Actively Controlled Retention Voltage of SRAMs

Ankush Mamgain

Indraprastha Institute of Information Technology

New Delhi

Guide

Dr. Anuj Grover

Submitted in partial fulfillment of the requirements

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

with specialization in VLSI and Embedded System July, 2018

Certificate

This is to certify that the thesis titled **"Actively Controlled Retention Voltage of SRAMs."** submitted by **Ankush Mamgain** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *VLSI & Embedded Systems* is a record of the bonafide work carried out by him under my guidance and supervision in the Security and Privacy group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Anuj Grover
Indraprastha Institute of Information Technology
New Delhi, 110020

Abstract

In advance technology nodes, static power consumption is a major component of total system power in systems that do not continuously operate at very high clock frequency. SRAMs not only contribute a major portion of SoC area but also of static power consumption. In this work, we propose an error amplifier based design to reduce retention leakage of a 4MB SRAM array. In 40nm LSTP technology, the amplifier consumes 81nW power. The overall memory subsystem leakage power reduces by 50% from no retention case and 33% from the conventional retention solution at TT (25°C) and by 75% from no retention & 69% from conventional solution at FNSP (140°C). Monte Carlo analysis shows the 3σ variations are within guard band limits.

Acknowledgments

The work for this thesis was carried out at STMicroelectronics, Greater Noida, India, during the year 2017-2018. First and foremost, I would like to express my special gratitude to my research advisor Dr. Anuj Grover for providing excellent guidance and encouragement throughout the journey of this work. Without his guidance, this work would never have been a successful one. My sincere thanks also goes to Mr. Kedar Janardan Dhori for his help and feedback whenever needed.

This work is dedicated to my parents. I am grateful to my brother Ashish Mamgain and brother-in-law Dr. Rajeev Nayan Bahuguna for encouraging me.

I am also thankful to my seniors Shashwat Kaushik, Ankush Singh, Vikas Shukla, Shrestha Bansal, Shivam Kalla and my friend Vaibhav Agarwal for their enthusiastic discussions whenever needed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

VLSI industry has advanced significantly over the decades and it has integrated systems onto single chips and enabled portable applications at reducing cost factors such that electronics is now pervasive and IoT is making everything smarter.

Memories are one of the important building blocks in such highly integrated circuits. In line with Moore's law, on-chip memory capacity doubles with every technology node. However, SRAM performance has not improved in line with logic performance and therefore performance gap between processor and memory has increased. This has led to hierarchical memory organization in processing units.

Figure 1.1 shows, how speed, cost, and density are related to different levels of memories. On-chip L1 and L2 memories are comprised of SRAMs. It is very important to decide the amount of caches on-chip and the number of cores in a system, which is a trade-off between area and performance. It is also important to use high-density SRAMs so as to improve the area and therefore, yield and product margins. SRAMs are being used in almost every electronics items. As a result, size and power consumption of SRAMs have become important figures of merit, which can be deciding factor in product's cost. Almost 70% to 90% area of a SOC is consumed by SRAMs [4]. Due to this, the leakage

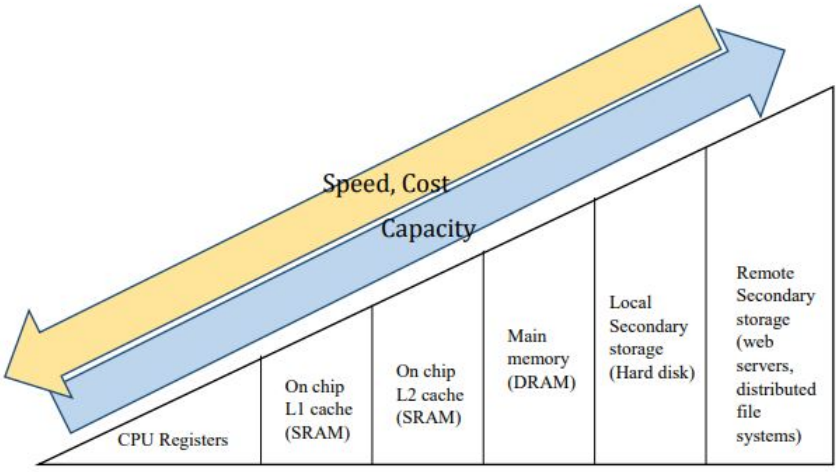

Figure 1.1: Memory Hierarchy In Advanced Processing Computer

current of embedded SRAMs dominates the standby current. In this dissertation, we have proposed a new architecture which promises to reduce leakage more efficiently for all PVT conditions as compared to conventional solutions when SRAMs are idle.

1.2 Thesis Organization

Thesis is organized as follows. Chapter 2 explains basic operation and figures of merit of 6T SRAM cell. It also explains the different types of leakage associated with the MOS transistor, leakage in the 6T SRAM cell & identify the devices responsible for leakage. In chapter 3, we review the conventional methods to modulate supply levels to achieve data retention voltage (DRV) in the SRAM followed by problems associated with these solutions. In chapter 4, we propose an actively controlled circuit to control retention voltage V_{RET} and then also propose a modification to alleviate its shortcomings. We present the results in chapter 5 and conclude in chapter 6.

Chapter 2

Figures of Merit of 6T SRAM Cell

A simplest SRAM array architecture is shown in Figure 2.1. It consists of memory cells, bitline conditioning (pre-charge circuit), row decoder, column decoder and column circuitry. Row decoder decodes the row address and a particular word line is activated. A column decoder and column MUX then selects a particular memory cell of a row which was activated by row decoder. The output of selected memory cell is then latched with the help of sense amplifier in column circuitry [1].

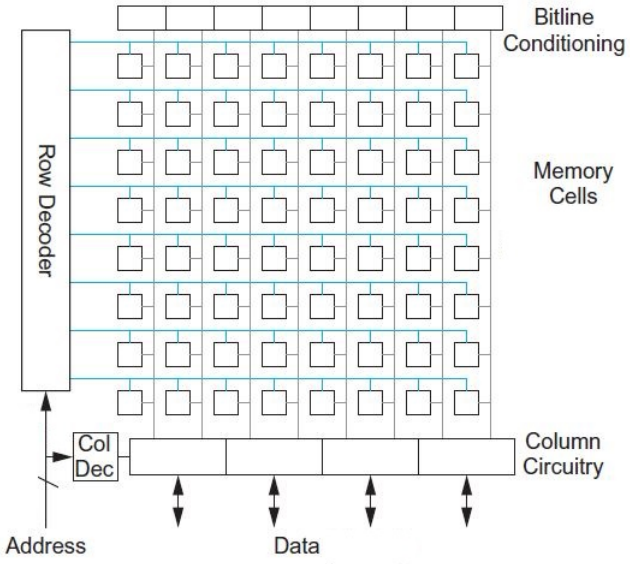

Figure 2.1: SRAM Architecture [1]

The 6T SRAM bit-cell contains two back to back inverters forming a latch so that it can hold a stable state and two pass gate (PG) transistors to read or write into the bit-cell. The positive feed-back ensures the state of cell such that any disturbance or noise should not affect the data. For writing into bit cell and reading from bit cell, sizing of transistors plays an important role. Before reading a bit cell both BLT and BLF

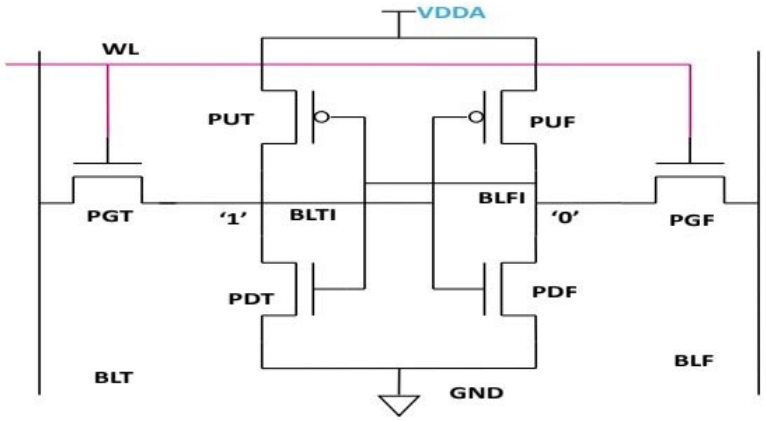

Figure 2.2: 6T SRAM Bit-Cell

are pre-charged to supply voltage using a pre-charge circuit. When the word line is high BLT and BLF interact with BLTI and BLFI. Suppose BLTI is at logic 1 and BLFI is at logic 0 then BLF will try to charge BLFI due to its large capacitance. To avoid this, size of PD device is larger than PG. During write operation if we want to write logic 0 and bit-cell stored with logic 1 then BLT will be discharged to ground and BLT will be charged to V_{DD}. BLT should be able to discharge BLTI to ground; this can be ensured if PU is weaker than PG [5]. Basic 6T SRAM bit cell is shown in Figure 2.2.

2.1 Cell Current

Cell current and its variations play a key role in defining the memory speed. If cell current is large, then the during read operation voltage difference between BLT and BLF creates in less time and sense amplifier can be triggered at lesser time. Pass gate (PG) (access transistor) and pull-down transistor (PD) both are responsible for cell current in SRAM cell. To improve cell current we can increase the size of PG and PD but on the other hand, there are some disadvantages of increasing sizes of PG and PD which are:

- The area will increase.
- Parasitic capacitance of the word line, BLT and BLF will increase. This can degrade the performance of SRAMs instead of improving.
- With increase in PG, bit line leakage also increases.

2.2 Bit-line Leakage

Increasing the size of PG will increase the leakage current from bit-line. For example in an 8-bit memory of single column; suppose bit-cell of first row is stored with logic 0 while others are stored with logic 1. When word line is selected to read a row, the remaining rows connected to same bit line leak. Due to this it takes longer to create the differential required to operate the sense amplifier. This degrades memory performance. It also poses limits on maximum number of bitcells connected to a bitline. This degrades SRAM density. This can be understood by Figure 2.3.

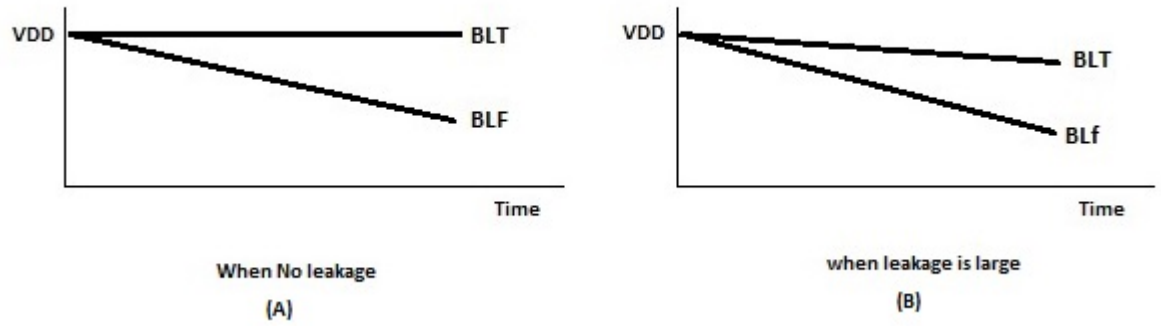

Figure 2.3: Bit Line Leakage

2.3 Write Margin

While writing into bit cell PG should be less resistive than PD. There can be mismatches and process variations which can affect the write operation. There are two ways by which

we define write margin.

- The highest level of Bitline to ensure cell flip with WL @ V_{dd}
- The lowest level of WL to ensure cell flip with BL @ gnd

2.4 Write Time

Write time is the time to write either logic 0 or logic 1 into a memory cell. Write time depends on the ratio of size of PU and PG transistors. This ratio is also known as pull up ratio. Increasing the pull up ratio will increase the voltage bump and this will increase the write time. If size of PU is greater than PD transistor then write operation will be difficult, because while PG is on BLT will try to discharge BLTI but at the same time PU transistor will be charging BLTI and due to this we use smaller size PU transistor.

2.5 Static Noise Margin

During a read operation, the data stored in memory cell should not be affected due to the interaction of BLT, BLTI and BLF, BLFI. This property is known as cell stability. When word line is selected for a particular row then PG and PD forms a type of potential divider network as shown in Figure 2.4. When word line is high some charge will transfer either from BLT to BLTI or vice versa. Due to this, there will be a voltage bump at BLTI. This sudden change in voltage should not be large enough to change the state of the memory cell. SNM of SRAM bit-cell can be defined as, the maximum noise voltage superimposed on bump voltage which the cell can tolerate without changing or flipping its logic is known as static noise margin (SNM). For this reason, PG should be weaker than PD or we can say that PG should be resistive than PD. Therefore the size of PD should be larger than PG. There is a limit to increasing size of PD because it can reduce the trip point.

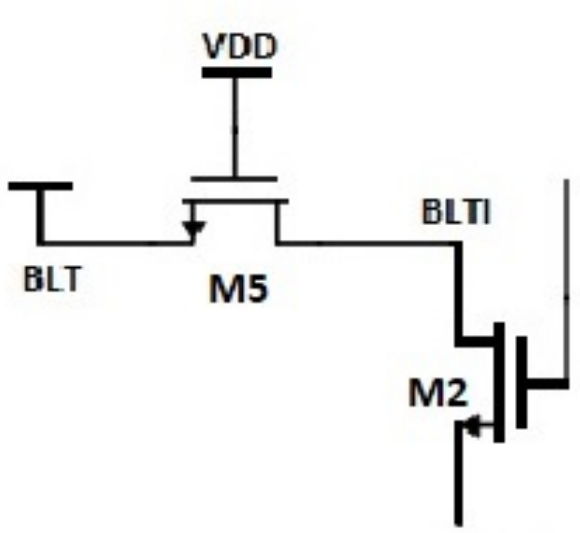

Figure 2.4: Potential Divider During Read Operation In SRAM

2.6 Retention Noise Margin

When memories are in idle mode, word line is kept at ground. The maximum noise voltage which the SRAM bit-cell can tolerate without changing or flipping its logic when they are in idle mode is known as retention static noise margin (RNM). When memories are in idle mode or whenever they are not being used, they consume power because of the leakage current from all the transistors. This power is being wasted without doing anything. Sub-threshold leakage contributes a major component of the power in SRAMs, especially at higher temperatures. Lowering supply voltage can help in reducing sub-threshold leakage. There is a limit to which the supply voltage can be reduced which is the retention voltage. The minimum voltage at which bit cell retains its data is known as the retention voltage. Sometimes we also say data retention voltage (DRV) instead of retention voltage.

2.7 Leakage

Static current flows in memories is responsible for power consumption, which has become a critical issue for designers. There are two components of leakage in CMOS circuits which can be understood by equation (2.1).

$$P_{Total} = P_{Dynamic} + P_{Static/Leakage} \tag{2.1}$$

$P_{Dynamic}$ consists of mostly switching power and short-circuit power. Dynamic switching power dissipation is caused by charging and discharging of capacitors of the circuit. Short circuit power dissipation occurs when both pull-up and pull-down networks are partially ON. Currents responsible for $P_{Static/Leakage}$ are known as leakage currents. Leakage currents are shown in Figure 2.5.

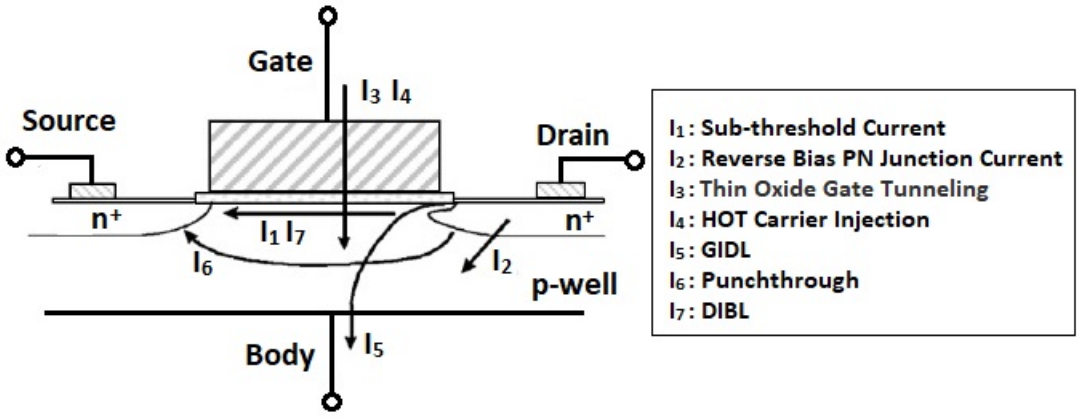

Figure 2.5: Leakage Currents in MOS [2]

2.7.1 Sub-threshold Current

When technology scales from one node to another node, to reduce dynamic power consumption supply voltage has also been scaled down. For same driving current capability or to maintain good overdrive voltage, threshold voltage should be reduced but reducing threshold voltage results in sub-threshold leakage current. When the applied gate to source voltage is less than the threshold voltage, weak inversion region forms in the channel which causes sub-threshold current between the source and drain. The I_{DS} constitutes the drift current and diffusion current. In strong inversion drift current is dominating due to a higher concentration of minority carriers in the channel while in weak inversion diffusion current is dominating because minority carrier concentration is very small and there is a small longitudinal electric field exist due to $|V_{DS}|$ and zero horizontal electric field [6].

$$I_{Sub-threshold} = \frac{W\mu C_{ox} V_T^2 e^{1.8}}{L} e^{\frac{(V_{GS}-V_{th})}{\eta V_T}} [1 - e^{\frac{-V_{DS}}{\eta V_T}}] \quad (2.2)$$

In short channel devices, drain interacts with source-channel junction by helping charges to pass through junction; this process reduces the threshold voltage of the device. This phenomenon is known as drain induced barrier lowering as the control over charge in the channel can also be controlled by a drain which was supposed to be controlled by the gate. This results in higher sub-threshold leakage in high V_{DD} conditions. When source and drain depletion region merge punch-through occurs.

2.7.2 Reverse Bias PN Junction Current

Source and drain junction of MOS device forms reverse bias PN junction. The minority charge carriers (holes in n side and electrons in p side) flow across the junction and this current is also responsible for leakage current.

2.7.3 Thin-oxide Gate Tunnelling

As technology is scaling down short channel effects are affecting the performance of the design. To control short channel effects, the thickness of oxide must decrease. This decrease in thickness gives rise to the high electric field which leads to direct tunneling current through gate of the transistor.

2.7.4 Hot Carrier Injection

When drain and gate are at the supply voltage, carriers pick up high energy from the electric field. As they move across channel they experience strong longitudinal as well as vertical electric field near $Si - SiO_2$ interface. Due to the electric field, they get attracted towards the gate. Their state is known as "hot". The higher the supply voltage, the hotter will be charge carrier. These "hot" carriers, when injected into gate oxide, shifts the threshold voltage of mos. These hot carriers are responsible for leakage current [7].

This injection is more likely to happen in case of electrons than holes because effective mass of electrons is less than holes. There is one more reason for this is that barrier height in bandgap diagram for electrons is 4.5eV less than that of holes [8].

2.7.5 Gate-induced drain leakage

(GIDL) is drained to substrate leakage current. The reason for this leakage current is very high field depletion region in drain gate overlap and flows from drain to substrate.

2.8 Summary

In this chapter we have discussed the SRAM array architecture in a broad sense and 6T SRAM bit-cell. We have also discussed about the figures of merit of 6T SRAM bit-cell which are cell current, bit-line leakage, write margin, write time, SNM, RNM and SRAM Leakage.

Chapter 3

Conventional Methods of Retention Voltage Generation

There are several methods which can be used for reducing rail to rail voltage. But it is important to understand the importance and parameters associated with leakage in 6T SRAM bit-cell. Section 3.1 explores the importance of reducing rail to rail voltage, such that it will be helpful in understanding the conventional solutions and proposed solution.

3.1 Importance Of Reducing Rail to Rail Voltage

If we closely look into the current equation in sub-threshold region equation (2.2), we can see that current has exponential dependency over V_{GS} and $|V_{DS}|$. Smaller the $|V_{DS}|$ smaller will be sub-threshold current. In Figure 3.1 SRAM memory is shown in the form of multiple bit-cells. During retention mode word line is kept at lower potential, BLT and BLF are pre-charged to supply. Assuming '1' is stored in bit-cell, we can see that $|V_{DS}|$ of transistors M5, M3 and M1 is zero so current flowing from these three devices will be negligible. $|V_{DS}|$ of transistors M2, M4 and M6 is not zero, so these devices will contribute towards leakage currents I1, I2 and I3.

Depending on technology and size of transistors of bit cell, we can reduce Rail

to Rail voltage by reducing V_{DD}, by raising VGND (source biasing), and by using both. Reducing V_{DD} reduces $|V_{DS}|$ of PU, PG, and PG transistors by lowering V_{DD} and similarly in raising V_{GND}. But in second case it also increases the source voltage of PD and PG transistors, and since their size is greater than PU transistor, they will contribute more leakage current, therefore we prefer to raise V_{GND} node instead of lowering supply voltage. Raising V_{GND} node will reduce $|V_{DS}|$ of M2, M4 and M6 and this will help in decreasing sub-threshold leakage from these devices.

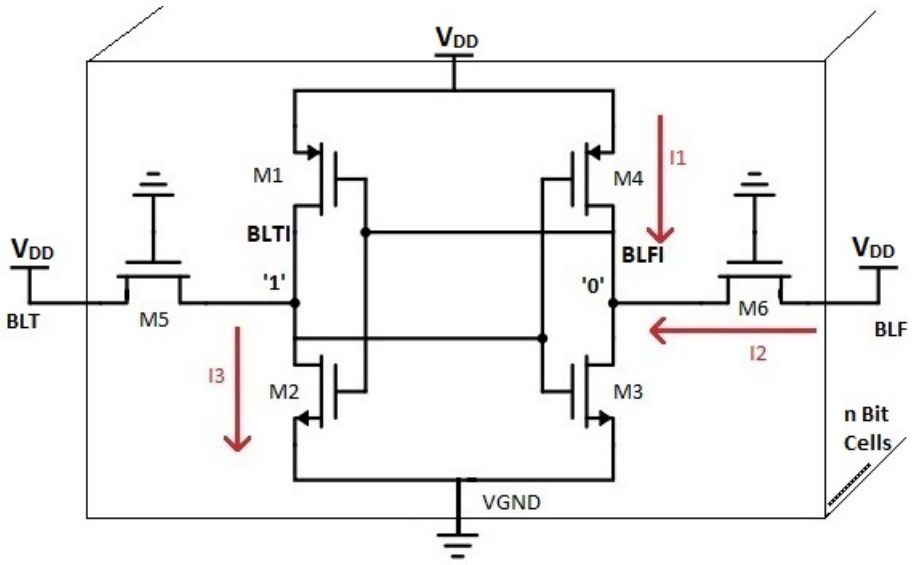

Figure 3.1: Leakage Components In 6T SRAM Bit Cell

3.2 Methods For Reducing Rail to Rail Voltage

In this section we will explore some solutions for reducing rail to rail voltage and problems associated with these solutions.

3.2.1 Reducing Supply Voltage of SRAM Array to Retention Voltage

One of the solutions is that if we can reduce the supply voltage provided by LDO voltage regulator to SRAM array. The output of voltage regulator is not only given to SRAM array, it is also given to other blocks of SOC, which means reducing the supply voltage directly can alter the performance of other blocks, which may not be desirable.

3.2.2 Diode Connected MOS for Reducing Rail to Rail Voltage

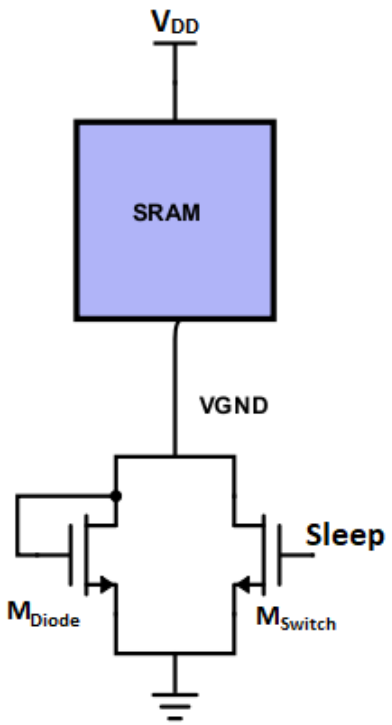

Figure 3.2: Conventional Diode Connected Architecture

Conventionally diode Connected MOS is used to reduce the rail to rail voltage of SRAM array [9]. We can see in Figure 3.2 that there are two transistors M_{Diode} and M_{Switch} connected in parallel at the footer. When SW is high, M_{Switch} turns on and it transfers ground to V_{GND} node and memory can be used for normal operations. When

SW is low M_{Switch} turns off, then a diode connected MOS M_{Diode} will adjust its $|V_{DS}|$ depending on current from SRAM array and V_{GND} will be equal to $|V_{DS}|$. This solution is simple and only (SW) switching power is consumed. V_{GND} raises gradually by leakage current. The importance of ground raising or source biasing is already explained in section 3.1, so diode connected at footer is preferred.

The disadvantage of using this architecture is that the voltage which appears at virtual ground (V_{GND}) is PVT dependent and therefore leakage gain is not consistent across PVTs. If the supply voltage is V_{DD} and $|V_{DS}|$ is drain to source voltage of diode connected MOS M_{Diode} used for retention mode, then V_{RET} should follow equation (3.1).

$$V_{RET} \geqslant V_{DD} - |V_{DS}| \tag{3.1}$$

If $|V_{DS}|$ of MOS is a larger value then instead of operating MOS in strong inversion region, it can be operated in the sub-threshold region by increasing its size such that it should satisfy equation (3.1).

Table 3.1 shows variations in virtual ground voltage (V_{GND}) for conventional approach with respect to all process and temperature. V_{GND} voltage is designed such that it should not violate equation (3.1). Let's take an example to understand this table, voltage at SNFP process at 140℃ temperature is 304mV, which violates the equation (3.1). This voltage can be reduce by increasing the size of diode connected MOS at the cost of reduced V_{GND} at other corners. For example at FNSP 25℃ the value of V_{GND} is only 156mV and at FNFP 25℃ it is 195mV, so in corners where leakage is expected to be high, the diode connected architecture limits to utilize the complete benefit.

	-40℃	25℃	100℃	140℃
FNFP	224	195	232	273
FNSP	198	156	171	195
SNFP	255	237	274	304
SNSP	233	185	202	226
TT	229	190	216	248

Table 3.1: V_{GND} When Conventional Method Is Used

3.2.3 Programmable Bias Transistor to Control (V_{GND})

We have seen that how V_{GND} is process and temperature dependent in diode connected retention scheme as shown in Figure 3.3. Another approach is presented in [3], in which instead of using a single transistor, four transistors are used at footer and they can be controlled by programming to ensure a required level of virtual ground voltage. The transistors are sized in a binary weighted sequence. Hence there can be 2^N virtual ground voltages. Once the process is known after the fabrication, one can program transistors for required virtual ground voltage. So, while this solution adapts to process changes, it is still dependent on temperature.

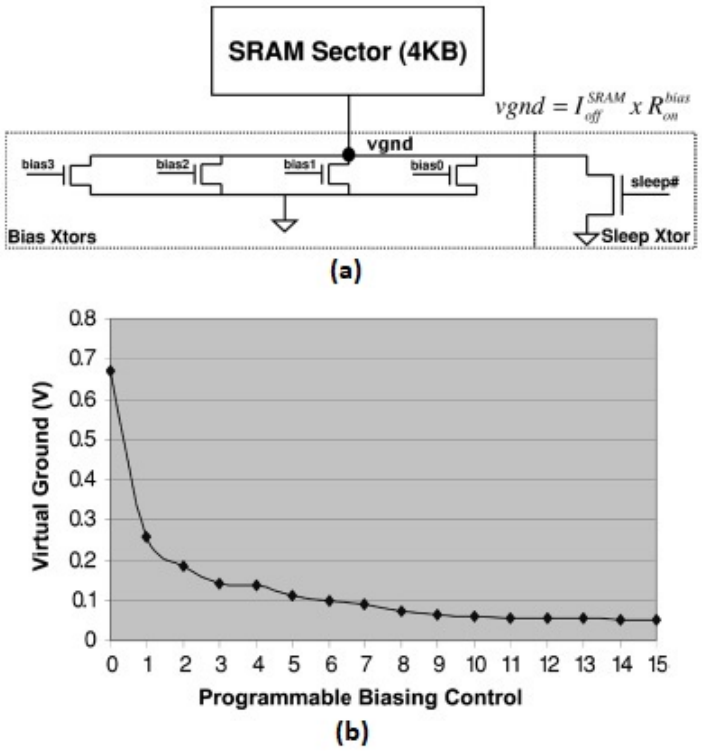

Figure 3.3: (a) SRAM sleep transistor with the programmable bias transistors to control the virtual-ground. (b) Virtual-ground voltages with respect to 4-bit bias settings [3]

3.3 Summary

In this chapter we have discussed the importance of reducing rail to rail voltage and analyzed the key components of leakage in 6T SRAM bit cell. We have also discussed the importance of raising the virtual voltage instead of reducing supply voltage of array. After understanding the importance of reducing rail to rail voltage we have discussed some architecture which can be used for reducing rail to rail voltage so that we can reduce static power consumption from a 6T bit cell. We have also discussed limitations of these solutions such as PVT sensitive nature of diode connected architecture, temperature sensitivity of programmable bias transistor base architecture.

Chapter 4

Active Control of Rail to Rail Voltage

In the last chapter we discussed solutions to reduce array leakage and their limitations. In this chapter, we will discuss the proposed solution. Our main aim is to reduce power consumption of memories by reducing rail to rail voltage. The reason for reducing rail to rail voltage is to reduce sub-threshold leakage which has been discussed in earlier chapter. If we can control the gate of MOS with the help of feedback from V_{GND} with the help of an error amplifier. This configuration resembles a voltage regulator and to understand the concept behind controlling the gate of MOS to maintain V_{GND} voltage lavel we will have to understand the basic fundamental of voltage regulator. The Design of Voltage Regulator, its functionality and its figures of merit are discussed in the following sections.

4.1 Low Dropout (LDO) Voltage Regulator

LDO regulators are in a class of linear regulators. These are used to provide a supply voltage to ICs which is independent of any variations in PVT, load, and change in current. There are other stable supplies which can be used to regulate voltage such as Digital regulator, switching regulator, and DC-DC converters. With respect to our application, these circuits are complex and consume a large amount of power. The main difference between a linear regulator and LDO is of the pass transistor. In LDO pass transistor is used in common source configuration (Common Emitter in BJT) while in linear regulators

it is in source follower (Emitter follower in BJT) configuration. Conventional LDO can be seen in Figure 4.1 below [10].

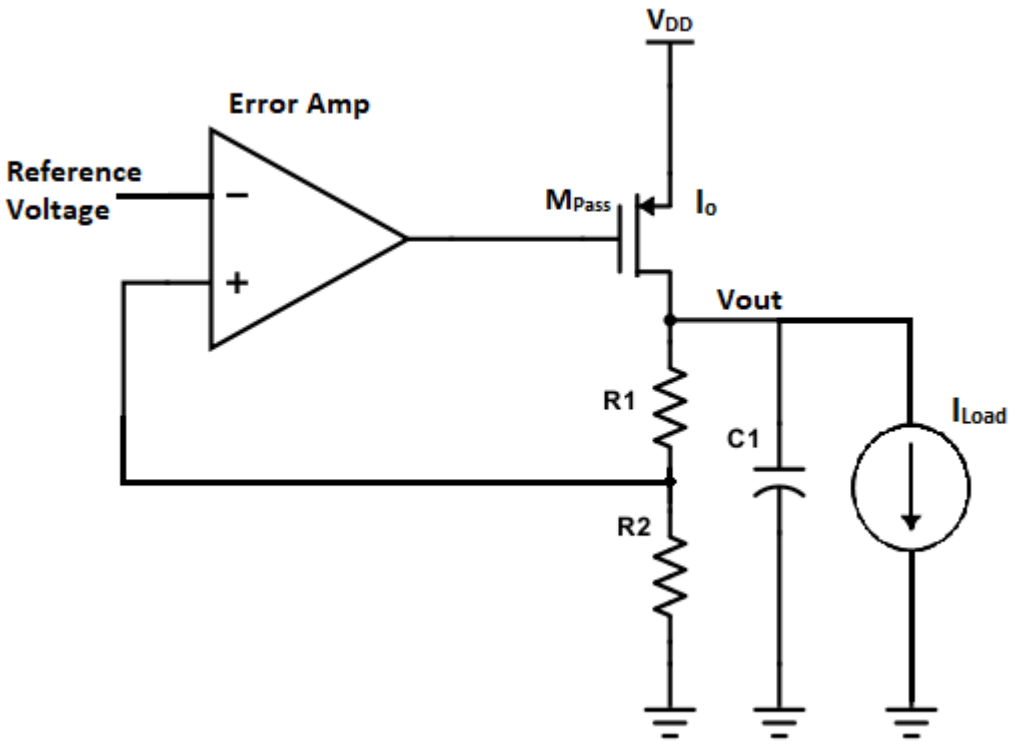

Figure 4.1: LDO Voltage Regulator

LDO consists of an error amplifier, pass transistor, and feedback network. The reference voltage can be taken from bandgap reference circuit which is PVT insensitive. The difference between bandgap circuit and LDO is, bandgap provides constant voltage irrespective of any variations at a very low current while LDO on other hand provides constant supply irrespective of the current sink by the load. In earlier technology BJT pass transistors were used which has been replaced by CMOS transistors. The reason that base current has dependency over emitter current and this can be in the order of few microamperes while in CMOS technology gate current is in the order of femtoampere which is negligible. Base current in BJT can be reduced by cascading BJTs but this implementation increases dropout voltage.

4.2 Functionality Of LDO and Its Figures of Merti

Error amplifier amplifies the difference between the reference voltage and scaled output voltage using resistor feedback. The amplified output then regulates the gate of pass transistor so that V_{OUT} can be adjusted to the required voltage. The pass transistor acts like a voltage-dependent current source which supplies load current. There can be situations when current can increase from very small value to a very large value; in this case, there will be larger dropout at pass transistor due to the limited bandwidth of error amplifier and it will take some time to change the gate voltage of pass transistor so to increase V_{OUT}. The decrease in V_{OUT} can be as large such that it can lead to malfunction in circuits where V_{OUT} is being supplied. Due to this reason, a large capacitor is used in parallel to V_{OUT} which supplies stored charge during such events and reduces output voltage of the LDO. Figures of merit of LDO are as follows.

4.2.1 Dropout Voltage

It is defined as the minimum voltage drop across the regulator where the device can no longer regulate the output voltage. If the difference between input supply voltage and output voltage becomes less than a certain value such that pass transistor operates in triode region; at this point output voltage cannot be regulated by the amplifier.

$$V_{Dropout} = I_{Load} \times R_{On} \tag{4.1}$$

R_{On} includes the output resistance of pass transistor, interconnects etc. For higher efficiency dropout voltage should be as low as possible.

4.2.2 Quiescent Current

Quiescent current or ground current is the combination of currents associated with reference voltage generator (BGR), sampling resistor network, biasing current of the error amplifier, the gate current of pass element and current which do not contribute to the

output current. In order to have good current efficiency, there should be low quiescent current.

$$I_Q = I@NoLoad \tag{4.2}$$

4.2.3 Efficiency

The efficiency of LDO depends on the quiescent current, output current, input supply voltage and output voltage. For high efficiency, the quiescent current and the drop out voltage must be small.

$$Efficiency = \frac{I_o V_o}{(I_o + I_Q)V_{DD}} \times 100 \tag{4.3}$$

4.2.4 Regulation

The DC Regulation performance of regulator is determined by line and load regulation. The line regulation defines the ability of LDO to maintain the specified voltage with varying input voltage.

$$LineRegulation = \frac{\Delta V_{Out}}{\Delta V_{In}} \tag{4.4}$$

While load regulation on the other hand defines the ability of regulator to maintain same output voltage while variation in load condition.

$$LoadRegulation = \frac{\Delta V_{Out}}{\Delta I_o} \tag{4.5}$$

4.2.5 PSRR

Power supply rejection ratio is the regulator's ability to prevent any ripples due to supply voltage. We can also define it as, ripples from supply should not be reflected in the regulated output or we can say the LDO should have the ability to suppress ripples from supply.

$$PSRR = 20 \log \frac{A_V(V_{DD} = 0)}{A_{DD}(V_{in} = 0)} \tag{4.6}$$

Where A_V is the gain from input to output when supply variations are zero while A_{DD} is the gain from supply to output when input is zero.

4.3 Proposed Solution

In Figure 3.2, instead of M_{Diode}, we propose to use a MOS M_{Pass}, in parallel with the M_{Switch}. We propose to control the gate of M_{Pass} from a low power Error amplifier, so that V_{GND} can be raised to target voltage (V_{Target}) irrespective of the and process and temperature. The proposed solution is shown in Figure 4.2. Since this configuration resembles to a voltage regulator which we have discussed in detail. Transistor M_{Switch} in parallel with M_{Pass} transistor will drive SRAM array either in retention mode or fully functional mode by SW signal. If supply voltage for a given technology is V_{DD} and retention voltage is V_{RET}, then condition for virtual ground is given by equation (4.7).

$$V_{GND} = V_{DD} - V_{RET} \tag{4.7}$$

A guard band of around 40mV is considered for any noise or supply fluctuations; so, instead of targeting virtual ground voltage to (V_{DD} - V_{RET}), we target 40mV below (V_{DD} - V_{RET}), we call it V_{Target}.

$$V_{Target} = V_{DD} - V_{RET} - 40mV \tag{4.8}$$

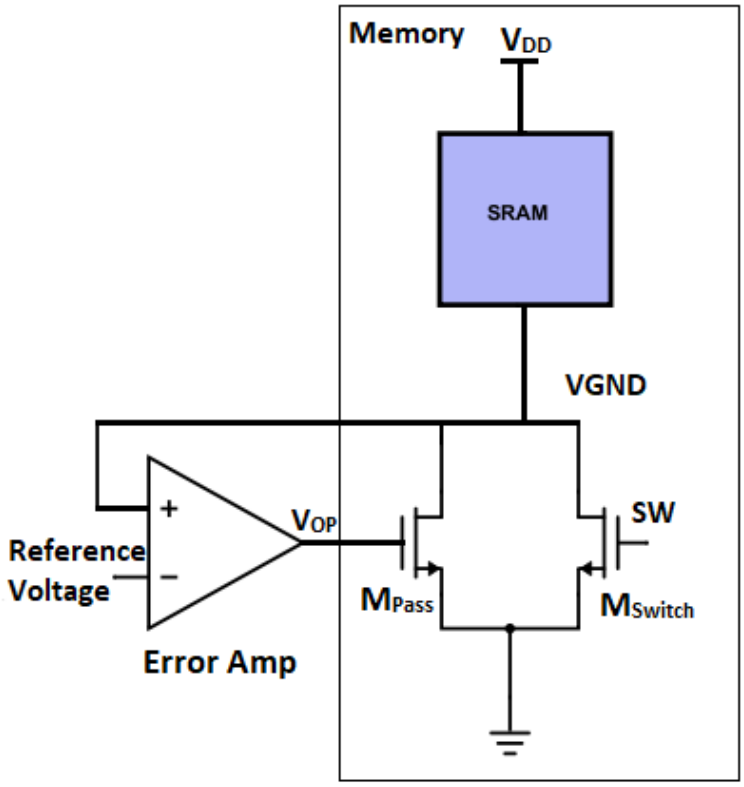

Figure 4.2: Proposed Architecture To Reduce Leakage Current

From equation (4.8), it is evident that V_{Target} should not be independent of V_{DD} to achieve desired leakage gain across PVT. So the choice of 'Reference Voltage' in Figure 4.2 has to be made carefully.

To identify the right reference voltage let us consider an example. If we take reference voltage from a bandgap reference then there can be consequences, which can be understood by the following example. Suppose the supply voltage is V_{DD} and there is a $\pm 5\%$ variation in supply voltage. When supply will be less than V_{DD} then supply to virtual ground voltage should be greater than or equal to retention voltage of memories. Since we want the reference voltage to be process and temperature independent but voltage-dependent, the reference voltage from BGR cannot be used. We could have used a resistor divider to implement voltage dependent reference voltage but due to larger area consumption of resistors we prefer to use MOS. Therefore, we designed a voltage-dependent reference voltage using stacked diode shown in Figure 4.3. Supply voltage of stacked diode should be same as SRAM array. This diode connected PMOS configuration acts as a voltage divider. Current consumption from this stack is less than 1nA to

minimize power overhead of our solution. Bulk and source are shorted in order to avoid V_{th} modulation.

The error amplifier is designed to operate in a sub-threshold saturation region and total power consumption of error amplifier is 81nW, which is negligible in comparison to the SRAM array leakage. V_{GND} for 40nm low voltage process can be at max 300mV. Since we have taken 40mV guard band therefore V_{Target} is 260mV. To get rid of the area and current consumed by resistor divider in Figure 4.1, feedback is directly given to error amplifier. As we have removed resistor divider, the reference voltage should be equal to V_{Target}. Since memories will be in idle mode there will never be a sudden change in current and load capacitor of memory is in the order on nanofarads so we have not used any capacitor in parallel to $|V_{GND}|$.

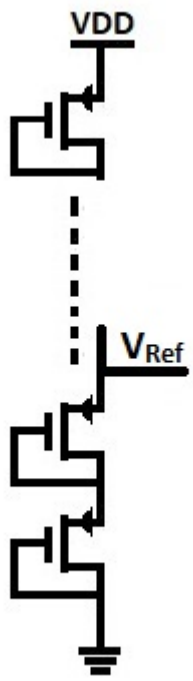

Figure 4.3: Reference Voltage Generation Circuit

There are two closed loop poles in proposed solution in Figure 4.2; one is at V_{OP} node (at a higher frequency) and the other one is at V_{GND} node. Pole associated with V_{GND} node depends on leakage current of the memory which is dependent on temperature. For low temperature, the current is very small and effective time constant is very large, so this pole is at a lower frequency. For higher temperature, the current is very large

and effective time constant is small in comparison to pole at low temperature. To ensure the stability of closed-loop system at the higher temperature we propose to introduce a feedback using $M_{Feedback}$ transistor as shown in Figure 4.4. This feedback is voltage current feedback; it reduces the effective output resistance of Error amplifier and shifts the pole at a higher frequency.

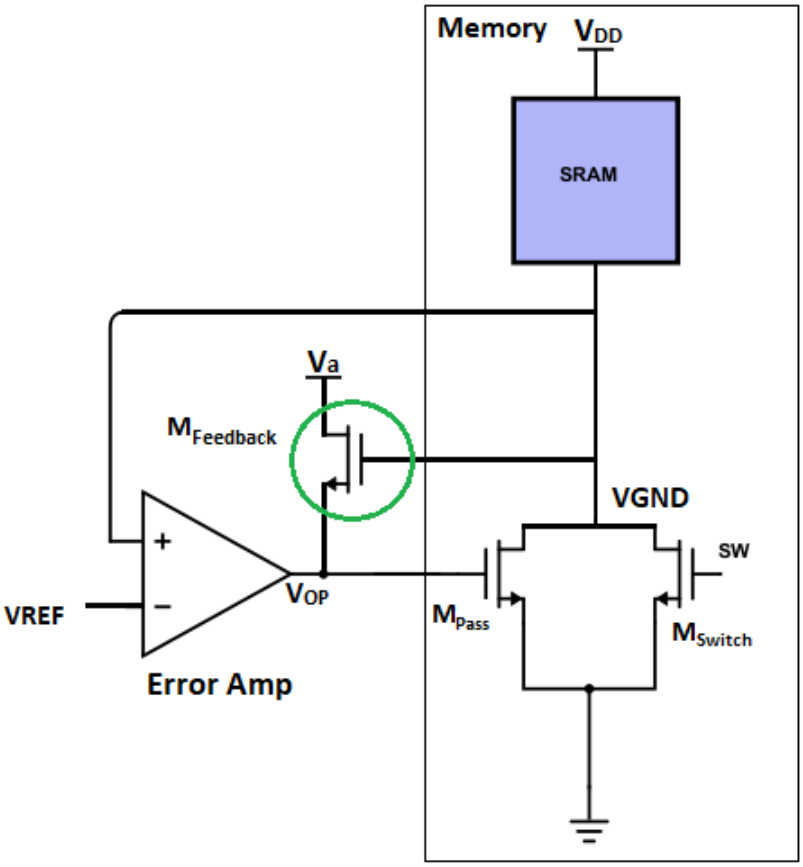

Figure 4.4: Modified Architecture of Voltage Regulator to Ensure Stability at Higher Temperature

4.4 Summary

In this chapter, we have discussed the proposed solution. This proposed solution resembles to an LDO voltage regulator, so we also discussed the functionality and figures of merit of LDO voltage regulator. We discussed the need for voltage dependent reference

voltage which should be process and temperature independent. We have also discussed how closed loop system could be unstable at a higher temperature, so we proposed a modified architecture which ensures the stability at a higher temperature.

Chapter 5

Results

In this chapter, simulation results are discussed. Schematics are designed in Cadence Virtuoso and simulated using Eldo and Spectre simulators in 40nm low power process. Characterization has been done for all corners which are FNFP, FNSP, SNFP, SNSP, and TT. Temperature range has been taken from -40 to 140 C.

5.1 AC Closed-Loop Analysis of Memory Subsystem

AC loop stability analysis is done to ensure that the closed loop operation of the system is stable. According to control system theory, the phase margin should be greater than 45 degree at UGB frequency to avoid the ringing effect. To ensure no ringing in the system, phase margin should be greater than equal to 60 degree. For AC analysis, feedback of loop is broken in such a way that AC signal should not see any loop and DC circuit is shorted. Phase margin at TT is 84.26 degree shown in Figure 5.1, for a worst corner (SNFP), it is 56.6 degree shown in Figure 5.2.

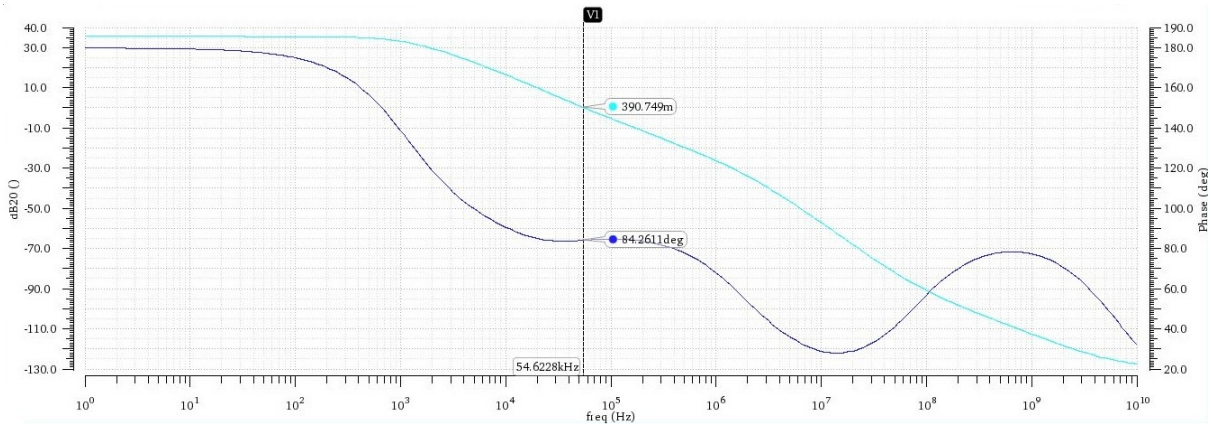

Figure 5.1: Close Loop Stability of SRAM Array During Retention Mode @TT (Best Case)

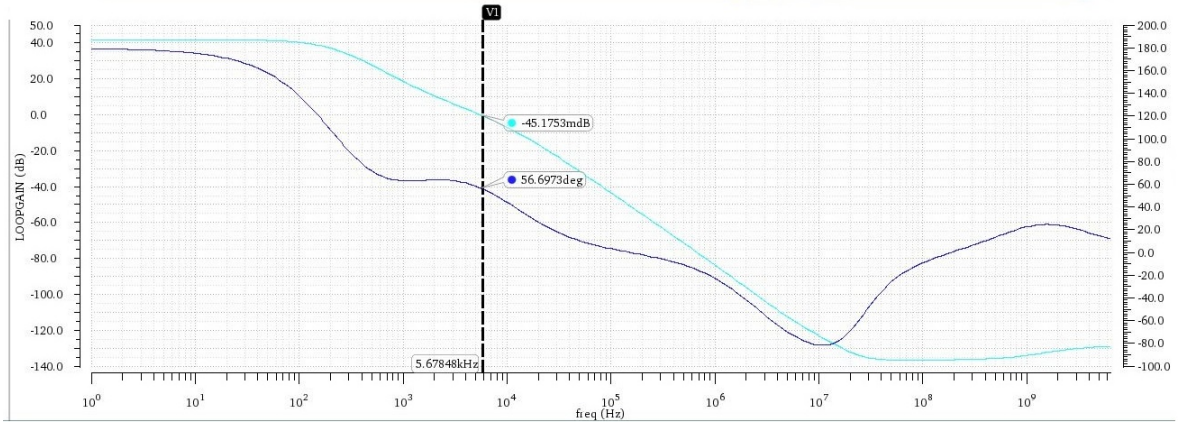

Figure 5.2: Close Loop Stability of SRAM Array During Retention Mode @SNFP (Worst Case)

5.2 Comparision of V_{GND} voltage between Diode connected and Proposed Architecture

Figure 5.3 shows V_{GND} for diode connected architecture (Figure 3.2). V_{GND} voltage is designed such that it should not violate equation (3.1) for all PVTs. The value of V_{Target} is equal to 257mV close to 260mV discussed in section 4.3. We can see that when we designed V_{GND} for V_{Target} at SNFP process at 140°C, then at FNSP (25°C) we are losing 144mV and at TT (25°C) we are losing 109mV similarly we are losing voltage at other PVTs. This limits us to utilize the complete benefit of this solution.

Figure 5.4 shows V_{GND} for proposed architecture (Figure 4.4) and it can be

seen clearly that due to negative feedback voltage at V_{GND} is almost constant irrespective of PVT variations. The maximum voltage we are losing is 24mV at FNSP (140℃).

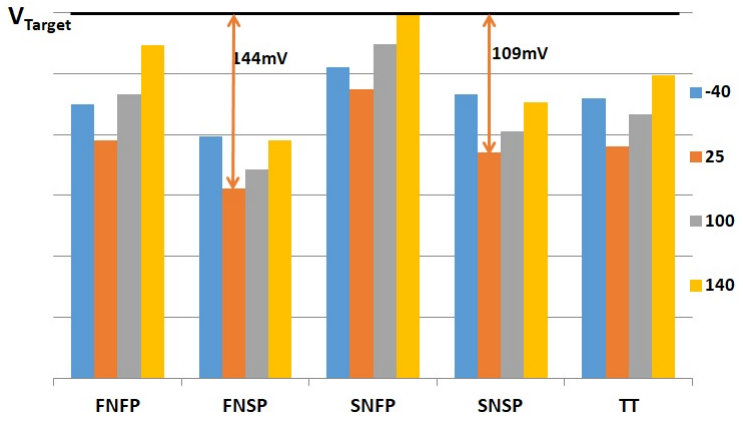

Figure 5.3: V_{GND}(mV) For Diode Connected Architecture (Figure 3.2)

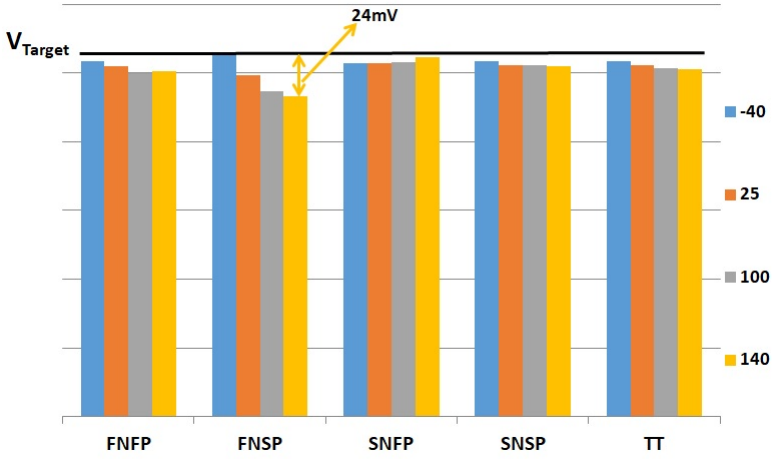

Figure 5.4: V_{GND}(mV) For Modified Proposed Architecture (Figure 4.4)

5.3 Transient Analysis of Memory Subsystem

Transient analysis is done to analyze the timing response of the system when the system is entering in retention mode. The analysis is done at -40℃ 25℃ 100℃ and 140℃ temperature values to validate the response across temperature and ensure stability of the system. Figure 5.5 shows the transient response of the system when feedback transistor

is not used, ringing behavior at higher temperature can be seen at the time when the system is entering in retention mode. Figure 5.6 shows the transient behavior of the system when feedback is used as proposed in Figure 4.2. It is clear from this Figure that the addition of $M_{Feedback}$ has improved the stability across temperature. In Figure 5.6, when 5% supply droop introduced at around 30msec, V_{GND} followed the supply droop. It has to be ensured that system doesn't violate equation (3.1) even in such condition. So V_{Target} has to be set carefully considering all voltages of operation.

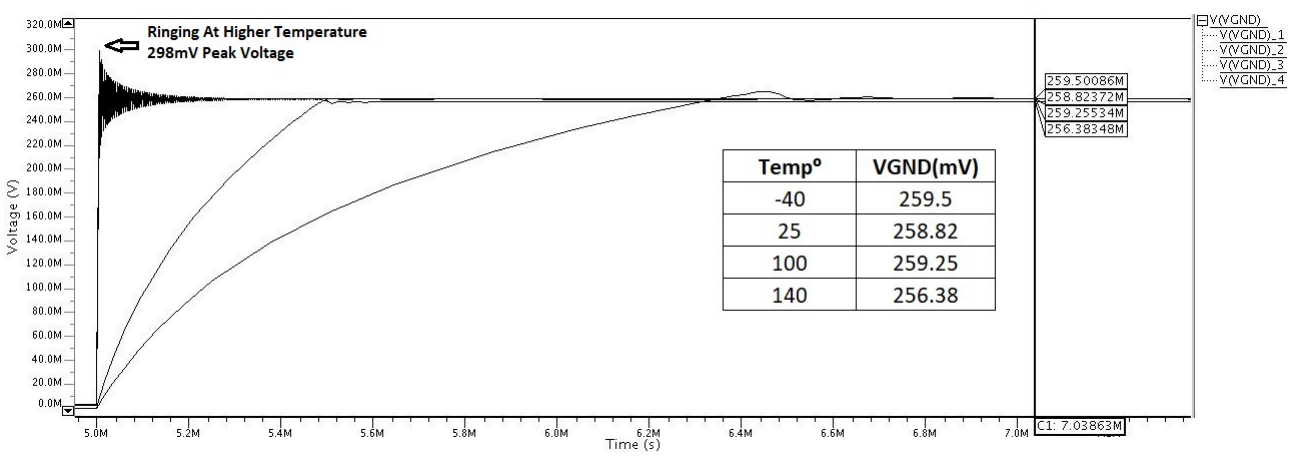

Figure 5.5: Transient Analysis When Feedback Is Not Used.

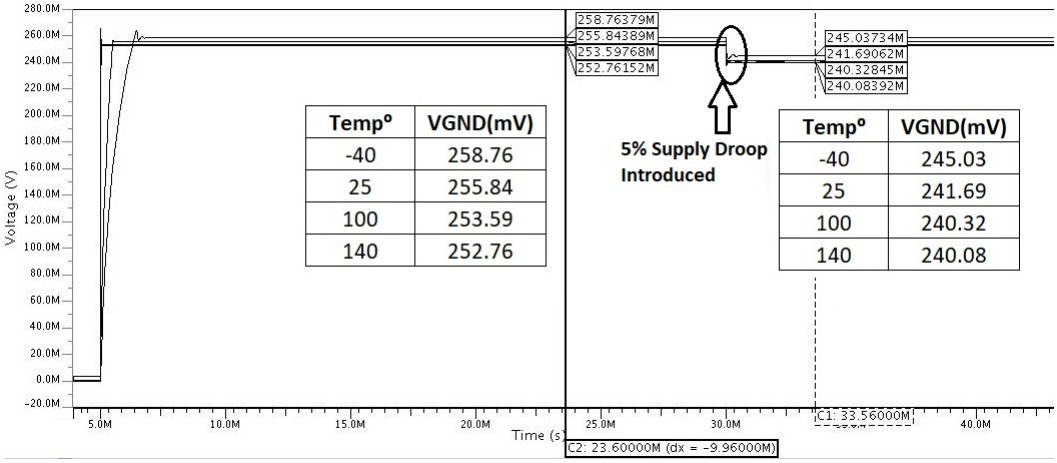

Figure 5.6: Supply Variation Analysis During Retention Mode

5.4 Leakage

Figure 5.7 and Figure 5.8 show the comparison of leakage reduced at temperature 25℃ and 140℃ respectively for diode connected and proposed architecture. Leakage of the error amplifier is also considered while calculating leakage of the system. It can be seen that at TT 25℃ using diode connected architecture leakage reduced by only 25.4% while using proposed architecture leakage reduced by 50.8%. This is approximate 33% reduction wrt conventional solution. Gains are highest at FNSP lot. Similarly in all the process' we can see improvements in leakage. Leakage at SNFP 140℃ is same in both architectures because voltage at V_{GND} tuned to V_{Target} at this lot and temperature condition. Figure 5.7 shows the comparison of leakage in conventional architecture and proposed architecture at 25℃Figure 5.8 shows the same comparison at 140℃

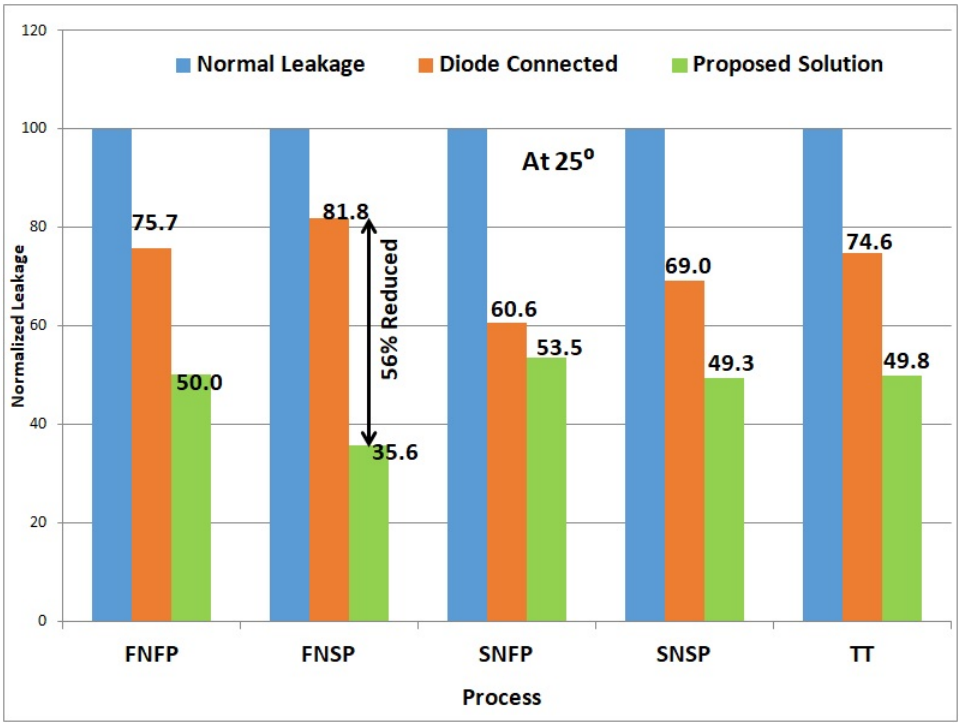

Figure 5.7: Leakage Comparison between Diode connected and Proposed Architecture at 25℃

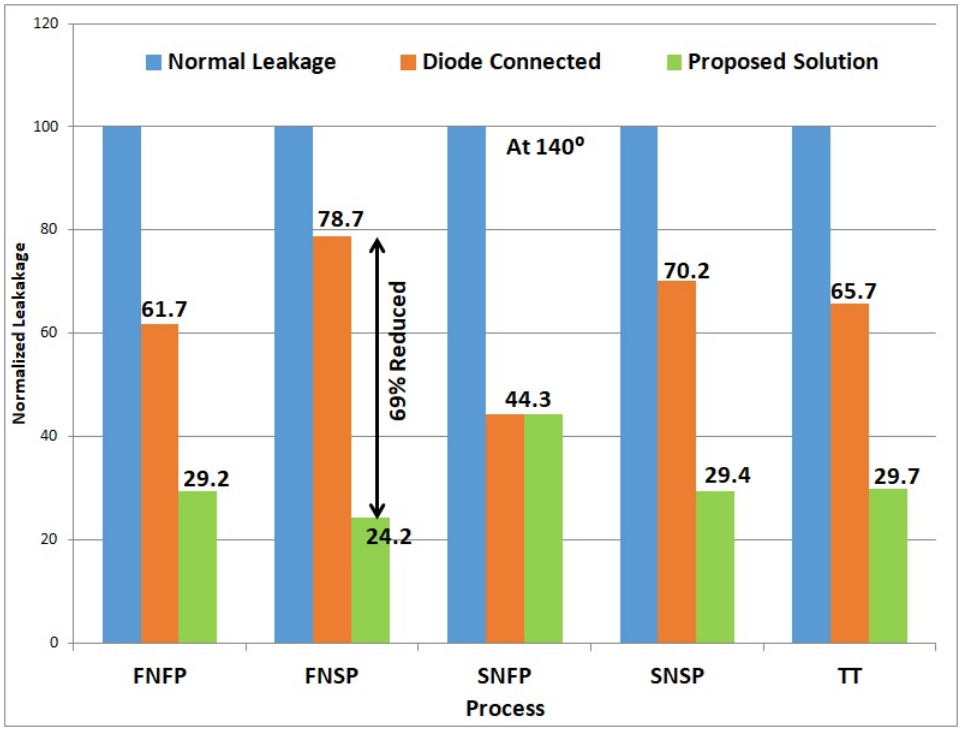

Figure 5.8: Leakage Comparison between Diode connected and Proposed Architecture at 140°C

5.5 Monte Carlo

To ensure that V_{GND} never violates equation (3.1), we introduced mismatch and did Monte Carlo analysis on the system. Figure 5.9 shows the Monte Carlo simulation result of V_{GND} voltage for 1000 samples. Sigma for V_{GND} is 4.047mV. Of this 1.68mV is contributed by reference voltage circuit and rest is contributed by the offset of Error amplifier. 3σ (12.14mV) value is well within the range of guard band considered in equation (4.8). The remaining portion of the guard band can be used for supply noise and other disturbances.

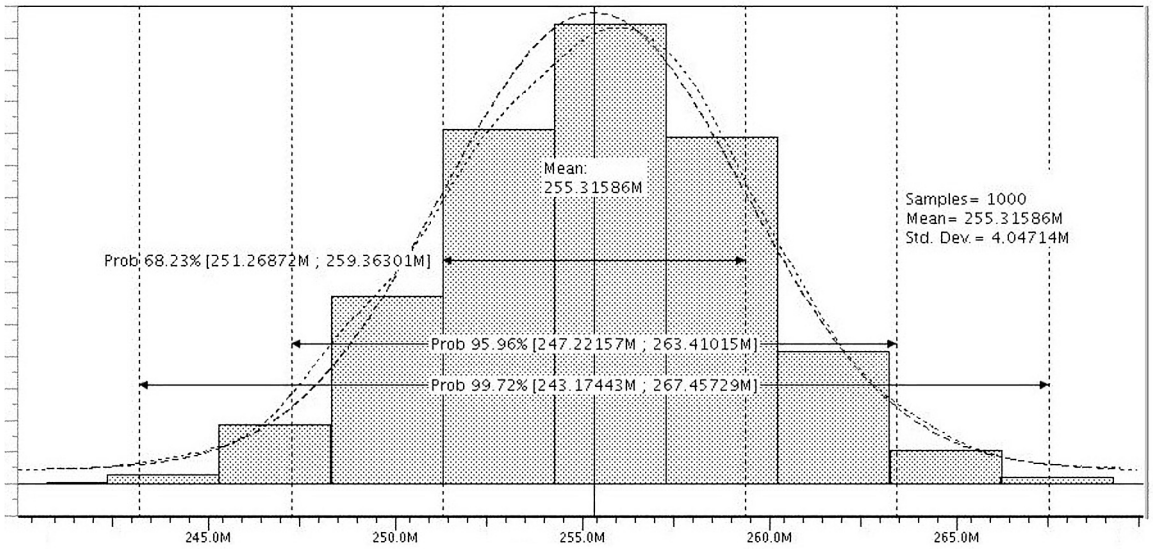

Figure 5.9: Monte Carlo Analysis of V_{GND}

Chapter 6

Conclusion and Future Work

In this dissertation, we have designed a low power error amplifier to maximize leakage gain with ensuring that data retention voltage is maintained constant at the array across all the PVTs. The amplifier consumes 81nW of power and the overall memory subsystem leakage power reduces by 50% from no retention case and 33% from the conventional retention solution at TT (25℃). Leakage gains are an additional 56% over the conventional design at FNSP lot at 25℃ and 69% at 140℃ Leakage gains are more significant at higher temperatures. Mismatch analysis on the system indicates that system is robust.

Retention voltage also depends on the process. We target retention voltage for worst process. In future we can work on process dependent reference voltage generation. One can also work to reduce V_{GND} by improving mismatches and the offset from error amplifier.

Bibliography

[1] N. H. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective*. Pearson Education India, 2015.

[2] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.

[3] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "Sram design on 65-nm cmos technology with dynamic sleep transistor for leakage reduction," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 895–901, 2005.

[4] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino, and S. Iwade, "A 90-nm low-power 32-kb embedded sram with gate leakage suppression circuit for mobile applications," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 4, pp. 684–693, 2004.

[5] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits*. Prentice hall Englewood Cliffs, 2002, vol. 2.

[6] P. F. Butzen and R. P. Ribas, "Leakage current in sub-micrometer cmos gates," *Universidade Federal do Rio Grande do Sul*, pp. 1–28, 2006.

[7] H. Sasaki, M. Saitoh, and K. Hashimoto, "Hot-carrier induced drain leakage current in n-channel mosfet," in *Electron Devices Meeting, 1987 International*. IEEE, 1987, pp. 726–729.

[8] T. Yuan and T. H. Ning, "Fundamentals of modern vlsi devices," *Cambridge, New York*, pp. 149–58, 1998.

[9] A. J. Bhavnagarwala, S. V. Kosonocky, M. Immediato, D. Knebel, and A.-M. Haen, "A pico-joule class, 1 ghz, 32 kbyte/spl times/64 b dsp sram with self reverse bias," in *VLSI Circuits, 2003. Digest of Technical Papers. 2003 Symposium on.* IEEE, 2003, pp. 251–252.

[10] V. Shirmohammadli, A. Saberkari, H. Martínez-García, and E. Alarcón-Cot, "Enhancing the performance of output-capacitorless ldo regulators by pass-transistor segmentation," in *Circuits and Systems (ISCAS), 2016 IEEE International Symposium on.* IEEE, 2016, pp. 490–493.