Sense Amplifier for Flash Memories: Architectural Exploration and Optimal Solution

Student Name: Jitendra Kumar Yadav

RollNumber: MT13156

June 05, 2015

Indraprastha Institute of Information Technology

New Delhi

Under the Supervision of

Dr. M.S. Hashmi (IIITD)

Mr. Vikas Rana (ST Microelectronics)

Submitted in partial fulfillment of the requirements

for the Degree of M.Tech. in Electronics and Communication Engineering,

with specialization in VLSI and Embedded Systems

This research has been done with the collaboration of ST Microelectronics Pvt. Ltd., Greater Noida.

Keywords: Flash Memories, NOR Flash, Read Path Optimization, Sense Amplifier, Latch Sense Amplifier, design and implementation, and 40nm CMOS

Certificate

This is to certify that the thesis titled **"Sense Amplifier for Flash Memories: Architectural Exploration and Optimal Solution"** submitted by **Jitendra Kumar Yadav** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *Electronics and Communication & Engineering* is a record of the bonafide work carried out by her / him under my / our guidance and supervision at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. M.S. Hashmi

Indraprastha Institute of Information Technology, New Delhi

Mr. Vikas Rana

ST Microelectronics, Greater Noida

Abstract

Nowadays, Non-Volatile Memories (NVM) is part of every electronic system which requires any data storage when power supply is off. Out of many available NVM solutions, Flash memories are the most powerful and cost effective solid state memory technology for portable embedded applications and mobile electronic devices. For these applications fast access time, low power and high density are critical objectives. To accomplish these, stringent design requirements are imposed upon read path of the memory. It is a well established fact that Sense Amplifiers (SA) is the heart of the read path. It is upon the SA to detect and decide the content stored in memory cell. Hence, design of the sense amplifier becomes crucial because any flaw will lead to erroneous bit at the output. Key performance metrics for SA are read access time, power consumption and offset. SA must also have robustness towards any variation in temperature, supply voltage and process. Therefore, to achieve desired performance for NVM, an optimally operated SA must be utilised in the read path.

In this research, we have designed and implemented frequently used industry standard sense amplifier topologies on 40nm STM40 triple well CMOS technology and elaborated upon the technical merits of these topologies. Also an effort has been made to segment these topologies according to the specific application areas.

Acknowledgement

I would like to express my special thanks to my guide Dr. M.S. Hashmi for his constant guidance and motivation. I would also like to extend my appreciation towards Mr. Vikas Rana for his valueable time and Mr. Ganesh Raj for mentoring me technically.

A special thanx to my family, their prayers have sustained me this far. I thank all my classmates for their valuable suggestions and fruitful discussions. I am also thankful to Pallavi Das and Pragya Sharma for all the love and care during this journey, Gaurav Narang for his brotherly support.

Jitendra Kumar Yadav

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Flash Memory Overview

In this era of System On Chip (SOCs), embedded memories are indispensable part of any electronic system [16]. Embedded semiconductor memories can be broadly classified into two categories Random Access Memory (RAM) and Read Only Memory (ROM). While RAMs can change their content virtually infinite number of times, ROMs either do not have the flexibility to alter their content or might do it for the very limited number of times in their lifetime. One more distinct feature between RAM and ROM is that the RAM can only retain its content until power supply is ON while ROM retains their content ideally forever. An Ideal Memory would contain write feature of RAM and retention feature of ROM. The memory which is closest to this Ideal memory is non-volatile memory (NVM) [29].
The first and one of its kind NVM is Erasable Programmable Read Only Memory (EPROM). This memory could be electrically programmable but needs to be exposed to UV light for erasure. EPROM has only one transistor per memory cell hence, it is cost effective and highly dense. Another kind of NVM is Electrically Erasable Programmable Read Only Memory (EEPROM). EEPROM is electrically programmable and erasable to the finest granularity of bytes but they are expensive due to its complex memory cell structure. The application area of above discussed NVM was limited due to the inherent limitations that they possess. But the introduction of Flash memories has extended the use of NVM for many other portable and personal electronic devices. Flash memories, on the one hand have electrical erasure property of EEPROM while having cost per bit comparable of EPROM. This impressive property of Flash memories have made it compatible with variety of applications varying from mass storage devices to high speed automotive chips [29].

1.1.1 Types of Flash Memories

Two dominant flash memory technologies are NAND flash and NOR flash technology [20]. Fig 1.1 shows summary of how these two flash technologies, NAND and NOR, are distinct on different design dimensions i.e. Read and Write Speed, Power, Capacity and Cost per bit. The choice of any particular flash technology is entirely dependent upon the application for which it is used. Both technologies are discussed here in brief.

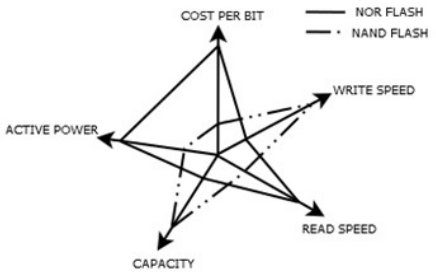


Figure 1.1: Comparison of NAND and NOR Flash Memory Technology

NAND Flash Technology

In NAND flash technology the basic memory cells are arranged in a similar fashion as of NMOS are in CMOS NAND gate implementation. This architecture helps to share source and drain diffusion areas of adjoining cells of same row which reduces the layout area for NAND cell. So generally, NAND Flash systems are designed to have very low cost per bit making it ideal for high density data storage consumer applications [19] [13]. Though, due to the same architecture, it has slow random read access. NAND flash technology finds its application as a sequential data storage element in file systems, video recorders and USB disk drives.

NOR Flash Technology

In the array configuration of NOR flash, basic memory cells are connected in parallel to achieve random access. The memory cells are organized in such a way that all the cells in the same sector have a common ground node and the bitlines are directly connected to the drains of memory cell. This enables short read times for applications, like microcontrollers, where fast random data access is needed [5]. For fast access time, NOR flash compromises with its array density. Generally, NOR flash memories are used for boot code storage in SOCs and Smart Cards. The feature, which enables use of NOR flash as an in-system commodity for the storage of code as well as data, are mainly due to NOR array organization [22]. The array organization is shown in Fig 1.2(b). This kind of array organization is suited for applications requiring high speed and noise immunity because there is direct access to the memory cell.

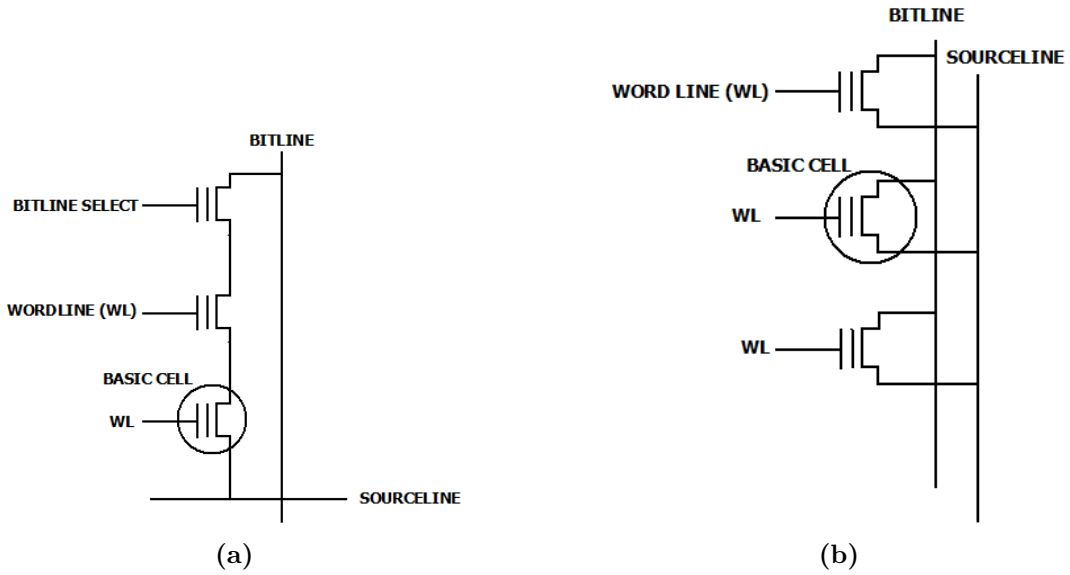


Figure 1.2: (a)NAND Flash Array , (b)NOR flash Array

1.1.2 Flash Memory Cell

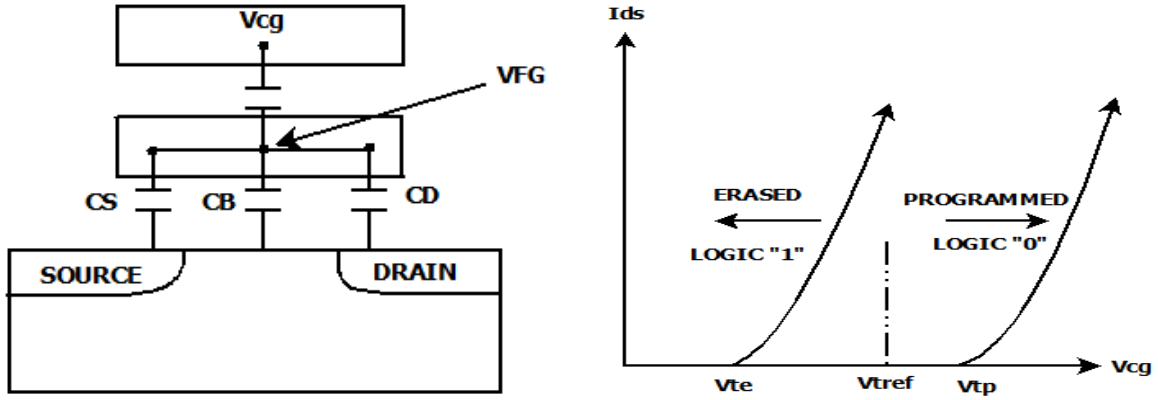
The memory cell needs to have means to alter the information electrically in non-destructive way. One of the solutions, and indeed the best one, is to alter the threshold voltage of the cell so that different threshold may represent different sates of the memory [5]. For two level flash cells these states are called as erased and programmed states for low threshold and high threshold devices respectively. The threshold voltage of MOS is related as

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

where K is a constant which depends upon gate and substrate material, channel doping and oxide thickness. C_{ox} is 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 MOS is Q i.e. charge trapped into the oxide. Fortunately there are charge injection techniques available to move charges in and out of the oxide. Normal MOS device cannot be used to retain the charges into its oxide so, accordingly modified version of MOS, known as Floating gate (FG) device is used for the same purpose. FG transistors can retain charge in their floating gate for extended period even after supply is turned off. The cross section schematic of generic floating gate device is shown in Fig 1.3.

1.1.3 Charge Injection/Removal Phenomena

It has been established that electrons trapped into floating gate modify the threshold voltage of the transistor and hence the state of the memory cell. There are many solutions available to transfer electron from and into floating gate. Most widely used phenomenon uses hot electron



(a) (b)

Figure 1.3: (a)Cross section of FG device, (b)Threshold voltage shift of FG device

Table 1.1: Cell node voltages required in different memory operations

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	Floating	0	0
Erase	-8.0	Float	8.0	8.0	0	Floating	0	0

and tunneling effects. More specifically, Channel Hot electron Injection (CHI) is used for charge trapping while Fowler-Nordheim (FN) tunneling can be used for both charge trapping as well as removal in Flash memories [24]. These phenomenas are diagrammatically shown in Fig 1.4. These process involved voltages as high as 18V to be generated on chip and applied across FG device.

These processes induce some reliability concerns in FG device used in NVM memory design. These concerns are dealt with technology by specially designed MOS which can withstand high voltages required by CHI and FN tunneling. Also, CHI is selective phenomenon that means it can be applied to bit and byte level of granularity while FN tunneling is a bulk phenomenon and can be applied to the whole sector or the array. For the same reason, flash memories are bit programmable and sector or bulk erasable.

1.1.4 Operations on Flash Memories

For the operations of read, program and erase we need to apply different voltages on the terminal of the flash cell. Table 1.1 consists of required voltages for selected as well as non-selected cells and sectors in case of erase [7]. These high voltages are generated on chip by means of charge pup circuits from a single power supply. For the purpose of supplying these voltages to the memory cell terminals various decoders along with switches and level shifters are used.

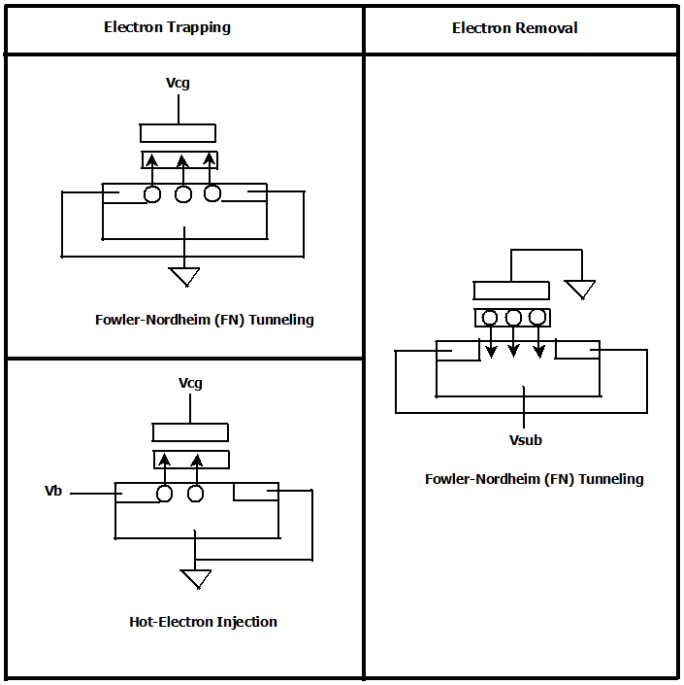


Figure 1.4: CHI and FN-tunneling phenomena, V_{CG} = High +ve voltage, V_B = Moderate +ve voltage, V_{SUB} = High -ve voltage

1.1.5 Test Modes in Flash Memories

Before moving to the further details it is important to have a look on different modes of operation in flash memories. Along with the user modes like read, program and erase there exists some test modes as well for analyzing the operating behavior of the device. These test modes are hidden from the end user and generally used for tasks like reference and memory cell characterization, silicon behavior characterization and testing of different blocks involved in memory architecture. Out of many test modes, two very significant ones, in terms of reading the memory, are Direct Memory Access (DMA) and Fast Direct Memory Access (FDMA) [6].

Direct Memory Access (DMA) Mode

The main purpose of DMA test mode is to be able to connect the cell terminals directly to the external I/O pads. Doing this enables device characterization, specially the matrix and the reference cells. Using DMA it is possible to analyze the matrix to the granularity of single cell, which is a major contribution. The setup path for DMA test mode is shown in Fig 1.5.

As shown in the setup path, DMA mode bypasses sense amplifier and the output latch and connects drain of the cell directly to the I/O pad. On this I/O pad an external supply is connected whose value is equal to that of forced one on drain node by the sense amplifier in read mode. In DMA mode gate voltage which is to be supplied to selected array cells is also supplied through an external pin. This way DMA enables measurement of cell current, trans-conductance and threshold voltage of cells by varying the bias conditions.

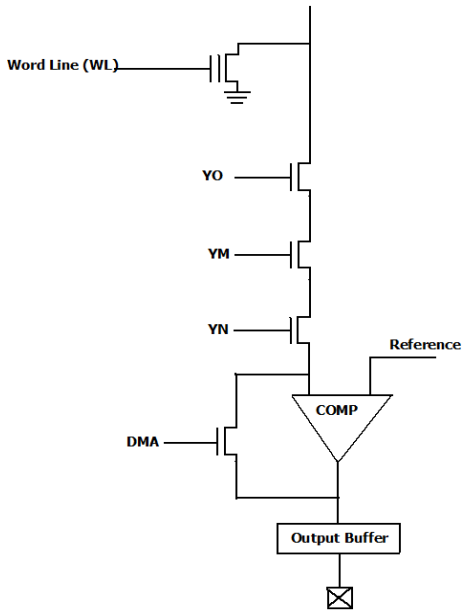


Figure 1.5: Setup for DMA operation

Fast Direct Memory Access (FDMA) Mode

Operating in DMA mode and measuring cell current at different biases is time consuming procedure. So another test mode to fasten this procedure is Fast DMA. FDMA mode is similar to the normal read mode with a slight difference. In FDMA mode a stable reference current is forced and cell current is compared against it in sense amplifier. This stable current can be internally generated or can be forced externally through DMA pin. Also in this mode as well gate voltage can be controlled through an external I/O pad. By changing this current and gate voltage in small increment we can plot the cell characteristics e.g. threshold voltage. Advantage of this mode is that, due to involvement of read operation, it is very fast.

1.2 Motivation and Aim of the Research

The demand of increasing system reconfigurability and exponential growth of computational complexity in modern electronic systems have given rise, the need to use on chip embedded Non Volatile Memory (eNVM). For embedded systems, non-volatile nature of on board flash memories is appealing part where flash memories is used for program and data storage. With the continuous thrust to improve system performance (speed, power consumption and area), flash memories will also have to cope up with this trend. This has lead to stringent design requirement imposed on read path of memory. Sense amplifier being the heart of memory read path, majorly contributes to the read performance of memory. Hence it is very important to choose an optimally designed sense amplifier topology.

In this research, we have designed and implemented most frequently used industry standard sense amplifier topologies and elaborated upon the technical merits of these topologies. Also an effort has been made to segment these topologies according to the specific application areas.

Chapter 2

The Readpath of Flash Memory

One of the main parameters to decide the performance of the memory is the access time i.e. how fast we can read from the memory for a given address. Considering the requirement of high speed embedded NVM, read path of the memory becomes critical [23]. For such high sped applications it is important to have speed and reliability both in place for read path elements. Fig 2.1 shows various building blocks involved in read path of the memory. To optimize the read performance of the memory every block involved in the read path must be analyzed.

2.1 Address Transition Detector (ATD)

ATD detects any change at the input of the memory address and start-off the operation by providing a pulsed signal of appropriate width. The ATD pulse also activates decoder circuits, which in turn biases the row and column of memory array with proper voltages.

2.2 Row and Column Decoders

In the memory array organization memory cells are arranged in the form of rows and columns. To read a particular memory location, we need to select a row and a column based upon the given address. This task is accomplished by the decoders. Decoders convert n input bits into 2^n output bits, only one of which is high at a time corresponding to the selected cell. In flash memories, decoding circuitry is not only meant to select desired memory cell from the array but also to pass high voltage required in different memory operations of read, write and program. For both row and column decoders, operating principles and design strategies are similar. Generally, for the read operation one row is selected and multiples columns are selected, depending upon the number of bits that we want to access at a time. For each selected column there is a sense amplifier to detect the content of the cell.

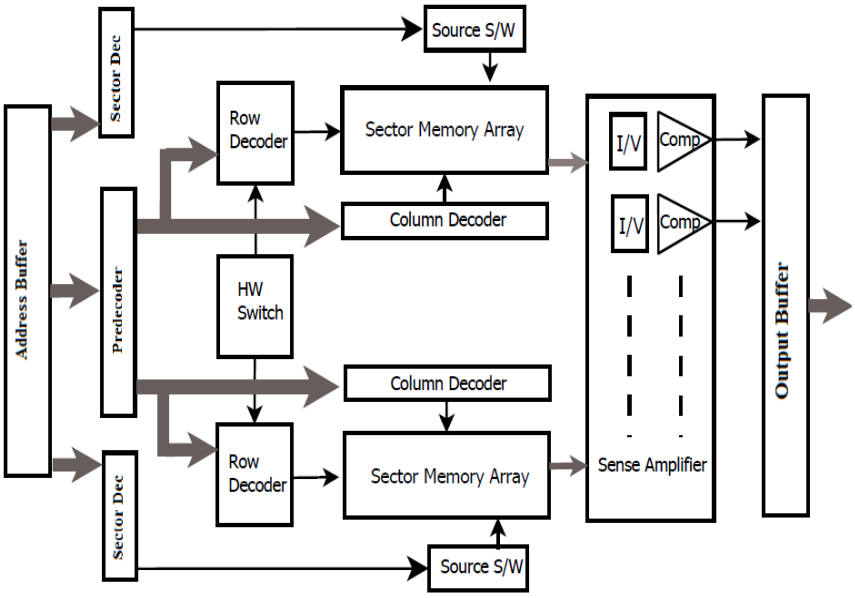


Figure 2.1: Readpath of Flash memory

2.3 Source and Bulk Switches

Gate of the selected cell is driven through word-line driver while drain terminal, in case of read, is controlled by sense amplifier. For the case of program drain voltage is passed through column decoder. We also need to bias the bulk and source node of the selected cell. For this purpose source and bulk switches are used. In flash array organization, source and bulk of all the cells are tied together for same sector. These switches drive source and bulk nodes of selected sector to different voltages depending upon operation.

2.4 Sense Amplifier (SA)

Sensing the content stored in the memory cell is perhaps the most important operation in the memory and the block which accomplishes that is known as sense amplifier. Sense amplifier is also known as heart of the readpath [23]. It decides the content of the memory cell by comparing the current drawn by selected memory cell from the matrix with the current drawn by reference cell, under same bias condition. Design of sense amplifier block is of paramount importance in the readpath of the memory because it is upon sense amplifier to decide the memory content. If, in case, sense amplifier does not interpret the memory cell current correctly then that leads to erroneous bit at the output. Hence, the requirement from sense amplifier is to have robustness towards any variation in temperature, supply voltage and process. Also there are other metrics to decide performance of any sense amplifier.

Broadly speaking all the sense amplifier topologies can be divided into two major categories which are static decision topologies and dynamic topologies. Static topologies, on one hand, rely on steady state value of the comparator nodes while on the other hand dynamic topologies

rely on the transient behavior.
The performance metrics of sense amplifiers under consideration is briefly explained here.

2.4.1 Performance Metrics of SA

Evaluation Time

There is a general trend that memory speeds are not catching up with that of processor speed. Memories have traditionally been on the slower side hence, there is always a constant force to have faster memories. For NVMs their speed requirements majorly depend upon the application for which they are designed. Embedded NVM application areas are very broad ranging from high speed microcontroller unit (MCU) to low power smart cards. To achieve high speeds, the access time of the memory must be as low as possible. Though this does not entirely depend upon sense amplifier block, but access time can be very much optimized from proper choice of sense amplifier topology according to the target application.
The total access time of the memory is divided into its sub-parts. In general, the these sub parts are Precharging Time T_{PRE}, Sensing Time T_{SENSE} and Latching Time T_{LATCH}. The total access time is sum of all of them and given by

$$T_{ACCESS} = T_{PRE} + T_{SENSE} + T_{LATCH} \tag{2.1}$$

Through the sense amplifier we can control the first two part of the above equation while latching time is generally fixed depending upon the output load and buffering time.

Offset

In embedded NVM along with the user mode operations like read, program and erase there are other test modes like Direct Memory Access (DMA) and Fast Direct Memory Access (FDMA) which needs to be performed before the NVM chip goes into end product. Among these, FDMA operation requires small current difference to be resolved by sense amplifier correctly. For this purpose sense amplifier must have as high resolution as possible because then that will lead to precise FDMA operations.
To measure the resolution, worst case random offset analysis is performed on sense amplifiers. For an ideal sense amplifier, its output voltage should be half the supply voltage when cell current is equal to the reference current. But due to various process mismatches in submicron technology, the sense amplifier shows offset from this ideal behavior. Offset for sense amplifier is measured in terms of current and defined as the difference of cell current and reference current when the output voltage crosses $V_{DD}/2$. For proper operation, the current difference between matrix cell and reference cell must always exceed this offset. Hence sense amplifier offset must be as low as possible.

Power

Number of sense amplifiers are decided upon the degree of parallelism that we want to achieve, which is generally varying from 16 bits to 256 bits. Each of the sense amplifiers used will burn power at the time of reading. Since the eNVM are going to be integrated with some another system, the power constraint are, generally put by the overall consumption of lets say, a SOC. Hence, the power consumption by the sense amplifier plays an important role in the overall consumption and must be minimized. In the simplest terms, the average power consumption of sense amplifier in read cycle is given by

$$P_{AVG} = V_{dd} I_{AVG} \tag{2.2}$$

Where, V_{dd} is supply voltage used I_{AVG} is the average current consumed by sense amplifier in a read cycle.

Area

Area occupied by sense amplifier is also important because of the degree of parallelism, as the number of sense amplifiers increases area overhead becomes more critical factor. Although in NVMs majority of the area is occupied by the memory array and the rest of the circuitry occupies relatively less area.

2.5 Reference Matrix

In flash memories the current drawn by the matrix cell is compared in sense amplifier with a reference current. This reference current is taken from the cells of reference matrix. Reference matrix contains cells of different thresholds, generally of four type program reference, erase reference, read reference and verify reference. These cells are used in different operations to compare against the matrix cells, hence their characteristics needs to be precisely set. These cells are arranged in a matrix. After fabrication boundary cells are deactivated by giving fixed electrical conditions so that they do not consume any current. Those cells which are in the center are clubbed with n bits each and their threshold is set by selective programming followed by FDMA. To minimize the effect of process variation on reference matrix, at any time average current of n cells of same group are taken out. Hence for reference cells there is a tradeoff between the reference matrix area and reference reliability.

2.6 Output Buffer

From the sense amplifier output to the I/O pad output buffer carries the sensed data. A simplest buffer contains series of inverters. Characteristics of a buffer include its ability to drive large capacitive loads and low transient delay. A more sophisticated scheme involves buffer controller and data latch control unit, all to prevent data to get corrupted.

Chapter 3

Design Constraints on Sense Amplifier

Interestingly along with the performance, reliability of the flash cell also depends upon the sense amplifier design. Reliability in NVM includes endurance and retention. Endurance defines for how many time we can reprogram the array and retention defines for how long the data can be retained into the flash cell, if not altered. As it has been discussed, the drain of the flash cell is connected to the sense amplifier through column decoder therefore, it is the job of sense amplifier to bias this drain node of the selected cell of the array. This drain bias defines an unavoidable phenomenon of flash memories known as read disturb. It is upon sense amplifier to avoid possible situation leading to read disturbs of the cells. This has been described here.

3.1 Read Disturbs

In the read operation of the flash memory, biases applied are of same kind as in the program mode but with a lower magnitude, as can be seen in Table 1 as well. The flash cell scaling has made it possible to generate spurious hot carriers even at low V_{DS} used for reading. This leads to read disturbs which is commonly known as soft programming [18]. As the read operation is most frequent operation, it can slowly change an erased cell into a programmed one by cumulative injection of charges in FG. For a programmed cell it can further increase its threshold voltage which makes it difficult to erase. For proper erasure i.e. threshold of every cell is below a certain value, an extra erase pulse might be needed which makes some of the cells over-erased also known as depleted cells. This situation is very harmful as it can upset the threshold distribution of the cells. It is particularly harmful for the reference cells of the sense amplifier as they are always expected to give fixed and stable current under predefined bias and any change in reference current may lead to incorrect output of current comparison. Hence to avoid read disturb cells drain voltage must be low but at the same time high enough to allow proper current to flow and guarantee quick reading.

So it is important to keep the drain voltage of flash cell in predefined limits, which is generally defined by the process and the choice of flash cell. For a designer this limit comes as a specification.

3.2 Need of Cascoding

Cascoding the bitlines is a necessity in flash memories to avoid read disturbs by keeping the V_{DS} of flash cell low. Fig 3.1 shows two basic cascoding schemes, constant bias and inverter based.

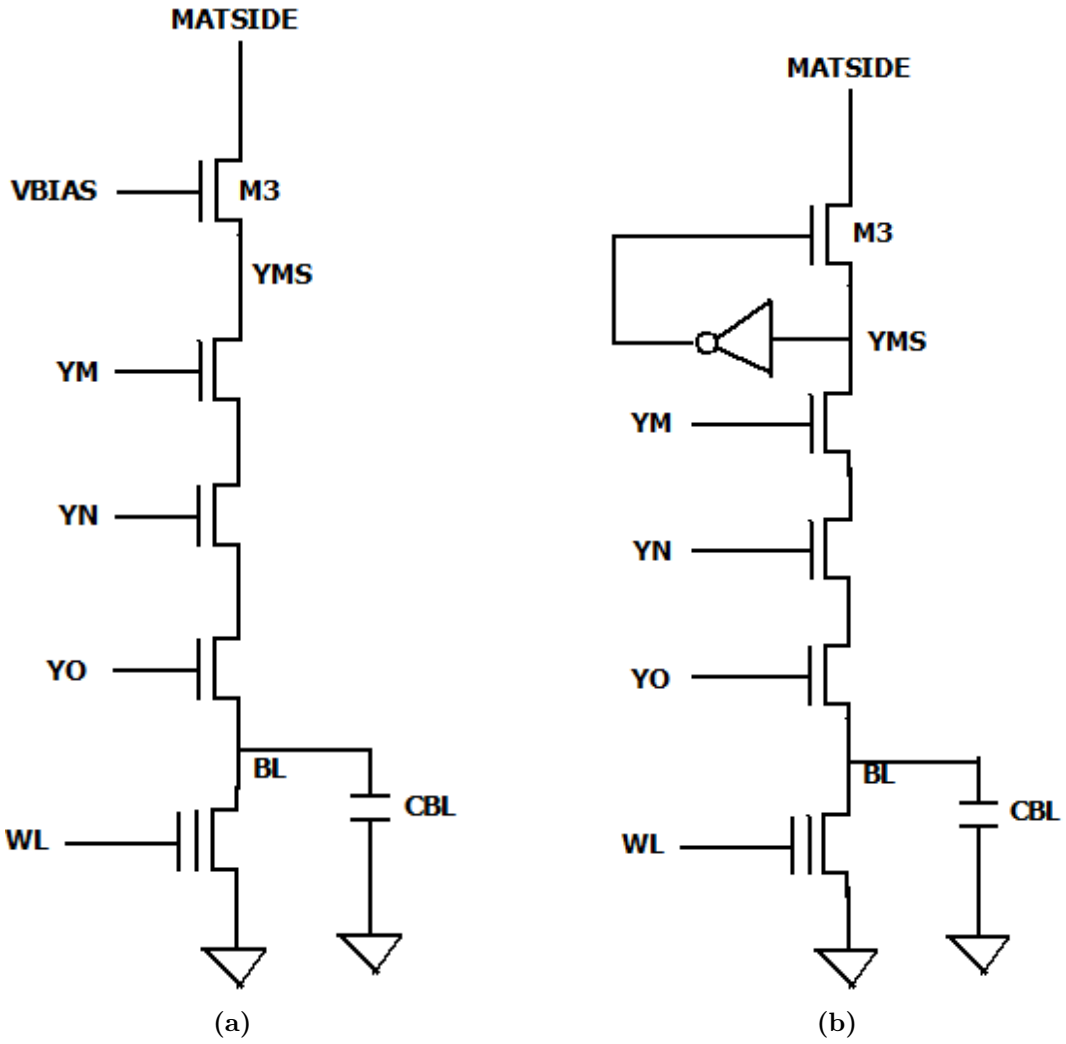


Figure 3.1: Typical cascade biasing schemes (a) Constant Bias (b) Inverter Based Approach

The YMS node shown reaches to the bitline of flash cell after column decoding. The capacitance on this node includes the capacitances from column decoder also drain-gate and drain-source capacitance of all the cells of same column. Hence the YMS node is highly capacitive node. On the other hand the OUT node only includes the parasitic of various MOS connected to this

node and hence is very less capacitive. Due to this differential nature, any small changes in YMS node will lead to high swing at OUT node. This decoupled behavior of OUT node is achieved by use of cascode device. Cascoding device also solves the problem of read disturb by keeping the voltage at YMS node below specified value. In typical conditions the drain voltage is constrained at 1V [17].

3.2.1 Constant Bias Approach

For constant bias approach, a highly stable fixed value voltage is connected to the gate of cascode device M1 [7]. This kind of scheme may be preferred in area constrained designs where V_{BIAS} is generated once and tapped to all the sense amplifiers resulting in less area than having local inverter for individual sense amplifier. But this scheme results in poor dynamic behavior of the circuit. While charging C_{BL}, M1 transistor will have a limited V_{GS} as YMS node is virtually at ground potential and node V_{BIAS} is fixed. Now the question is to choose proper value of V_{BIAS} to prevent read disturb. For this purpose we need to see worst case threshold variation of cascade device M1, shown in Fig 3.2, and then equation 3.1 must be satisfied.

$$V_{BIAS} < V_{TH}\textbf{(WORST CASE)} + V_D \tag{3.1}$$

$$V_{BIAS} < (1.5 + 1)V \tag{3.2}$$

$$V_{BIAS} < 2.52V \tag{3.3}$$

Hence in our case, we have considered V_{BIAS} of 2.4V.

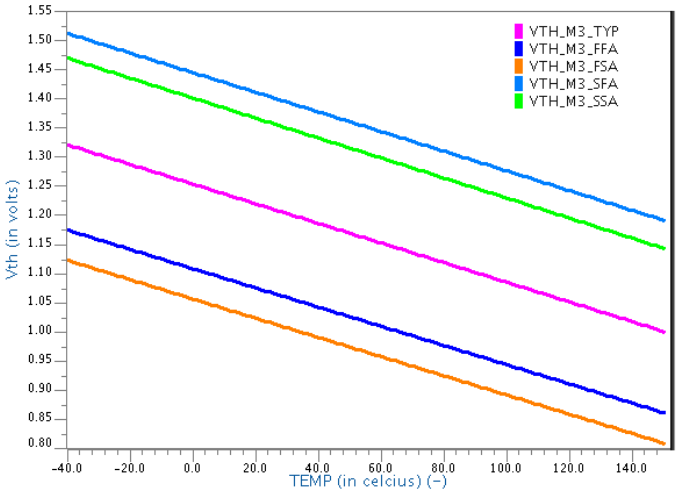


Figure 3.2: Worst case threshold variation of cascode device

3.2.2 Inverter Based Approach

In inverter based approach a negative feedback loop consisting CMOS inverter is used for the purpose of cascoding [7]. This has better dynamic behavior because when YMS is at ground, M1 can see an V_{GS} equal to the supply voltage. Also, by keeping $(W/L)_{NMOS} >> (W/L)_{PMOS}$ we can achieve a switching threshold closer to 1V to prevent any read disturbs.

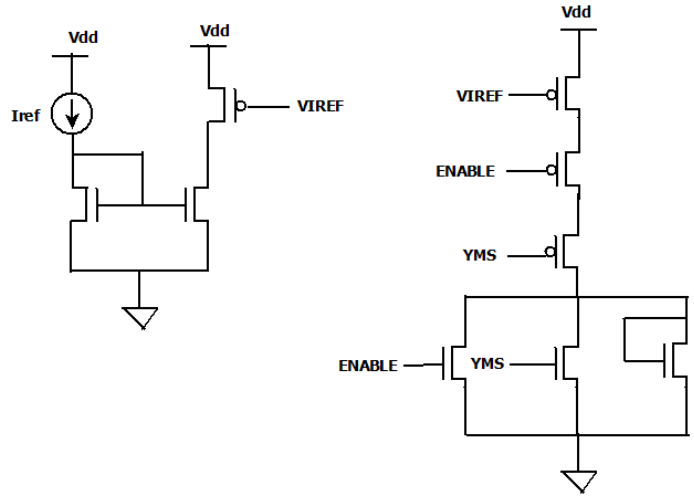


Figure 3.3: Constant current based NOR Design

Only problem in this kind of approach is high power consumption. Inverter in the switching region may consume high currents to settle the feedback loop. To solve this problem we have designed this inverter with a constant current so that its consumption can be controlled. The value of this current must be properly chosen so that settling time of the loop is not degraded. The designed inverter is actually a NOR gate with an enabling signal. When enabling signal is asserted, it behaves like an inverter. The design is shown in Fig 3.3.

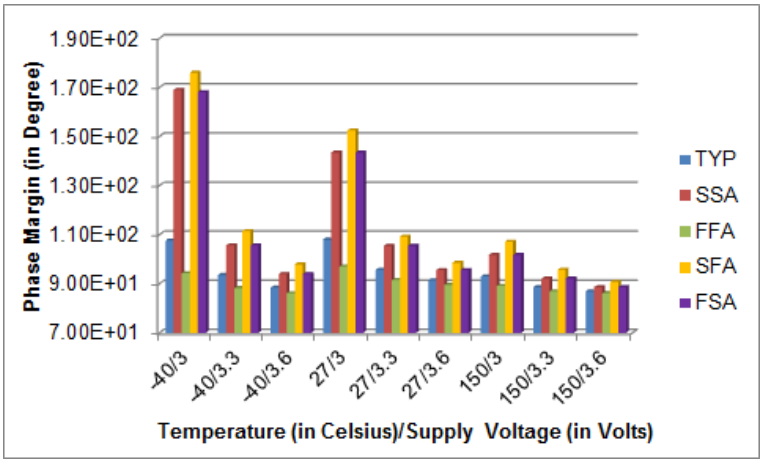


Figure 3.4: Loop Phase margin for inverter based cascoding

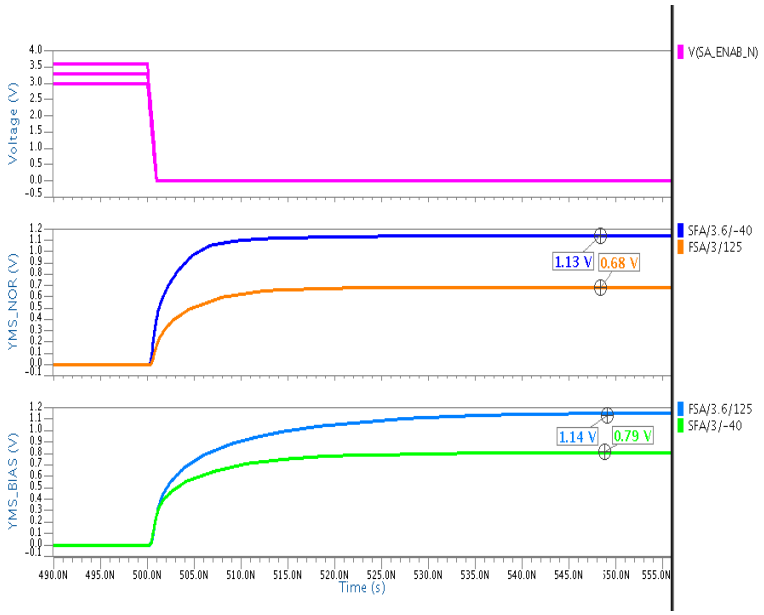


Figure 3.5: Spread of YMS node across PVT

Now, since this NOR gate is in connection with the cascade transistor and they together form a negative feedback loop, its important to analyze the stability of the loop. In this regard, Fig 3.4 shows the phase margin of the feedback loop. The result shows the feedback loop is well stable since the worst case phase margin is 87 degree.

Fig 3.5 shows the spread of the voltage at YMS node across all PVT conditions for both the approaches. Maximum value at YMS node, the bitline of memory cell, is within 1.15V. Also, this can be observed that spread in case of constant bias approach is less while in inverter based approach YMS settles faster. For the further cases of sense amplifier design we have implemented inverter based approach since speed is of major concern in NOR based flash design.

Chapter 4

Static Type Sense Amplifier Topologies

Static type sense amplifier topologies have the legacy to make their decisions on the basis of steady state value of the cell current. Static topologies are not sensitive to the transient behavior of the circuit and can wait for the transients to settle down. Due to any spurious behavior in the circuit, even if the initial decision made by sense amplifier is wrong but eventually after some time it settles in the correct direction given the steady state value of the currents sunk by cell is undistorted.

There is extensive work which has been done in the area of static sense amplifier topologies. The most basic topology which is also considered as the conventional one is reported in [3]. This topology uses I-V conversion branches and a differential amplifier to make the decision. A lot of improved differential structures have also been proposed [28] [26]. In [32] the modifications have been made keeping the low voltage supply requirements into prospective. Papaix, Caroline et.al. have designed a single ended sense amplifier reported in [27] but this sense amplifier cannot meet the high resolution requirement of current flash memories. An effort towards reducing the sense amplifier offset has been made in [4] [21]. Liu et.al. in [21] have proposed asymmetrical voltage biasing at sensing nodes to compensate the offset while offset detection was done through specially designed reference cells. This solution for offset detection is customized and hence this topology cannot be used as a replacement of currently used sense amplifier. Here, in this work we have designed three topologies including the conventional SA topology. The major focus was to explore different kinds of behavior of sensing nodes and conclude its impact on the overall sense amplifiers performance metrics.

4.1 Comparator

Normally sense amplifier itself is divided into different sections, in this case it is current to voltage conversion of matrix and cell currents and then comparison. The idea here is to convert the

currents drawn by the reference cell and matrix cell into voltage and then compare them into a comparator which can produce the output of logic levels. The current drawn by the reference cell is I_{REF} is equal to 8 µA. Under the same bias conditions, current drawn by the programmed cell is less than I_{REF} and that by the erased cell is greater than I_{REF}. This difference in current is to be converted to a sufficient differential voltage which is recognizable by the comparator used. Before discussing the I-V conversion, the comparator is explained since same comparator is used in further static sensing techniques.

To choose a suitable comparator for the sense amplifier we need to see the gain requirement from the minimum input differential voltage that is needed to be sensed after I-V conversion. A small setup has been created to evaluate the minimum voltage difference to be used as resolution of comparator. This setup is shown in Fig 4.1. As discussed under the section of TEST MODES, under FDMA operation we need to resolve small current difference which here is taken as 1 µA from the reference current of 8 µA. So I_{REF} is fixed and I_{CELL} is taken 7 µA in one case and 9 µA in another case. This has been run through all PVT corners and the difference in voltage between REFSIDE and MATSIDE has been plotted in Fig 4.2.

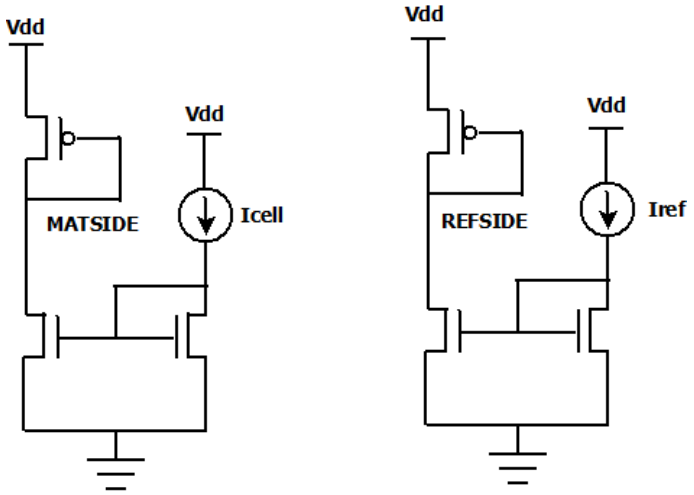


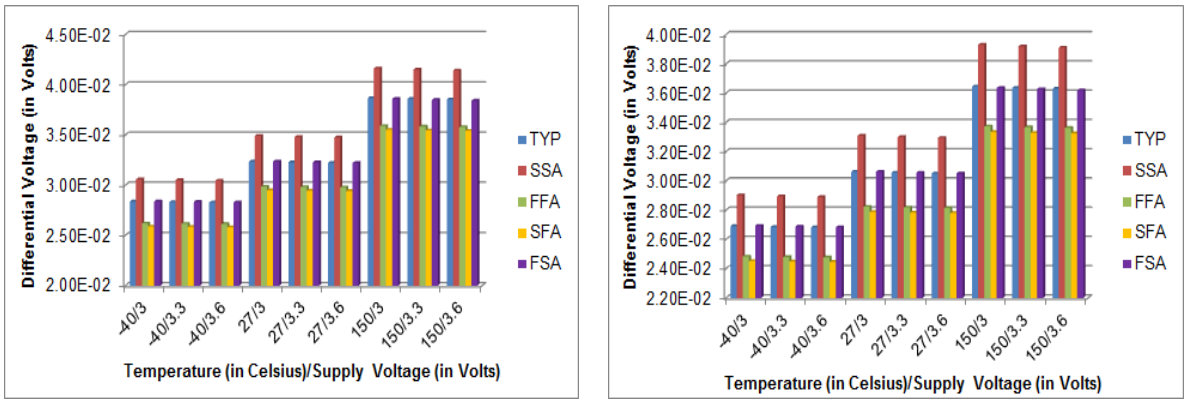
Figure 4.1: Setup to measure Differential voltage at two sensing nodes

From the graphs of Fig 4.2 it can be seen that minimum differential voltage between two branches of I-V conversion is coming out to be 24.5 mV in SFA corner at -40/3 with current of I_{CELL} of 9 µA. Typical value of supply voltage is 3.3 V. Hence we need a small signal differential gain given by equation 4.3.

$$Gain = \frac{(V_{OH} - V_{OL})}{\Delta V_{IN}} \tag{4.1}$$

$$Gain = \frac{(3.3 - 0)}{0.0245} \tag{4.2}$$

$$Gain = 134.69$$



(a) **(b)**

Figure 4.2: Differential voltage between REFSIDE and MATSIDE for (a)I_{CELL}=7μA, (b)I_{CELL}=9μA

$$Gain(dB) = 42.58dB \tag{4.3}$$

Though there are many options available to realize a comparator for gain of $42dB$. The simplest of which is a single stage differential amplifier working as a comparator. Fig 4.3 shows the schematic of such differential amplifier.

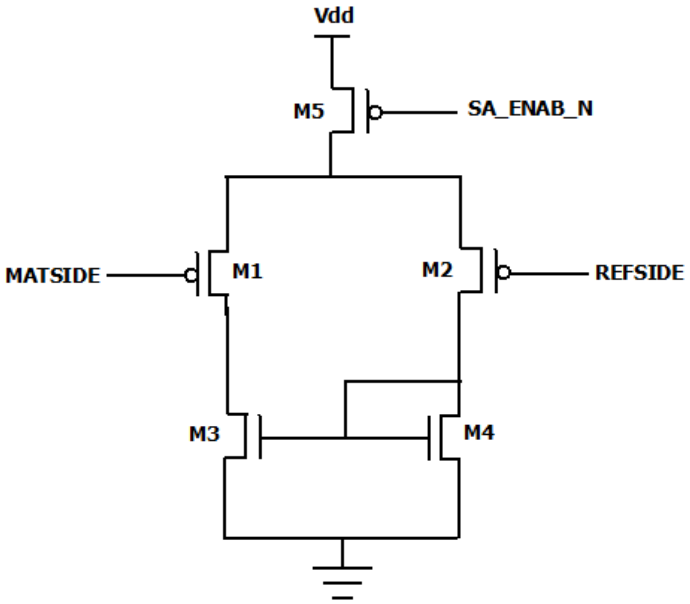


Figure 4.3: Differential amplifier used as a Comparator

The steps taken to design the differential amplifier are discussed in brief. The small signal gain of differential pair is given by

$$A_V = g_m R_{out} \tag{4.4}$$

Where,
g_m = Transconductance of input device of differential pair
R_{out} = Resistance seen at the output terminal OUT.
Further, g_m and R_{out} is given by

$$g_m = \mu_n C_{ox}(W/L)(V_{GS} - V_{TH}) \tag{4.5}$$

$$R_{out} = r_{01} \parallel r_{03} \tag{4.6}$$

For a given minimum overdrive voltage $(V_{GS} - V_{TH})$, ratio (W/L) decides the transconductance of the input device. Also to achieve higher gain R_{out} is maximized by keeping high length load devices which increases r_{03} and minimize the degradation of intrinsic gain of input device i.e. g_{m1} r_{01}. But R_{out} can not be increased indefinitely since increasing it reduces 3dB bandwidth and hence the speed. This explains typical gain bandwidth tradeoff of an amplifier. 3dB bandwidth of differential amplifier is given as

$$\omega_{-3dB} = \frac{1}{R_{out}C_L} \tag{4.7}$$

Along with the small signal parameters, large signal DC operating point is also to be set which will ensure all the devices operating in saturation region. This is defined by Input Common Mode Range (ICMR). For the design shown in Fig 4.3 maximum and minimum value of ICMR is given by

$$ICMR_{max} = V_{DD} - V_{SG1} \tag{4.8}$$

$$ICMR_{min} = V_{DSsat3} + V_{TH3} - |V_{TH1}| \tag{4.9}$$

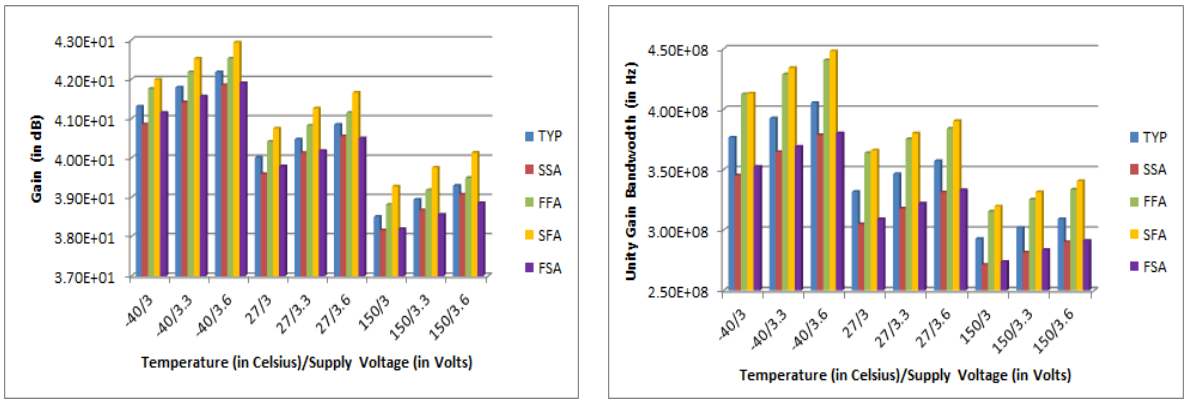
Also slew rate is limited by the current passing to the load capacitor. Maximum current that can flow through C_L, at any time, is I_{ref} hence SR is given by

$$SR = \frac{I_{ref}}{C_L} \tag{4.10}$$

Finally power dissipation of comparator is given by

$$P_{diss} = V_{DD}(I_{M1} + I_{M2}) \tag{4.11}$$

The variation of two important parameters of comparator, differential gain and unity gain bandwidth (UGB), has been shown across PVT in Fig 4.4(a) and (b) respectively. Finally, the rest of performance parameters achieved by the design of single stage differential amplifier operating as open loop comparator is summarized in Table 4.1.



(a) (b)

Figure 4.4: (a)Comparator gain across PVT, (b)Comparator UGB across PVT

Table 4.1: Worst case performance parameters for comparator

Performance Parameter	Value
Differential Gain	38.2 dB
Resolution	37 mV
Unity Gain Bandwidth	271 MHz
3-dB Bandwidth	2.43 MHz
Slew Rate	249 V/μSec
Offset	7.26 mV
Propagation Delay	7.14 ns

4.2 Conventional Sense Amplifier (CONV SA)

In our case the current values are

$$I_{REF} = 8\mu A$$
$$I_{CELL_P} = 2\mu A$$
$$I_{CELL_E} = 16\mu A$$

While in the case of FDMA I_{REF} is same but I_{CELL_P} and I_{CELL_E} is modified to

$$I_{CELL_P} = 7\mu A$$
$$I_{CELL_E} = 8\mu A$$

Where,
I_{CELL_P} = Programmed cell current
I_{CELL_E}= Erased cell current

4.2.1 Idea

The key idea for conventional differential sensing approach is to convert the reference as well as cell current into its corresponding voltage values and then amplify the difference between them by means of differential amplifier [3]. The differential amplifier used here is the one explained in previous section. This approach can be simply viewed as passing a known difference of current through a resistor to generate a differential voltage, which then is high enough to be sensed by following comparator circuitry. The resistors here are realized with active loads represented with transistors M1 and M2 in Fig 4.5.

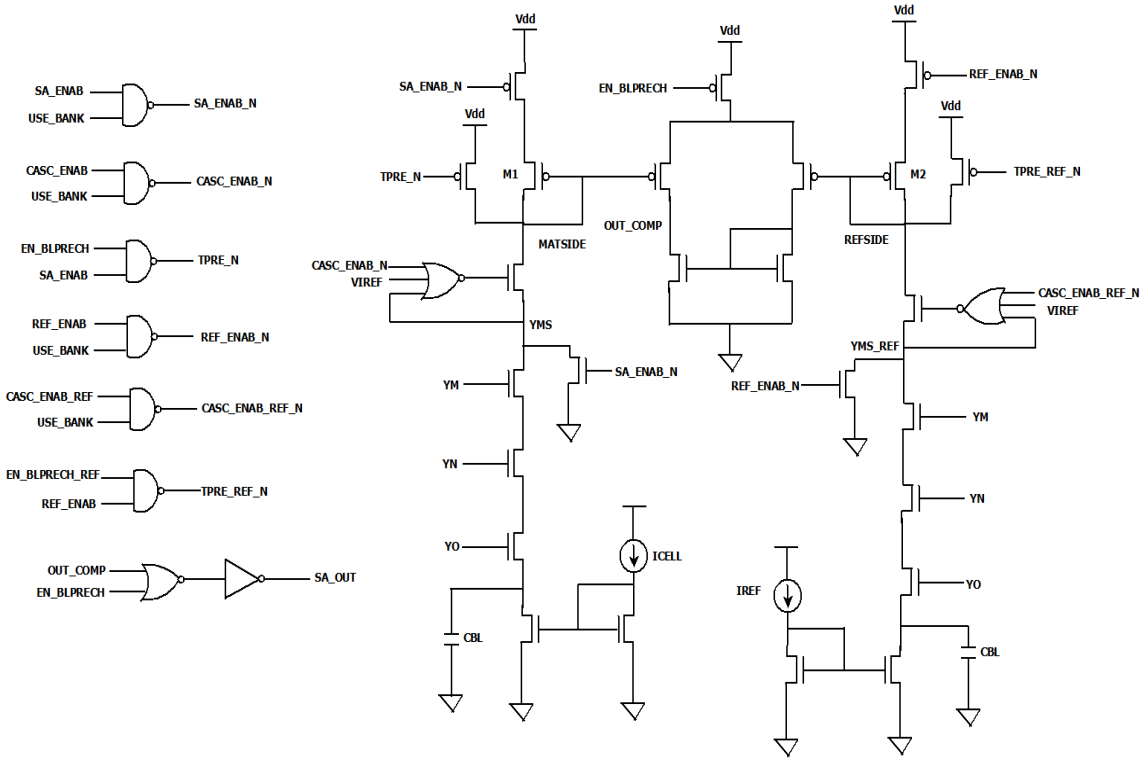


Figure 4.5: Full Schematic of Conventional Sense Amplifier

4.2.2 Working

Figure 4.5 shows the full schematic of conventional SA and Fig 4.6 shows sense phases used. This topology has two branches of current-voltage converter as identical, both in electrical properties and layout deposition. If MM, the matrix cell, is ideally programmed, no current is sunk in the matrix side and potential at MATSIDE (V_M) is at V_{CC} while potential at REFSIDE (V_R) is at a comparatively lower value due to constant reference current. However, in the actual case

programmed cell also sink some small current due to tail of its threshold variation which causes V_R to be lower than V_{CC}. On the other hand if MM is erased then it will draw more current than reference cell which will make V_M to go at lower potential than V_R. Hence in the former case comparator output is 0 while in the latter it is 1, representing programmed and erased bits.

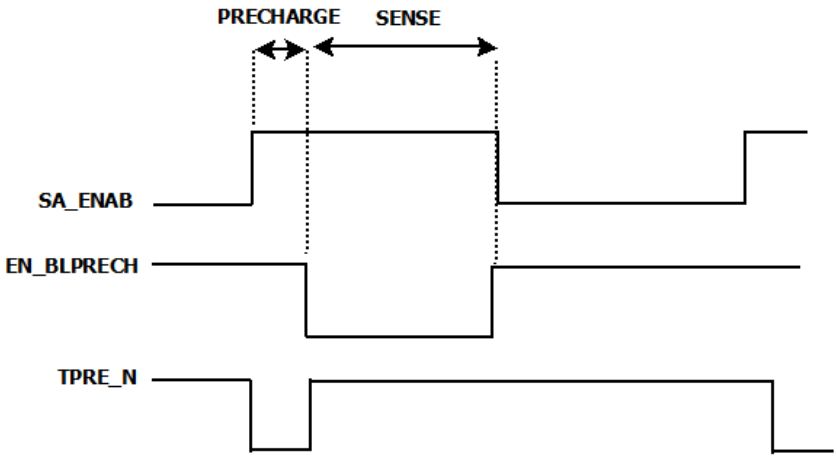


Figure 4.6: Phases of sense amplifier

4.2.3 Design Guideline

There are two conditions which have to be intrinsically satisfied to guarantee correct reading of matrix cell data. One is that the potential V_R should be in between the potentials of V_M node for programmed and erased case. This will ensure comparator flipping in opposite directions for cases of programmed and erased cell.

$$V_{M_E} < V_R < V_{M_P} \tag{4.12}$$

Where,
V_{M_E}= Potential at V_M node when matrix cell is erased
V_{M_P}= Potential at V_M node when matrix cell is programmed

The other condition is that the minimum difference between V_M and V_R at under worst case condition must be greater than the resolution of the comparator used, to flip the comparator in correct direction. This condition is given by.

$$\min\left(V_R - V_{M_E}, V_{M_P} - V_R\right) > \text{Resolution of Comparator} \tag{4.13}$$

While designing this topology first condition is taken care by ensuring same electrical conditions at both the sides of I-V conversion to generate proper differential current. The second condition, on the other hand, should be taken care while designing the comparator.

The active load is diode connected PMOS, which are always in saturation. The potential at V_R

is $V_{DD} - (V_{GS})_{PMOS}$. Going by the standard current equation of MOS

$$I_{SD} = \frac{1}{2}\mu_P C_{OX}\frac{W}{L}\Big(V_{SG} - |VTH|\Big)^2\Big(1 + \lambda V_{SD}\Big) \tag{4.14}$$

This after rearranging comes to,

$$V_{SG} = |V_{TH}| + \frac{2I_{SD}}{\mu_P C_{OX}\frac{W}{L}\Big(1 + \lambda V_{SD}\Big)} \tag{4.15}$$

This equation can define the common mode voltage level of V_R and V_M node. Proper care must be taken while deciding the voltage levels to keep cascode device in saturation.

Also,

$$r_0 = \frac{1}{\lambda I_D} \tag{4.16}$$

This equation suggests that keeping high L devices helps in obtaining sufficient channel resistance, which helps in having larger differential voltage i.e. $V_R - V_M$. But by doing so we are also increasing the capacitance at the sensing nodes, which will increase the charging/discharging time of the associated capacitance of sensing node and hence the sensing time. So appropriate choice of the aspect ratio for input devices must be made.

4.2.4 Results

Conventional SA is simplest in design since the only constraint is on the proper sizing of active loads. This topology has the advantage that there is minimal risk of incorrect reading and also there is no additional glue logic for sense phase generation is needed. On the other hand this topology suffers with high power consumption since the I-V conversion branch will consume full cell current and for highly erased cell this current consumption is even more severe. Also, in case of FDMA, for the currents which are very close, delay of the sense is very high.
Results of parameters access time, power consumption and sense amplifier offset is shown in this section. To report power consumption, worst case erased cell is taken whose cell current is 40μA. All the results are across all process voltage and temperature variation (PVT) condition. The temperature is varied from -40degree to 150degree centigrade. Typical value of supply voltage is 3.3V which is varied from 3V to 3.6V. Design has been run on standard set of 5 process corners.

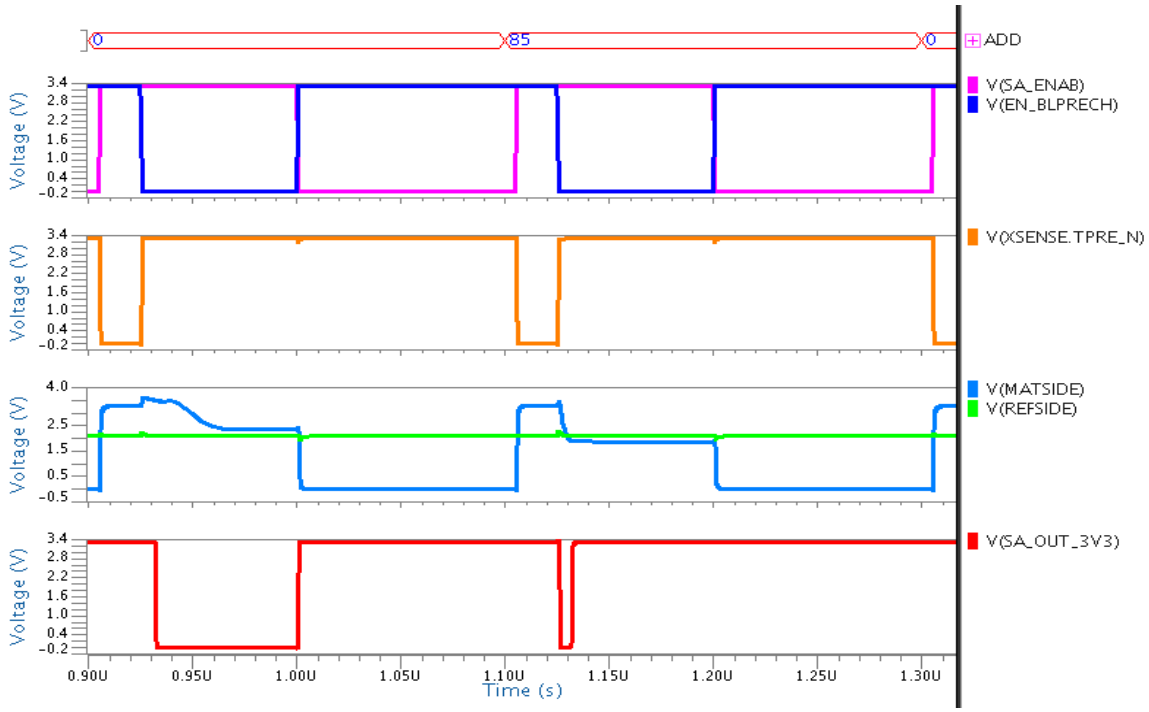
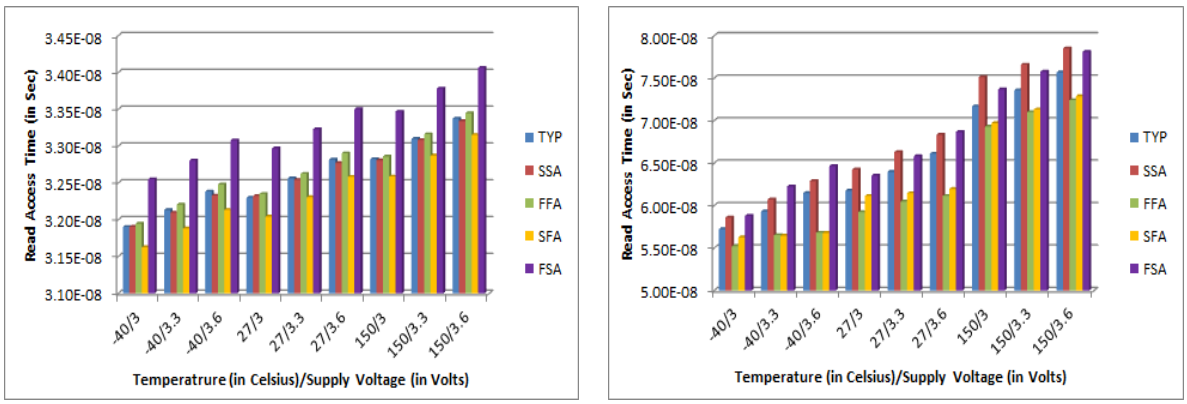
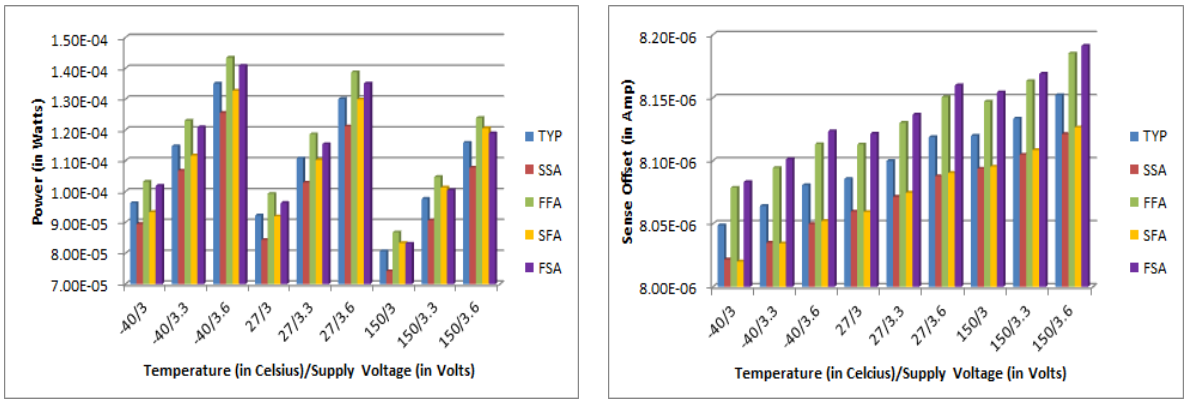


Figure 4.7: Read Waveform of Conventional SA



(a) (b)

Figure 4.8: Conventional SA (a)Worst case User mode Read Access time, (b)Worst case FDMA mode Read Access time



(a) (b)

Figure 4.9: Conventional SA (a)Power consumption for worst case erased cell, (b)Offset across PVT

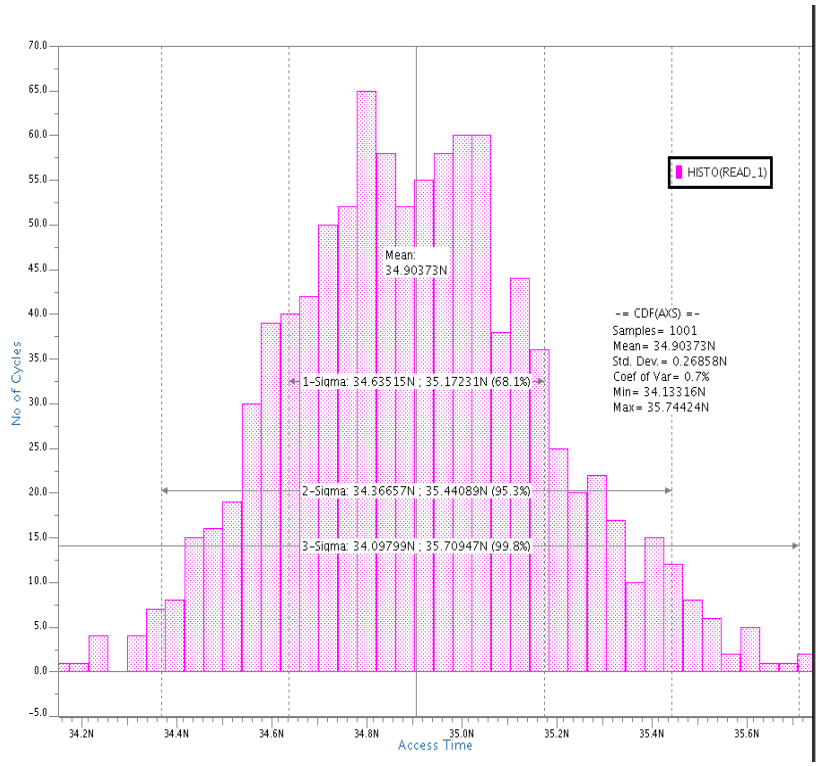


Figure 4.10: Monte-Carlo variation of user mode Access time for Conventional SA

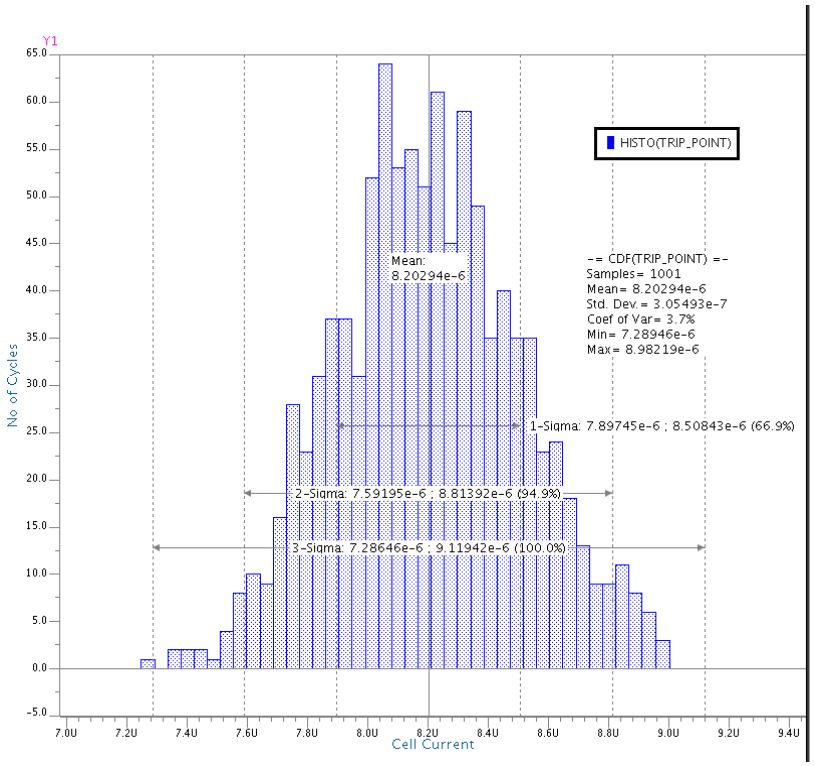


Figure 4.11: Monte-Carlo variation of Offset for Conventional SA

4.3 Mirror Sense Amplifier

4.3.1 Idea

To mitigate the disadvantage of Conventional SA topology that its power consumption is high, a change has been made in this topology which proves to be very effective as far as current consumption in the cell branch is concerned, without much of the area increment. The idea here is to limit the current in the cell branch by using the current mirror connection from the reference branch and collect the difference of I_{REF} and I_{CELL} into capacitor C_{PAR} [2]. This will make sure that the final value of current in the branch never exceeds I_{REF} itself. This is pictorially represented in Fig 4.12. Here the node $MATSIDE$ will adjust itself to limit the current in cell branch. This may result in significant power reduction.

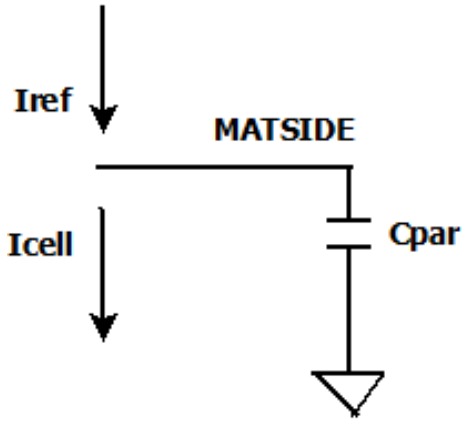


Figure 4.12: Idea behind Mirror SA

4.3.2 Working

Full schematic of this topology is shown in Fig 4.13. This is also a two phase sense amplifier 1st being the precharging phase and second being the sensing phase. During precharging phase signal $EN\_BLPRECH\_N$ is low which turns ON the precharge PMOS device. A high peak current flows to charge the bitline parasitic capacitance.

Here in this topology the parasitic capacitance of $MATSIDE$ node C_{PAR} is used to integrate the current difference of $I_{CELL} - I_{REF}$ thus giving rise to a voltage which is compared against a constant REFSIDE voltage V_R generated by reference I-V conversion branch. So, when $I_{CELL} > I_{REF}$ the parasitic capacitor C_{PAR} is charged and voltage at node $MATSIDE(V_M)$ is high. Then the result of comparison of V_M with V_R is 0. In the other case when $I_{CELL} < I_{REF}$ the parasitic capacitor C_{PAR} starts discharging from V_{DD}. Initially the comparator output might be 0 since $V_M > V_R$. But once C_{PAR} discharges below V_R the comparator flips in the other direction which results in the correct output data i.e. 1.

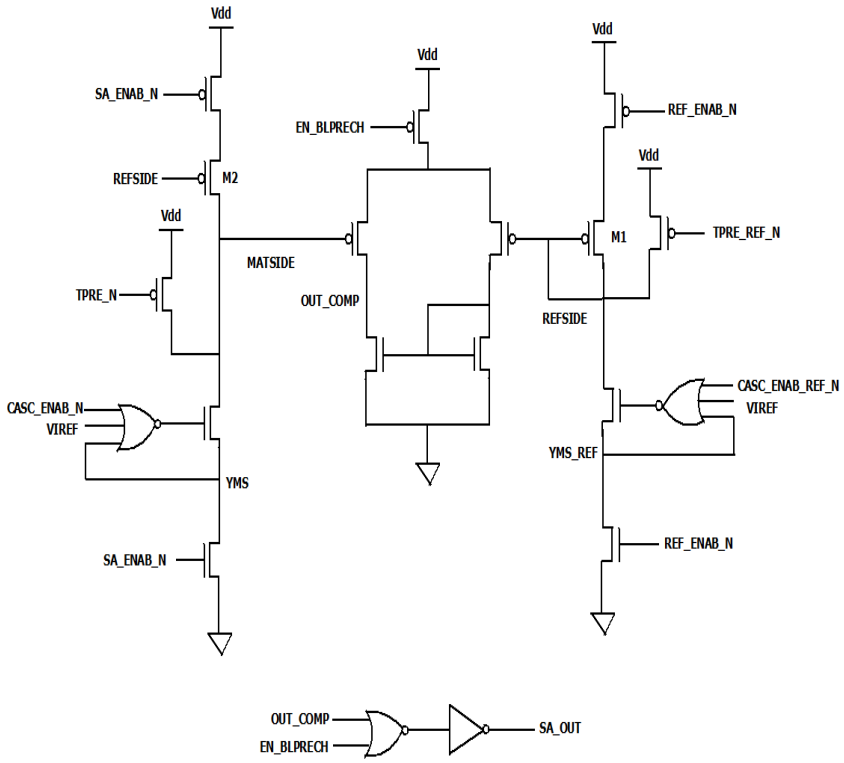


Figure 4.13: Full schematic of Mirror SA

4.3.3 Design Guidelines

Since the basic working of this topology is dependent upon the charging and discharging of parasitic capacitance C_{PAR}, the governing equation is given by

$$V_C = \frac{1}{C}\int Idt \tag{4.17}$$

Hence the time it takes to discharge the capacitor from V_{CC} to below V_R is

$$T_{DISC} = \frac{V_{DIFF}C_{PAR}}{I_{DISC}} \tag{4.18}$$

Where,
V_{DIFF} = absolute differential voltage $V_R - V_M$
C_{PAR} = parasitic capacitance at MATSIDE
I_{DISC} = Discharging current

For faster sensing this time I_{DISC} must be as low as possible. For doing so, C_{PAR} must be as low as possible, since I_{DISC} is fixed at I_{CELL} and is not at designers hand while V_{DIFF} is the required minimum voltage that C_{PAR} has to be discharged and governed by reference branch of I-V conversion. Also C_{PAR} is directly proportional to the length and width of the MOS used. Hence, keeping low W and L of MOS at MATSIDE helps in keeping the sensing time low but at the same time making design prone to channel length modulation.

4.3.4 Calculation of the systematic offset

Mirror SA suffers from a systematic offset which is caused by the current mirror involved. The current which is supplied to the parasitic capacitor C_{PAR} is given by

$$I_c = \gamma_1 I_{REF} - \gamma_2 I_{CELL} \tag{4.19}$$

Where γ_1 and γ_2 are the current mirror factors which occur due to channel length modulation effect. These are given by

$$\gamma_1 = \frac{1 + \lambda V_{DS2}}{1 + \lambda V_{DS1}} \tag{4.20}$$

$$\gamma_2 = 1 \tag{4.21}$$

Hence even though I_{REF} and I_{CELL} are equal, due to the channel length modulation factors there will be some amount of I_C present which then otherwise should be 0. Though this offset is not very high and within specified limits in our case but for the application having to sense very low values of cell current, this may possess a serious problem.

4.3.5 Results

Mirror SA is having less power consumption which is more evident for highly erased cells. Also the ramping behavior of node MATSIDE helps in making the differential voltage larger with time. This helps in having faster response from comparator. But the delay of the circuit is still high which is more severe for lower difference of currents.
The simulation results of read access time in user mode and $FDMA$ mode is shown in this section. From the read waveform shows the expected behavior of $MATSIDE$ node in Fig 4.14. Also the simulation result for power confirms the clear advantage of this topology.

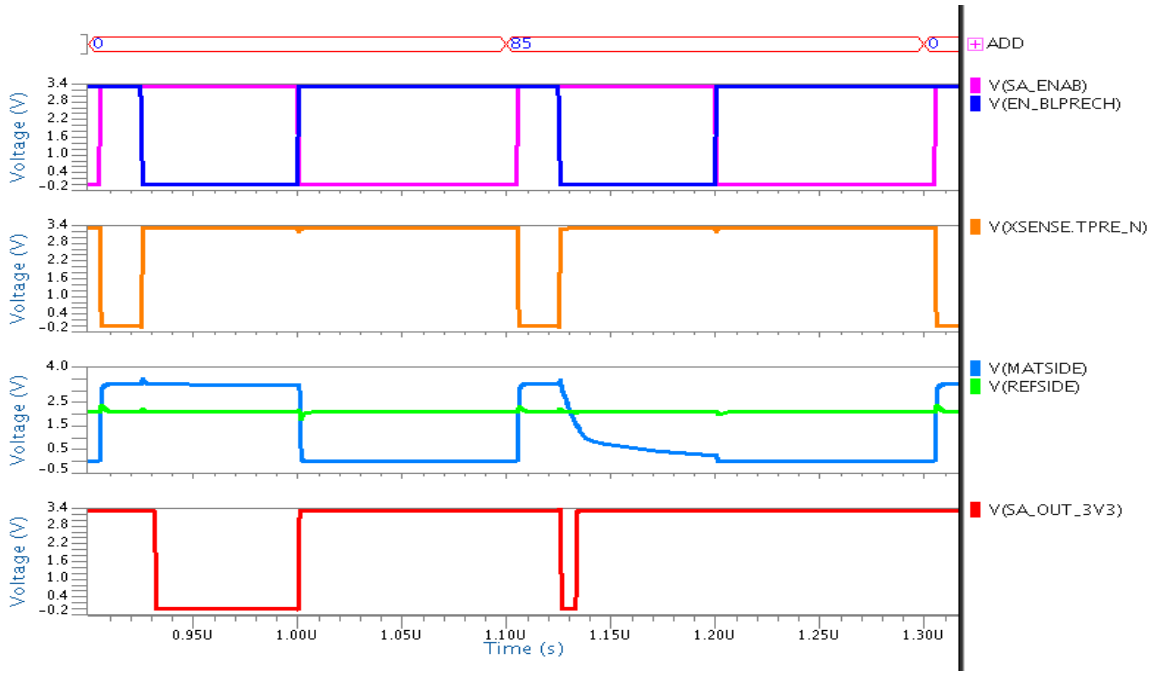
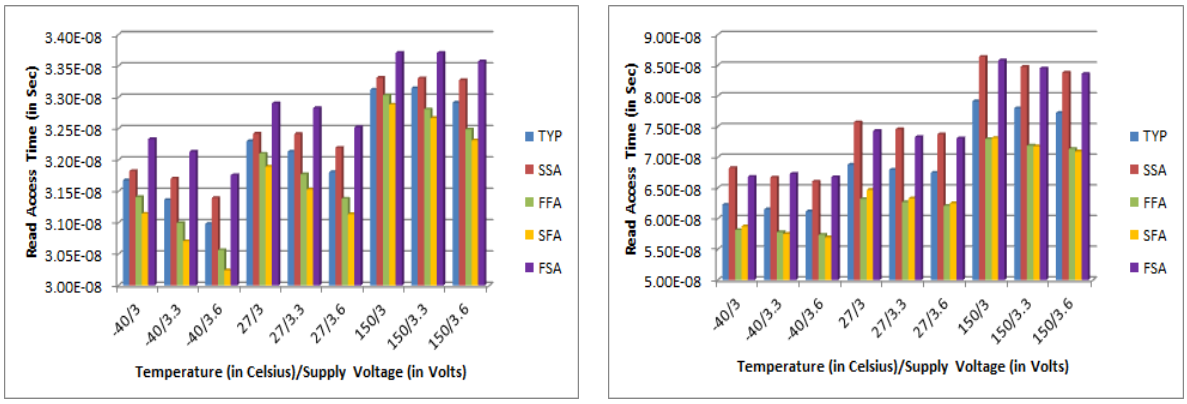
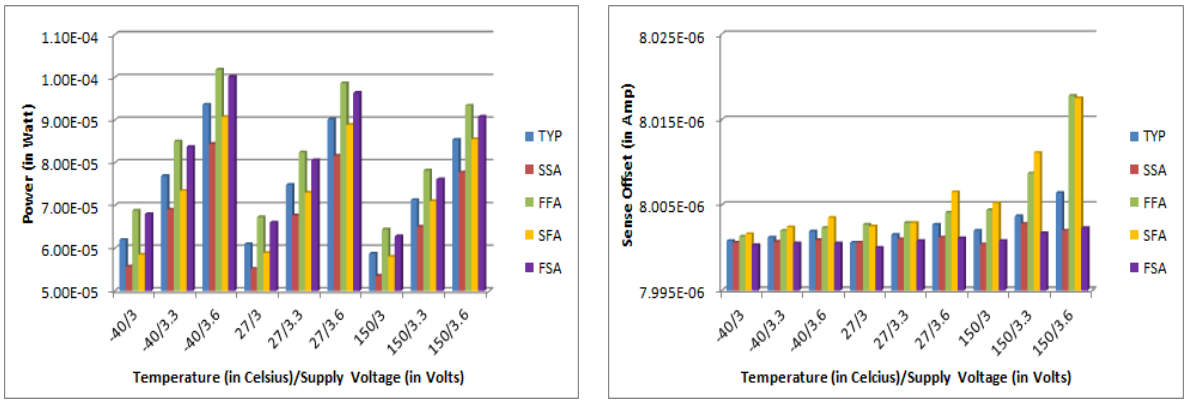


Figure 4.14: Read waveform of Mirror SA



(a) (b)

Figure 4.15: Mirror SA (a)Worst case User mode Read Access time, (b)Worst case FDMA mode Read Access time



(a) (b)

Figure 4.16: Mirror SA (a)Power for worst case erased cell, (b)Offset across PVT

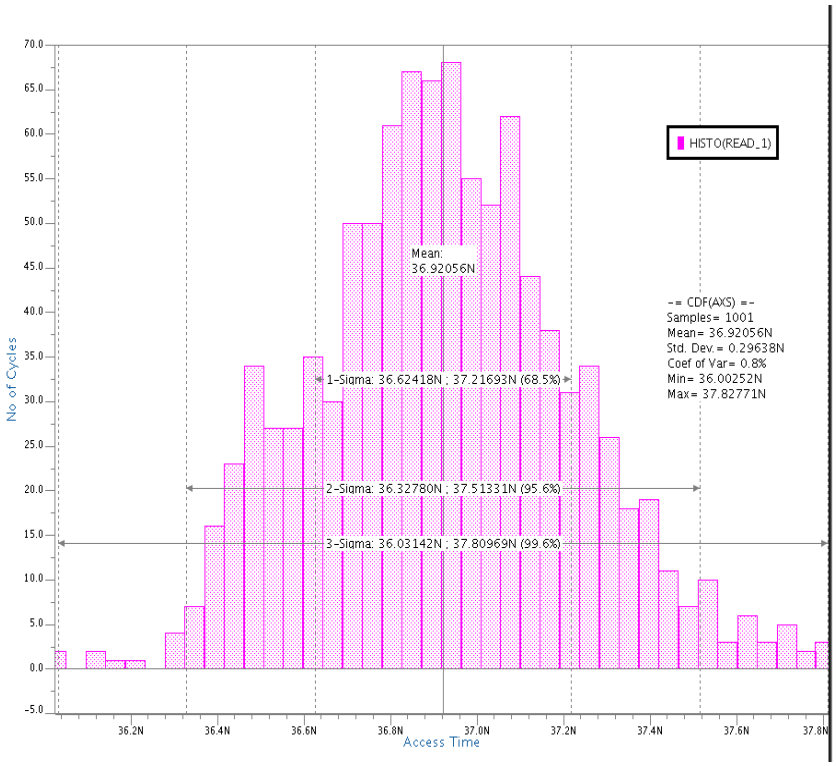


Figure 4.17: Monte-Carlo variation of user mode Access time for Mirror SA

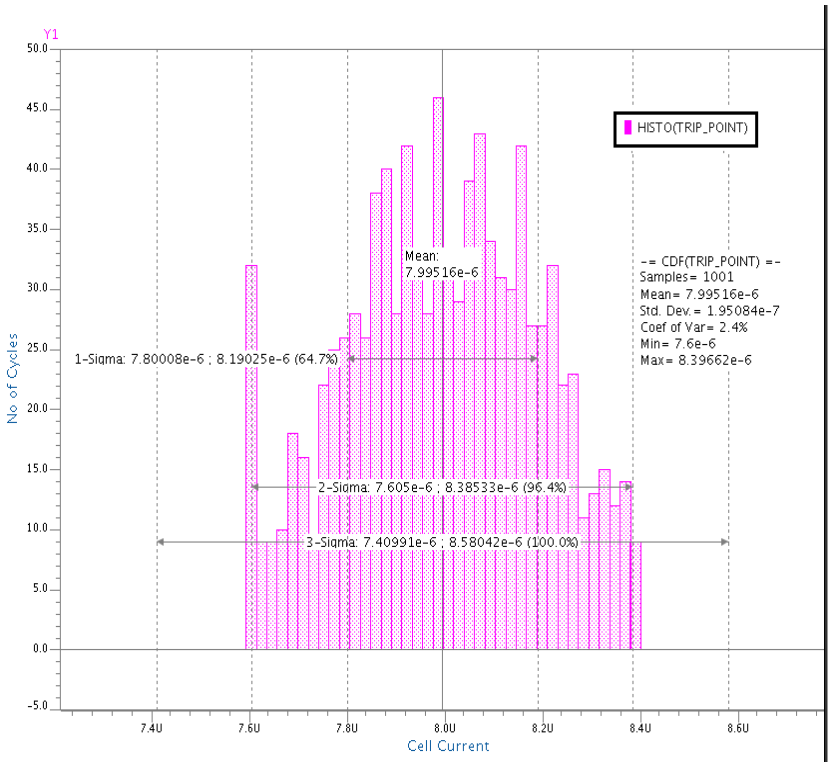


Figure 4.18: Monte-Carlo variation of Offset for Mirror SA

4.4 Fully Symmetric Sense Amplifier (FS SA)

4.4.1 Idea

Though Mirror SA topology works fine but suffers with a systematic mismatch at the input of the comparator due to the inequality of two I-V conversion branch [4]. Hence to minimize the effect of asymmetry while preventing the ramping nature of the sensing nodes, which helps in having faster sensing, a new topology has been developed. This topology has two same branches for I-V conversion, both of which are of ramping nature. This is in contrast to the Mirror SA which had reference branch at fixed voltage and the cell branch was of ramping nature. The two parasitic capacitors at REFSIDE (C_{REF}) and MATSIDE (C_{REF}) is used here to integrate the current difference of $I_{REF} - I_{CELL}$ and $I_{CELL} - I_{REF}$ respectively. Doing this helps in generating the differential voltage faster which in turn results in faster sensing time. The idea of designed topology is shown in Fig 4.19.

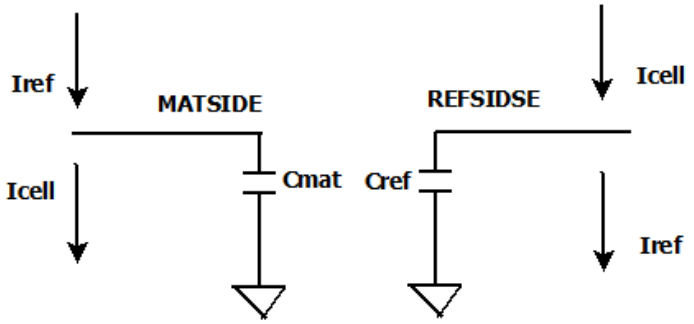


Figure 4.19: Idea behind Fully Symmetric Sense Amplifier

4.4.2 Working

Full schematic of this topology is shown in fig 4.20. This topology also works in two phase first is precharge and second is sensing phase. During the precharge phase both MAT_MIRR and REF_MIRR node are directly connected to V_{DD} through the precharge $PMOS$. And also nodes MATSIDE and $REFSIDE$ is precharged high through EQ signal. After sufficient precharging time, when bitlines of the memory cell are setteled, sensing phase starts. Before starting of sensing phase the cell current as well as reference currents are also settled to its correct value. Once precharge is released, MATSIDE and REFSIDE starts to move according to the I_{REF} and I_{CELL}. If the matrix cell is programmed i.e. $I_{CELL} < I_{REF}$ then capacitor C_{REF} is charged up and maintain itself at high value while capacitor C_{MAT} starts to discharge. Thus REFSIDE is high and $MATSIDE$ is ramping low thus output of the comparator is logic 0. In the other case when the matrix cell is erased one i.e.$I_{CELL} > I_{REF}$ then C_{MAT} is charged to high value and C_{REF} discharges giving rise to a high $MATSIDE$ while REFSIDE is ramping low. In this case output of comparator is logic 1.

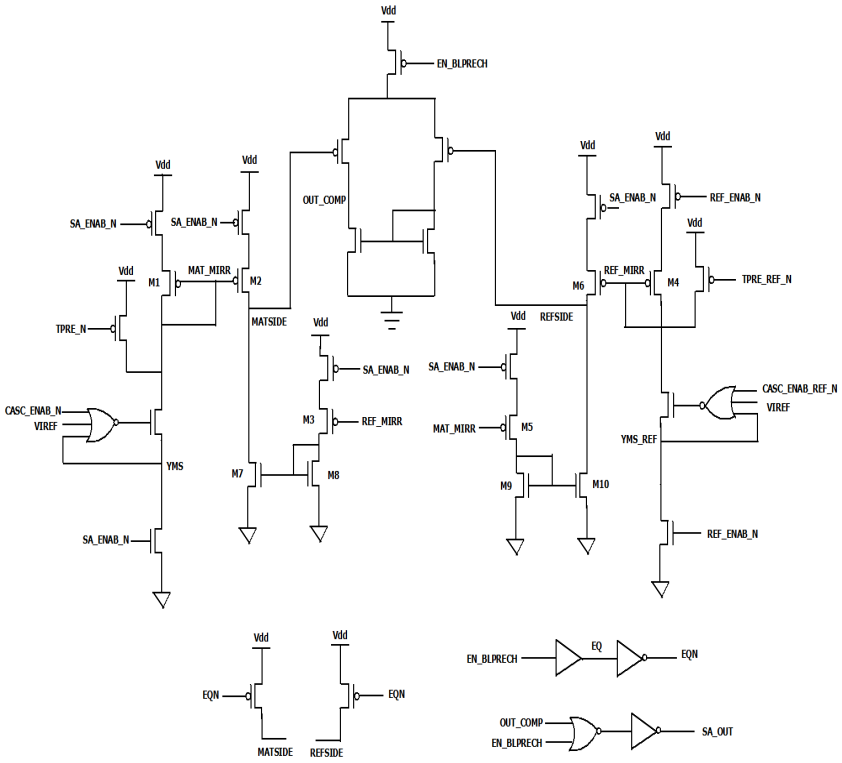


Figure 4.20: Complete Schematic of Fully Symmetric Sense Amplifier

4.4.3 Design Guideline

Since there are current mirrors involved to route the reference and cell currents, keeping high length devices helps in reducing the channel length modulation effect of individual current mirror pair. Also cascoding can be considered if sufficient headroom for mirror MOS is available to keep it in saturation.

4.4.4 Systematic offset Analysis

Here in Fully Symmetric SA systematic offset of previous topology is compensated. Also, Static nature of this SA is useful to settle the mirrored currents.

$$I_M = \alpha_{M1} I_{CELL} - \alpha_{M2} I_{REF} \tag{4.22}$$

$$I_R = \alpha_{R1} I_{REF} - \alpha_{R2} I_{CELL} \tag{4.23}$$

$$\alpha_{M1} = \frac{1 + \lambda_P V_{SD2}}{1 + \lambda_P V_{SD1}} \tag{4.24}$$

$$\alpha_{R1} = \frac{1 + \lambda_P V_{SD6}}{1 + \lambda_P V_{SD4}} \tag{4.25}$$

$$\alpha_{M2} = (\frac{1 + \lambda_P V_{SD5}}{1 + \lambda_P V_{SD1}})(\frac{1 + \lambda_N V_{DS10}}{1 + \lambda_N V_{DS9}}) \tag{4.26}$$

$$\alpha_{R2} = (\frac{1+\lambda_P V_{SD3}}{1+\lambda_P V_{SD4}})(\frac{1+\lambda_N V_{DS7}}{1+\lambda_N V_{DS8}}) \tag{4.27}$$

Where,
α_{M1} = current mismatch introduced by mirror $M_1 - M_2$
α_{R1} = current mismatch introduced by mirror $M_4 - M_6$
α_{M2} = current mismatch introduced by mirror $M_1 - M_5$ and $M_9 - M_10$
α_{R2} = current mismatch introduced by mirror $M_4 - M_3$ and $M_8 - M_7$

Subtracting equation 4.24 from 4.23

$$I_M - I_R = K_1 I_{CELL} - K_2 I_{REF} \tag{4.28}$$

Where,

$$K_1 = \alpha_{M1} + \alpha_{M2}$$
$$K_2 = \alpha_{R1} + \alpha_{R2}$$

After substituting the values we get,

$$\alpha_{M2} = \frac{(1+\lambda_P V_{SD2})(1+\lambda_N V_{DS9}) + (1+\lambda_P V_{SD5})(1+\lambda_N V_{DS10})}{(1+\lambda_P V_{SD1})(1+\lambda_N V_{DS9})} \tag{4.29}$$

$$\alpha_{M2} = \frac{(1+\lambda_P V_{SD6})(1+\lambda_N V_{DS8}) + (1+\lambda_P V_{SD3})(1+\lambda_N V_{DS7})}{(1+\lambda_P V_{SD4})(1+\lambda_N V_{DS8})} \tag{4.30}$$

For the same process, channel length modulation parameter λ_N and λ_P are going to be same. Also for the case when $I_{REF} = I_{CELL}$ the drain source voltages of the two current mirror branches will be same hence giving $K_1 = K_2$, Eliminating any systematic offset at the input of the comparator.

4.4.5 Results

Power consumption for this topology is high since there are two mirror branches involved which will consume power for half of the read cycle. Also the added branches will account for area increment. Simulation results for access time, power consumption and offset is shown in this section. Increase in the power consumption per read cycle can be observed and also there is significant reduction in systematic offset of the sense amplifier. Though after considering random offset, there is not much advantage over mirror SA topology.

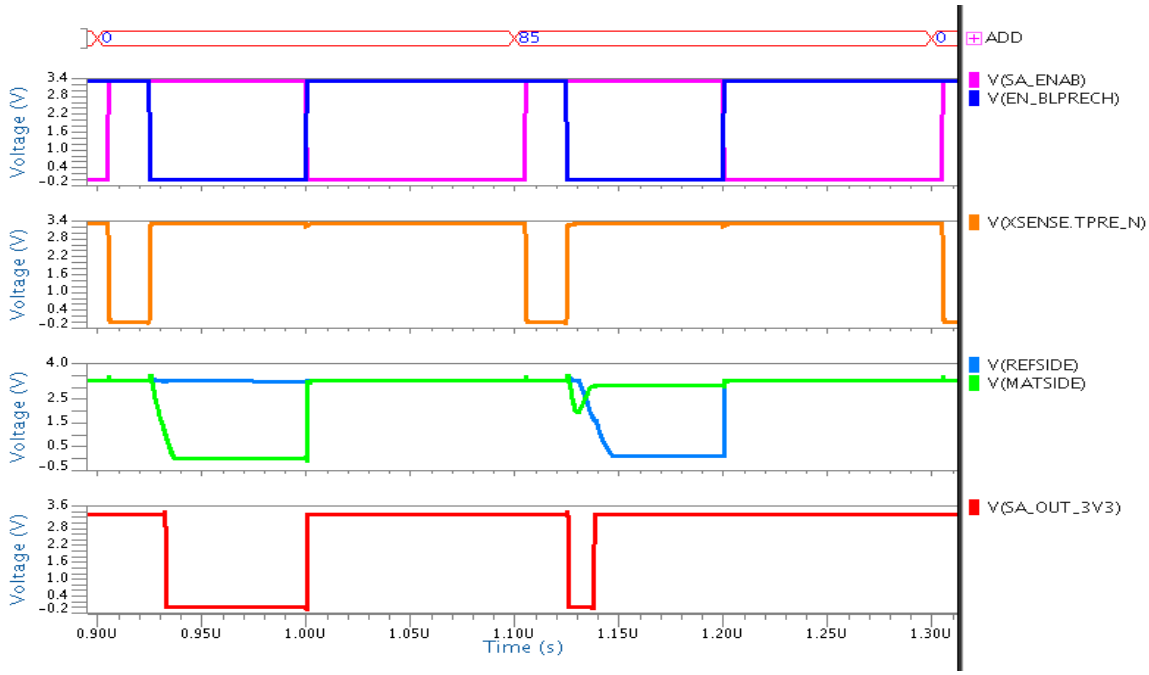
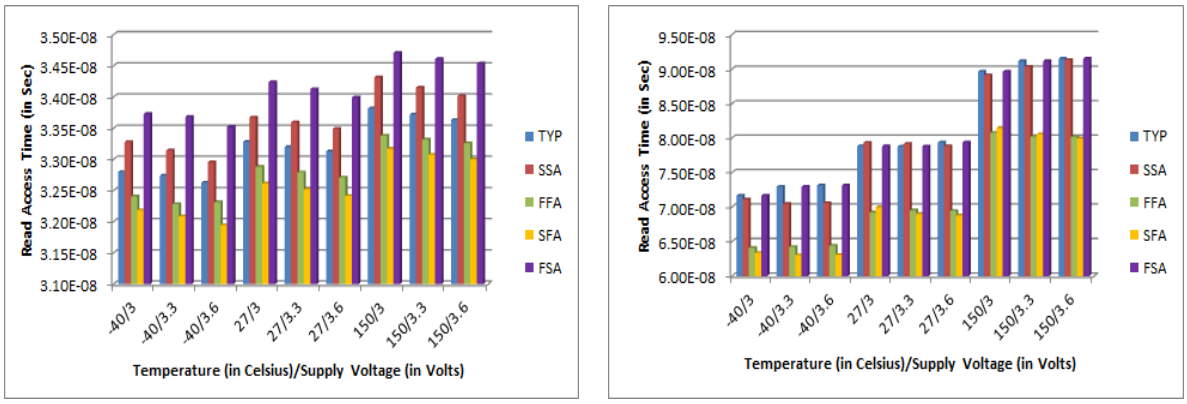
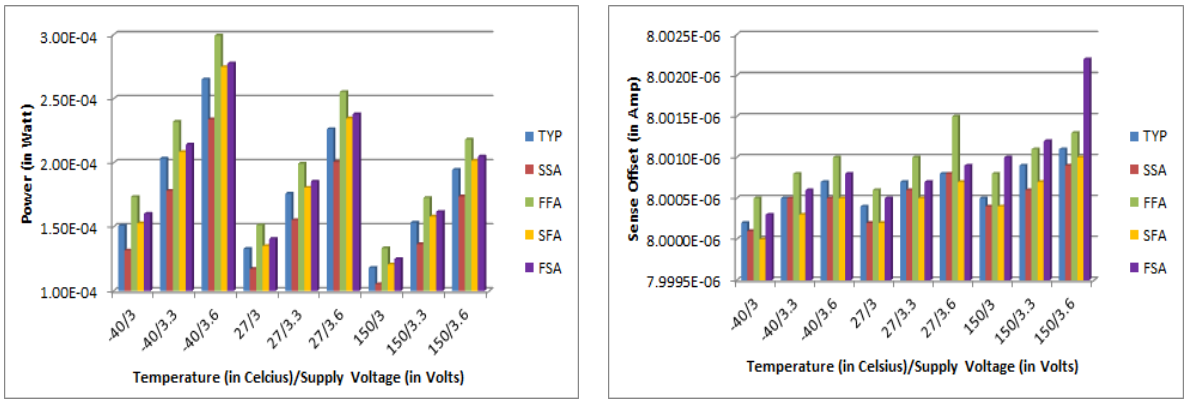


Figure 4.21: Read waveform of Fully Symmetric SA



(a) (b)

Figure 4.22: Fully Symmetric SA (a)Worst case user mode Read Access time, (b)Worst case FDMA mode Eead Access time



(a) (b)

Figure 4.23: Fully Symmetric SA (a)Power for worst case erased cell, (b)Offset across PVT

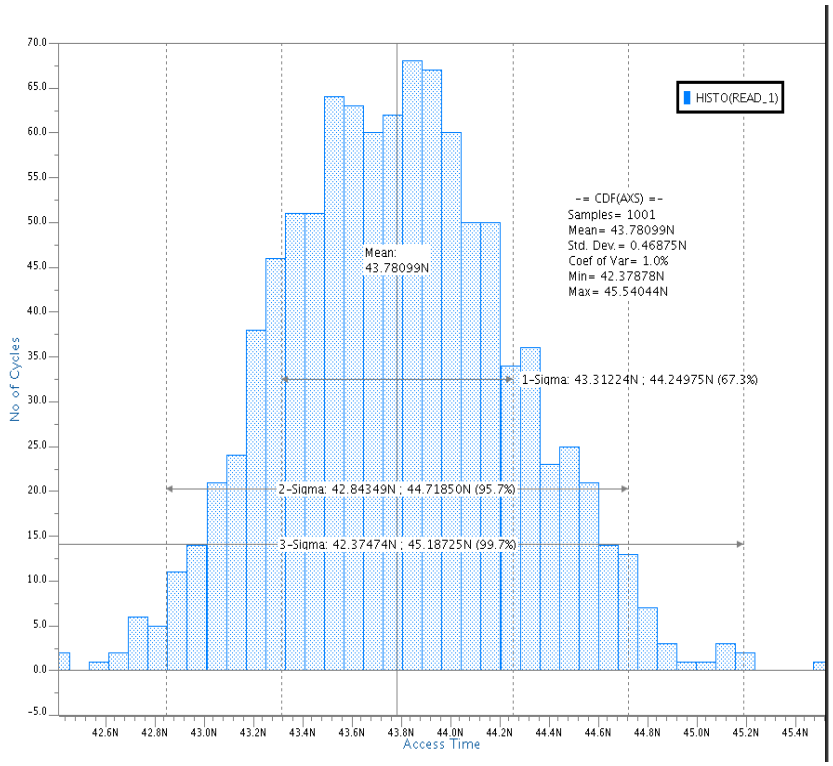


Figure 4.24: Monte-Carlo variation of user mode Access time for Fully Symmetric SA

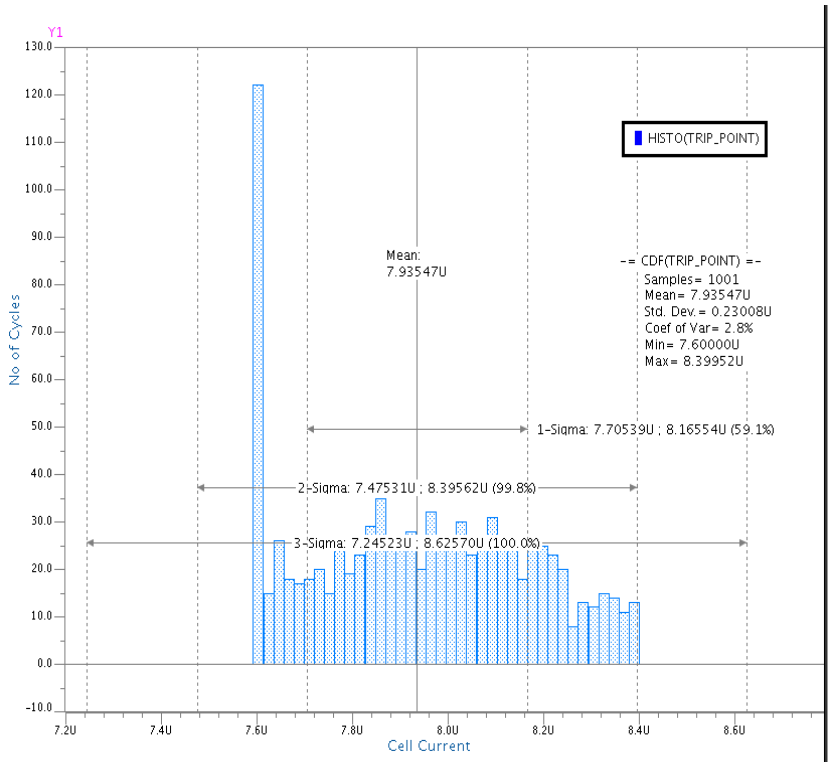


Figure 4.25: Monte-Carlo variation of Offset for Fully Symmetric SA

Chapter 5

Dynamic Sense Amplifier Topologies

Dynamic sense amplifier topologies make use of regenerative positive feedback action of cross coupled MOS. Such topologies, in general sense, are sensitive to the transients of the circuit and once they start to slew in one direction they cannot rebound. Designing of such sensitive structure, across PVT condition is a challenge. Fortunately, with proper design strategies we can avoid any chances for dynamic sense amplifier to make wrong decision. Also, dynamic topologies are fast in nature as compared to their static counterparts making it a necessity for speed critical applications.

There are a number of dynamic sense amplifiers found in literature which researchers have proposed. Uetake et.al. in [31] have proposed a novel latch type sense amplifier for SRAM, over which sensing time improvement is reported in [11]. But still topology presented in [23] suffers with static power consumption and also has higher delay for small cell currents. A novel sensing solution was reported in [12] for Bi-Nor kind of flash memories which with proper modification can be applied to Nor type flash memories as well. For the case of very small cell current an offset tolerant small cell current sense amplifier for flash memories is reported in [9]. Finally, a comparative study of various latch type sense amplifiers is also presented in [25]. Here in this work I have developed and designed three topologies of sense amplifiers all of them uses cross coupled MOS latches for faster sensing. Working of the topologies and their technical merits are also discussed along with.

5.1 Analysis of Regenerative Latch

Fig 5.1 shows the simplified schematic of a regenerative cross-coupled latch type comparator. They are bi-stable in nature, of which the two stable states are V_{o1} being high and V_{o2} being low or V_{o2} being high and V_{o1} being low. Here I_1 and I_2 are the DC currents of respective MOS. For the proper operation of the latch the DC currents must be well settled since any spurious transients in the DC currents may lead the latch to settle in the wrong direction. Time constant and propagation delay of such latch is reported in [1] and are given by equation 5.1 and 5.2 respectively.

Latch time constant ς_L,

$$\varsigma_L = \frac{C}{g_m} = 0.67 C_{ox} \sqrt{\frac{WL^3}{2K'I}} \tag{5.1}$$

Propagation delay time t_p,

$$t_p = \varsigma_L ln\Big(\frac{V_{OH} - V_{OL}}{2\Delta V_i}\Big) \tag{5.2}$$

Where,
C = Capacitance at the latching nodes
g_m= Transconductance of the latching device
K' = Technology dependent constant
W = Width of the latching device
L = Length of the latching device
I = DC current of the latch
V_{OH} = High voltage level of output
V_{OL} = Low voltage level of output
ΔV_i = input differential voltage before the latching begins

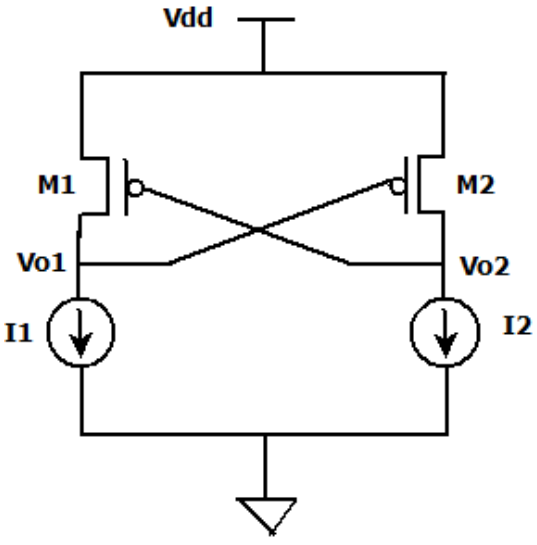


Figure 5.1: Regenerative Latch

From the above mentioned equation it is evident that latch time constant is strongly dependent on channel length of input devices. Low channel length will lead to higher process variation while high channel length will make the latch time constant high which in turn make the response of latch slow. Hence proper care must be taken to make balance between the two.

5.2 Half Latch Sense Amplifier (HL SA)

5.2.1 Idea

The basic idea behind this topology is to explore the possibility of having a half latch kind of configuration to generate full scale output voltage depending upon comparison of reference and cell current. For this purpose the DC currents I_1 and I_2 of latch shown in Fig 5.1 will be replaced by I_{REF} and I_{CELL}. Properly settled values of differential current will give rise to a differential voltage ΔV_i at the latching nodes, represented by $MATSIDE$ and $REFSIDE$ in Fig 5.2. This differential voltage then will be amplified by the latch till the full scale output voltage.

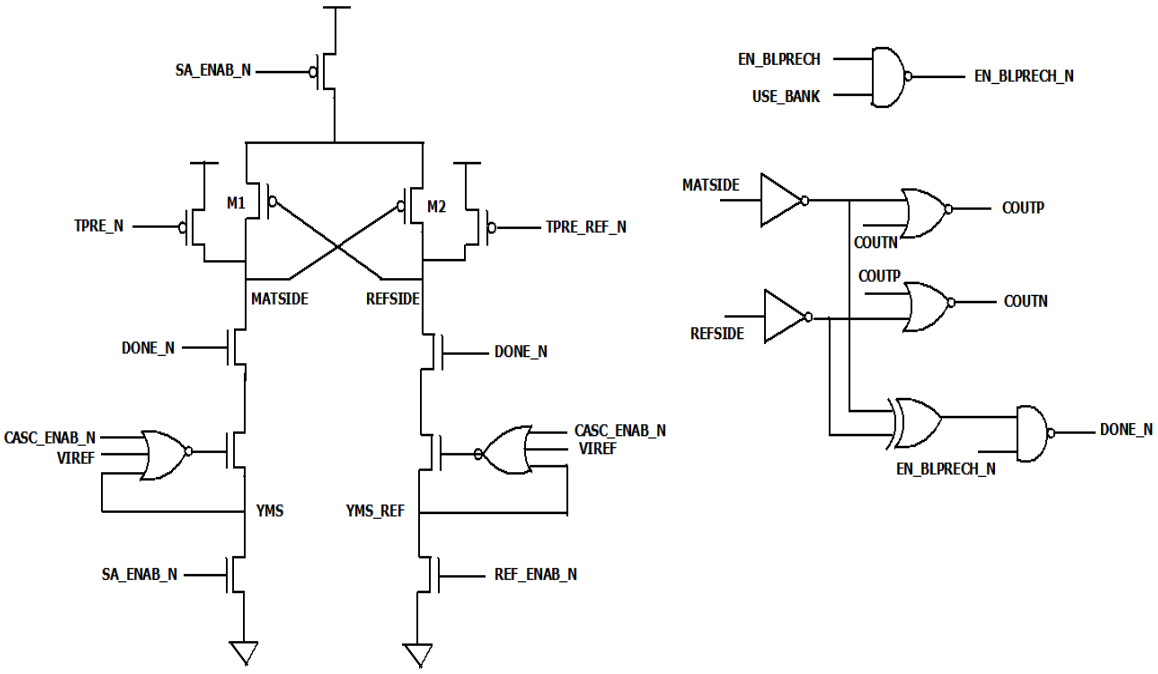


Figure 5.2: Full schematic of Half Latch Sense Amplifier

5.2.2 Working

Full Schematic of half latch based topology is shown in Fig 5.2. This sense amplifier has similar two phase sensing cycle as of static topologies, precharge and evaluation. During the precharge cycle, along with bitline precharging, the latching nodes are also precharged high to V_{DD} which will make latching transistor M_1 and M_2 to go in cut-off and it will avoid any latching action to begin. After the release of precharge, the parasitic capacitance of latching nodes will start to discharge with a differential current of $I_{REF} - I_{CELL}$ which will cause one node to discharge faster than the other. For the proper explanation let us assume that $I_{CELL} > I_{REF}$. In this case node MATSIDE will discharge faster than REFSIDE and this discharging will continue till the time MATSIDE discharges up to$V_{DD} - V_{TH_P}$ At this instant transistor M_2 will turn ON in saturation and will re-accomplish REFSIDE node to V_{DD} potential. This will now turn M_1

into cut off and hence node $MATSIDE$ will be discharged to its lowest potential. To avoid any capacitance mismatch at the latching nodes, a NOR based S-R latch is used to latch the sensed data. Also the same circuitry is used to generate an additional signal $DONE\_N$ which goes low once the data is captured in S-R latch. $DONE\_N$ is used in the sensing branch at NMOS switch to cut off the sensing path once the latching operation is complete. This avoids any static current to flow and results in power saving.

5.2.3 Design Guidelines

Here, in this topology the mismatch between M1 and M2 is very critical as this mismatch may cause reliability issues. Hence the latching devices should not be made on minimum length and also their aspect ratios should be kept high. High aspect ratio may cause higher capacitance at latching nodes and hence larger delay but this is the tradeoff to avoid process mismatch.

5.2.4 Results

The regenerative action of cross coupled latch has faster sensing which is more remarkable in the case of lower difference of cell and reference currents. Power consumption of the topology is also limited due to the cut off action provided by $DONE\_N$ signal Simulation graphs of access time, power and offset is shown in this section. Read waveform in Fig 5.3 shows the behavior of $MATSIDE$ and $REFSIDE$ node. The reduction in access time can be seen for the cases of user mode read as well as $FDMA$ read shown in Fig 5.4.

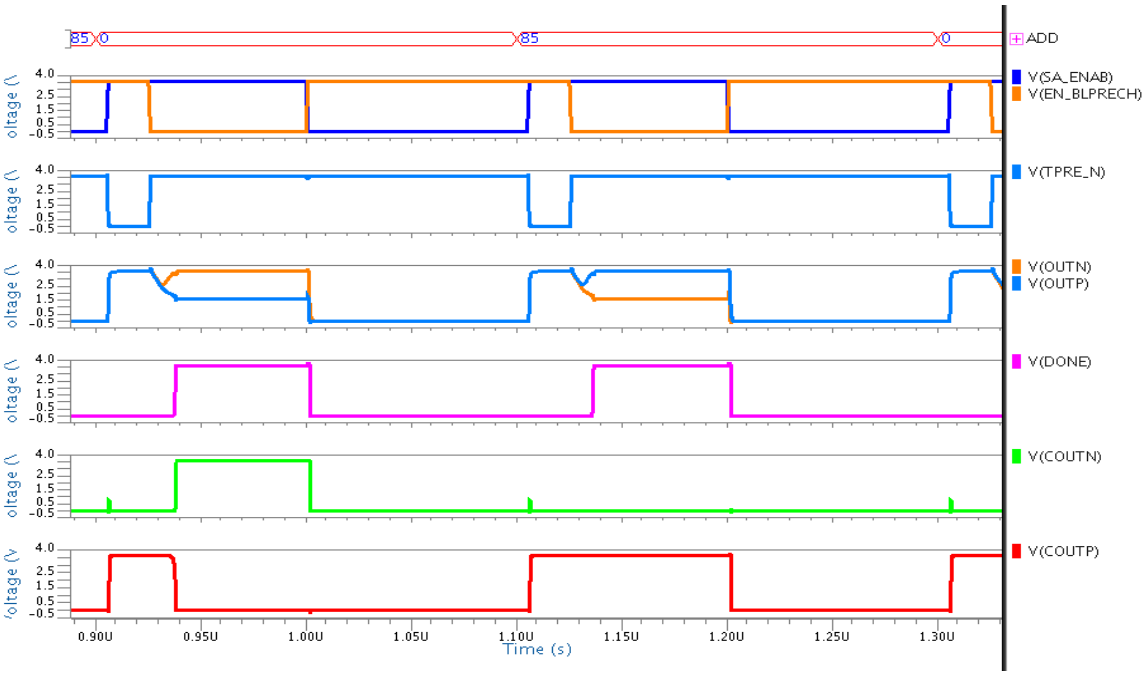
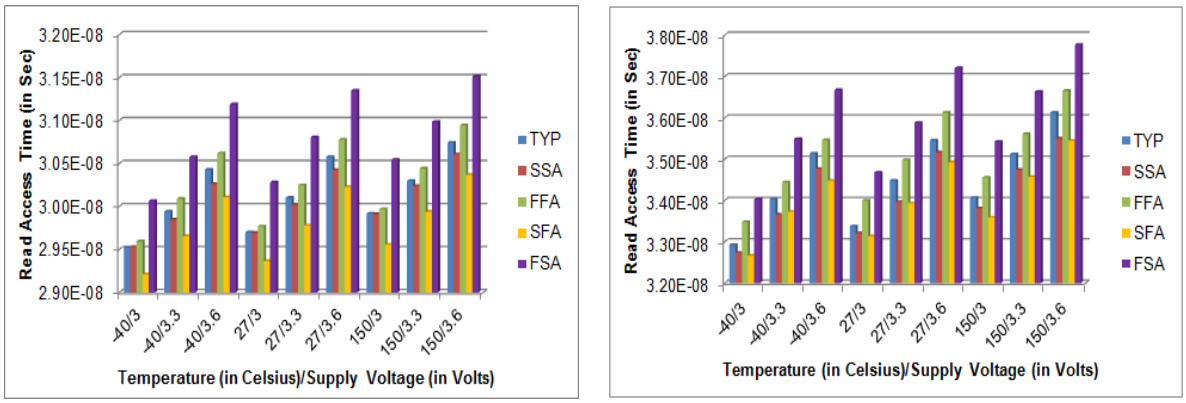
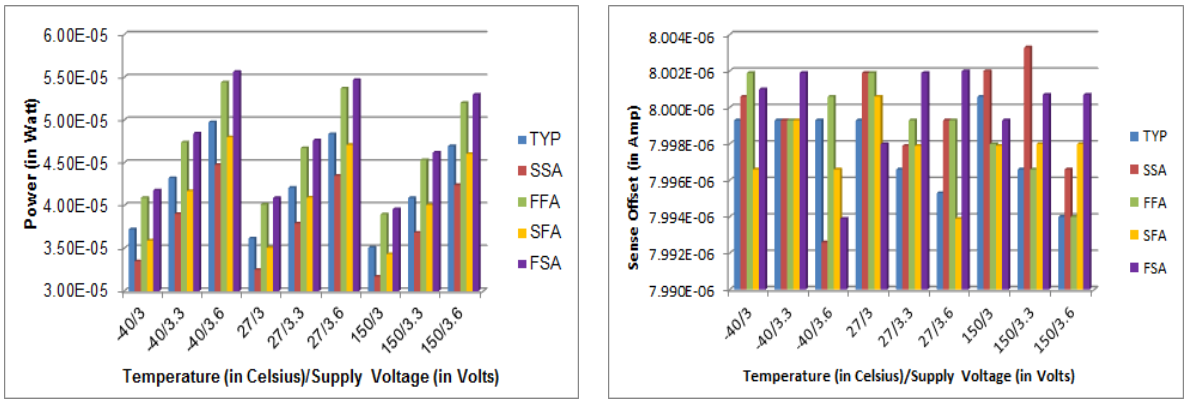


Figure 5.3: Read waveform of Half Latch SA



(a) (b)

Figure 5.4: Half Latch SA (a)Worst case user mode Read Access time, (b)Worst case FDMA mode Read Access time



(a) (b)

Figure 5.5: Half Latch SA (a)Power for worst case erased cell, (b)Offset across PVT

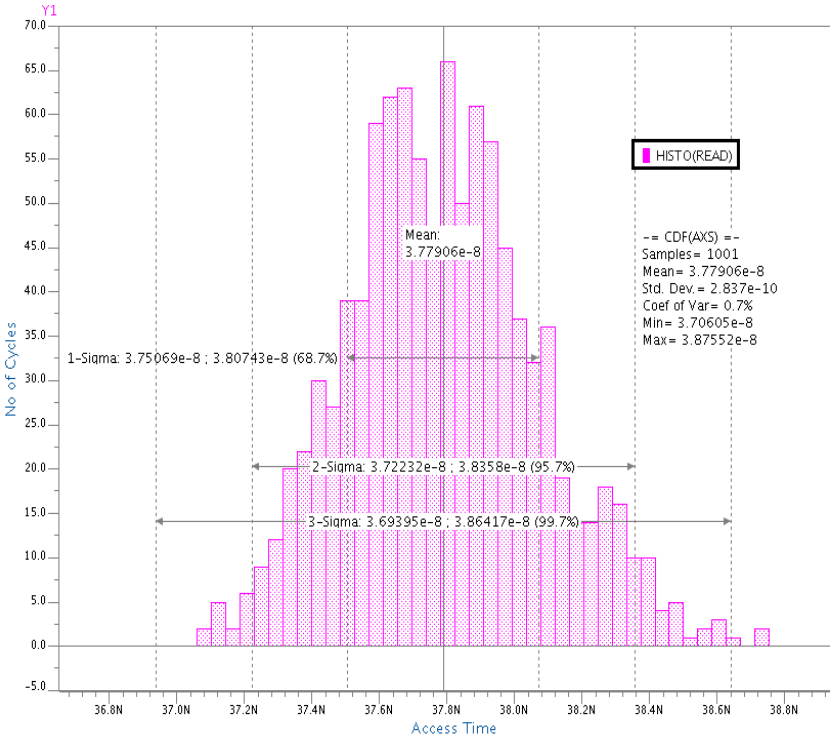


Figure 5.6: Monte-Carlo variation of user mode Access Ttime for Half Latch SA

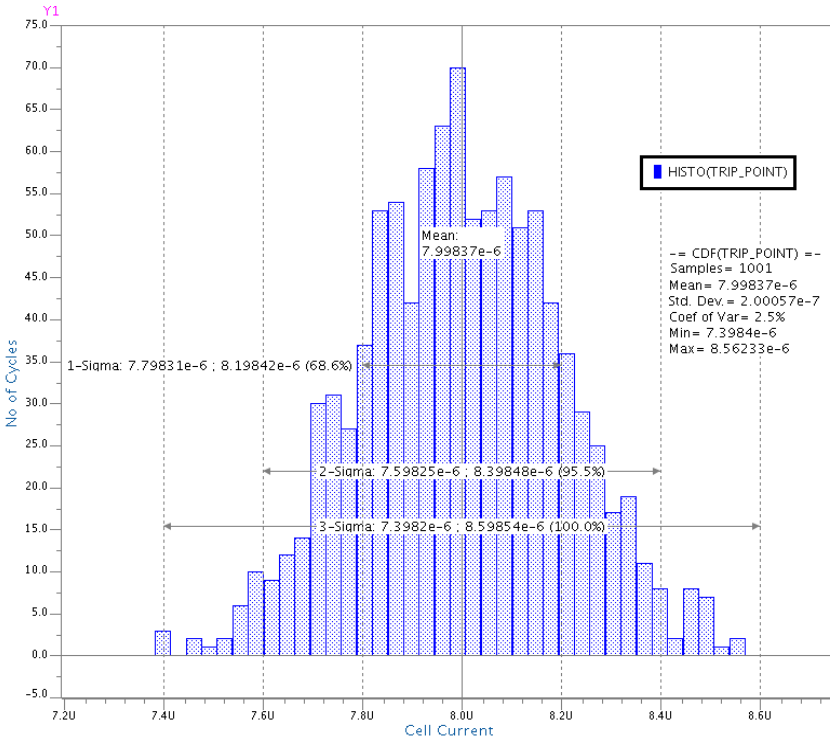


Figure 5.7: Monte-Carlo variation of Offset for Half Latch SA

5.3 Half Latch and Comparator based Sense Amplifier (HLC SA)

5.3.1 Idea

If we see the read access time for half latch SA topology, it is in the order of 37ns in worst case PVT variation. This is due to having high aspect ratio of latching devices. This access time can be further improved if we can have assist schemes to further fasten the process of latching. In this topology we have used a PMOS differential pair with $NMOS$ cross coupled load as the 2nd amplifying stage of the previous sense amplifier topology.

5.3.2 Working

As can be seen in the fig 5.8, the 1st of the sense amplifier, which is used here as the pre-amplification stage, is similar to the half latch SA. In addition to that the 2nd amplification stage is used for faster sensing. An additional signal T_{PRE} is taken out after passing T_{PRE}_N through an inverter. This signal is used to prevent $NMOS$ latch of the 2nd stage comparator to slew during the precharge phase.
During precharge phase, node $MATSIDE$ and $REFSIDE$ is precharged high and at the same time node OUTP and OUTN are grounded. The slewing behavior of $MATSIDE$ and

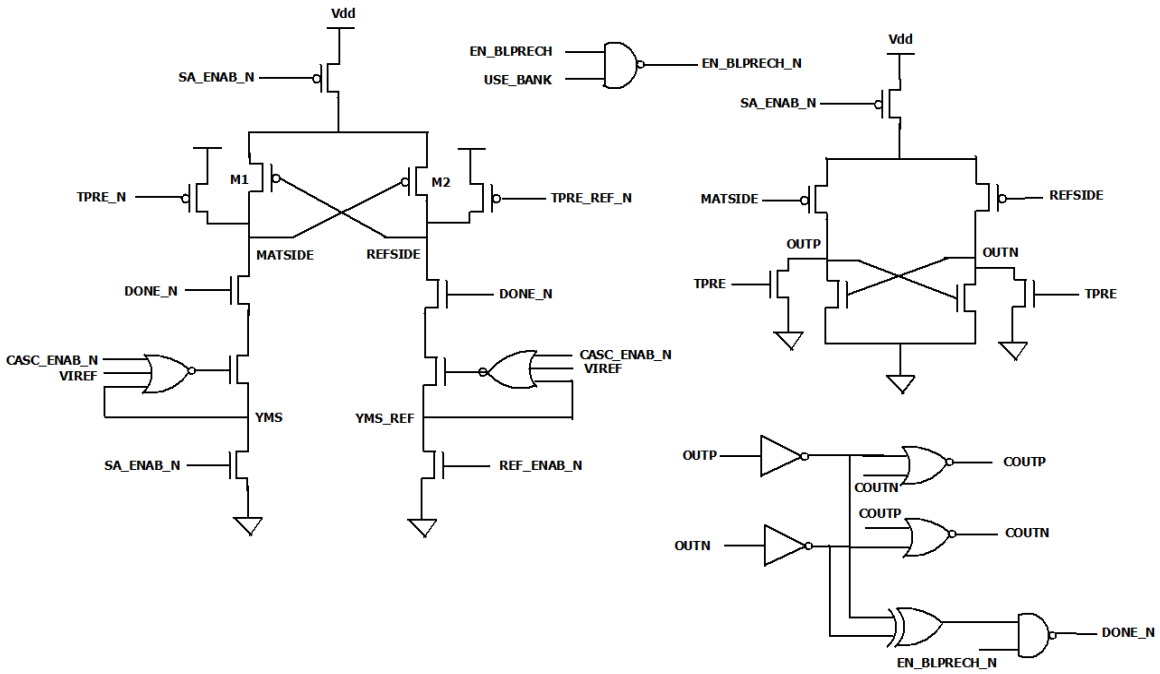


Figure 5.8: Full Schematic of Half Latch and Comparator based SA

$REFSIDE$ is similar to previous topology, only difference is now the differential voltage of these nodes will be amplified by the 2nd stage and final full scale voltage will be produced by the $NMOS$ latch.

5.3.3 Design Guideline

In this topology along with the sizes of $PMOS$ latch devices, the sizes of $PMOS$ input differential pair w.r.t. the $PMOS$ latch is also critical. If $PMOS$ cross couple is very conductive it will switch on early during the linear discharging of $MATSIDE$ and REFSIDE node. If it is very resistive, it will delay the re-accomplishment of slower discharging node. Hence the sizes must be chosen properly. Also for the $NMOS$ cross couple minimum channel length devices must be avoided to minimize process variation.

5.3.4 Results

This topology has better access time as compared to the half latch SA but it might show a higher offset since the addition of 2nd stage comparator. This 2nd stage will add the power consumption to the sense amplifier. The results of access time of programmed and erased cell across PVT, power for the worst case erased cell and offset is shown in this section. The advantage of using another amplification stage is evident from Fig 5.10(a) and (b).

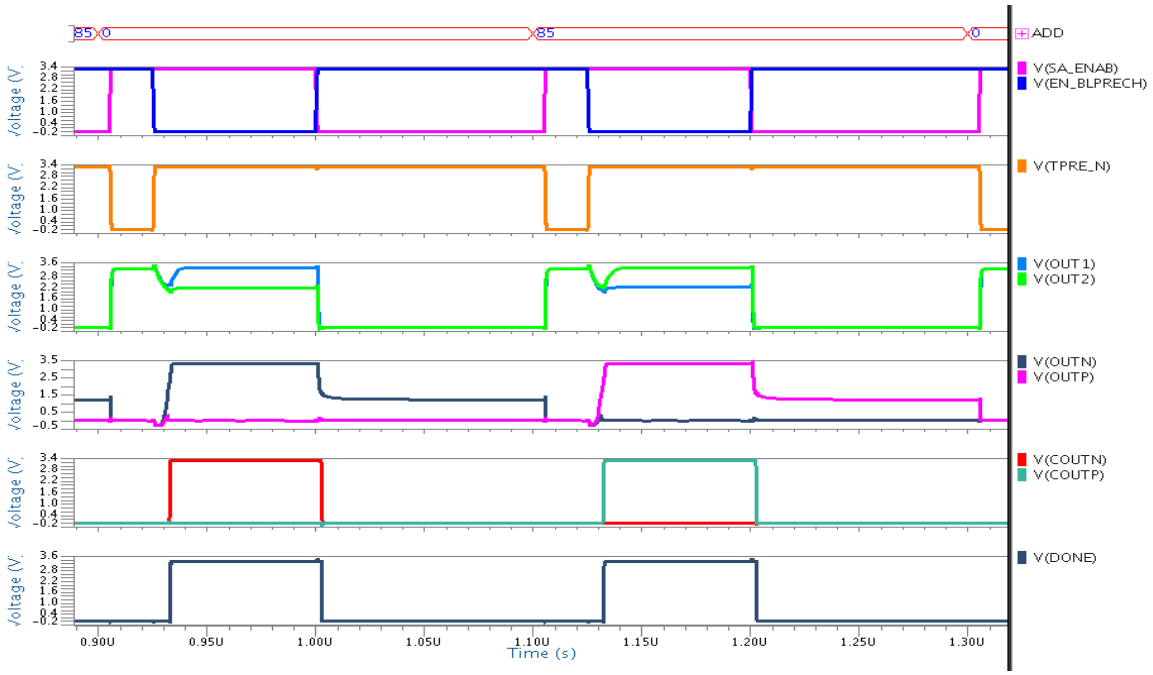
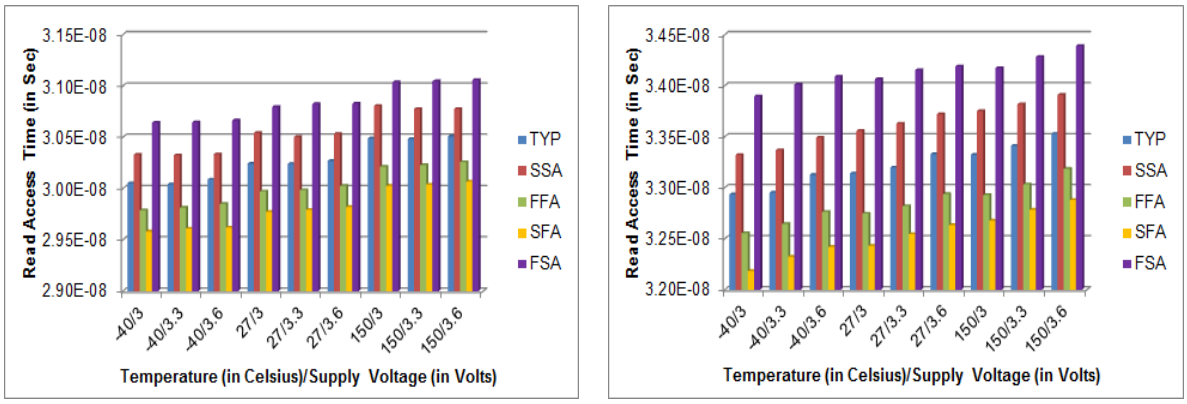
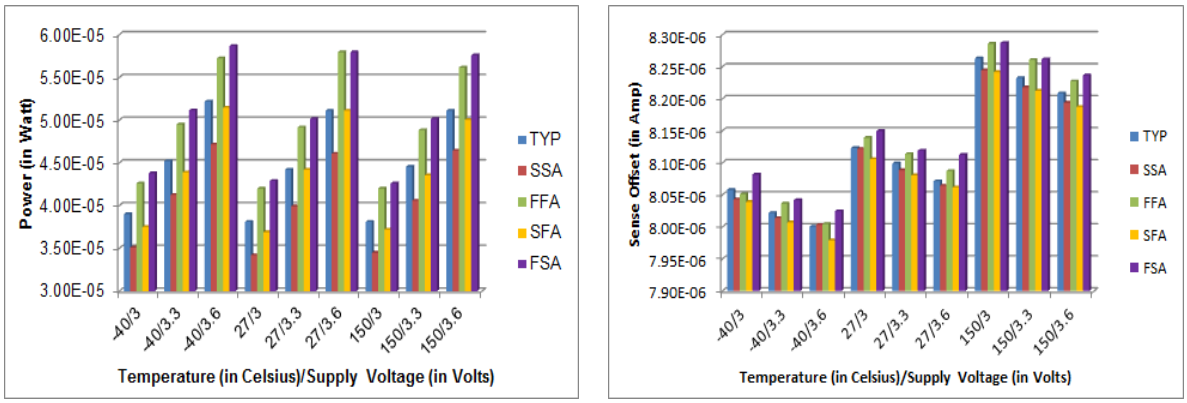


Figure 5.9: Read waveform of Half Latch and Comparator SA



(a) (b)

Figure 5.10: HLC SA(a) Worst case user mode Read Access time, (b)Worst case FDMA mode Read Access time



(a) (b)

Figure 5.11: HLC SA (a)Power for worst case erased cell, (b)Offset across PVT

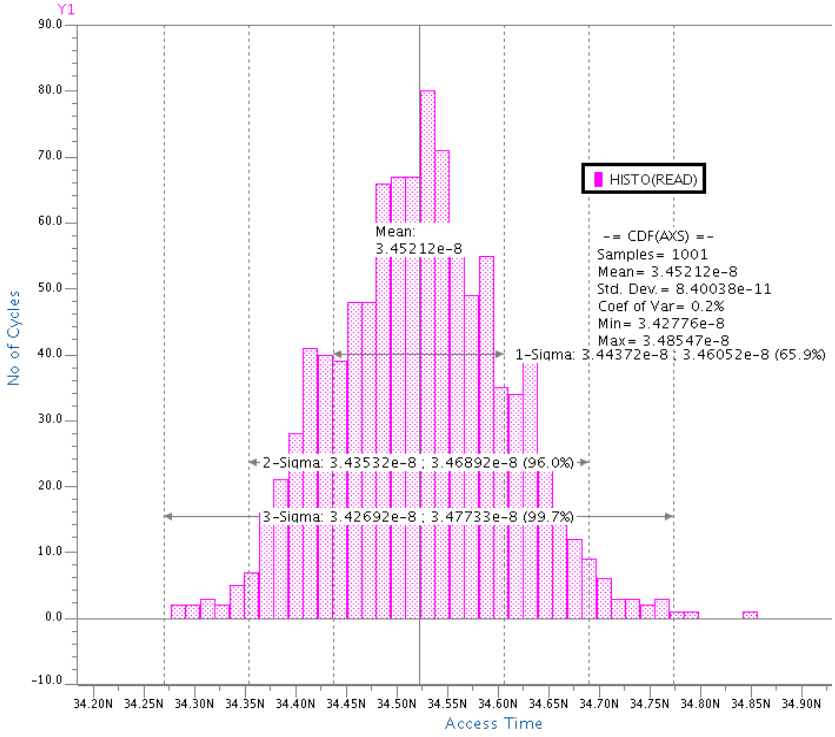


Figure 5.12: Monte-Carlo variation of user mode Access time for HLC SA

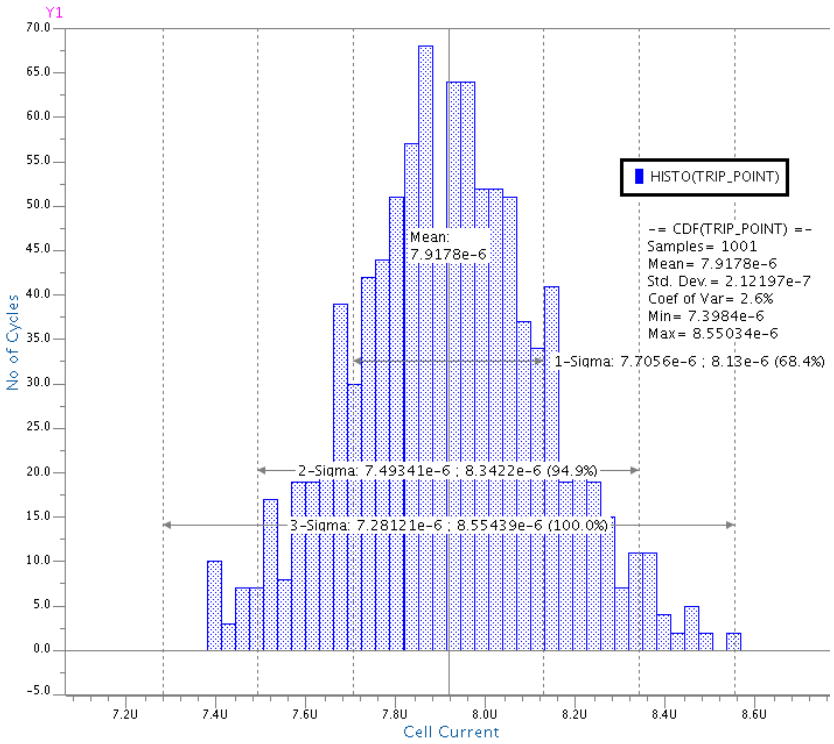


Figure 5.13: Monte-Carlo variation of Offset for HLC SA

5.4 Full Latch Sense Amplifier (FL SA)

5.4.1 Idea

For the two previously discussed topologies, still the sensing nodes $MATSIDE$ and $REFSIDE$ has to discharge at least up to $V_{DD} - V_{THP}$ to start the sensing action. For small cell currents this discharging might be slow resulting in higher access time. So in this topology we have tried to reduce this differential voltage till which sensing nodes should be discharged for even better access time. This is accomplished by having an NMOS latch in cascade with PMOS cross coupled latch.

5.4.2 Working

In this topology an additional signal $EN\_BLPREECH\_BUFF\_N$ is introduced which is the delayed version of $EN\_BLPRECH\_N$. This is shown in timing diagram of fig 5.14.

During the precharge phase $TPRE\_N$ is low and hence the bitlines were clamped to its final value through the precharge device. During this phase$MATSIDE$ and $REFSIDE$ are at V_{DD} level, making M_1 and M_2 to be in cut-off. At the same time $EN\_BLPRECH\_BUFF\_N$ is low, which cuts off the source connection of M_3 and M_4, that does not allow $NMOS$ cross coupled to begin regenerative action even though their drains are at high potential. After the precharging, the

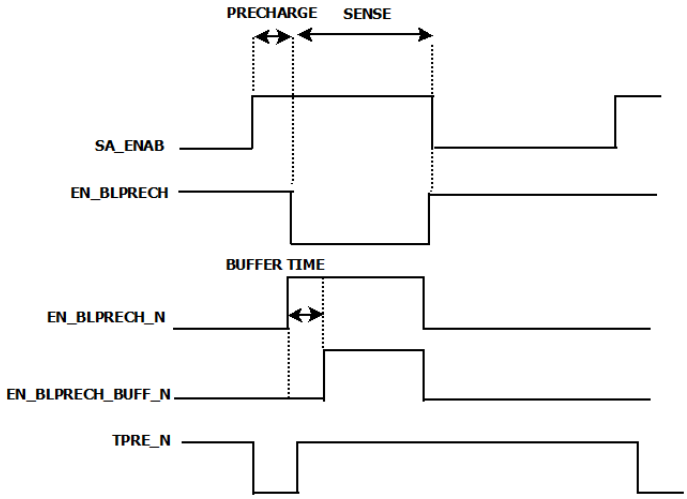


Figure 5.14: Timing Phases for Full Latch SA

linear discharging of nodes MATSIDE and REFSIDE starts which continues till the NMOS latch is activated by turning $EN\_BLPRECH\_BUFF\_N$ high. The buffering delay added between the $EN\_BLPRECH\_N$ and $EN\_BLPRECH\_BUFF\_N$ decides the initial voltage difference that $NMOS$ cross couple will resolve. This differential voltage can be empirically set to make the sense amplifier work across all PVT. The full schematic of this topology is shown in fig 5.15.

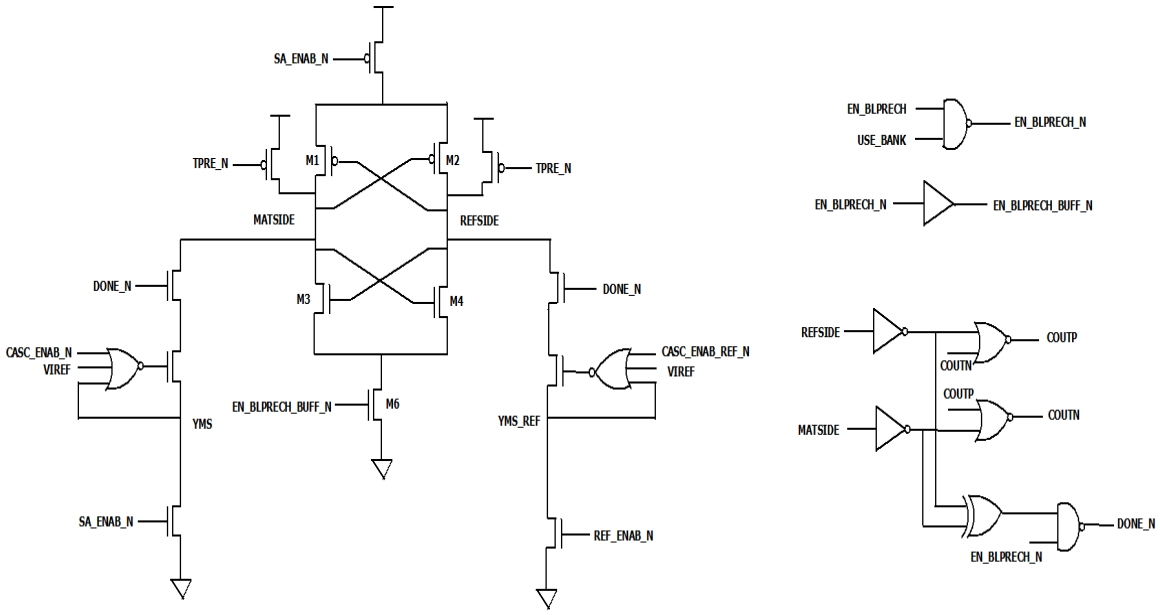


Figure 5.15: Complete Schematic of Full Latch Sense Amplifier

5.4.3 Design Guidelines

In this topology the most important parameter to set is the delay between $EN\_BLPRECH\_N$ and $EN\_BLPRECH\_BUFF\_N$ signal. If the delay is too less then initial voltage difference

ΔV_i for the $NMOS$ cross coupled will be less which can cause the $NMOS$ latch to work in dead zone for some PVT conditions which will raise the reliability concerns for the sense amplifier. If this delay is high then before the NMOS cross couple starts its action, PMOS cross couple will act making the access time higher.

5.4.4 Results

Full Latch SA has better access time than any other topology which is discussed in this work which is evdent from Fig 5.17. To limit the power consumption same S-R latch is used to generate $DONE\_N$ signal. But for the currents which are as close as 10nA, this sense can show some reliability concerns.

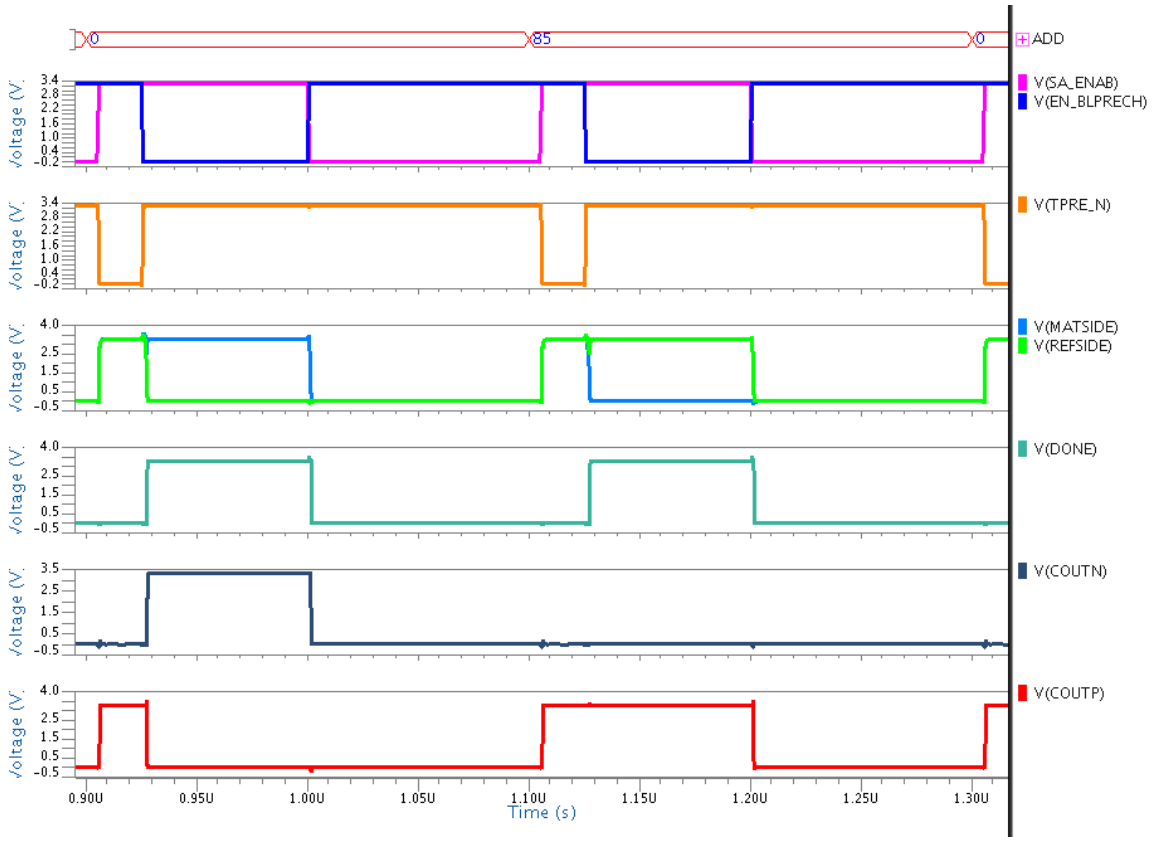
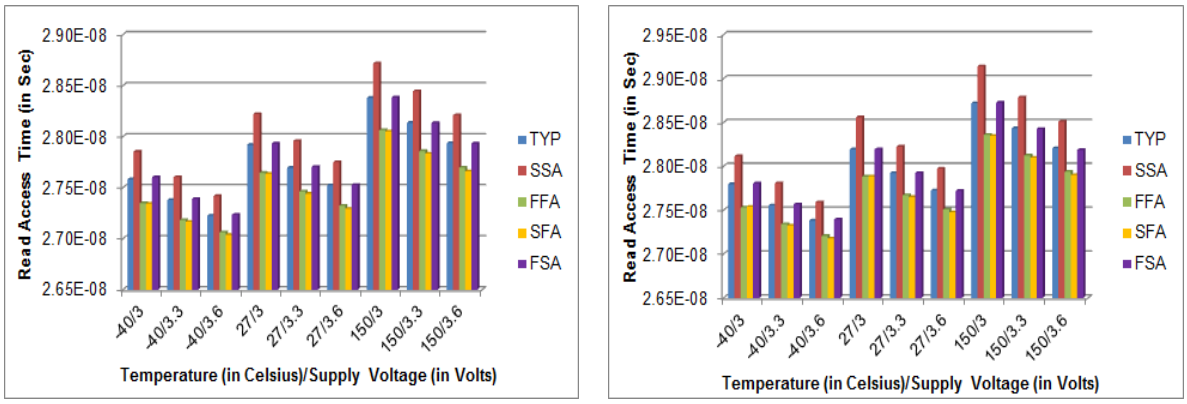
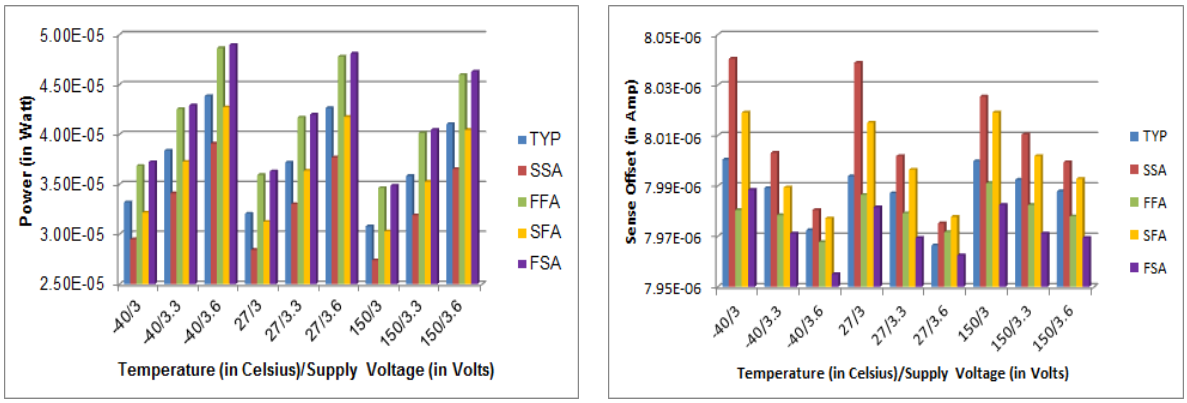


Figure 5.16: Read waveform of Full Latch SA



(a) (b)

Figure 5.17: Full Latch SA (a)Worst case user mode Read Access time, (b)Worst case FDMA mode Read Access time



(a) (b)

Figure 5.18: Full Latch SA (a)Power for worst case erased cell, (b)Offset across PVT

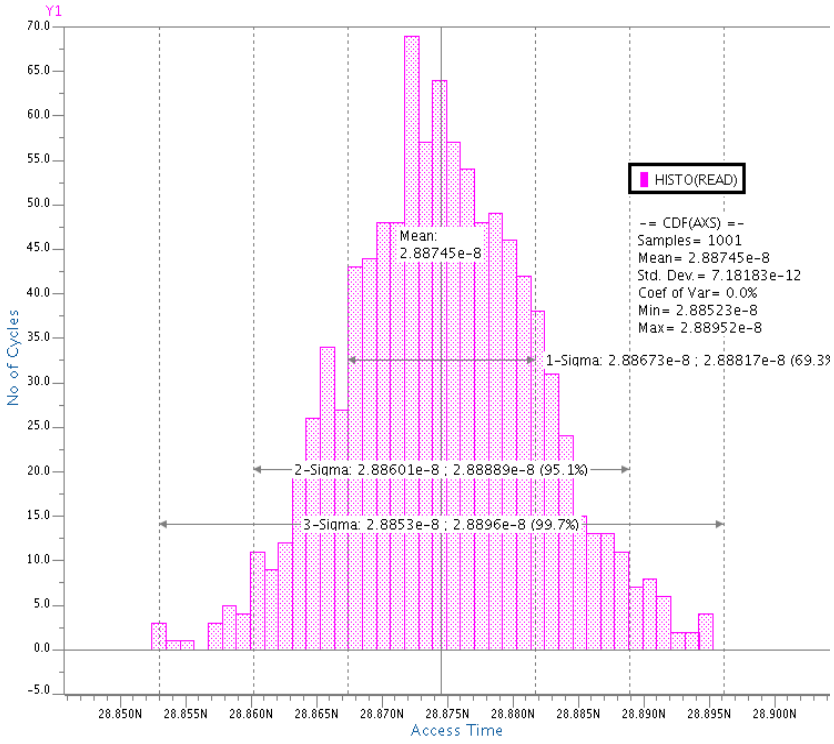


Figure 5.19: Monte-Carlo variation of user mode Access time for Full Latch SA

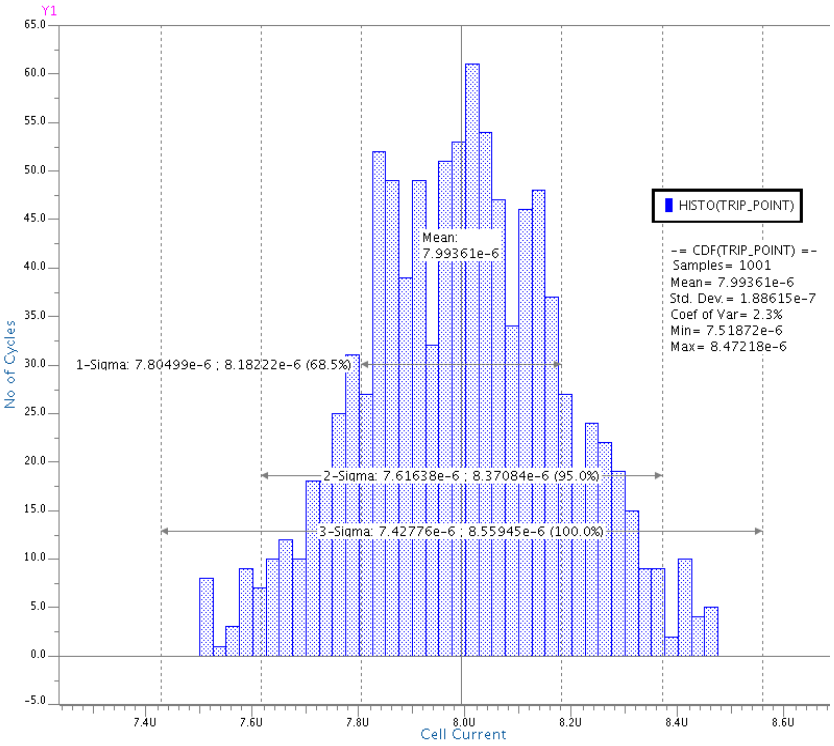


Figure 5.20: Monte-Carlo variation of Offset for Full Latch SA

Chapter 6

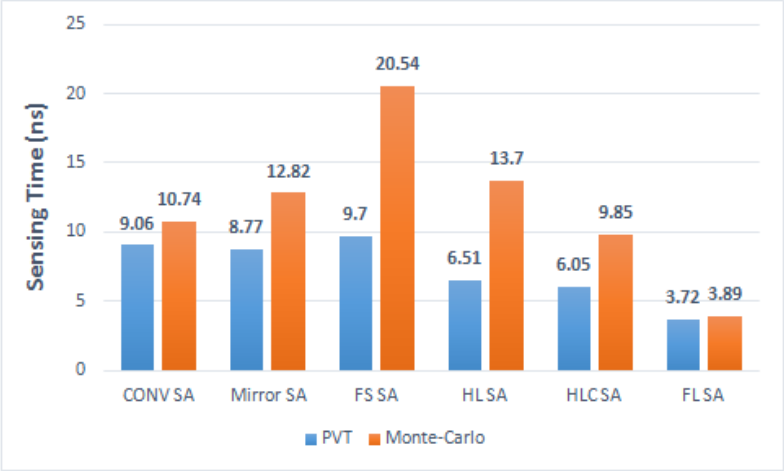
Comparison of Performance Metrics

For the sake of fair comparison all the sense amplifier topologies are designed and simulated in same technology, 40nm STM40 triple well CMOS. All the circuits are simulated with 10% variation in supply voltage, temperature ranging from -40 to 150 degree centigrade. Also the design is run through five standard process corners which are TYP/TYP, SLOW/FAST, FAST/SLOW, FAST/FAST, SLOW/SLOW. In this section, the comparative results of the worst case measured performance parameters are shown and discussed. Firstly the worst case access time in the case of user mode read and FDMA read is presented followed by the power and offset results.

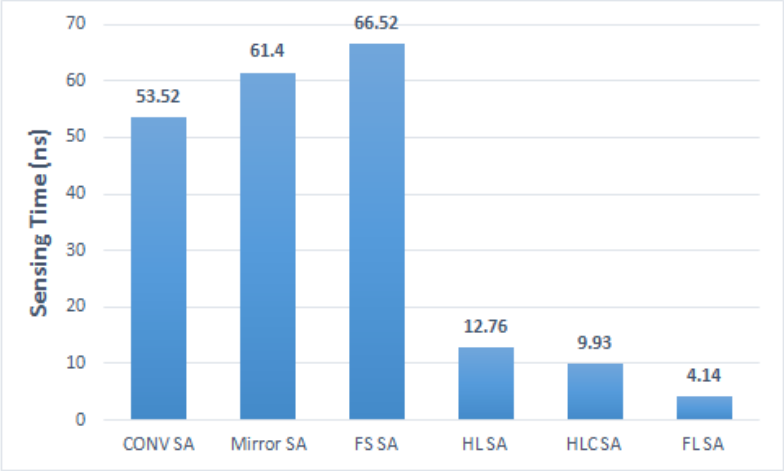
6.1 Comparison of Access time

Acheiving lower read access time is one of the major motivating factor behind choosing any sense amplifier topology. The reported access time comprises of precahrge time and sensing time. In all the discussed topologies according to the worst case precharge time across PVT, T_PRE was fixed at 25ns.

In the case of conventional SA worst case user mode read sensing time acheived was 9.06ns. This is reduced to 8.77ns by using the discharging behaviour of sensing node in mirror SA. In the case of both ref and sense nodes discharging, used in fully symmetric SA, the sensing time is 9.7ns. This increase in sensing time is due to time taken by involved current mirrors to settle. In the case of latch based dynamic sense amplifier, with half latch SA we have acheived sensing time of 6.51ns which is furhter reduced to 6.05ns by introducing comparator in HLC SA. Finally with full latch sense ampllifeir the worst case sensing was 3.72ns. Comparison of sensing time in user mode read is shown in Fig 6.1(a). By using full latch based sense amplifier we are able to acheive 58.94% reduction in user mode read access time. This reduction in sensing time is more prominent for lower difference of currents in FDMA read mode as shown in Fig 6.2(b). For the case of FDMA, by using full latch based topology instead of conventional topology, the reduction in sensing time acheived is 92.26%. This is a major improvement for speed critical applications. With the Monte-Carlo results, we can conclude the insensitivity of Full Latch SA towards read current.



(a)



(b)

Figure 6.1: Comparison of Sensing time for (a) User mode (b) FDMA mode

6.2 Comparison of Power

Power consumption is another critical aspect of sense amplifier which is even more important with higher degree of parallism in modern memories. For conventional sense amplifier power consumption for worst case erased cell, $I_{CELL} = 40\mu A$, was $143\mu W$. This consumption is reduced to $102\mu W$ in mirror sense amplifeir by limiting the current in cell branch. But for fully symmetric sense amplifier, the power consumption is very high due to static power consumed in aditional current mirror branches. For the case of dynamic sense amplifier the power consumption is gretely reuced by eliminating any static current consumption after the data is latch into S-R flip flop. In conclusion maximum power reduction acheived is 65.74 % in the case of full latch sense amplifier. The comparison result is shown in Fig 6.2.

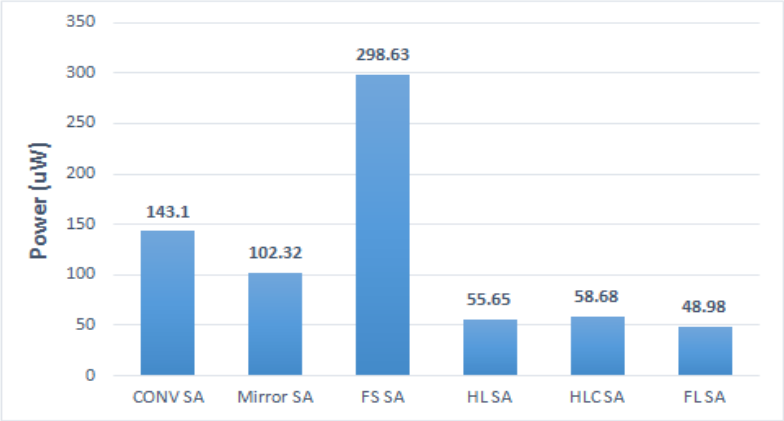


Figure 6.2: Comparison of Power Consumption for different SA

6.3 Comparison of Sense Offset

Offset, systematic as well as random, is of concern while reading with close cell and ref currents and also in FDMA test mode. The total offset in any case should not increase $1\mu A$ when measured around reference current. Lowest offset across all PVT was $2nA$, acheived in fully symmetric sense amplifier. In the same PVT variation, the highest offset was $191nA$ for conventional SA. This is due to the higher dependence of overdrive voltage of diode connected MOS towards its threshold variation.

The random offset in the case of dynamic SA topologies were contained by having higher device sizes. Fig 6.3 shows the comparison of offset across PVT and with 1000 Monte-Carlo simulations.

Finally Table 6.1 summarizes the performance parameters of all the discussed sense amplifier topologies.

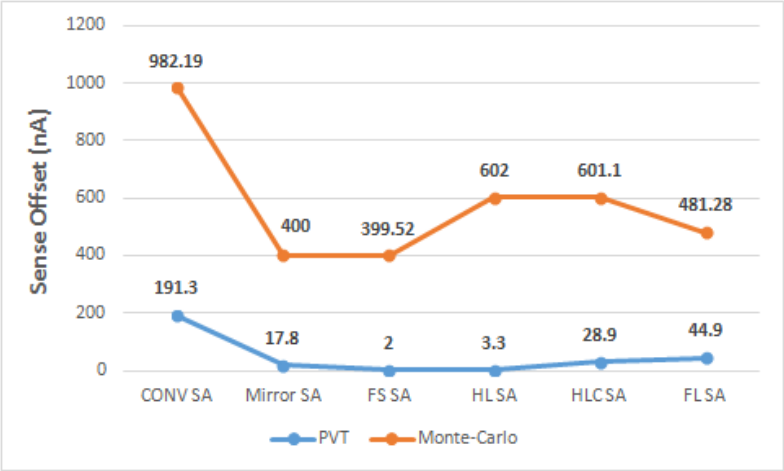


Figure 6.3: Comparison of Offset for different SA

Table 6.1: Comparison of Performance Metrics for Various Sense Amplifiers

Performance Metric	**Conv SA**	**Mirror SA**	**FS SA**	**HL SA**	**HLC SA**	**FL SA**
Worst Case User Mode Sensing Time (ns)	34.06	33.77	34.70	31.51	31.05	28.72
Worst Case FDMA Access Time (ns)	78.52	86.4	91.52	37.76	34.93	29.14
Monte-Carlo of User Mode Access Time (ns)	35.744	37.827	45.540	38.755	34.5	28.89
Worst Case Power Consumption (μW)	143.5	102	299	55.65	58.68	48.98
Sense Offset Across PVT (nA)	191.3	17.8	2	3.3	28.9	44.9
Monte-Carlo of Sense Offset (nA)	982	400	399	602	601.1	481

Chapter 7

Conclusion and Future Work

7.1 Summary

In this work, different static and dynamic kind of sense amplifier topologies were designed and their peformance was evaluated on different metrics. Due to the latest flash technology node, 40nm, used for designing, there are concerns of reliability which will be eliminated for stable process nodes.
It has been concluded that for speed critical applications, it is a necessity to adopt dynamic sense amplifier topology. More specifically, Full Latch type SA has shown signifcant reduction in sensing time. If the fabrication process is not stable, for speed critical applications, at some perticular technology node then Mirror SA will be the best choice since it will not have any reliability concerns. In case of Multilavel flash technology where sense offset is of major concern, Fully Symmetric SA must be considered but its power consumption should be examined.

7.2 Future Work

Design of Dynamic type sense amplifier has been crucial in this dissertation. Topological exploration has been covered in this research while specific issues must be dealt acccording to application requirements. If required, techniques for offset reduction for dynamic sense amplifier topologies should be examined. Finally, this work assists NVM designers to choose the most important building block of memory read path i.e. sense amplifier, chosen from available options according to the system level specifications.

Bibliography

[1] Allen, P. E. *CMOS analog circuit design*. Oxford University Press, New York, 2012.

[2] Amin, A. Design and analysis of a high-speed sense amplifier for single-transistor non-volatile memory cells. *IEE Proceedings G (Circuits, Devices and Systems) 140*, 2 (1993), 117–122.

[3] Amin, A. A., and Emoto, B. High speed differential sense amplifier for use with single transistor memory cells, May 26 1992. US Patent 5,117,394.

[4] Bedeschi, F., Bonizzoni, E., Khouri, O., Resta, C., and Torelli, G. A fully symmetrical sense amplifier for non-volatile memories. In *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on* (2004), vol. 2, IEEE, pp. II–625.

[5] Bez, R., Camerlenghi, E., Modelli, A., and Visconti, A. Introduction to flash memory. *Proceedings of the IEEE 91*, 4 (2003), 489–502.

[6] Campardo, G. *VLSI-design of non-volatile memories*. Springer, Berlin New York, 2005.

[7] Cappelletti, P. *Flash memories*. Kluwer Academic Publihsers, Boston, Mass, 1999.

[8] Chang, M.-F., and Shen, S.-J. A process variation tolerant embedded split-gate flash memory using pre-stable current sensing scheme. *Solid-State Circuits, IEEE Journal of 44*, 3 (2009), 987–994.

[9] Chang, M.-F., Shen, S.-J., Liu, C.-C., Wu, C.-W., Lin, Y.-F., King, Y.-C., Lin, C.-J., Liao, H.-J., Chih, Y.-D., and Yamauchi, H. An offset-tolerant fast-random-read current-sampling-based sense amplifier for small-cell-current nonvolatile memory. *Solid-State Circuits, IEEE Journal of 48*, 3 (2013), 864–877.

[10] Chimenton, A., and Olivo, P. Impact of high tunneling electric fields on erasing instabilities in nor flash memories. *Electron Devices, IEEE Transactions on 53*, 1 (2006), 97–102.

[11] Chrisanthopoulos, A., Moisiadis, Y., Varagis, A., Tsiatouhas, Y., and Arapoyanni, A. A new flash memory sense amplifier in 0.18 μm cmos technology. In *Electron-*

ics, Circuits and Systems, 2001. ICECS 2001. The 8th IEEE International Conference on (2001), vol. 2, IEEE, pp. 941–944.

[12] Chung, C.-C., Lin, H., and Lin, Y.-T. A novel high-speed sense amplifier for bi-nor flash memories. *Solid-State Circuits, IEEE Journal of 40*, 2 (2005), 515–522.

[13] d'Abreu, M. Nand flash memory: The driving technology in digital storage-overview and challenges. In *VLSI (ISVLSI), 2013 IEEE Computer Society Annual Symposium on* (2013), IEEE, pp. 1–1.

[14] Hajimiri, A., and Heald, R. Design issues in cross-coupled inverter sense amplifier. In *Circuits and Systems, 1998. ISCAS'98. Proceedings of the 1998 IEEE International Symposium on* (1998), vol. 2, IEEE, pp. 149–152.

[15] Jefremow, M., Kern, T., Backhausen, U., Peters, C., Parzinger, C., Roll, C., Kassenetter, S., Thierold, S., and Schmitt-Landsiedel, D. Bitline-capacitance-cancelation sensing scheme with 11ns read latency and maximum read throughput of 2.9 gb/s in 65nm embedded flash for automotive. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International* (2012), IEEE, pp. 428–430.

[16] Larcher, L., Pavan, P., and Maurelli, A. Flash memories for soc: an overview on system constraints and technology issues. In *System-on-Chip for Real-Time Applications, 2005. Proceedings. Fifth International Workshop on* (2005), IEEE, pp. 73–77.

[17] Lee, Y.-H., McMahon, W., Lu, Y.-L., and Freidin, Z. Reliability tradeoffs and scaling issues of read drain bias in nor flash memory. *Electron Devices, IEEE Transactions on 56*, 9 (2009), 2045–2051.

[18] Lee, Y.-H., Meng, Q., Jiang, L., et al. Drain read disturb assessment of nor flash memory. In *2008 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)* (2008), pp. 83–84.

[19] Lim, S.-H., and Park, K.-H. An efficient nand flash file system for flash memory storage. *Computers, IEEE Transactions on 55*, 7 (2006), 906–912.

[20] Lin, J.-H., Chang, Y.-H., Hsieh, J.-W., Kuo, T.-W., and Yang, C.-C. A nor emulation strategy over nand flash memory. In *Embedded and Real-Time Computing Systems and Applications, 2007. RTCSA 2007. 13th IEEE International Conference on* (2007), IEEE, pp. 95–102.

[21] Liu, Y.-C., Chang, M.-F., Lin, Y.-F., Wu, J.-J., Yeh, C.-J., Shen, S.-J., Chen, P.-C., Tsai, W.-C., Chih, Y.-D., and Natarajan, S. An embedded flash macro with sub-4ns random-read-access using asymmetric-voltage-biased current-mode sensing scheme. In *Solid-State Circuits Conference (A-SSCC), 2013 IEEE Asian* (2013), IEEE, pp. 241–244.

[22] Lu, C.-Y., Lu, T.-C., and Liu, R. Non-volatile memory technology-today and tomorrow. In *2006 13th International Symposium on the Physical and Failure Analysis of Integrated Circuits* (2006), pp. 18–23.

[23] Micheloni, R., Crippa, L., Sangalli, M., and Campardo, G. The flash memory read path: building blocks and critical aspects. *Proceedings of the IEEE 91*, 4 (2003), 537–553.

[24] Motta, I., Ragone, G., Khouri, O., Torelli, G., and Micheloni, R. High-voltage management in single-supply che nor-type flash memories. *Proceedings of the IEEE 91*, 4 (2003), 554–568.

[25] Na, T., Woo, S.-H., Kim, J., Jeong, H., and Jung, S.-O. Comparative study of various latch-type sense amplifiers. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 22*, 2 (2014), 425–429.

[26] Otsuka, N., and Horowitz, M. A. Circuit techniques for 1.5-v power supply flash memory. *Solid-State Circuits, IEEE Journal of 32*, 8 (1997), 1217–1230.

[27] Papaix, C., and Daga, J. M. A new single ended sense amplifier for low voltage embedded eeprom non volatile memories. In *Memory Technology, Design and Testin, IEEE International Workshop on* (2002), IEEE Computer Society, pp. 149–149.

[28] Pathak, B., Cabrera, A., Christensen, G., Darwish, A., Goldman, M., Haque, R., Jorgensen, J., Kajley, R., Ly, T., Marvin, F., et al. A 1.8 v 64 mb 100 mhz flexible read while write flash memory [in cmos]. In *Solid-State Circuits Conference, 2001. Digest of Technical Papers. ISSCC. 2001 IEEE International* (2001), IEEE, pp. 32–33.

[29] Pavan, P., Bez, R., Olivo, P., and Zanoni, E. Flash memory cells-an overview. *Proceedings of the IEEE 85*, 8 (1997), 1248–1271.

[30] Razavi, B. *Design of analog CMOS integrated circuits*. McGraw-Hill, Boston, MA, 2001.

[31] Uetake, T., Maki, Y., Nakadai, T., Yoshida, K., Susuki, M., and Nanjo, R. A 1.0 ns access 770 mhz 36 kb sram macro. In *VLSI Circuits, 1999. Digest of Technical Papers. 1999 Symposium on* (1999), IEEE, pp. 109–110.

[32] Zhang, H., and Lu, L. A low voltage sense amplifier for embedded flash memories.