SENSE AMPLIFIERS FOR HARDWARE IMPLEMENTATION OF NEURAL NETWORKS

A Project Report

submitted by

BENSON M SUNNY

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

MAY 2019

THESIS CERTIFICATE

This is to certify that the thesis titled **SENSE AMPLIFIERS FOR HARDWARE IMPLEMENTATION OF NEURAL NETWORKS**, submitted by **BENSON M SUNNY**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Janakiraman Viraraghavan
Project Guide
Assistant Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date: 12th May 2019

ACKNOWLEDGEMENTS

I would like to thank the following people who have profoundly helped me in my project work.

First and foremost, I express my sincere gratitude to Dr. Janakiraman Viraraghavan for his guidance and support through the project.

I would like to thank Ashwin and Lakshya, for providing support and innovation throughout my project related activities. I also thank everyone else at the Embedded Systems lab for their support.

Last, but not least, I thank my family and friends who have stood by me through thick and thin, without whom I could not have completed this project.

ABSTRACT

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. They are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found.An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets. Hence they are used in image processing, forecasting, speech recognition, medical diagnosis.

Implementing ANNs in hardware while expending as little energy as possible is challenging. The heart of an ANN evaluation is a multiply-accumulate (MAC) operation. In memory computing (IMC) is one such low power technique that performs the MAC operation within the memory, avoiding Von-Neuman type fetch and execute cycles. The sense amplifier (SA) is critical in IMC and in this thesis we assess the feasibility of adopting three different kinds of SAs. We compare the SAs on the basis of simulated performance and power numbers.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Neural networks are a set of algorithms, modeled loosely after the human brain, that is designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

Recently there have been a number of important advances that have allowed neural networks to become much more widely adopted, moving away from shallow to deep architectures. The rise of the internet means that we now live with an abundance of data. ANNs learn by example and so, with the widespread availability of billions of text documents, images, videos and audio recordings it is possible to train these algorithms reducing uncertainty on unknowns.

The ability to ingest this increased cache of data is largely due to advancements in hardware. Recently, a research group at Stanford University led by Andrew Ng found it possible to increase the training speed of a neural net by up to 100-fold when employing graphics processing units (GPUs) compared to traditional CPUs. The employ of GPUs allows for exploitation of the parallelism in the GPU architecture to be leveraged when performing mathematically intensive operations that a CPU is not designed to handle as optimally.

Finally, there have been important advancements in the techniques employed to train the ANN. Traditional gradient-based optimization tends to get stuck in poor solutions when trying to optimize deep architectures when randomly initialized. Strategies like greedy layer-wise unsupervised training helped progression from more traditional shallow architectures to the deep-learning frameworks that tend to work well for these big data applications.

1.2 Objective and Scope

The broader objective of this project is to implement a neural network in hardware. Since training a network is a complex process that requires large datasets and powerful computing, this is done offline. Then the trained parameters are loaded into the device. These parameters(weights) are read-only, so their values do not change. Repetitive fetching of these weights from the memory for operations is a wastage of energy. So, we switch to an in-memory computing method, where we try to implement the neural network within the memory itself. This can result in saving the energy required for repetitively fetching the weights.

For this, we implement a neural network trained in the MNIST database. MNIST contains images of handwritten digits from 0-9. The trained neural network has 784 inputs, 100 neurons in layer 1, and 10 output neurons.

For implementing this in hardware, we store the weights in SRAM columns. Inputs are provided through word lines. This is shown in Fig 1.1.

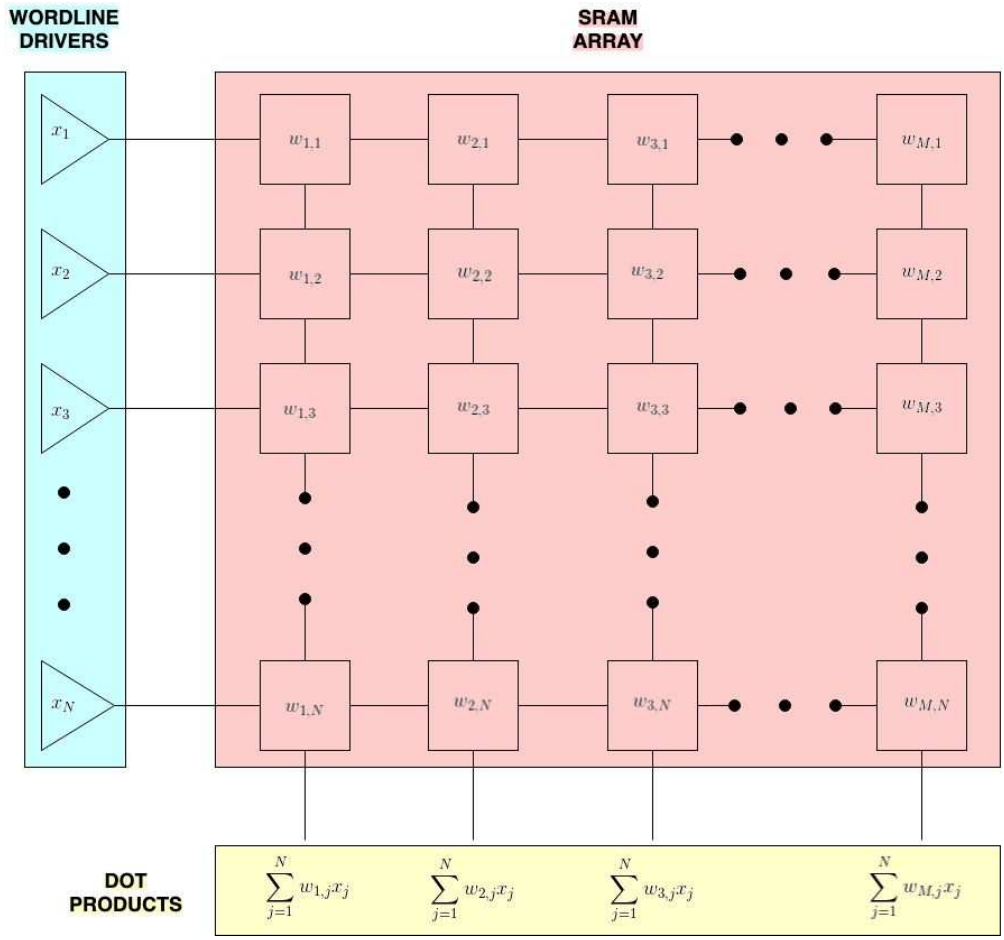

Figure 1.1: Implementation of a layer within memory columns

To save energy consumption, word line voltages are kept at subthreshold levels. In the subthreshold region, the current has an exponential relation to the input voltage. To make the current linearly proportional to the input, the word line drivers are designed such that the word line voltages have an appropriate logarithmic dependence to the input, thus cancelling out the exponential relation provided by the transistors. All the SRAMs in a column share a common bit line, which is precharged. Once the word lines are turned on for some time, the bit line voltage settles to a voltage which is proportional to the product of the input and the weight stored.

Now we need to convert the bit line voltage to a neuron output. One way to do this is to place ADCs at the column base but the area overhead for this is huge. An alternate way is to use Sense Amplifiers, which can act as comparators. This comparison capability is used to estimate a higher and lower bound on the voltage of the Bit-Line (comparable to a Flash ADC). Once this is done, the task of measurement is approximated to an interpolation of those bounds.

For this purpose, three sense amplifiers were designed and simulated.

- Voltage Mode Sense Amplifier.
- Current Controlled Latched Sense Amplifier.
- Slew Sense Amplifier.

All these sense amplifiers were tested for SRAM read operations as well as SRAM columns containing 784 SRAM cells. All designs were done in TSMC 28nm technology. Multiple parameters of these sense amplifiers were measured and compared.

1.3 Organization of the thesis

The remaining of this thesis is organized as follows.

Chapter 2 deals with Voltage Mode Sense Amplifier(VSA). This deals with the design, offset measurement, delay, power consumption, sizing and working of the VSA.

Chapter 3 deals with Current Controlled Latched Sense Amplifier(CSA). This also deals with the same parameters as that of the VSA. Comparisons are done with VSA parameters.

Chapter 4 deals with Slew Sense Amplifier(SSA). The parameters are measured and are compared with the other sense amplifiers.

Chapter 5 is a conclusion chapter that summarizes all the various parameters of the sense amplifiers. Recommendations are made based on the requirements.

CHAPTER 2

VOLTAGE MODE SENSE AMPLIFIER(VSA)

2.1 Introduction

Figure 2.1 shows a voltage mode sense amplifier. This sense amp uses a latch mechanism which is composed of two cross-coupled inverters(M8, M5, and M0, M4) which use positive feedback to cause latching behavior. This amplifier has two stable points where OUT and OUTB are at opposite rails and one metastable where OUT and OUTB are at the same voltage.

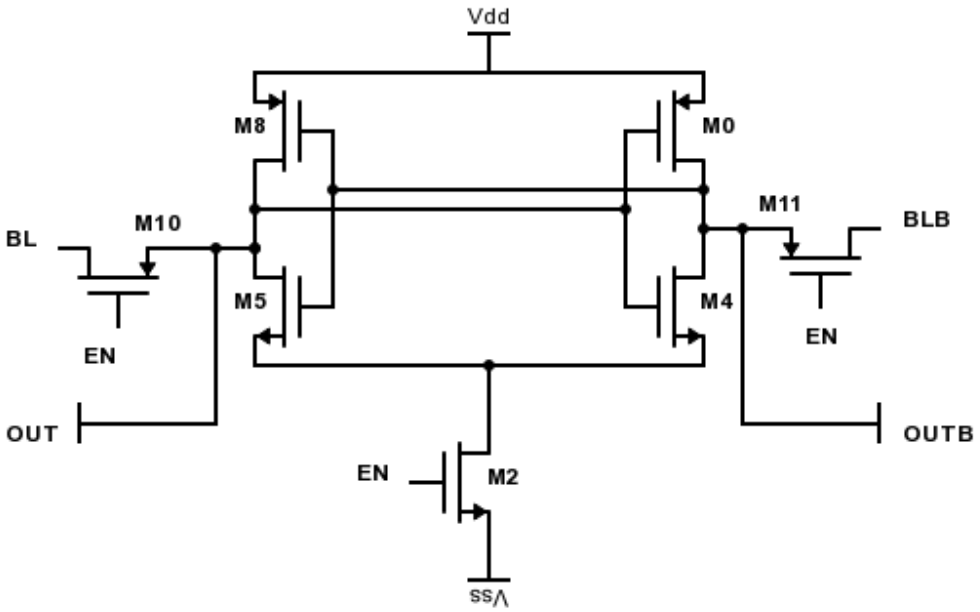

Figure 2.1: Voltage Mode Sense Amplifier

Before a read cycle, the bit lines which are connected to the amplifier inputs through pass transistors are precharged to the supply voltage. If the amplifier were to be turned on right after the precharge stage, there would not be a sufficient voltage difference between the bit lines, which could cause the amplifier to latch incorrectly. To solve

this issue, the bit lines are allowed to discharge until they have created a large enough voltage difference for the sense amplifier to correctly identify the value stored in the memory cell. The discharging bit lines are connected to the inverter inputs, allowing a small bias to develop on the MOSFET gates before the sense amplifier is enabled and the inverters are turned on. To turn on the amplifier, the signal EN is brought high, which turns on the NMOS current source(M2). After being enabled, the bias allows the sense amplifier to latch the correct output value.

The pass transistors connect the input and output of the sense amplifier to the bit lines and are active when the sense amplifier is disabled. This allows the bias to still develop on the gates while the sense amplifier is off and the bit lines are discharging. Once the sense amplifier is enabled, the pass transistors turn off and the bit lines are decoupled from the sense amplifier inputs. Once the amplifier is enabled, the bit lines can no longer discharge.

2.2 SRAM Read Operation

To evaluate the working, the sense amplifier was connected to an SRAM cell through column switches. Fig 2.2 shows the simulation waveforms when the memory cell stores a 1.

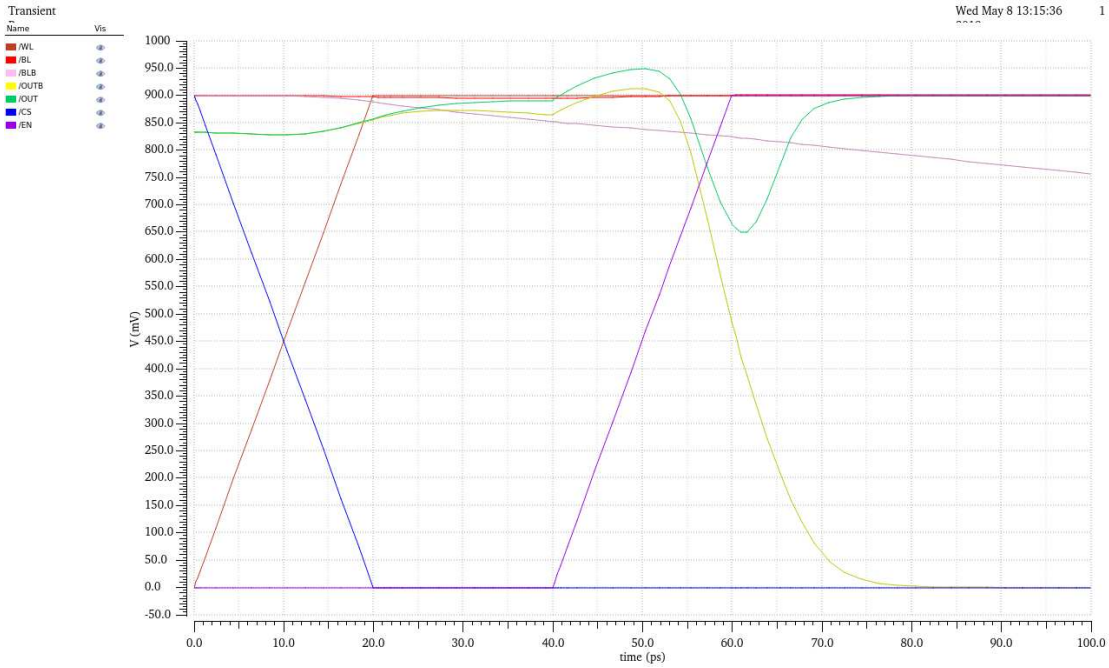

Figure 2.2: VSA SRAM read waveforms

From the figure, once the WL is turned on, BLB starts discharging. Once the BL and BLB have developed a sufficient differential voltage across them, EN is turned on at 40ps. Once EN is turned on, the latching begins and OUT and OUTB attains opposite rail voltages.

2.2.1 Power Dissipation

From the simulations, the average power dissipation for one read operation was found to be 57.67 uW.

2.2.2 Sensing Delay

Sensing delay is the time taken by the sense amplifier to latch to the output from the moment the WL is turned on. This delay depends upon a number of factors, one of them being the bit line capacitance. With an increase in bit line capacitance, the BLB takes longer time to discharge, hence the time required for the necessary differential voltage to appear across the bit lines also increases which in turn increases the sense amp delay.

A plot of the sense amplifier delay for a read operation in relation to the bitline capacitance is shown in Fig 2.3

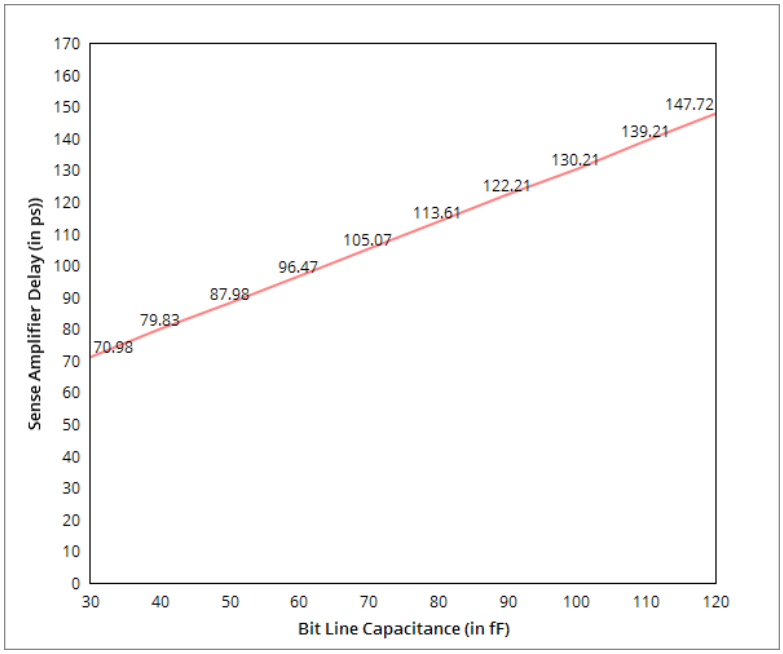

Figure 2.3: Sense amplifier delay versus Bitline capacitance

2.3 Offset Voltage

To measure the minimum bit line differential voltage required for the sense amp to latch the output correctly, the following input configuration was used. BL was kept at 0.9V and BLB was swept down from 0.9V. EN was given as a pulse with a period of 1ns. The offset can be measured from the time at which OUT switches from 0V to 0.9V. A sample input simulation is shown in Fig 2.4.

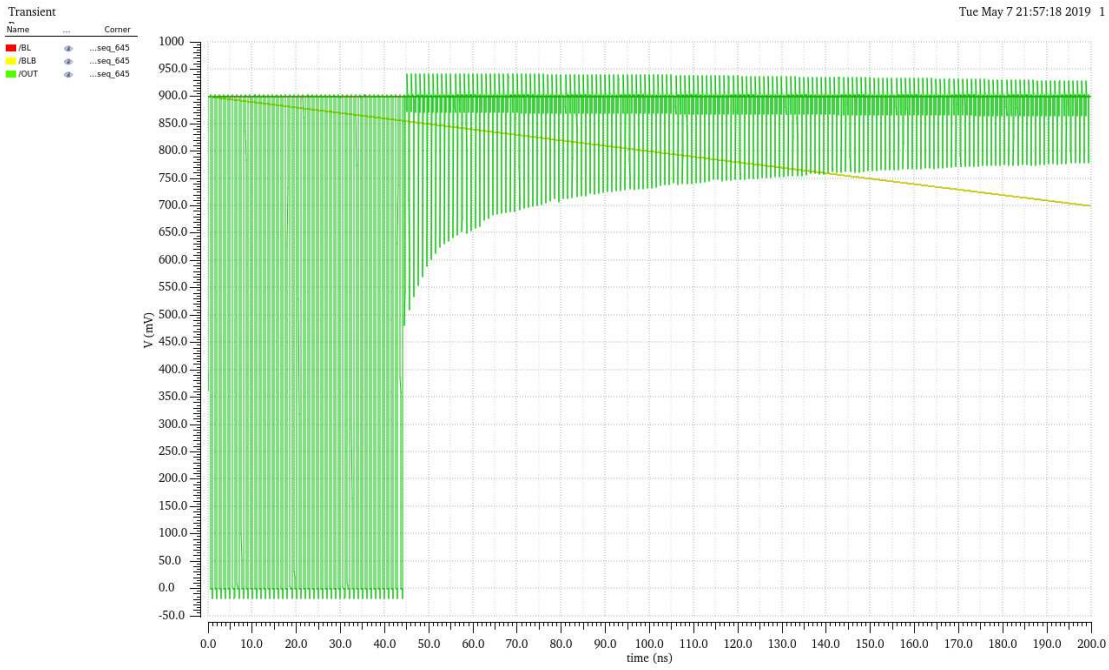

Figure 2.4: Sample Input Configuration For Offset Measurement

1000 Monte Carlo simulations were run, taking into consideration mismatch as well as process variations. From these simulations, the maximum offset was found to be 45 mV. So, the EN signal must be made high only after the bit line differential voltage has reached at least 45 mV for fully accurate estimations across all corners.

2.3.1 Sense Amplifier Sizing

The major source of offset in the sense amp is the mismatch variations among the transistors. The major contribution is from the mismatch in the pass transistors (M10, M11) and the NMOS pulldown transistors (M5, M4). Mismatch among the PMOS mosfets (M8, M0) does not have a considerable impact on the offset. This is because when the sense amplifier is turned on, the NMOS mosfets are immediately in saturation.

On the other hand, the PMOS mosfets are either in cutoff or in deep triode mode. Hence, the decision making in the sense amp is largely influenced by the NMOS transistors.

To reduce the offset due to the mismatch in the pass transistors and the NMOS pulldown transistors, M10, M11, M5, and M4 were sized to W/L = (600nm/30nm).

To size the PMOS transistors, the Static Noise Margin(SNM) of the inverter pair was taken into consideration. To maximize the SNM, the characteristics of the inverters in the pair should pass exactly through the center in the plot shown in Fig 2.5.

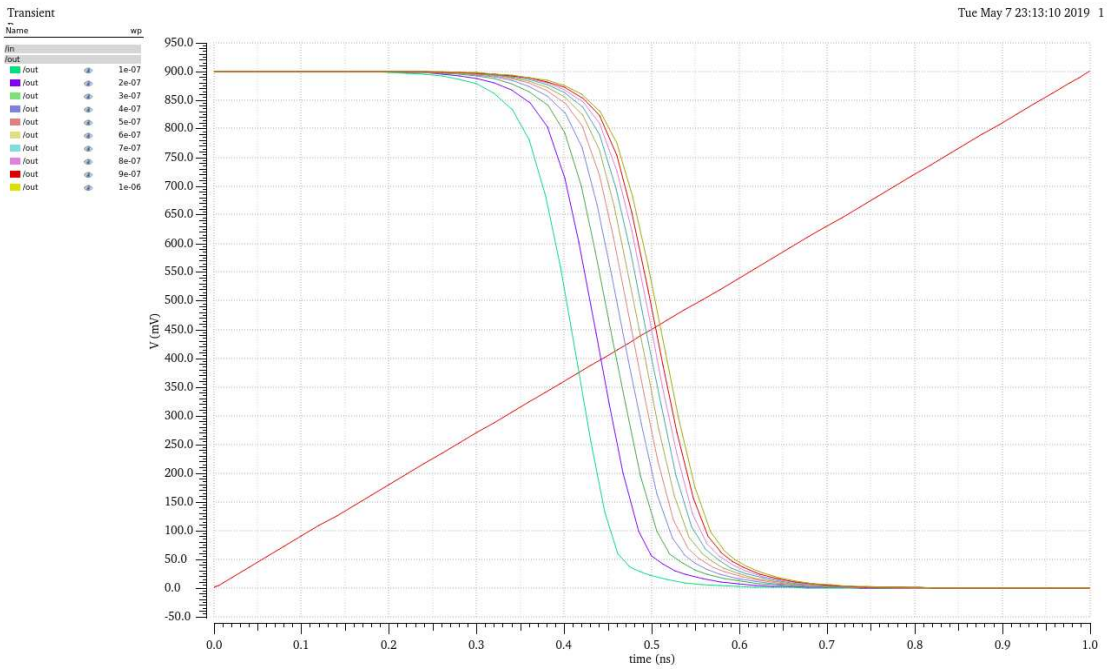

Figure 2.5: Inverter Parametric Analysis

In Fig 2.5, the characteristics of an inverter are plotted. The width of the NMOS is kept at 600nm and the width of the PMOS is swept from 100nm to 1um. From the plot, it is found that for PMOS size = 800nm, the characteristics are closest to the midway point. Hence both the PMOS transistors M8 and M0 are sized to W/L=(800nm/30nm).

2.4 Integration With SRAM Columns

The sense amplifier was connected to SRAM columns containing 784 cells through pass transistors. The simulation waveforms are shown in Fig 2.6.

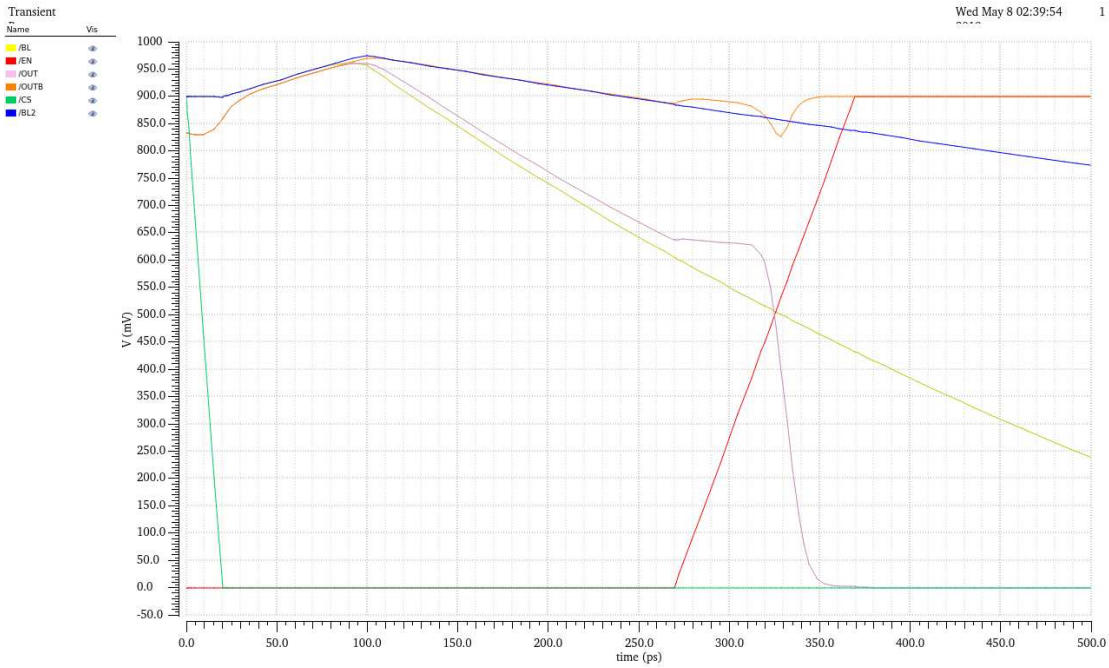

Figure 2.6: Simulation Waveforms On Integration With SRAM Columns

In this simulation, BL has the fastest possible discharge because the SRAM column to which it is connected stores the least amount of 1 in them. Again as in the case of an SRAM read operation, once BL and BL2 have sufficient differential voltage across them, the sense amp is turned on.

The timing of the EN signal is crucial. EN should be delayed such that the differential voltage is at least 45mV for 100% success rate. But this is not possible as EN should be high at least before the fastest discharge (BL in Fig 2.6) falls to low voltages. This can cause errors in the sense amp result, depending on the process and mismatch variations.

CHAPTER 3

Current Controlled Latched Sense Amplifier (CSA)

3.1 Introduction

This design combines aspects from the latch based voltage mode sense amplifier and a differential pair amplifier. The schematic is shown in Figure 3.1.

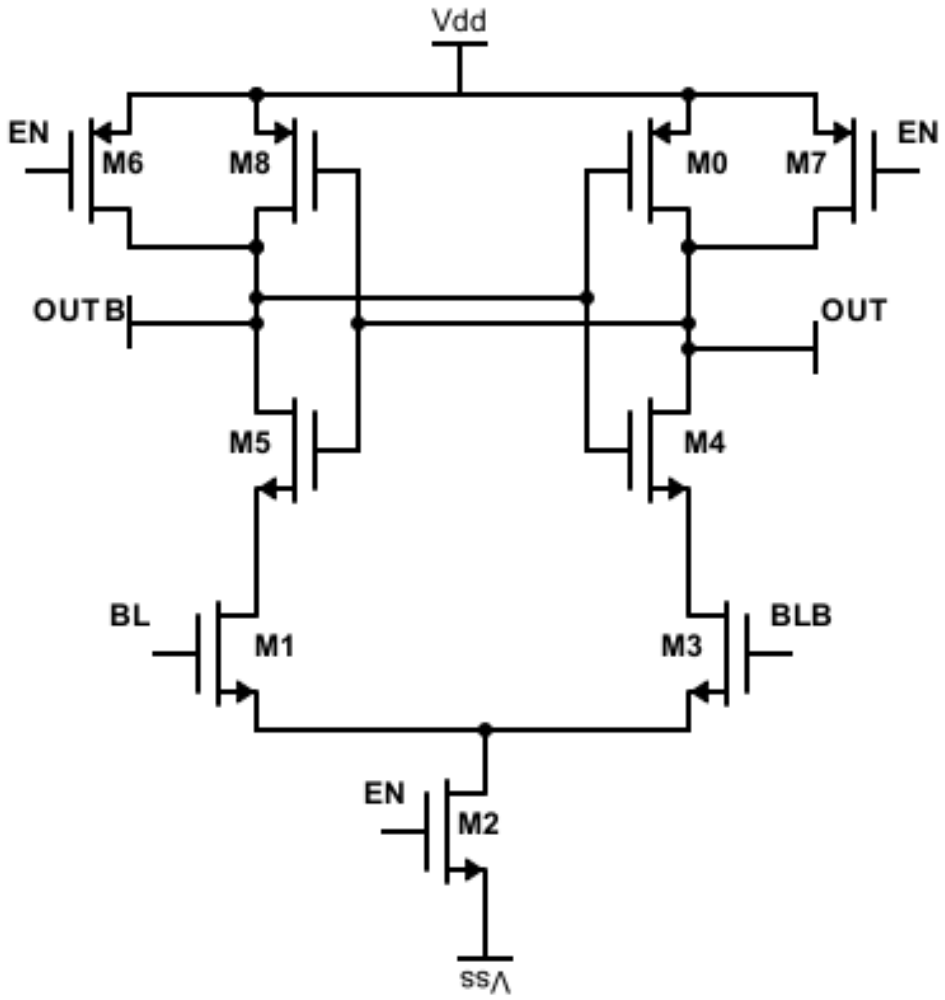

Figure 3.1: Current Mode Sense Amplifer

The amplifier uses the same latch based design as the voltage mode sense amp that was described above in 2.1 but has the bit lines attached to the gates of transistors M1

and M3. This causes an extremely high impedance to be seen at the input, and the bit lines to be decoupled from the output. The two inverters are made from the MOSFET pairs M8, M5 and M0, M4. When the amplifier is off, the signal EN is at a logic low. This turns on the reset transistors M6 and M7, clearing the previous latched value and holding the latch at its metastable point. The bit line inputs are connected to the gates of MOSFETs M1 and M3 . When the sense amplifier is enabled, the reset transistors turn off, and the current source M2 turns on, causing an even current to flow through each half of the amplifier. The bit line that is discharging causes the gate voltage to drop, which decreases the current flowing through that side of the amplifier and causes the voltage to rise. This will cause the amplifier to latch, which will stop the static flow of current and display a valid value at the output.

3.2 SRAM Read Operation

To evaluate the working, the sense amplifier was connected to an SRAM cell through column switches. Fig 3.2 shows the simulation waveforms when the memory cell stores a 1.

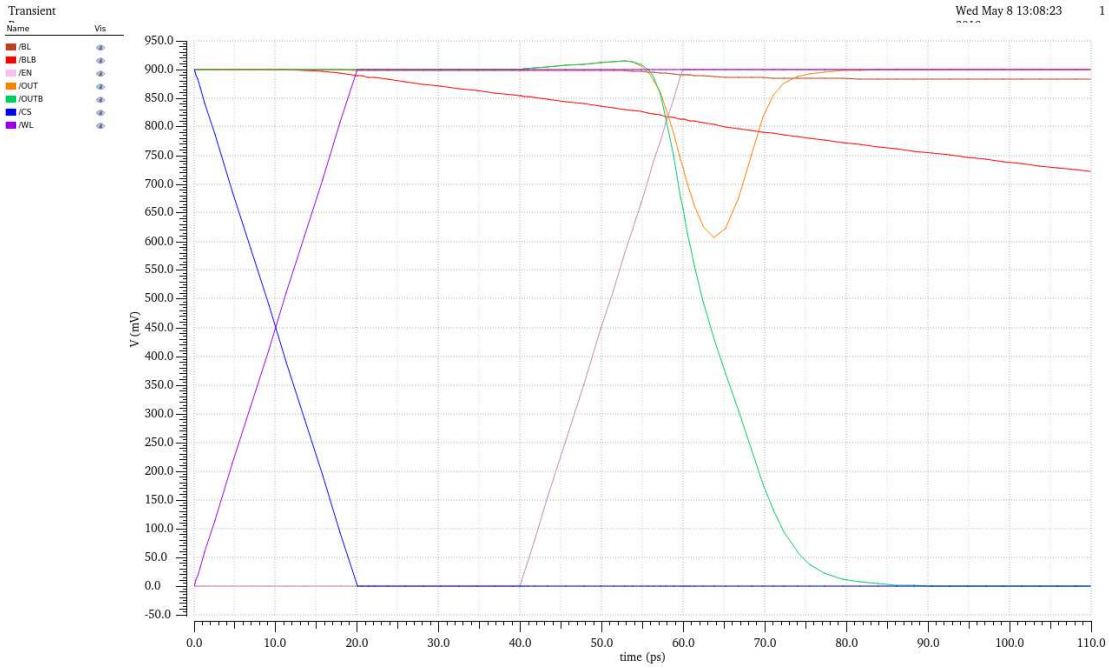

Figure 3.2: CSA SRAM read waveforms

From the figure, once the WL is turned on, BLB starts discharging. Once the BL and BLB have developed a sufficient differential voltage across them, EN is turned on.

Once EN is turned on, the latching begins and OUT and OUTB attains opposite rail voltages.

3.2.1 Power Dissipation

From the simulations, the average power dissipation for one read operation was found to be 52.34 uW, close to the power dissipation in VSA.

3.2.2 Sensing Delay

As in the case of VSA, with an increase in bit line capacitance, the sense amp delay increases. But since the bit line is decoupled from the output, the effect of increasing bit line capacitance on the sensing delay is not as profound as is the case in VSA. This is evident from the simulations and is displayed in Figure 3.3.

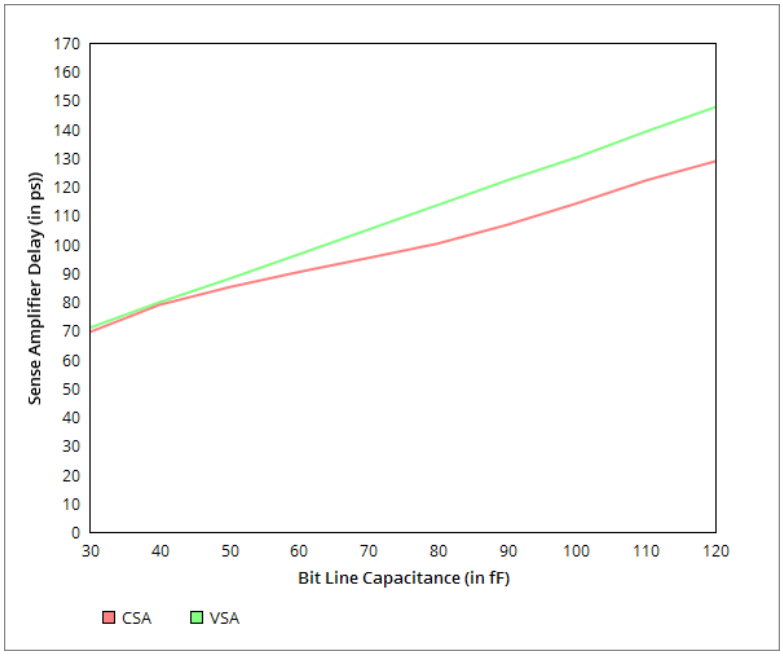

Figure 3.3: Sense amplifier delay versus Bitline capacitance

3.3 Offset Voltage

To measure the minimum bit line differential voltage required for the sense amp to latch the output correctly, the same setup explained in section 2.3 and displayed in Figure 2.4 is used. 1000 Monte Carlo simulations were run, taking into consideration mismatch as well as process variations. From these simulations, the maximum offset was found to be 41 mV. So, the EN signal must be made high only after the bit line differential voltage has reached at least 41 mV for fully accurate estimations across all corners.

3.3.1 Sense Amplifier Sizing

As is the case for VSA, the major source of offset in the sense amp is the mismatch variations among the transistors. The major contribution is from the mismatch in the NMOS transistors (M1, M3, M4, and M5). Mismatch among the PMOS mosfets (M8, M0) does not have a considerable impact on the offset. This is because when the sense amplifier is turned on, the NMOS mosfets are immediately in saturation. On the other hand, the PMOS mosfets are either in cutoff or in deep triode mode. Hence, the decision making in the sense amp is largely influenced by the NMOS transistors.

To reduce the offset due to mismatch, M1 and M3 were sized to W/L= (1um/30nm), M4 and M5 were sized to W/L = (500nm/30nm).

Again, to size the PMOS transistors, the Static Noise Margin(SNM) of the inverter pair was taken into consideration. To maximize the SNM, both the PMOS transistors M8 and M0 are sized to W/L=(700nm/30nm).

3.4 Integration With SRAM Columns

The sense amplifier was connected to SRAM columns containing 784 cells through pass transistors. The simulation waveforms are shown in Fig 3.4.

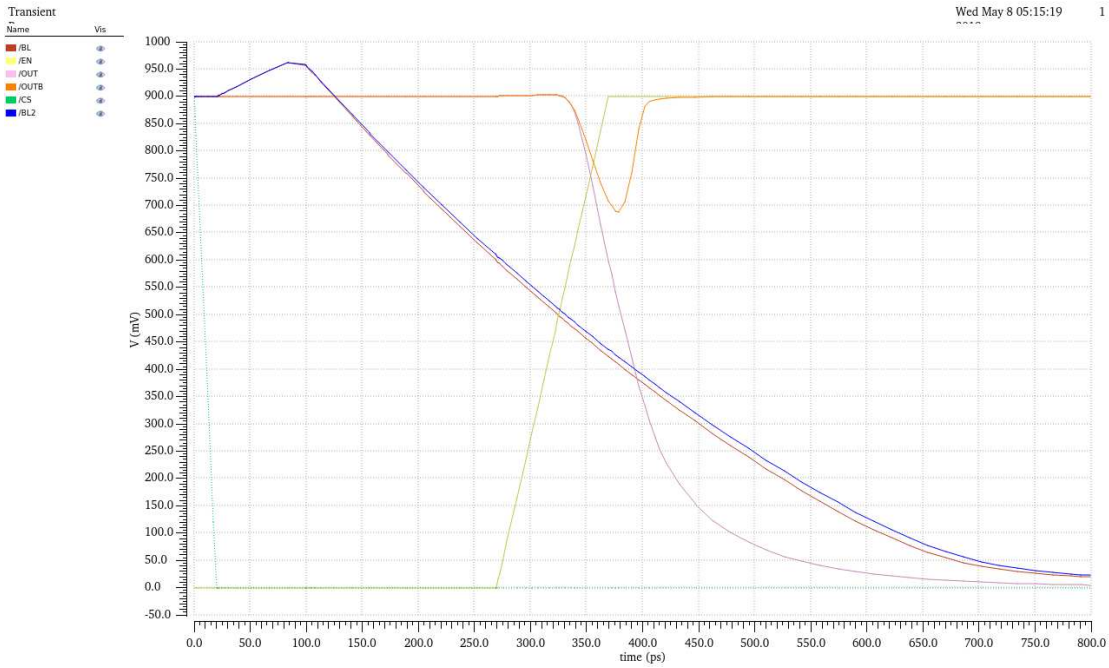

Figure 3.4: Simulation Waveforms On Integration With SRAM Columns

In this simulation, BL has the fastest possible discharge because the SRAM column to which it is connected stores the least amount of 1 in them. Again as in the case of an SRAM read operation, once BL and BL2 have sufficient differential voltage across them, the sense amp is turned on.

As is the case with VSA, the timing of EN signal is crucial. EN should be delayed such that the differential voltage is at least 41mV for 100% success rate. But this is not possible as EN should be high at least before the fastest discharge (BL in Fig 3.4) falls to low voltages. This can cause errors in the sense amp result, depending on the process and mismatch variations. The errors are likely to be lesser than VSA since the maximum offset value is slightly lower.

3.5 Layout

The DRC and LVS clean layout of the sense amplifier is shown in Figure 3.5.

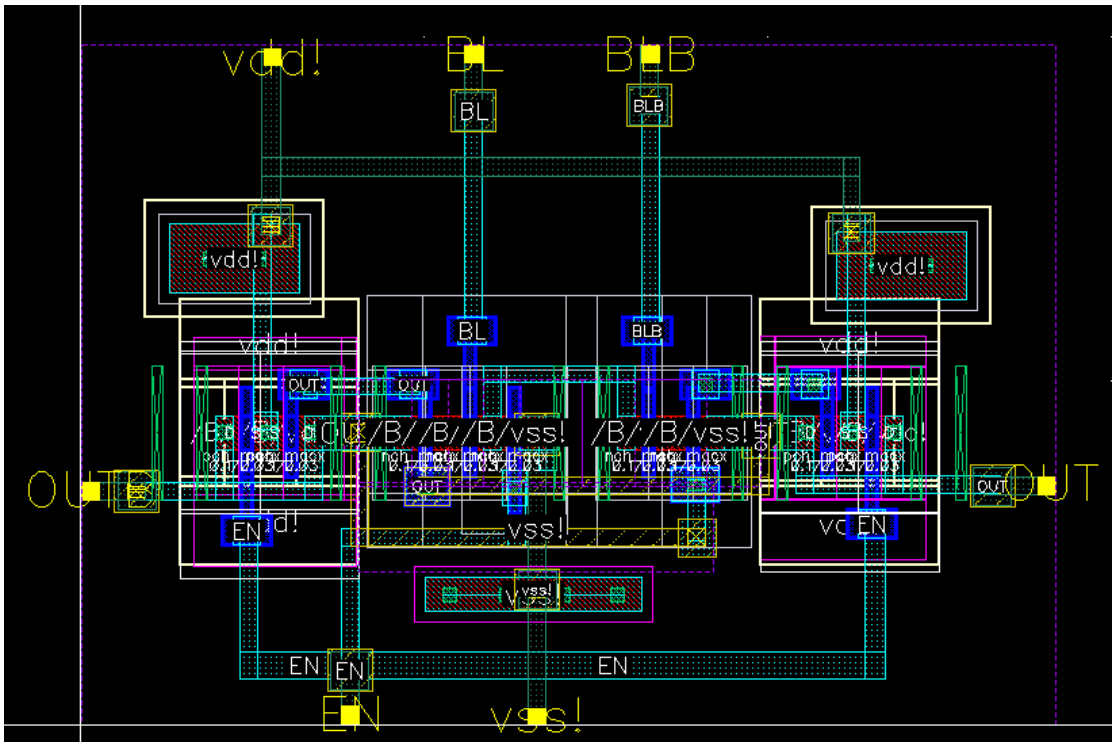

Figure 3.5: Layout of CSA

CHAPTER 4

Slew Sense Amplifier (SSA)

4.1 Introduction

The slew sense amplifier differs from the previously mentioned sense amplifiers in that it is a self-timed sense amplifier. There is no need for a properly timed enable signal in this case. Also, the circuitry for pre-charging the bit lines is built into the sense amplifier.

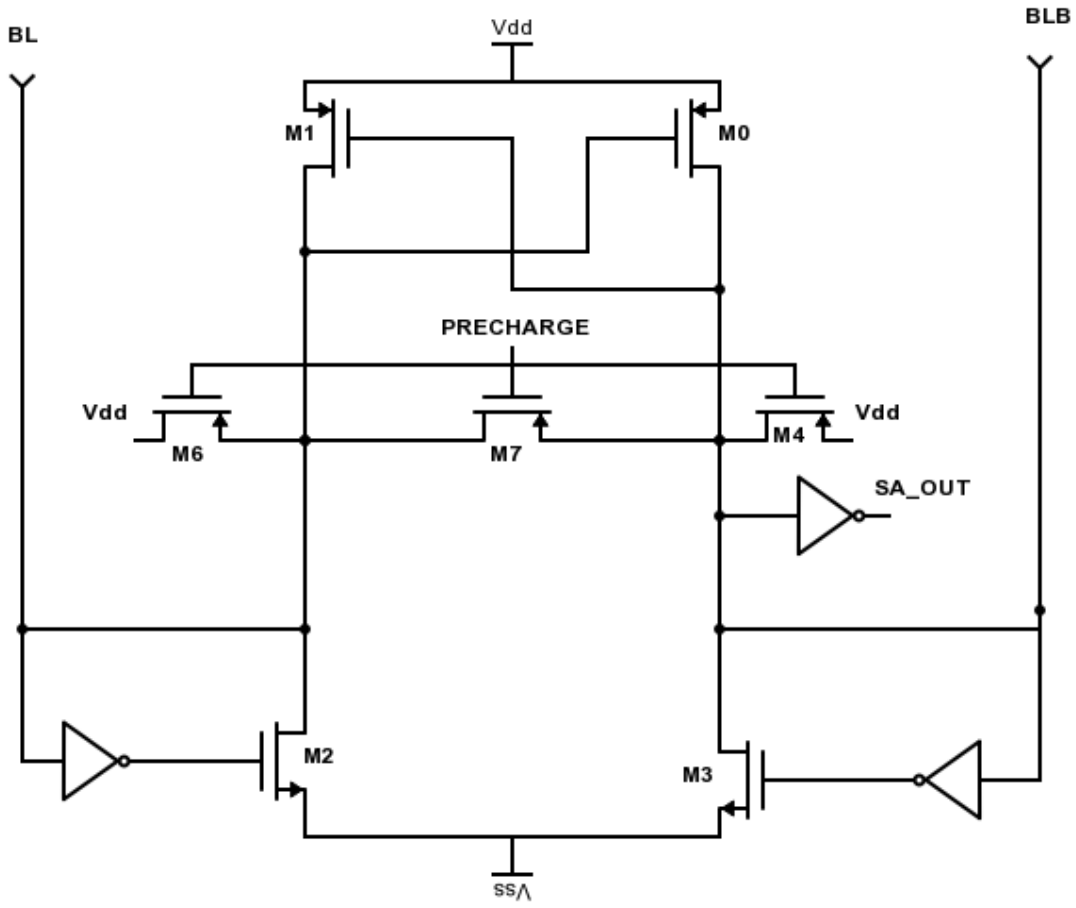

Figure 4.1: Slew Sense Amplifier

Before a read operation, the precharge transistors(M6, M7, and M4) are turned on. When the column switches are turned on, the bit lines are precharged. Also, the outputs of the inverters are low, thus M2 and M3 are off. Then the word lines are turned

on. Once the precharge transistors are turned off, the bit lines start discharging. At one point, one of the inverters switches depending on whether BL or BLB discharges faster. For example, if BL were to discharge faster, the inverter output connected to M2 switches state and M2 turns on. This results in an abrupt increase in current, and hence, M0 turns on. Since M1 and M0 form a cross-coupled PMOS stack, M1 turns off. The output is taken through an inverter and is labeled as SA_OUT.

4.2 SRAM Read Operation

To evaluate the working, the sense amplifier was connected to an SRAM cell through column switches. Fig 4.2 shows the simulation waveforms when the memory cell stores a 1.

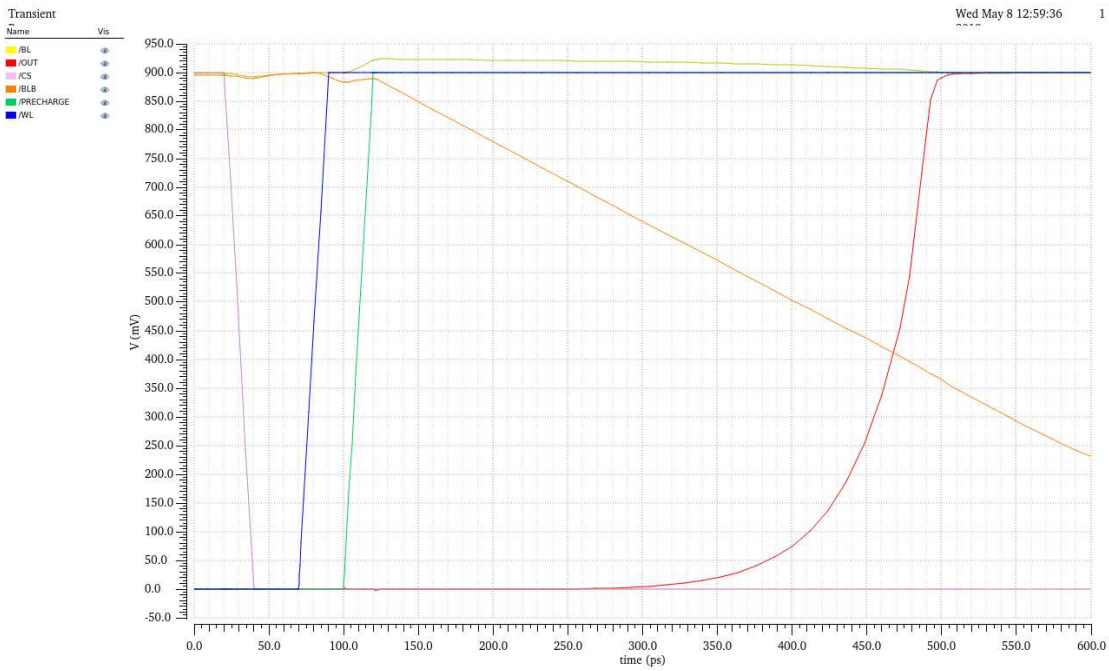

Figure 4.2: SSA SRAM read waveforms

From the figure, once WL is turned on and precharge transistors are turned off, BLB starts discharging. Once BLB falls to voltages low enough to switch the inverter, the inverter starts latching and SA_OUT goes high.

4.2.1 Power Dissipation

From the simulations, the average power dissipation for one read operation was found to be 39.23 uW. But since the delay is higher and WL needs to be turned on for a longer time, the energy consumption of this sense amplifier would be higher.

4.2.2 Sensing Delay

The variation of sensing delay with bit line capacitance is shown in Fig 4.3.

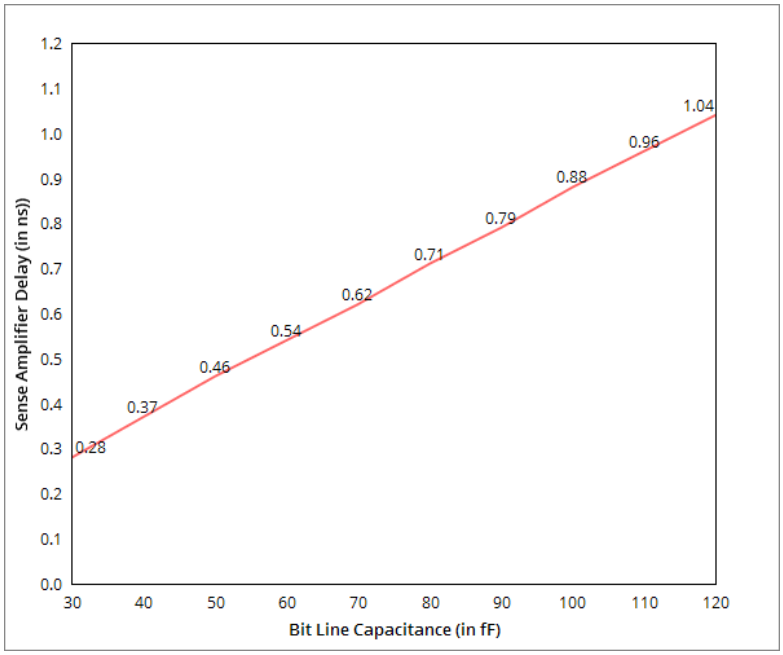

Figure 4.3: Sense amplifier delay versus Bitline capacitance

The delay is higher compared to the previously described sense amplifiers.

4.3 Integration With SRAM Columns

The sense amplifier was connected to SRAM columns containing 784 cells through pass transistors. The simulation waveforms are shown in Fig 4.4.

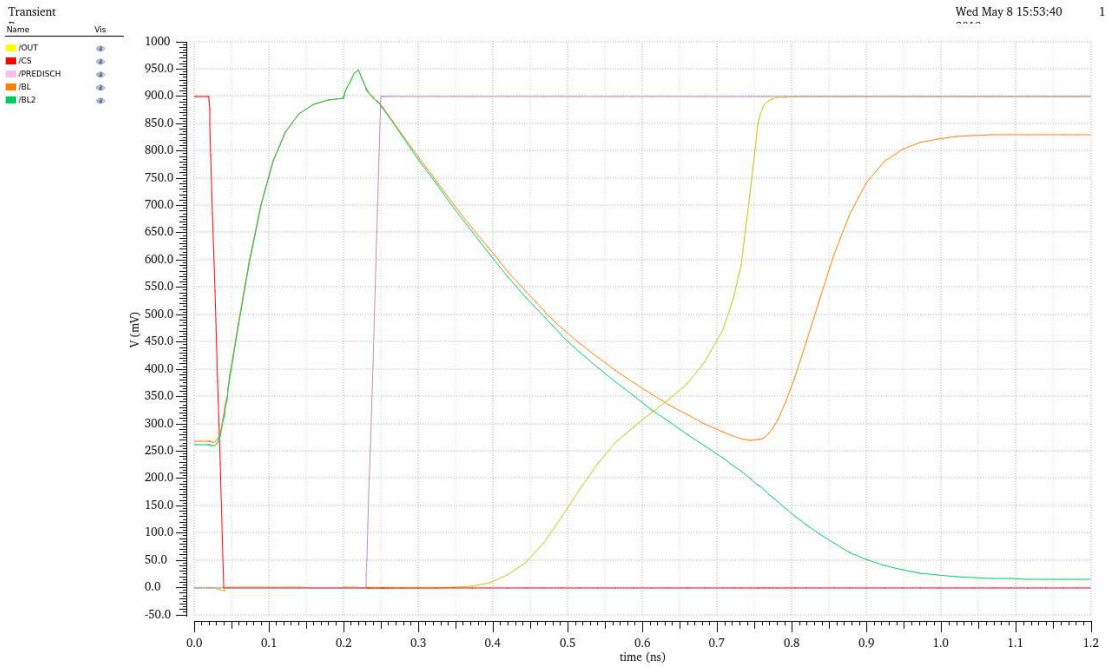

Figure 4.4: Simulation Waveforms On Integration With SRAM Columns

Due to the absence of needing a properly timed enable signal, the SSA is less prone to errors compared to VSA and CSA. On the flip side, the sensing delay and energy consumption are higher in SSA.

CHAPTER 5

Conclusion

The major constraints for the sense amplifier operation are its area, sensing delay, accuracy and energy consumption.

If the area is the primary concern, VSA is the most suitable as it has the least number of transistors among the three. CSA would take up more area and SSA would take up the most area.

If the delay is the primary concern, CSA is the most suitable as it produces the fastest output among the three and is more resilient to increasing bit line capacitance. VSA has slightly more delay values while the delay in SSA is much higher.

If the accurate prediction of the output is the primary concern, SSA would come in handy as the errors created in VSA and CSA due to the timing of the enable signal can be avoided. But since the word line needs to be turned on for more time for the output to latch, the energy consumption of this sense amplifier would be higher

REFERENCES

[1] Patrick G. Drennan, Colin C. McAndrew. *Understanding MOSFET Mismatch for Analog Design* . IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 38, NO. 3, MARCH 2003.

[2] Joseph F. Ryan and Benton H. Calhoun. *Minimizing Offset for Latching Voltage-Mode Sense Amplifiers for Sub-threshold Operation* . 9th International Symposium on Quality Electronic Design.

[3] Sarah Brooks. *Design of a Low Power Latch Based SRAM Sense Amplifier* . WORCESTER POLYTECHNIC INSTITUTE.

[4] Bernhard Wicht, Thomas Nirschl, Doris Schmitt-Landsiedel. *Yield and Speed Optimization of a Latch-Type Voltage Sense Amplifier* . IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 39, NO. 7, JULY 2004.

[5] A. Pulla Reddy, G. Sreenivasulu. *Design and Implementation of High Speed Sense Amplifier for Sram* . American-Eurasian Journal of Scientific Research 12 (6): 320-326, 2017.

[6] Disha Arora. *Design and Analysis of Sense Amplifier Topologies for Volatile and NonVolatile Memories* . Indraprastha Institute of Information Technology, New Delhi.

[7] Balaji Jayaraman et al. *80-kb Logic Embedded High-K Charge Trap Transistor-Based Multi-Time-Programmable Memory With No Added Process Complexity* . IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 53, NO. 3, MARCH 2018.

[8] Travis N. Blalock, Richard C. Jaeger *A High-speed Sensing Scheme for 1T Dynamic RAMâĂŹs Utilizing the Clamped Bit-Line Sense Amplifier* . IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 27. NO. 4, APRIL 1992.

[9] Travis N. Blalock, Richard C. Jaeger *A High-speed Clamped Bit-Line Current-*

Mode Sense Amplifier . IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 26, NO. 4, APRIL 1991.

[10] Jaspal Singh Shah, David Nairn, Manoj Sachdev *An Energy-Efficient Offset-Cancelling Sense Amplifier* . IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMSâĂŤII: EXPRESS BRIEFS, VOL. 60, NO. 8, AUGUST 2013.