOS latch/flip-flop." *IEEE Journal of solid-state circuits* 25.4 (1990): 942-951.

[10] Shih, Yi-Chun, et al. "Logic Process Compatible 40-nm 16-Mb, Embedded Perpendicular-MRAM With Hybrid-Resistance Reference, Sub-$\mu $ A Sensing Resolution, and 17.5-nS Read Access Time." *IEEE Journal of Solid-State Circuits* 54.4 (2019): 1029-1038.

[11] Grover, Anuj, et al. "Low standby power capacitively coupled sense amplifier for wide voltage range operation of dual rail SRAMs." *2015 International Conference on IC Design & Technology (ICICDT)*. IEEE, 2015.

[12] Zhang, Hua, and Ling Lu. "A low-voltage sense amplifier for embedded flash memories." *IEEE Transactions on Circuits and Systems II: Express Briefs* 62.3 (2014): 236-240.

[13] Razavi, Behzad. "The StrongARM latch [a circuit for all seasons]." *IEEE Solid-State Circuits Magazine* 7.2 (2015): 12-17.

[14] M. H. White, D. A. Adams, and J. K. Bu. On the go with SONOS. IEEE Circuits and Devices, 16(4):22–31, 2000

[15] Yadav, Jitendra Kumar, and Mohammad S. Hashmi. *Sense amplier for flash memories: architectural exploration and optimal solution*. Diss. 2015.

[16] Athmanathan, Aravinthan, et al. "Multilevel-cell phase-change memory: A viable technology." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6.1 (2016): 87-100..

[17] Bedeschi, Ferdinando, et al. "A fully symmetrical sense amplifier for non-volatile memories." *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No. 04CH37512)*. Vol. 2. IEEE, 2004..