LOGIC-IN-MEMORY COMPUTING USING RRAM BASED NV-SRAM

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of
Master of Technology

by

VARUN BHATNAGAR

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
JUNE 2022

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **LOGIC-IN-MEMORY COMPUTING USING RRAM BASED NV-SRAM** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF TECHNOLOGY** and submitted in the **DEPARTMENT OF ELECTRICAL ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from August 2020 to June 2022 under the supervision of **Prof. Santosh Kumar Vishvakarma, Professor, Department of Electrical Engineering, Indian Institute of Technology Indore.**

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Varun 1/6/22

Signature of the student with date

VARUN BHATNAGAR

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

09/06/2022

Signature of the Supervisor of M.Tech. thesis (with date)

Prof. Santosh Kumar Vishvakarma

Mr. Varun Bhatnagar has successfully given his M.Tech. Oral Examination held on **June 7th, 2022**.

Signature of Supervisor of M.Tech. thesis
Date: 09/06/2022

Convener, DPGC
Date:09/06/2022

Signature of PSPC Member #1
Date: 09.06.2022

Signature of PSPC Member #2
Date: 9/06/2022

ACKNOWLEDGEMENTS

I would like to take this moment to convey my appreciation to everyone who helped make this period learnable, enjoyable, and pleasant. First and foremost, I'd want to express my gratitude to my supervisor, **Prof. Santosh Kumar Vishvakarma**, who was a continual source of inspiration during my research. This research activity was conducted with his ongoing mentoring and research suggestions. His unwavering support and encouragement have inspired me to stay focused on my studies.

I am grateful to **Dr.-Ing. Marc Reichenbach** for providing me with the chance to work on the NV-SRAM project which also aided me in conducting extensive study and learning about a bunch of new topics. I would also like to thank **Prof. Mukesh Kumar** and **Dr. Anirban Sengupta**, members of my research progress committee, for taking the time to assess my progress throughout the semesters. Their insightful feedback and ideas assisted me in improving my work at various levels.

I would want to express my gratitude to Indian Institute of Technology Indore for allowing me to test my research talents. My sincere acknowledgement to all members of the Nanoscale Devices, VLSI Circuit System Design Lab (NSDCS) research group, particularly **Mr. Gopal Raut**, **Mr. Narendra Singh Dhakad**, and **Ms. Sumiran Mehra**, for their discussions and assistance throughout my thesis study. I'd want to convey my heartfelt gratitude for their unwavering love, care, and support throughout my life.

Varun Bhatnagar

This Thesis is Dedicated

to

The Almighty GOD,
My Family, and Friends

ABSTRACT

Semiconductor Industries are dealing with the continual influx of accessible data and the never-ending requirements for better and quicker performance to maintain a competitive advantage and fulfil today's needs for an ideal user experience.

In-Memory Computing(IMC) is gaining traction as a result of this. Because IMC is more about how much data can be absorbed and evaluated in a short amount of time. Incorporating alternate memory technologies, such as nonvolatile memory, is an important trend for IMC since it allows for multilayer storage, higher density, and low-power duty-cycle operation. Static Random Access Memory (SRAM), Resistive RAM (RRAM), and phase-change memory(PCM) are examples of trending memory technologies. RRAM-based architectures are proving themselves as a well-established storage method.Because of its simple structure, compatibility with existing CMOS technology, high switching speed, and ability to scale to the tiniest dimensions, RRAM is one of the most intriguing memory technologies.

In this thesis Logic-In-Memory designs have been proposed using IHP 130nm RRAM technology. Combinational circuits like NOR, NAND, XOR have been successfully simulated and optimized in terms of operating power and speed of operation. These designs are hybrid as it is incorporated with 130nm CMOS transistor technology to achieve better functionality. Further, Non-Volatile SRAM(NV-SRAM) cell is proposed which is a combination of 6T-SRAM and RRAM to incorporate non-volatility feature. Using this 8T1R NV-SRAM, a 4-cell array has been designed with all the peripherals such as Sense Amplifier, Decoder, Bitline Driver and others.

Lastly, A PMOS-based voltage reference generator is proposed which makes the reference generation stage less susceptible towards small variations in incoming voltage (in the order of uV) as the Loading Effect (LE) grows at the input of the comparator stage in ADC in any memory(RRAM) array. The circuit stability is evaluated for process variation and device mismatch. Further, sleep mode is applied using the power-gating (PG) technique to minimize power dissipation.

Contents

Abstract **i**

List of Figures **vi**

List of Tables **ix**

List of Abbreviations **xi**

1 Introduction and Related Work **1**

1.1 Overview 1

1.2 Memristors 2

1.2.1 Resistance Switching Modes 5

1.3 Applications of RRAM 7

1.3.1 Non-volatile Logic 7

1.3.2 Neuromorphic Computing 8

1.3.3 Security Application 9

1.3.4 Non-volatile SRAM 9

1.4 Organization of Thesis 10

2 Simulation Methodology **11**

2.1 SG13 Design Kit Model 12

2.2 Circuit Entry and Simulations 13

2.2.1 ADE L Simulation 14

2.2.2 Corner Simulation 14

2.2.3 Monte-Carlo and Parametric Simulation 15

3 Logic-In-Memory Implementation using IHP-130nm RRAM Technology 17

3.1 Non-Volatile In-Memory Computing Architecture 17

3.2 Combinational Logic Design using IHP 130nm RRAM. 19

3.2.1 NOR Logic using RRAM 19

3.2.2 NAND Logic using RRAM 20

3.2.3 XOR Logic using RRAM 20

3.3 Summary 21

4 Non-Volatile SRAM 23

4.1 Conventional 6T-SRAM 23

4.2 Major Design Challenges of Conventional SRAM 25

4.2.1 Inter-Die (Global) Variations 25

4.2.2 Intra-Die (Local/Mismatch) Variations 26

4.2.3 Read/Write Conflict 26

4.2.4 Temperature Dependence of Read and Write Ability 27

4.2.5 Trade-off between Performance and Stability 27

4.2.6 Soft-Error Issue 28

4.3 Low-Power SRAM Cell Design 28

4.3.1 Use of Multi-V_{TH} Device 28

4.3.2 Use of Stacking Effect 29

4.3.3 Supply voltage scaling 30

4.3.4 Single-ended approach 31

4.4 Low-power SRAM memory array 32

4.4.1 Sense Amplifier 32

4.4.2 2-bit D Latch 33

4.4.3 1X2 Decoder 34

4.4.4 Bitline Driver 35

4.4.5 Write Switch 35

4.4.6 Equalizing Circuit 36
4.5 Non-Volatile SRAM 37
4.5.1 8T1R NVSRAM memory cell 38
4.6 Results and Discussion 39

5 Loading Effect Free MOS-only Voltage Reference Ladder 43
5.1 Introduction 44
5.2 Related Work and Motivation 47
5.3 Robust Voltage Reference Generation Circuit in RRAM 49
5.3.1 Parallel PMOS transistors Analysis in reference ladder circuit 50
5.3.2 Pareto analysis for finding parallel MOSs in reference circuit . 52
5.3.3 Small signal analysis of the two parallel MOS in node voltage 54
5.3.4 Low power and robust voltage reference ladder circuit 56
5.4 Simulation Results and Discussion 57
5.4.1 Process-Voltage-Temperature (PVT) variations impact 57
5.4.2 Monte-Carlo for Process-Variation and Mismatch Analysis . . 59
5.4.3 Physical performance parameters evaluation of proposed MOS-based resistive ladder and comparison with the state-of-the-art 60
5.5 Summary 61

6 Conclusion 63
6.1 Future scope of work 64

References 65

Publications 71

List of Figures

1.1 Semiconductor Memory Overview. 3
1.2 MIM Structure of RRAM device. 4
1.3 1T1R configuration. 5
1.4 Resistance Variation during SET-RESET Operation 6

3.1 Logic In-Memory architecture. 18
3.2 IHP-130nm RRAM-based NOR Logic Design. 19
3.3 IHP-130nm RRAM-based NAND Logic Design. 20
3.4 2-input hybrid XOR logic. 21
3.5 2-input XOR output waveform. 21

4.1 Conventional 6T-SRAM. 24
4.2 Stacking of two NMOS devices. 29
4.3 2X2 6T SRAM Array with Peripherals. 32
4.4 Single Sense Amplifier(SA). 33
4.5 Single bit D Latch. 33
4.6 1X2 Decoder. 34
4.7 Bitline Driver. 35
4.8 Write Switch. 35
4.9 Equalizing Circuit. 36
4.10 Proposed 8T1R NV-SRAM Cell. 38
4.11 8T1R NV-SRAM Cell Write Operation. 39
4.12 8T1R Bit-line behaviour. 40
4.13 Restore Operation 40

4.14 Hold Power Analysis. 41

5.1 RRAM crossbar array and reference generation stage 45
5.2 ON current and Delay variation with respect to technology node. . . 48
5.3 Output voltage variation by changing W/L ration of MOS 51
5.4 Output generation scheme in crossbar array 53
5.5 Single section of the proposed 2-level MOS based circuit. 55
5.6 Parallel generated reference voltage output from all 16 nodes. 56
5.7 Proposed 2-level reference generator MOS ladder. 57
5.8 PVT Variation Analysis 58
5.9 Monte-Carlo Simulation for Process variation and Mismatch. 59

List of Tables

2.1 Recommended Software versions. 13
2.2 Standard Corners. 15

5.1 Performance comparison with the state-of-the art 60

List of Abbrevations

ADC	:	Analog to Digital Converter
ADE	:	Analog Design Environment
CDF	:	Cadence Design Framework
CMOS	:	Complementary Metal-oxide-semiconductor
DRC	:	Design Rule Check
EXT	:	Device Extraction
GDSII	:	Graphic Data System II
LE	:	Loading Effect
MC	:	Monte Carlo
MOS	:	Metal-oxide-semiconductor field-effect transistor
NVSRAM	:	Non-Volatile SRAM
PDK	:	Process Design Kit
PG	:	Power gating
RRAM	:	Power gating
SRAM	:	Static Random Access Memory
XL	:	Layout Accelerator (layoutXL)

Chapter 1

Introduction and Related Work

1.1 Overview

Integrated electronic circuits made using semiconductor devices have increased the amount of processing components and memory bits accessible to system engineers for over fifty years. This expansion has resulted in orders of magnitude increases in speed, reliability, and power consumption, as well as considerable cost per device reductions. These developments are a direct result of regular device shrinking in the semiconductor production process, as predicted by Gordon Moore in 1965 ("Moore's Law"), who predicted the expansion and spread of digital computing and its uses. The basic structure of digital computing units has been built on the traditional stored-program machine architecture, which is characterised by a split between functional units for data storage and instruction execution, as established by von Neumann in the 1940s ("von Neumann architecture"). Moore's law, on the other hand, can't be maintained permanently. Nanoscale CMOS transistor dimensions are expected to reach crucial physical boundaries within the next ten years [1]. Even before Moore's law comes to an end due to technological restrictions, the area of computing is grappling with other fundamental issues that necessitate novel answers. The first issue is known as "the memory wall" [2], and it is related to the time and bandwidth needed to access memory. Another issue is the power shortage caused by computer energy dissipation [3]. These difficulties are presently viewed

as important roadblocks in computing's evolutionary path, necessitating considerable research investments to build new architectures for next-generation computing systems. Microelectronic technology will require breakthroughs "beyond Moore" to accommodate fresh applications in the future, when device dimensions will no longer be scalable [4]. These improvements might include cutting-edge new gadgets like carbon nanotubes. Over the next 2-3 decades, a less radical hybrid strategy combining traditional CMOS with technological advances is likely to give a more feasible and quick growth path. Multi-layered integrated circuits are an example of a "more than Moore" technique that is becoming commercially viable. Memristive devices are another emerging technology that will expand the possibilities of CMOS [4]. The influence of memristive technology on computers is the subject of this thesis.

1.1.0.1 Semiconductor Memory overview

The semiconductor memory can be categorized widely into two types, volatile memory and non-volatile memory (NVM). Volatile memory, as the name suggests, can retain the data as long as the power is maintained. When the power supply is turned off, it loses the data stored in it. On the other hand, non-volatile memory does not lose the stored data even when the power is turned off and is retained by non-electrical states.

The most widely used volatile memories are Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM). The programmable read only memory (PROM) is an example of non-volatile memory where the data is stored by means of fuses.

1.2 Memristors

The semiconductor industry has seen aggressive scaling in last 25 years in flash memory segment which is based on the charge trapping in the floating gate in MOS transistors, even crossing the limits put by Moore's law [5]. However, as the devices

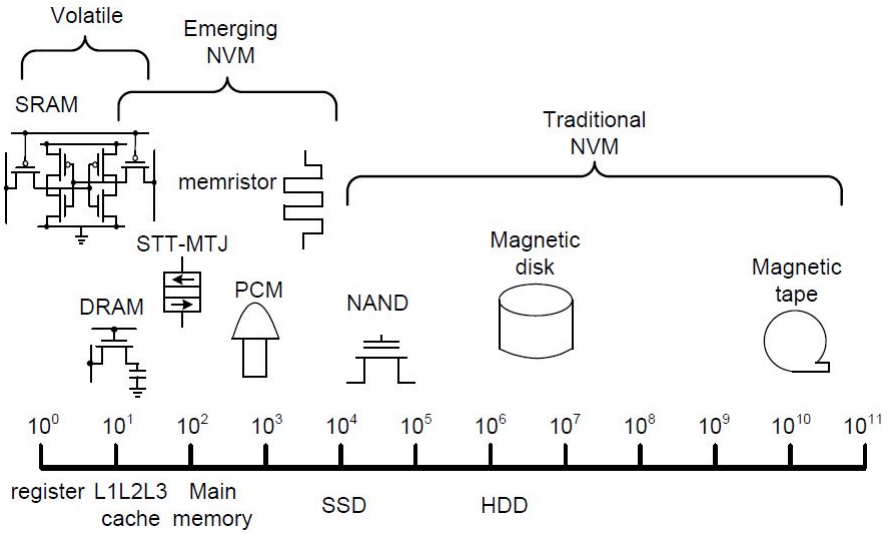


Figure 1.1: Semiconductor Memory Overview.

are scaled down in nm regime, specially below 20 nm, the challenges for smooth operation have increased drastically which in turn has caused increased bit error rate and reduced write endurance (the maximum number of write cycles a memory can handle before it becomes unreliable). When flash process technology node scales below 15 nm, these problems become severe [6]. Several alternative technologies have been investigated in recent years in order to develop a substitute for flash memory. The stored data is represented as a resistance in most of these potential technologies, and the storage device is fabricated within the metal layers itself. Nonvolatility, reasonably high write durability, high density, decent scalability beyond 10 nm, and quick read and write are all characteristics shared by these technologies. Certain potential memory technologies are fast enough and long enough to be considered SRAM and DRAM replacements, allowing universal memory to be used. Thus, memristors, or more technically, memristive devices, are a type of nonvolatile memory technology that is gaining traction in recent times.

The electrode material used in traditional electrical devices is of critical importance as it serves as a transit path for the charge carriers. In RRAM, the material used for electrodes has a significant impact on the device's switching behavior. A stable resistive switching behaviour was reported in the

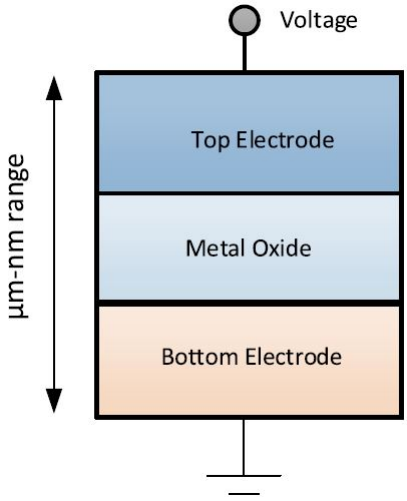
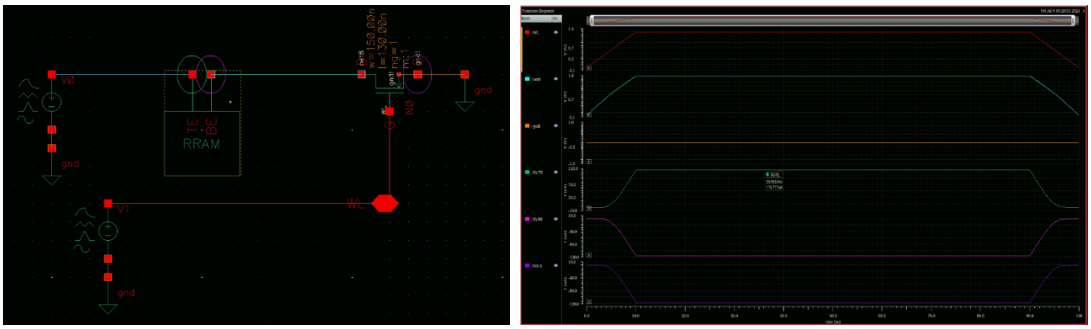


Figure 1.2: MIM Structure of RRAM device.

copper/poly(3-hexylthiophene): [6,6]-phenyl-C61-butyric acid methyl ester/indium-tin oxide (Cu/P3HT: PCBM/ITO) structure. However, it vanished when the Cu electrode was replaced with a Pt electrode. RRAM electrodes have been made from different types of materials. On the basis of their composition, the materials for electrode can be divided into five categories. Elementary substance Electrodes made of Cu, W, Ag, Pt, Ti, Al, graphene and carbon nanotubes are the most prevalent and widely utilized in elementary substance electrodes. However, for Si-based electrodes, the typical electrodes that are employed are mainly p-type Si and n-type Si. Allow electrodes such as Cu-Ti, Cu-Te, and Pt-Al are generally used to stabilize resistive switching behavior of the device. TiN and TaN are the most popular nitride-based electrodes. On the other hand, Al-doped ZnO, Ga-doped ZnO, and ITO are some of the most common oxide-based electrodes that are used. In many different types of oxides, non-volatile resistance switching has been seen, based on the large diversity of materials used. Thus, RRAM has an advantage in terms of material selection since metal-oxide-metal (MOM) devices can be easily produced using oxides currently employed in semiconductor technology.



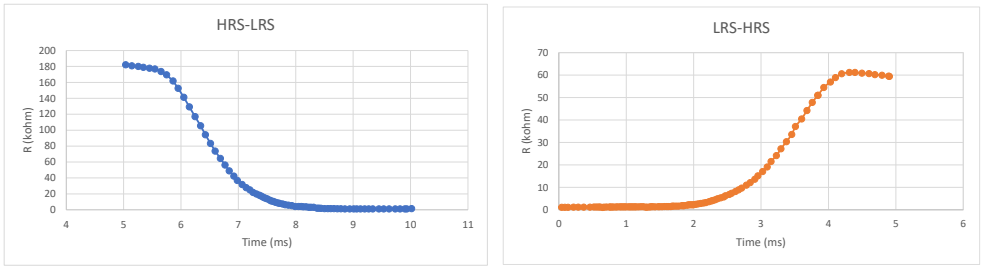
(a) Schematic view (b) Simulation waveform

Figure 1.3: 1T1R configuration.

1.2.1 Resistance Switching Modes

A resistive random access memory (RRAM) is made up of resistive switching memory cells with a metal-insulator-metal structure, or MIM structure. An insulating layer (I) is layered in between two metal (M) electrodes in this arrangement(Fig 1.2) . The schematic view of an RRAM cell is shown in figure whereas figure represents the cross-sectional view of the same cell, respectively.

A voltage pulse applied externally across the RRAM device allows it to change from OFF state (or the high resistance state, HRS) to ON state (or low resistance state, LRS), and vice versa. The OFF state and the ON states are logically represented by '0' and '1' respectively. This shift in resistance values in an RRAM cell is thought to be caused by the resistive switching (RS) phenomenon as shown in fig 1.3(b). Schematic of the setup used is also shown in Fig 1.3(a). A pristine RRAM is generally in the high resistance state (HRS) at first. In order to facilitate the transition of the device from the HRS to the LRS, a high voltage pulse is applied which helps in the creation of conductive path in the oxide layer, and the RRAM cell is switched into an LRS. This procedure is known as 'electroforming'. The voltage which facilitates this phenomenon is known as forming voltage (V_f). This phenomenon happens because of the soft breakdown of the MIM structure. To turn back from LRS to HRS, in the RRAM cell, a voltage pulse termed as the RESET voltage (Vreset) is applied, enabling this shift in the states, and the procedure is



(a) SET Operation (b) RESET Operation

Figure 1.4: Resistance Variation during SET-RESET Operation

known as the 'RESET' process as shown in Fig 1.4(a). On application of the voltage pulse, the RRAM's HRS can be switched to LRS. In the 'SET' process, a voltage is applied to change the states from HRS to LRS. This voltage is known as SET voltage (Vset)Fig 1.4(b).

In order to read data from an RRAM memory unit efficiently, a small read voltage is used to detect if the unit is in the logic low (HRS) or logic high (LRS) state without disturbing the present state of the cell. RRAM thus acts as a non-volatile memory as both logic high and logic low states retain their values even after supplied power is removed. The RRAM can be categorized into two types of switching modes based on the applied voltage polarity: (i) unipolar switching and, (ii) bipolar switching [7]. The transition (set and reset operation) of the memory unit between different resistance states in unipolar switching is not dependent on the applied voltage polarity, i.e., transition can occur when a same type voltage but of various magnitude is applied as shown in Figure. In bipolar switching, however, the device's transition (set and reset process) between different resistance states is dependent on the applied voltage polarity, i.e., a change from an HRS to LRS happens at one polarity (either of positive or negative type) of voltage, and the transition in the opposite direction i.e., back to HRS is caused by opposite type of applied voltage as shown in Figure. The physical mechanism, Joule heating causes a conducting filament to burst during a reset operation in unipolar switching. In bipolar switching, charged species migration is the predominant factor for conductive filament breakdown, but Joule heating helps in speeding up the process. A compliance current

(I_{cc}) is imposed on the device to ensure that the dielectric switching layer does not permanently break down during the forming/set operation. During off-chip testing, the I_{cc} is commonly maintained by a device which is used for cell selection such as a transistor, diode or resistor or a parameter analyzer used for semiconductor devices.

1.3 Applications of RRAM

Because of its high speed, non-volatile data store capabilities, increased storage density, and logic processing function, RRAM is viewed as one of the most promising candidates among upcoming memory technologies which has the potential to reorganize the memory hierarchy. This section discusses the different innovative RRAM applications.

1.3.1 Non-volatile Logic

Because of the distinct processing and memory unit in a computer system with von Neumann architecture, instruction codes and data are exchanged using buses between various units. The 'von Neumann bottleneck' is the result of this data transfer process, which results in higher energy consumption and time delay. The computing procedure that uses RRAM crossbar array to adjust the memory and computational activities in the same core is recommended for lowering the effect of von Neumann constraints. Furthermore, RRAM which consists of only two-terminals has a compact device structure, the $4F^2$ array design is particularly useful to achieve high density of integration and reduced cost. For example, very commonly used Boolean logic functions like 'logic NOT', 'logic AND', and 'logic OR' need numerous transistors, each of which occupies 8-10F^2 of space, whereas, only two or three RRAM cells can be used to realize these logic operations, leading to an overall approximate area of around $10F^2$.

1.3.2 Neuromorphic Computing

One of the most promising approaches to avoid the 'von Neumann bottleneck' has been to use brain-inspired neuromorphic computing, which has shown amazing potential in a variety of complex and cognitive tasks such as visual/audio recognition, self-driving, and real-time big-data analytics. RRAM-array which are based on neuromorphic computing has advantages over CMOS-based neuromorphic networks in terms of on-chip weight storage, online training, and scalability to a considerably higher array size. Furthermore, the speed of processing in RRAM increases by three orders of magnitude while energy consumption decreases by almost four orders of magnitude.

Two methodologies are proposed for achieving hardware-implemented neuromorphic computing paradigms: one resembles the form and working methodology of biological neural networks, and the other one operates by accelerating existing artificial neural network (ANN) algorithms. A synapse in a neural network is utilized to transport spikes between neurons as well as store information about the transferring weights. Different learning rules, such as spike-time-dependent plasticity (STDP) and spike-rate-dependent plasticity (SRDP), can be used to obtain information about weights.

Although various works in the research field have attempted to imitate such learning principles on RRAM devices, extending these rules based on bioinspired learning to perform a complex task is still difficult because of the inadequate theoretical methodology. An ANN can be immediately mapped to an RRAM-based neuromorphic network, which is a feasible approach. This approach has been used to show complex tasks like as speech and pattern recognition. RRAM-based synapse, while promising, is still a long way from being widely used because a number of difficulties, including optimization of materials, variation suppression, design of control circuit, architectural, and design methodologies needed for analog computing are required to be addressed in an adequate manner.

1.3.3 Security Application

With the rapid advancements in the world of information technology, the security component has grown more prominent, necessitating the use of hardware-based security integrated circuits. The security designs (at the circuit level) based on RRAM are highly resilient to attacks of various types than the same which are designed with CMOS logic, which leverages the random nature of the semiconductor fabrication process. RRAM based security circuits are more reliable to attacks of various types than security circuits based on CMOS logic, which leverages the random nature of the semiconductor manufacturing process. Larger fluctuation in device parameters of RRAM, for example, random telegraph noise (RTN), changes in the resistance value, and probabilistic switching are ideal for security applications, as opposed to memory applications, which require a smaller degree of variation among various factors. For device authentication (strong PUF) and key generation (weak PUF) applications, a novel security feature based on RRAM known as physical unclonable function (PUF) is presented. Strong PUF necessitates a much greater number of input and output combinations, whereas weak PUF necessitates only a limited number of CRPs with extremely high dependability. Despite the fact that RRAM-based PUFs have demonstrated excellent performance, more practical examples and analyses are required to be investigated to establish the viability of this unique primitive in the domain of hardware security.

1.3.4 Non-volatile SRAM

SRAM and DRAM, which are volatile memory technologies, may require more than 50% of the static power in today's mobile SoC chips. RRAM-based NV-SRAM was developed to accomplish fast simultaneous memory operations, decreased size, and low power consumption. To form an 8T2R structure, 2 RRAM units are layered on 8 transistors. Also, nonvolatile ternary content-addressable memory (TCAM) with a 4T2R unit structure and non-volatile flip flops with lowered stress time and reduced write power designed with RRAM has been recently reported.

1.4 Organization of Thesis

The rest of the thesis is laid out as follows:

Chapter 2: In this chapter we discussed the simulation methodology of importing the libraries of different technology such as IHP130nm, 45nm. We have also discussed how to run several types of simulations in cadence-virtuoso environment.

Chapter 3: Logic-in-memory computing background and related architectures are discussed in this chapter. Furthermore, we explain how RRAM cell can be used to design logic-in-memory cells and then be used for other combinational logics.

Chapter 4: Non-volatile SRAM is discussed in this chapter along with the detailed discussion of 6T-SRAM array with peripheral circuitry.

Chapter 5: In this chapter, Loading effect has been discussed and how it can be minimized using efficient designing of voltage reference ladder circuit. It's pareto analysis has also been discussed.

Chapter 6: This chapter brings the effort to a close and lays the groundwork for future projects..

Chapter 2

Simulation Methodology

To ensure repeatability, prominent tools and procedures were used in this study. All simulations were performed using the Cadence Virtuoso tool. It was decided to use the most recent version accessible. A CMOS component library, the ability to assess power, delay, and component counts, and a memristor model were all met by Virtuoso. There have been a variety of memristor models suggested and utilised, each with its own set of characteristics. The IHP RRAM model was chosen for this project because it can represent several models, whereas other models are limited to a single memristor fabrication. Various memristor settings must be carefully chosen in order to employ the IHP model. Resistances R_{high} and R_{low} are two such parameters, with $R_{high} >> R_{low}$ indicating that a memristor with Rhigh resistance is in the 0 state while a memristor is in the 1 state with R_{low} resistance. Other features of the memristors, such as internal thresholds and doping width, influence the specific values of R_{low} and R_{high}. At t=0ns, all memristors and other pertinent values in the designs are set to 0 or high resistance. In the Spectre simulator and Virtuoso ADE L, all simulations were run using transient analysis. It is not realistic to compare memristor-based designs to classic CMOS designs, and occasionally to other memristor designs, while implementing them. Second, RRAM device themselves are still poorly understood in terms of interference and other circuit practical limits. As a result, the RRAM model's real dimensional area needs are unavailable, making genuine area comparisons across designs difficult.

The Virtuoso Editor was used to create all of the schematics. At t=0ns, all memristors and other pertinent values in the circuits are set to 0 or high resistance. In the Spectre simulator and Virtuoso ADE L, all simulations were run using transient analysis. It is not realistic to compare memristor-based designs to classic CMOS designs, and occasionally to other memristor designs, while implementing them. Second, RRAM device themselves are still poorly defined in terms of interference and other circuit practical limits. As a result, the RRAM model's real dimensional area needs are unavailable, making genuine area comparisons across designs difficult.

2.1 SG13 Design Kit Model

This guide describes the SG13 Cadence™ Design Kit family that currently contains the SG13S. SG13 is a state-of-the-art 0,13 µm BiCMOS process with tungsten local interconnection, aluminum metallization, high-speed npn SiGe: C-HBTs, MIM capacitors, poly-Si resistors, inductors and more. The standard backend option offers seven metal layers including thick Top-Metal1 and Top-Metal2. Drawings in SG13 are made in 1:1 scale. The kit offers an Assura runset for Cadence and is optimized for the development of analog and high-speed circuits such as for fiber communication or wireless applications. The default simulation environment is Spectre (/RF), additionally an ADS link is provided. Non-Cadence users can ask for a special "GDS-Kit" which contains device layout samples, models and documentation as well as rule files and ASCII technology file. General: The grid resolution in the SG13 process is 5 nm. All layout components implemented in the Design Kit are on grid. Following this, your entire design must be on grid. If you draw a metal object by using the path command, the path width should be an even number times the grid resolution to avoid off-grid DRC errors. Process options: Please note that users can not change technology and/or module options on their own within the PDK installation. Resultant, you can not handle the circuits for different process options within the same Cadence program instance. This is done to reduce error probability. For convenience, the Design Kit will write a prompt in the bottom left

Table 2.1: Recommended Software versions.

Cadence Virtuoso	IC6.1.6 or above
Assura	Assura sub-version 4.1_HF4
MMSIM	mmsim72 or above
IUS	IUS82 or above
Encounter	soc913
QRC	ext914

of all windows which contains the technology, frontend and backend option.

Recommended Software Versions Cadence virtuoso IC6.1.6 or above Assura Assura sub-version 4.1 HF4 MMSIM mmsim72 or above IUS IUS82 or above Encounter soc913 QRC ext914 as shown in Table 2.1

2.2 Circuit Entry and Simulations

Circuit Entry For a new design, select the Library manager → File → New → Library in the library browser menu. Enter a name for your new library → click apply and the TechnologyFile window appears → select Attach to an existing techfile → and then OK. Select SG13_dev as technology library. Now it is possible to select a library by clicking on it. To create a new schematic, select Library → File → New → Cellview in the library browser menu. Enter "schematic" in the name line and press TAB. The composer will be selected as the schematic entry tool. It's good to keep the schematic separate from the test environment and the supply. In order to do so, you must create a symbol for your design. The best time to do this is afer defining all I/O pins as symbolic pins (create → pin), by selecting Design → Create → CellviewByCellview. Doing this again after the changes have been made to the pin, the definitions will allow you to modify the symbol.

2.2.1 ADE L Simulation

To start simulations, select Tools → AnalogEnvironment in the composer window. A new window will pop up which allows to define simulations, variables, environment, and models. The default Setup is to use Spectre (or Spectre/RF) Simulator/Directory/Host to display the dialog box → Spectre(/RF) as the simulator at 27° environment temperature with typical mean models. If you want to view the results, use the Waveform viewer from the Tools menu and the Calculator to select transient (V_T), AC (V_F), DC Sweep (V_S) or operating point (V_{DC}) voltages in the schematic. To probe currents, you must first select the currents to be probed by using either Outputs → ToBeSaved or Outputs → SelectAll. In some cases this may cause a significant slow-down or convergence problems.

2.2.2 Corner Simulation

Analog circuits have to work across wide supply voltages, temperatures and process parameter ranges. To simulate the process parameter range, best case and worst case simulation models have been characterized. These are implemented into the Design Kit while each case has its own model file that is stored in the path /revn.n.n/tech/spectre These files only contain relative numbers to vary the common models which reside in the models.all file. Thus it is easy to define different process corners which may be useful for the design. Typically model file, supply voltage and temperature will be varied in the corner simulations (PVT variation = process, voltage, temperature). Table 2.2 shows an example range for those independent parameters. For three corners, three temperatures and three supply voltages you will run 27 simulations.

Own corners can be defined by creating a new corner file. All models are kept in the models.all file. The parameters vary with variables whose default value is equal to 1.0 (such as models.typ). Therefore, the corner files only change those predefined variables by a relative value, e.g. 0.8 for a –20% deviation. Any simulation with the Spectre Simulator will be done with models.typ model file unless it is changed.

Table 2.2: Standard Corners.

Corner Simulation	**Modelfile**	**Temperature** ($^{\circ}C$)	**Supply Voltage**
Typical Mean	models.typ	27	Vdd
Worst Case	models.wcs	125	Vdd-10%
Best Case	models.bcs	-40	Vdd+10%

To change according to model file in the Analog Artist tool select Setup → Model Libraries. The Model Library Setup form will appear. Select with left-clicking the model path and it displays automatically into the Model Library File box. Change the model file to, e.g. 'models.wcs' and click Change. Select OK to continue. To change the simulation temperature, select Setup → Temperature in the Analog Design Environment tool. Change the temperature and click OK to continue.

2.2.3 Monte-Carlo and Parametric Simulation

In many cases the corner simulations are not sufficient to characterize a circuit. Many analog designs rely on a good matching between individual elements of a circuit. You can perform statistical simulations by the method of a Monte-Carlo (MC) analysis and include either process tolerance, device mismatch or both. MC will run for instance for a hundred times and calculate each time, individual parameters for every element, based on process statistics and limited to a deviation of 3σ. In fact, process tolerance must be simulated with a typical mean model setup. Before starting a MC run, you should always force a recreation of the netlist by selecting Options → Netlist → Recreate. Parametric simulations are useful to combine, for example, a DC Sweep of temperature which runs at different operating voltages. To do this, choose Variables → CopyFromCellview in the Analog Environment window → and edit the default value and finally Variables → CopyToCellview to store the value in the schematic. Before doing a parametric analysis, you should always force Simulation → Netlist → Recreate.

Chapter 3

Logic-In-Memory Implementation using IHP-130nm RRAM Technology

3.1 Non-Volatile In-Memory Computing Architecture

In a traditional computing system, the data is stored in the memory architecture only. I/O connections are required since it is segregated from the general CPUs. As a result, all data must be sent out to an external processor and then stored back in during information processing. This architecture, on the other hand, will cause substantial I/O congestion in data-oriented applications like machine learning acceleration, resulting in a significant degradation in overall performance. Furthermore, considerable static power will be required as all of the information must be stored even when they are not being used.

Adding extra I/O pins or using them at a greater frequency might theoretically solve the bandwidth problem at present. In practice, however, the delay in signal propagation and integration of signal related difficulties limit I/O frequency, and I/O quantity is constrained by technology associated with packing, thus bandwidth

can only be increased so far. It is also possible to reduce the volume of information transferring between the processor and the memory, in addition to enhancing memory bandwidth. The CPU typically just reads the raw information from the main memory. The I/O communication demand can be greatly decreased if some of the operations can be executed in the memory itself before delivering data. For eg., if we need to perform addition of 8 numbers, we must feed all 8 values into the CPU in the traditional manner. If the 2 numbers can be added as an **in-memory logic operation**, the pre-processing of addition can be performed inside the memory itself and only have to read out four values. In order to execute in-memory computing, we must implement logic operations within memory in order to perform pre-processing. Logic-in-memory architecture is the name for this type of architecture. Figure 3.1 summarizes the requirements of the ideal logic-in-memory architecture in general. Through large bandwidth and power efficient reconfigurable I/Os, a large non-volatile memory sea is connected to thousands of small accelerator cores.

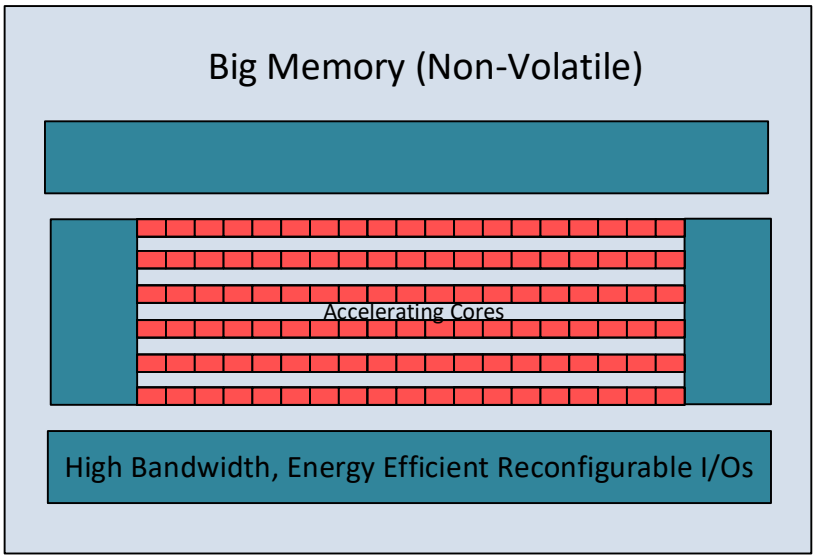


Figure 3.1: Logic In-Memory architecture.

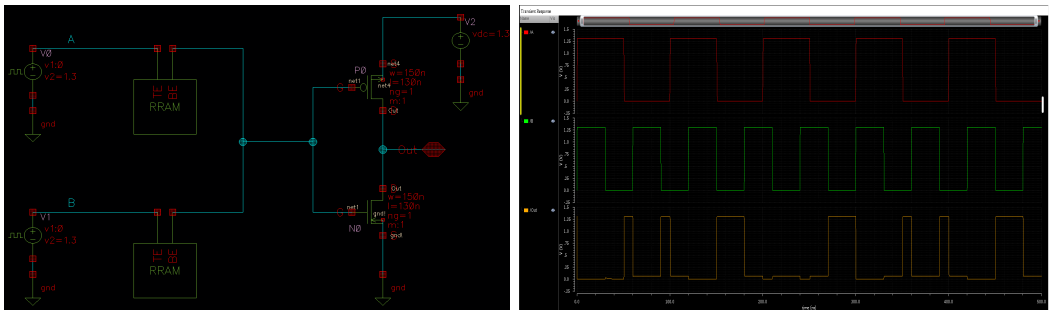
Logic-in-memory designs connected with non-volatile memory are shown in the coming section, taking into account leakage power minimization at the same time. By combining CMOS circuits with storage devices, Figure 3.2 depicts a logic-in-

memory architecture. However, there are two fundamental drawbacks to this very cell-level in-memory architecture. To begin with, logic implemented through the standard CMOS process, is stored in memory, making it tough to reconfigure. Second, it can only execute basic logic. In this context, the memory complexity will be considerably increased.

3.2 Combinational Logic Design using IHP 130nm RRAM.

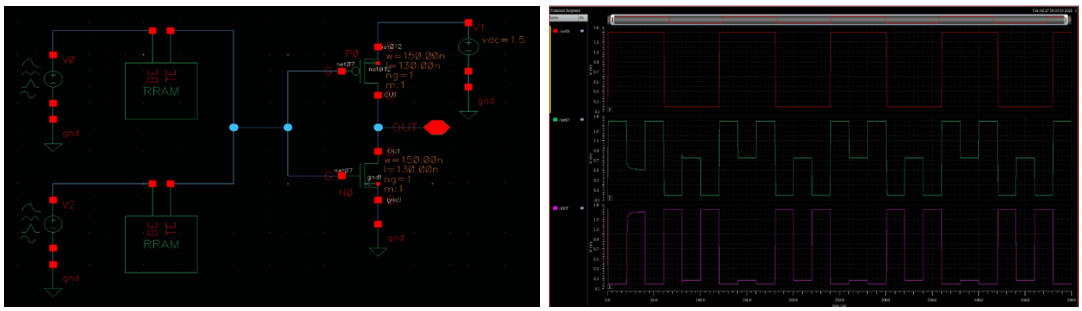
3.2.1 NOR Logic using RRAM

This is illustrated in Figure 3.2 for a NOR gate, where the beginning resistances of memristors in1 and in2 are the gate's inputs, and the ultimate resistance of memristor out is the gate's output for a NOR gate as illustrated in Fig. 3.2(a). The operation of this gate is divided into two sections. The output memristor is initially set to a known logical state in the first step. A voltage V0 is put across the logic gate in the second stage of operation. The voltage across the output memristor is determined by the logical state of the input and output memristors when V0 is applied. To maintain correct functioning, the memristor's nonlinear features, notably the threshold currents or voltages, are used. Similarly, NAND logic can also be designed which is explained in the next subsection.



(a) Schematic (b) Output Waveform

Figure 3.2: IHP-130nm RRAM-based NOR Logic Design.



(a) Schematic (b) Output Waveform

Figure 3.3: IHP-130nm RRAM-based NAND Logic Design.

3.2.2 NAND Logic using RRAM

NAND logic can also be designed in similar manner as NOR logic by justing reversing the polarity of the RRAM devices(Fig.3.3(a)). The ristances of both the devices varies accordingly. The voltage is strong enough to change the logical state of the output memristor for specific input input combinations, i.e., the memristor voltage/current is higher than the threshold voltage/current, whereas the output remains in the initialise state for other input combinations, i.e., the RRAM voltage/current is below the threshold voltage level. It's worth noting that complete switching isn't always possible with RRAM with a threshold current.

3.2.3 XOR Logic using RRAM

Extending from the basic logic cell design, the XOR logic designing is explained in this section. The XOR logic is designed using RRAM-based NOR gate. NOR stages are repeated to achieve the functionality as shown in Fig. 3.4. The schematics for NAND, NOR and XOR are hybrid in nature i.e., the overall design consists both RRAM (IHP 130nm) and CMOS (IHP 130nm) technology. Output waveform is also shown (Fig. 3.5). As we move on to extend the design further for other complex logic circuit, problem of Loading effect(LE) arises which is due to the dual polarity structure of the RRAM. The loading effect is also discussed in detail in chapter 6.

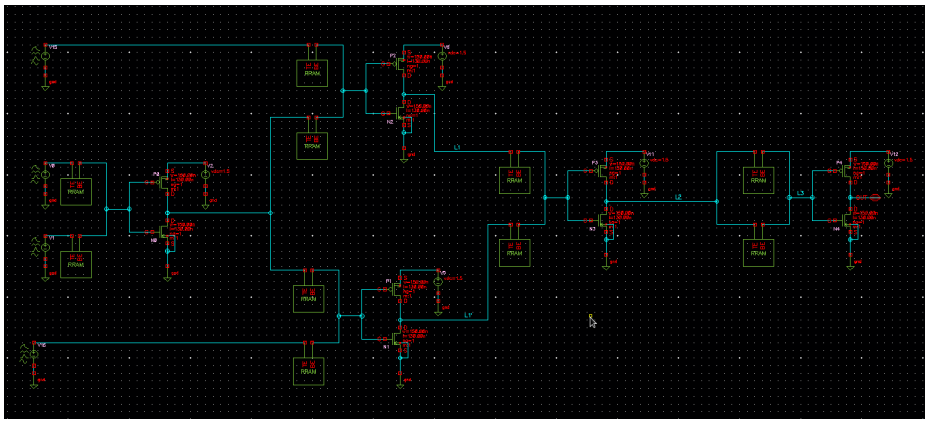


Figure 3.4: 2-input hybrid XOR logic.

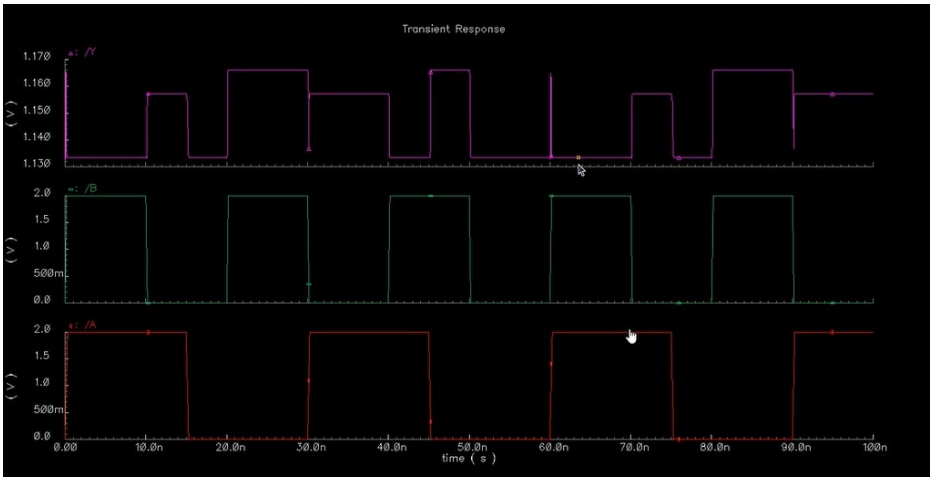


Figure 3.5: 2-input XOR output waveform.

3.3 Summary

In this chapter logic-in-memory cells have been proposed using IHP130nm technology. Depending on the switching mechanism (unipolar or bipolar) of the RRAM device, input voltage pulse needs to be applied. In these designs unipolar switching is used 1.3v square wave pulse is applied and the output waveforms are plotted. To use bipolar switching the applied input pulse needs to be varied in both positive as well as negative cycle. Basic logics such as NOR, NAND and XOR are proposed

and similar technique can be used to design other complex combinational circuits as part of the future work.

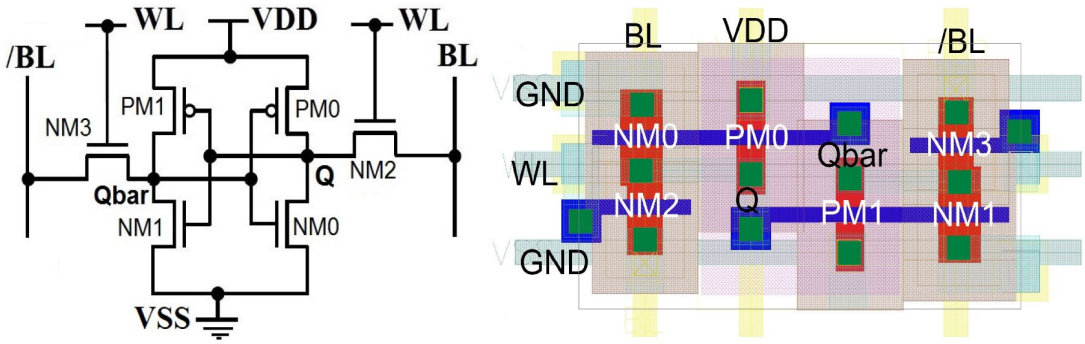
Chapter 4

Non-Volatile SRAM

Static random access memory (SRAM) is categorised as volatile memory since it loses its contents when turned off. It is volatile since there is no data when power comes back onto the device. The dynamic random access memory (DRAM) found in all modern computers and laptops is another form of volatile memory. The NV-SRAM is a type of non-volatile memory which combines the benefits of SRAM with nonvolatility feature. When compared to competing options such as huge capacitors and batteries to maintain data on devices when power is stopped, the nvSRAM offers significant benefits for applications where high speed and nonvolatile storage are needed at a low cost. Smart metres, servers, programmable logic controllers (PLCs), games, multifunction printers, and storage units are among these uses. It is a novel hybrid structure that combines SRAM and RRAM technologies (NVSRAM).

4.1 Conventional 6T-SRAM

Fig.4.1(a) shows the fundamental construction of a standard 6T SRAM cell. As shown in the diagram, it comprises of a cross-coupled inverter pair made up of NM0-PM0 and NM1-PM1 devices, as well as two NMOS access devices NM2 and NM3 linked to the complementary bit lines BL and /BL. The bit lines voltage is removed from the cell storage nodes Q and Qbar during standby mode because the word line signal WL remains at logic 0. The inverter pair's intrinsic feedback loop



(a) Schematic (b) Layout

Figure 4.1: Conventional 6T-SRAM.

retains its latching property and saves the stored data in this circumstance. When both complimentary bit lines are driven by the data to be written, the cell's write operation is enabled by activating the WL at logic 1. Data is always written into a normal 6T SRAM cell by first writing a '0' into one of its storage nodes. For example, assuming Q and Qbar carry logic 1 and 0 respectively, and logic 0 has to be written at node Q, which is referred to as the cell's write 0 operation. As a result, 0 and 1 drive write 0, BL, and /BL, respectively. As a result of this condition, the node Q discharges through the activated access device NM2 and discharges BL. Initial logic 0 at Qbar, on the other hand, maintains PMOS device PM0 active and prevents node Q from being discharged by delivering a charging current from VDD. As a result, a successful write operation necessitates access devices that are more powerful than the pull-up devices.

The WL stays active to switch on the access devices during the read 0 operation, and the bit lines are precharged to VDD. As a result, the node Q storing 0 supplies a discharge channel to the corresponding bit line BL, and the sensing amplifer senses the voltage difference of both bit lines to provide the read output. However, in this case of BL discharge, the access and pull-down devices (NM2 and NM0, respectively) form a potential divider, resulting in a non-zero potential at node Q. Furthermore, if NM2's driving current is greater than NM0's, a voltage increase at Q greater than the switching voltage of the opposite inverter may occur. It will flip the data,

resulting in read failure. As a result, a cell's effective read operation necessitates access devices that are weaker than the pull-down devices. 6T SRAM layout is also shown in Fig 4.1(b).

4.2 Major Design Challenges of Conventional SRAM

SRAM has been the chosen cache memory technology by VLSI designers for many years. It's because SRAM's operating speed is comparable to that of a basic logic circuit, and it uses less static power. Furthermore, SRAM may be manufactured using the same method as conventional logic circuits, resulting in no additional processing costs throughout the manufacturing process. SRAM is a better choice than other existing memory options such as DRAM or Flash memories because of such advantages [8]. SRAM is an essential building component in today's SoCs, and optimising its low power, huge bandwidth, high density, or dependable functioning as per the specified application is a real concern. Memory, on the other hand, is the most vulnerable component of any system when it comes to the ever-increasing process variations that come with technology downscaling. It is related to higher V_{TH} fluctuations with lower device geometry for contemporary scaled technologies, which increases the risk of memory bit cells failing owing to difficulty in maintaining their scaling ratio [9]. Furthermore, due to the lower voltage swing required to tolerate the effect of noise voltage, voltage scaling adds to the issue of bit cell stability. In the context of higher process variability at lesser technology, the scenario at low supply becomes much worse.

4.2.1 Inter-Die (Global) Variations

Variances in the average value of different device parameters for numerous dies are referred to as inter-die variations. The average NMOS/PMOS threshold voltage, dielectric thickness, and poly width are among the factors. Systematic processing

changes impacting individual dies cause a global variance, and all devices on a single die are impacted in the same way.

4.2.2 Intra-Die (Local/Mismatch) Variations

The discrepancy in parameter values among nominally matched devices on the same die is referred to as intra-die variations. The difference in numbers of NMOS/PMOS channel-adjust doping ions, poly line-edge roughness, lithographic loss, and transient phenomena like Negative Bias Temperature Instability all contribute to mismatch fluctuations (NBTI) When it comes to the growing prevalence of design failures reported in advanced technology nodes, local variances are more concerning than global variations. These intra-die fluctuations are inversely proportional to channel area, and so grow exponentially as technology downscales. At low supply voltage, such changes significantly decrease memory metrics such as read current (IRead) and Read and Write Static Noise Margin (RSNM and WSNM, respectively) of SRAM cells. The significant variability of device ON current (I_{ON}) induced by V_{TH} fluctuations degrades read performance.

4.2.3 Read/Write Conflict

A trade-off exists between read and write operations in a conventional 6T-SRAM cell, making it challenging to concurrently enhance both read and write performance. The beta ratio, which is the ratio of pull down to access device sizes, and the switching voltage of cross-coupled inverters must both be high to increase read stability. To avoid read upset, the necessary ratio is typically in the region of 1.2-3. On the other hand, raising the pull-up ratio (ratio), which is the ratio of access to pull-up device sizes, improves writeability. The normal value is generally required to be less than 1.8 for excellent write ability.

A read decoupling strategy, in which storage nodes are isolated from the read channel, is extensively employed for the different SRAM cells to eliminate this read/write trade-off.

4.2.4 Temperature Dependence of Read and Write Ability

Slow PMOS and fast NMOS (SF corner) are identified as the worst corners for read stability when considering inter-die differences represented by different process corners. Slow NMOS and Fast PMOS (SF corner) are the worst process corners in terms of writeability. The rapid NMOS (pull down and access transistors) result in poorer cell read stability when the threshold voltage (V_{TH}) of NMOS is dropped at high-temperature levels. Reduced temperature, on the other hand, increases the VTH of NMOS devices and enhances cell stability. With increasing/decreasing temperature values, the read performance (I_{Read}) of 6T SRAM cells improves/degrades. When working with low supply voltage and small devices in advanced technological nodes, the temperature sensitivity of SRAM cells becomes critical.

4.2.5 Trade-off between Performance and Stability

Supply voltage scaling is a typical method of reducing current SoC devices' power usage. The switching power of the architecture is lowered quadratically by voltage scaling, but the reduced driving current of transistors degrades the operating frequency by many orders of magnitude. As a result, the leakage power of such slow clock cycles dominates the system's overall power, putting a restriction on low-voltage operation in power-limited applications. Although high-VTH devices appear to be an alternative for controlling leakage, the lower driving current slows down the system speed. Furthermore, the traditional 6T has the issue of a difficult trade-off between stability (SNMread) and performance (I_{read}). Because the read stability (RSNM) is determined by the ((W=L)pulldown/(W=L)access) ratio, raising its value improves read stability and eliminates the danger of read failure. However, for a bigger I_{read} and increased read performance, a greater value of (W=L)access (i.e. a lower ratio=(W=L)pulldown/(W=L)access) is required. As a result, it has to choose between memory read stability and read performance.

4.2.6 Soft-Error Issue

Due to lower critical charge (Q_{crit}) at continuous scaling of supply voltage VDD and smaller storage node-capacitances for tiny feature sized devices of advanced technology nodes, the effect of soft error is becoming crucial for 6T SRAM cells. It's because a soft error happens when an alpha particle or other cosmic rays collide with a memory node and transfer energy to a storage node, causing the memory node to lose its stored data. Because the sensitivity to soft errors is determined by the charge stored at the storage node, a minimum quantity of storage node capacitance is required to tolerate the occurrence of soft errors. As a result, every ten percent drop in VDD raises the soft error rate (SER) by 18 percent. When numerous bits of memory are affected by a soft mistake, the situation becomes quite serious. Bit-interleaving array design is a frequent solution for removing such multi-bit errors. For traditional SRAM cells, although, the use of bit-interleaving causes a half-select problem.

4.3 Low-Power SRAM Cell Design

Low-power SRAM cell can be designed by modifying the design structure of the conventional 6T cell. The implementation can be done in a several manners.

4.3.1 Use of Multi-V_{TH} Device

The multi-V_{TH} approach may be used to build an SRAM cell with a dedicated read port. For this design, high-V_{TH} devices are used in the core latch section, which includes the write access devices and cross-coupled inverter pair, to reduce leakage, but low-V_{TH} devices are required in the read port to preserve read performance. This approach might be a suitable solution for power optimised SRAM cell design because leakage is the biggest power source for the SRAM at low supply voltages. Due to the deterioration in write performance, this is not true for energy efficiency. The total SRAM energy (E_{total}) can be written using Eq. 4.1:

$$
\begin{aligned}
E_{total} &= E_{switching} + E_{leakage} \\
&= C_{switching} \times V_{DD}^2 + I_{leakage} \times V_{DD} \times T
\end{aligned}
\tag{4.1}
$$

where the operation time T = 2 $\times max(\mathrm{t}_{read}, t_{write})$.

When T is decided by the write time t_{write}, the slower write speed caused by high-V_{TH} devices along the route may result in a higher E_{total}. As a result, these power-saving multi-V_{TH} approaches should include some extra write-assist techniques to boost write performance while simultaneously decreasing energy consumption.

4.3.2 Use of Stacking Effect

Another efficient way to reduce leakage current is to use the stacking effect. According to this, a stack is formed when two devices are linked in series, as shown in Fig. 4.2. If the higher device is turned on and the lower one is turned off, the intermediate node will grow to a positive voltage V_y.

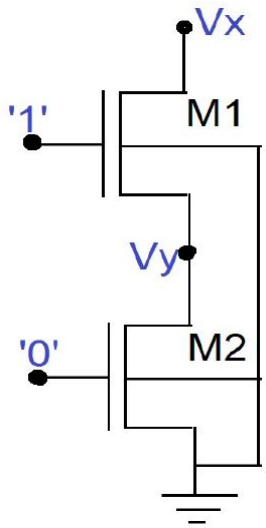


Figure 4.2: Stacking of two NMOS devices.

The effect of this non-zero voltage V_y is as follows:

1. M1's gate to source voltage (V_{GS}) decreases, and the drain current decreases as well.

2. The non-zero value of source to body voltage for M1 increases body biassing, which raises the threshold voltage. The leakage current is lowered by a given amount as a result.

3. With a positive source voltage (V_y), the reduced drain induced barrier lowering (DIBL) effect for M1 raises the threshold voltage and hence reduces leakage current down the circuit.

For PMOS devices, a similar effect may be achieved. As a result, using stacking devices in either the SRAM cell's pull-up or pull-down paths may assist to minimise total leakage power.

4.3.3 Supply voltage scaling

A CMOS circuit's power consumption in active mode is a mix of dynamic and static power. The static leakage current via each transistor is the only source of power in standby mode. The dynamic power of CMOS is split into two categories: (i) switching power to charge and discharge the load capacitor; (ii) short circuit power due to the tiny current flowing via the V_{DD} to V_{SS} channel due to the input signals' non-zero rise and fall times. The simple equations for the dynamic and leakage power are given as:

$$P_{dynamic} = \alpha \cdot f \cdot C \cdot V_{DD}^2 \tag{4.2}$$

$$P_{leak} = I_{leak} \cdot V_{DD} \tag{4.3}$$

where, f is the operation frequency, C is the load capacitance, VDD is supply voltage, is switching activity factor and I_{leak} is the total leakage current. The dynamic and leakage power have a quadratic and linear relationship with the supply voltage VDD, as shown by equations 4.4 & 4.5. Supply voltage scaling is thought to be an effective way to lower the power of SRAM cells based on these relationships. Furthermore, while VDD reduction degrades performance, scaling must be applied

to the non-critical route of the SRAM cell, while the critical path must receive an unscaled supply voltage.

4.3.4 Single-ended approach

The complementary bit lines (BL and /BL) in an SRAM array are significantly loaded by the node capacitance of many SRAM cells. When a read or write operation is conducted, these bit lines waste a lot of power during their switching. It's easier to grasp by looking at the read and write power equation for an 6T-SRAM memory, which is shown below.

$$P_{write} = \alpha_{write} \cdot C_{BL} \cdot V_{DD}^2 \cdot f \tag{4.4}$$

$$P_{Read} = \alpha_{read} \cdot C_{BL} \cdot \Delta V_{BL} \cdot V_{DD} \cdot f \tag{4.5}$$

here, α is the switching activity factor, C_{BL} is the bit line capacitance and V_{BL} is the bit line swing produced during the read operation.

As a result, by minimising one half of the active power owing to bit line switching, a single-ended approach, is viable for the power efficient SRAM cell design. The single-ended SRAM cell also helps to cut the cell's leakage power consumption in half. Instead of using a differential technique to execute the read or write operation, this scheme uses only a single bit line. The fundamental issue with the single-ended technique is that while executing write 1, the storage node Q suffers from V_{TH} loss through the access transistor, necessitating write enhancement.

4.4 Low-power SRAM memory array

A 2x2 (4-cell) array of 6T-SRAM has been design using IHP-130nm technology. The array includes all the peripherals such as Sense amplifier, 2-bit D latch, 1x2 decoder, bitline driver, write switch, 2-bit ATD, and Equalizing Circuit. All the circuitries are discussed in detail in subsequent chapters.

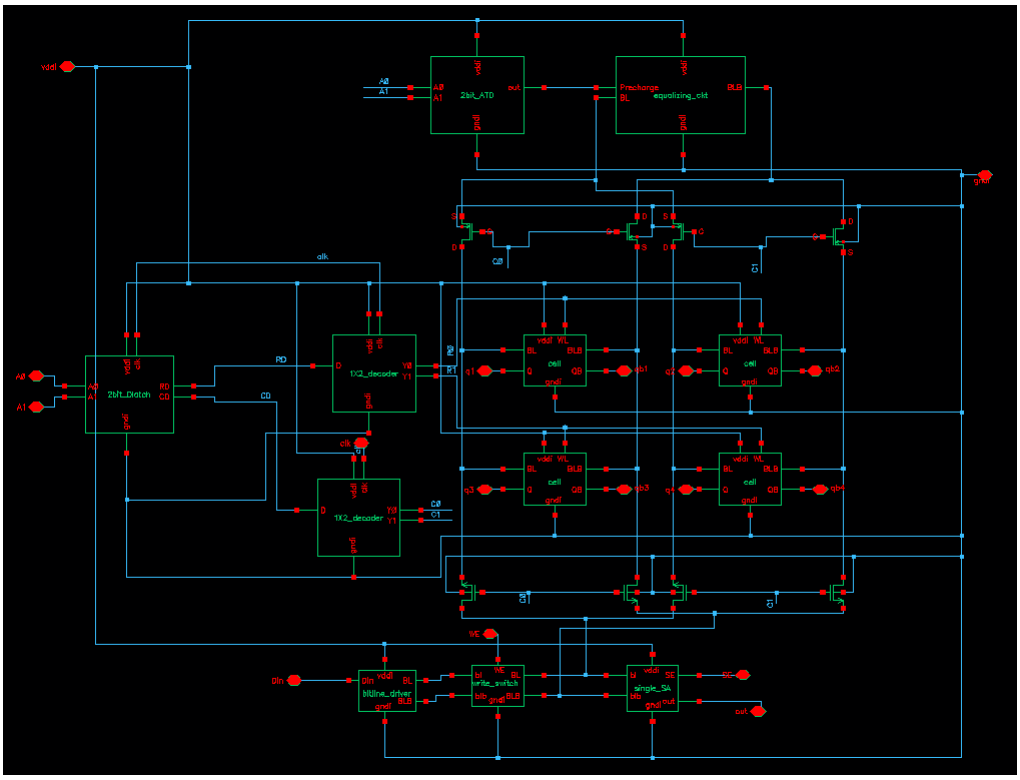


Figure 4.3: 2X2 6T SRAM Array with Peripherals.

4.4.1 Sense Amplifier

A sense amplifier is one of the circuitry elements of a semiconductor memory chip (integrated circuit) in modern computer memory. The word itself dates back to the period of magnetic core memory. A sense amplifier is a component of the read circuitry that is used to read data from memory. Its job is to detect low-power signals from the bitline that represents a data bit (1 or 0) stored in a memory cell

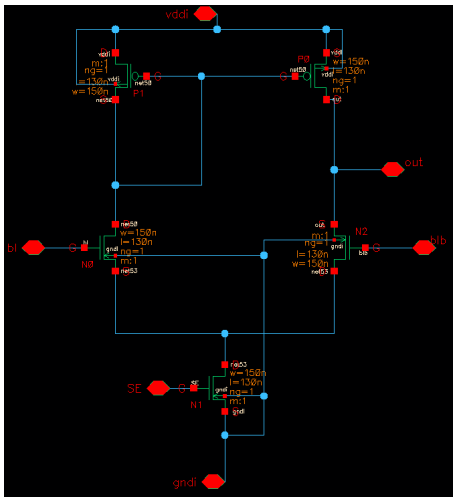


Figure 4.4: Single Sense Amplifier(SA).

and amplify the small voltage swing to recognisable logic levels, allowing the data to be properly interpreted by logic outside the memory.

4.4.2 2-bit D Latch

We may omit one of those inputs to produce a latch circuit with no "illegal" input states since the enable input on a gated S-R latch allows us to latch the Q and not-Q outputs regardless of the status of S or R. A 1-bit D-latch is a circuit with internal logic that looks like this Fig. 4.5.

A gated S-R latch with an inverter applied to make R the complement of S is

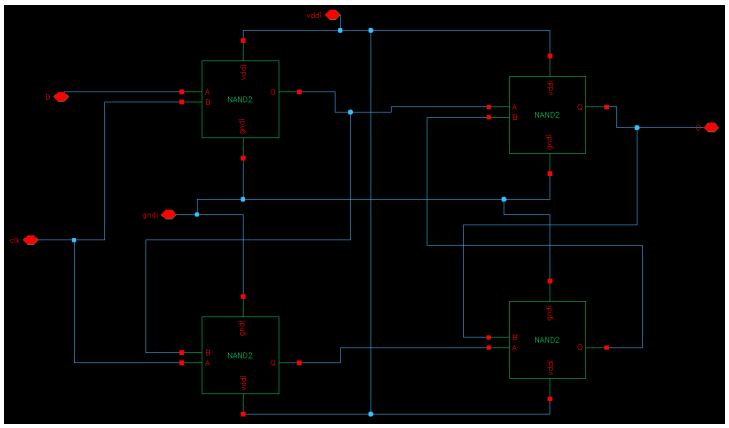


Figure 4.5: Single bit D Latch.

what the D latch is. A multi-bit memory circuit is one use for the D latch. Setting the enable input high (1) and adjusting D to any value the stored bit to be, allows us to "write" (store) a 0 or 1 bit in this latch circuit. When the enable input is set to low (0), the latch ignores the D input's state and keeps the stored bit value, sending at Q and its inverse on output not-Q.

4.4.3 1X2 Decoder

Rows and columns make up the memory cell matrix. The word line is used to access each row, whereas the bit line is used to access each column. Bit line and complementary bit line are both used to latch a bit in a memory cell. Depending on the write and read operations, digit lines or data lines are responsible for transporting data to and from bit lines. The complete word is usually attained by making the word line high. To discriminate between columns, a column decoder is employed. The decoders must be given row and column addresses in order to choose a word. Row and column decoders, on the other hand, are not designed since each bit is accessible individually for testing purposes.

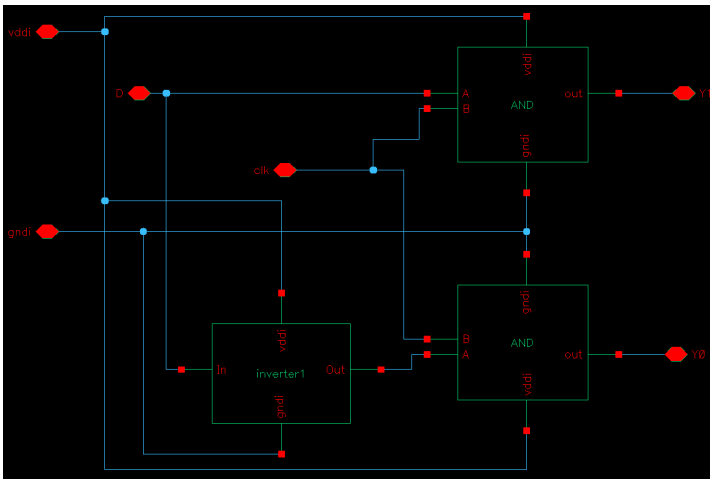


Figure 4.6: 1X2 Decoder.

4.4.4 Bitline Driver

Bitline driver is a circuit which controls the charging of bitlines. Based on the voltage present on the bitlines, it gets sensed by the sense amplifier and corresponding output bit value gets generated.

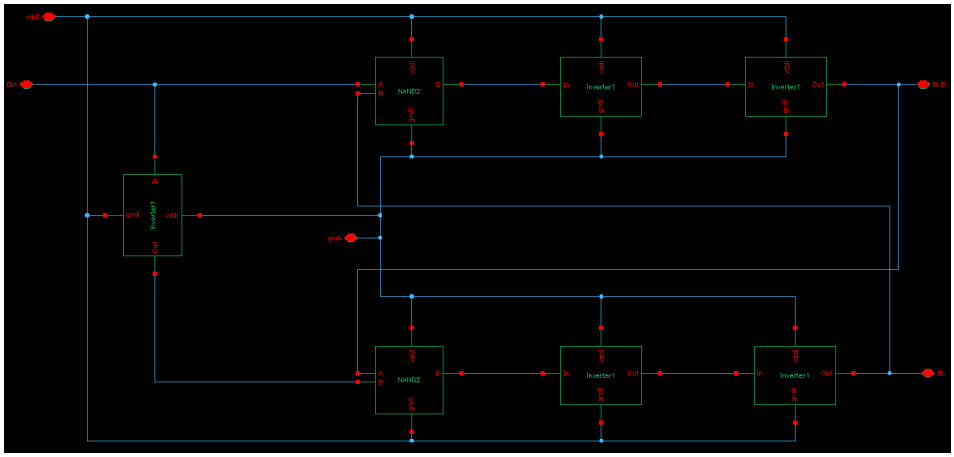


Figure 4.7: Bitline Driver.

4.4.5 Write Switch

Write switch schematic is shown in Fig. 4.8. It enables writing on bl and blb.

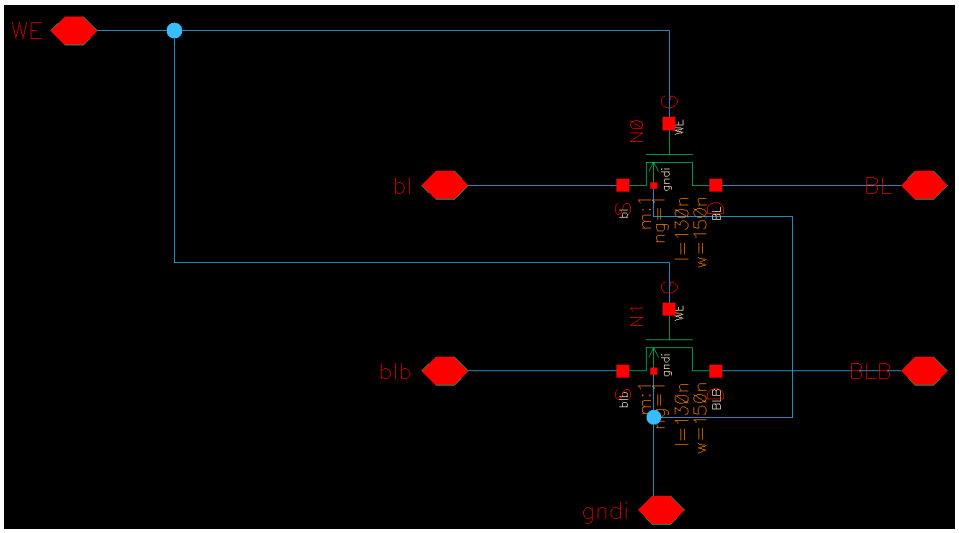


Figure 4.8: Write Switch.

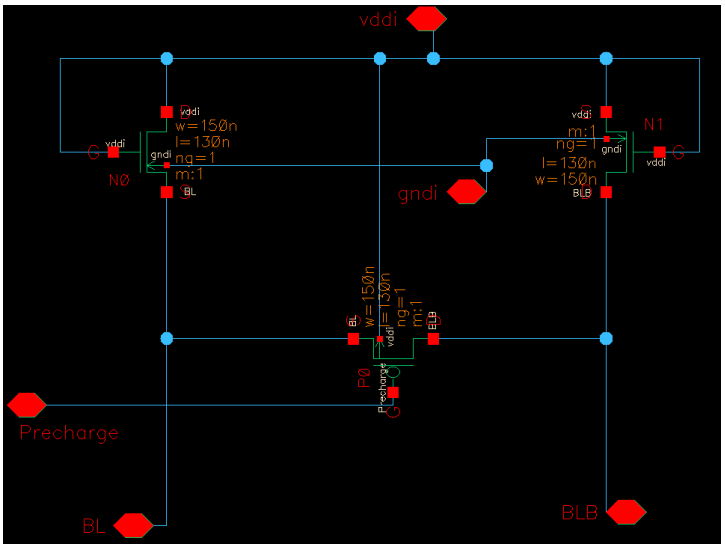


Figure 4.9: Equalizing Circuit.

4.4.6 Equalizing Circuit

Pre-charge is increased before each read and write operation, and both bit lines are charged to Vdd/2. Both bit lines are shorted during the pre-charge and equalizer cycles, resulting in a voltage differential of zero. When the sensing amplifier is turned on after the pre-charge is turned off, it detects a slight voltage difference between two-bit lines and saves a bit in the memory cell. Before writing, the pre-charge and equalizer circuits are turned on. 1 is provided by the data in input signal. At the same time, the sense amplifier is turned on and the write enable is turned on. When all inputs are stable, the word line is enabled. The pre-charge circuit is then engaged again to read the stored 1. The read enable, sense amplifier, and word line signal are all turned on at the same time. The data out signal is now high. Writing 0 follows the same procedure. Before performing a write 0 operation, the pre-charge and equalizer circuits are turned on. The data in signal is lowered, and the sensing amplifiers and word line are turned on. The value 0 is kept in the bit cell. Only during read operations will that 0 be reflected in the data out signal. As a result, data out drops following a read operation.

4.5 Non-Volatile SRAM

Nonvolatile memory is memory that keeps its data without the use of electricity. Nonvolatile memory includes nonvolatile SRAM (nvSRAM), ferroelectric RAM (Fe-RAM), electrically erasable programmable ROM (EEPROM), and flash memories, among other technologies. This type of memory is utilised in situations where essential data must be stored after the power is turned off or when the power is interrupted while the system is running. Hot plugging of cards in servers, medical equipments and industrial computers, is an example of power disruptions. Furthermore, because SRAM cells are made using the same manufacturing method as logic, there is no additional expense. When compared with other memory technologies such as Flash or DRAM memory, SRAM has unique features. Lowering the supply and threshold voltages, on the other hand, degrades SRAM cell performance and increases leakage currents in the standby state [5]. The main trends driving the rapid growth of emerging memories have been power supply reduction and down scaling in CMOS technology. In this reference, SRAM technology faces challenges as its leakage current rises resulting in a significant increase in power consumption, which is incompatible with battery-powered applications [11]. To address these challenges, RRAM is seen as a viable approach that has the potential to take SRAMs to the next level in storage technology [12]. ReRAM is a good choice for creating neuromorphic applications [15], non-volatile logic gates [16], and innovative SRAM designs [17,18] beyond storage applications because to its comparatively low access latency, high density, and analogue feature [13,14]. In SoC applications, hybrid non-volatile SRAM (NVSRAM) mainly based on upcoming memories is predicted to replace classic SRAM.

The nvSRAM is a type of non-volatile memory which combines the benefits of SRAM with the nonvolatility of nonvolatile memory. When compared to competing options such as huge capacitors and batteries to maintain data on devices when power is stopped, the nvSRAM offers significant benefits for applications where high speed and nonvolatile storage are needed at a low cost. Smart metres, servers,

programmable logic controllers (PLCs), games, multifunction printers, and storage units are among these uses. It is a novel hybrid structure that combines SRAM and RRAM technologies (NVSRAM). At low supply voltages, the proposed NVSRAM cell is meant to be electrically stable.

4.5.1 8T1R NVSRAM memory cell

The suggested 8T1R structure is shown in Figure 4.10. In comparison to state-of-the-art NVSRAMs, this innovative cell features a small amount of control signals. The two NMOS transistors N4 and N5 are coupled to the node Q and Qb.

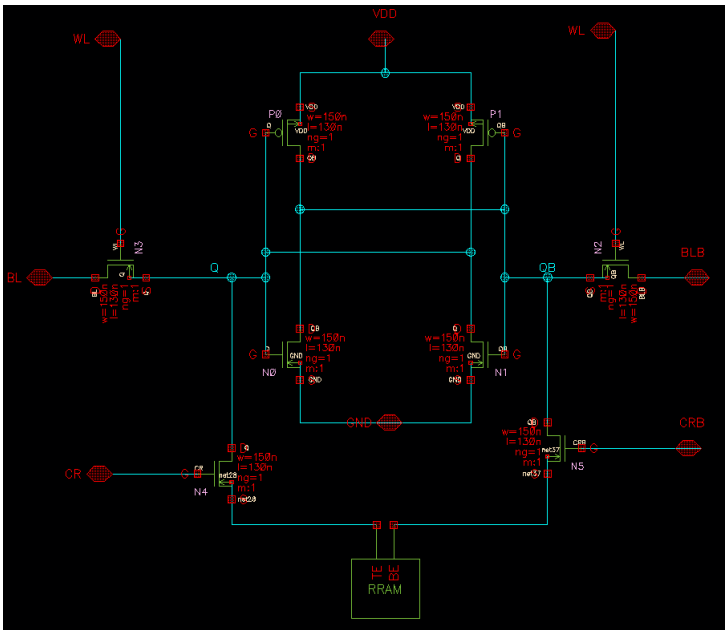


Figure 4.10: Proposed 8T1R NV-SRAM Cell.

The first phase in the RRAM process is FORMING, which involves applying a high voltage across the memory cell to transition it from HRS (high resistance state) to LRS (low resistance state). The FORMING procedure is only done once over the device's lifetime.

Following FORMING, the SET and RESET procedures are carried out by delivering a specified voltage to the OxRAM cell's electrodes (i.e. VSET and VRESET). The I-V curve, which magnifies low current levels, is the traditional depiction of the

OxRAM I-V hysteresis. N2 and N3 are the standard access transistors, whereas N4 and N5 are extra transistors utilised for the STORE and RESTORE non-volatile storage operations. SET and RESET are the two operations that make up STORE. In the OxRAM device, SET saves '1' and RESET stores '0.' BL and WLb are set high during FORMING, whereas WL and BLb are set low. The OxRAM may be constructed in one step at this point (M6 is ON: direct connection between the BL and the OxRAM). The access transistors (N2 and N3) are triggered by delivering a pulse to WLb, which is set to 5 V while BLb is grounded. A WRITE operation comes before the STORE operation. BL and WLb are set low during the WRITE "0" operation, whereas WL and BLb are put high. RESET is the stage following WRITE "0," in which the WL and BL are both set to zero. Write "1" operation is also shown through waveform in Fig. 4.12.

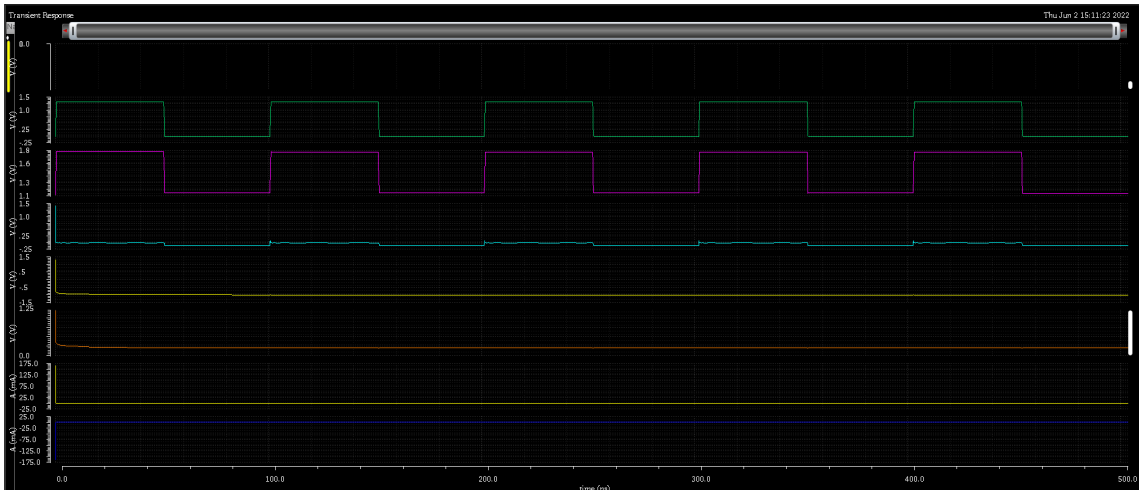


Figure 4.11: 8T1R NV-SRAM Cell Write Operation.

4.6 Results and Discussion

A traditional SRAM is discussed in detail in this section. Complete array has been proposed along with the peripherals such as Sense amplifier, 2-bit D latch, 1x2 decoder, bitline driver, write switch, 2-bit ATD, and Equalizing Circuit.

Novel 8t1r NV-SRAM cell is also proposed in this section. Cell is verified using

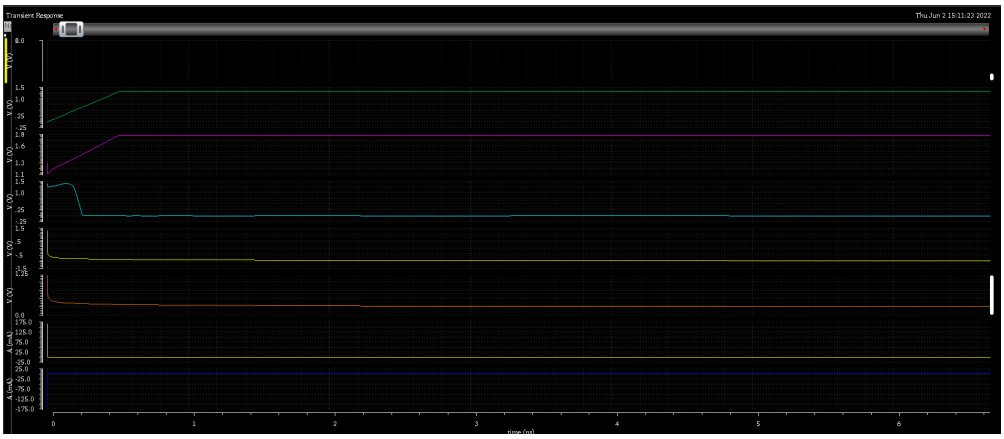


Figure 4.12: 8T1R Bit-line behaviour.

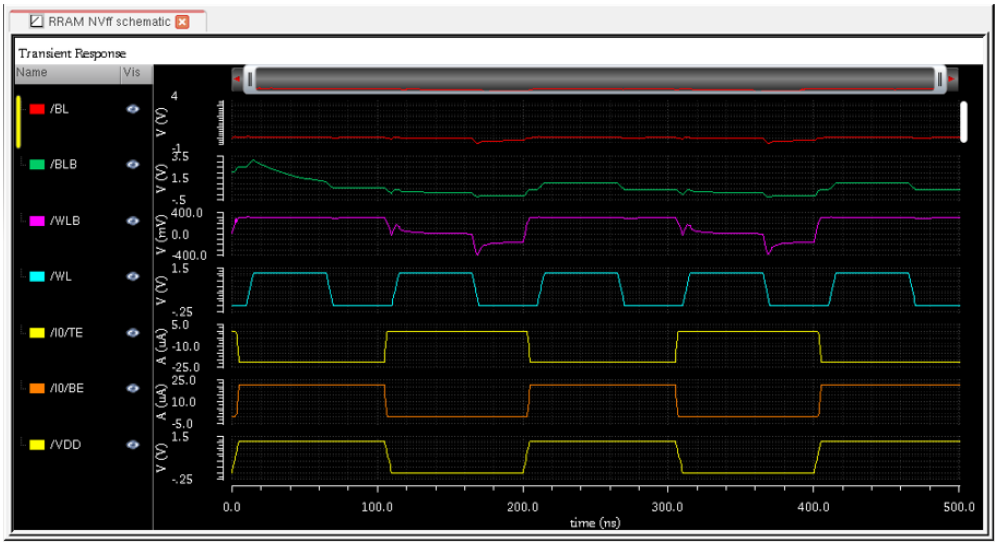


Figure 4.13: Restore Operation .

the output waveforms via blb discharge phenomena. Another waveform is also shown to verify the restore operation when VDD is turned off. Lastly, hold power analysis is performed as it is one of the crucial analysis parameter in any memory cell. It can be seen from the curve (Fig 4.14) that the overall power reduction is significant when complete array has been used as compared with the non-array (individual) design.

BL and WLb are set high during FORMING, whereas WL and BLb are set low. The OxRAM may be constructed in one step at this point (M6 is ON: direct connection between the BL and the OxRAM). The access transistors (N2 and N3)

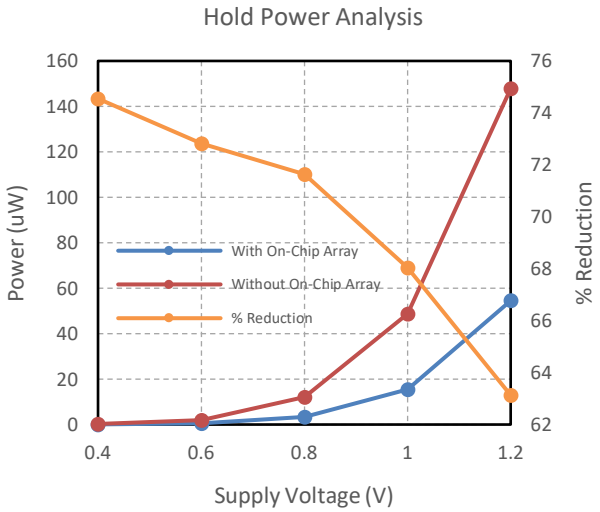


Figure 4.14: Hold Power Analysis.

are triggered by delivering a pulse to WLb, which is set to 5 V while BLb is grounded. A WRITE operation comes before the STORE operation. BL and WLb are set low during the WRITE ”0” operation, whereas WL and BLb are put high. RESET is the stage following WRITE ”0,” in which the WL and BL are both set to zero. Write ”1” operation is also shown through waveform

Chapter 5

Loading Effect Free MOS-only Voltage Reference Ladder

A deeper neural network is required with the growing requirements, variants, features, and complexities. For designing an efficient DNN accelerator, ReRAM-based In-memory computing architecture is well matured. In the analog domain, with the increase in ReRAM $m \times n$ crossbar array, the Loading Effect (LE) seems to grow at the input of the comparator stage in analog to digital converter (ADC). The reference voltage generating ladder nodes for ADC are susceptible to design parameters due to small input voltages. We used the `PMOS` transistor for the design of ladder circuitry. The circuit stability is evaluated for process variation and device mismatch. Further, sleep mode is applied using the power-gating (PG) technique to lower power dissipation. The more parallel transistors with ladder have shown more stability but consume a much larger on-chip area and energy. Therefore, in this article, a Pareto study has been performed to evaluate robust and stable circuitry with minimum loading effect in reference voltage ladder for ADC. The analysis is assessed that two parallel transistors have significant reliability. Further, we analyzed Process, Voltage, and Temperature (PVT) variation impact on proposed circuitry. An NMOS based Current mirror is also designed and used along with the proposed reference voltage ladder to achieve much better stability in terms of power supply and reference voltage variations. Finally, at the `180nm` technology

node, the proposed ladder have consumed $0.7uW$ and showed less variability for parametric variations. Therefore, the circuit supports the power-gating technique in sleep mode, saving 43% of total power. Our circuit's Monte-Carlo simulation for node voltage variation shows the minimum mean and σ deviation.

5.1 Introduction

The semiconductor industry is reluctant to adopt memristor technology primarily due to its high read and write latencies. Recent evolution in RRAM-based In-Memory-Computing circuits is considered a breakthrough in Non-Volatile processing technology. RRAM-based crossbar memory structures offer high operating speed, high reliability, high integration density, and low power consumption, making them quite effective in heavy data generation applications [10]. Energy efficiency has become a significant task as the scaling of transistors on-chip increases continuously. As in conventional Von-Neumann Architecture, the traffic between memory and computing units is heavy, especially for deep neural network (DNN) accelerators [11, 12]. It adversely affects the area and power requirements of the entire chip. Further, when physical interfaces are involved, digital can never outperform the best analog systems, but it usually outperforms all-analog methods in practice. RRAM-based memory architecture has been considered as the next generation of main memory because of its high density in the crossbar array and is likely to replace DRAM [13]. Furthermore, RRAM has achieved read latency close to DRAM while the write latency is less than 10% by optimization as compared to DRAM [13].

Consider the RRAM crossbar array of 1T1R configuration shown in Figure 5.1, which depends upon the resistance-state of the RRAM where the different configurations mode can be defined. Theoretically, the resistance of an RRAM device tunes into an arbitrary state by changing the tunneling gap to a specific length [14]. However, current work has shown RRAM device in crossbar can represent up to 5-bit data precision [15, 16]. The array performs the computation in analog mode (i.e., Volt/Reg). Therefore, up to 5-bit ADC and DAC are the essential blocks for

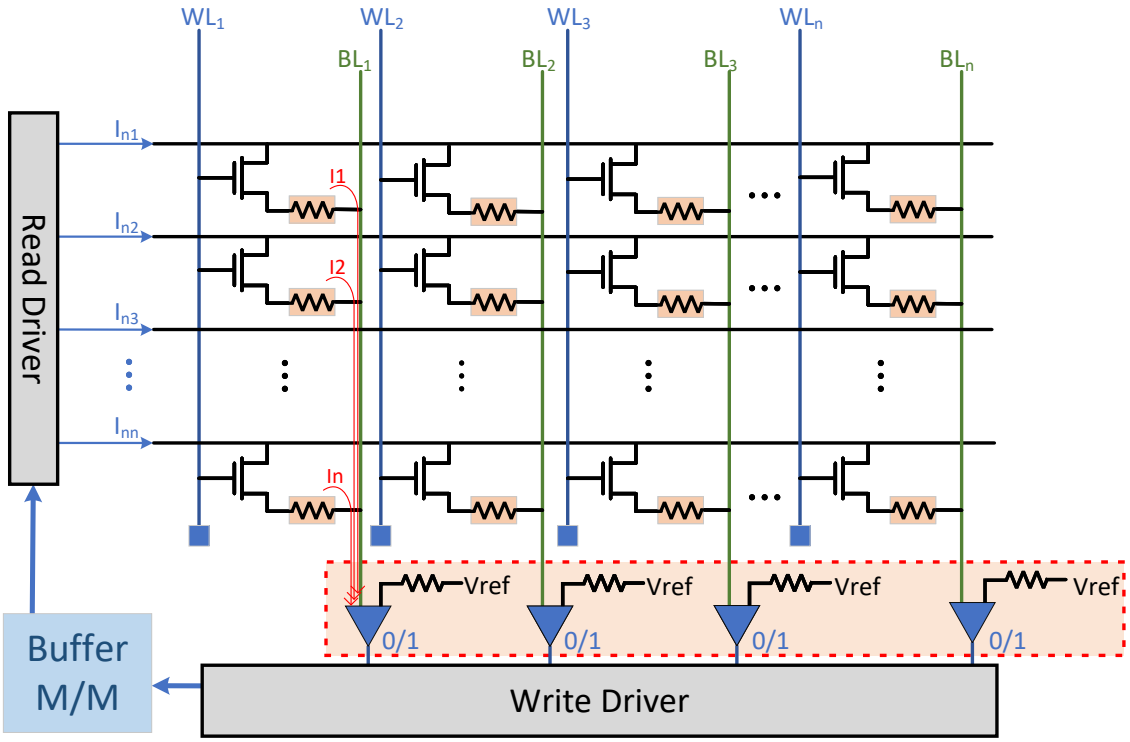


Figure 5.1: RRAM Crossbar Array to show Loading Effect on the final accumulated current output. Reference is taken out as voltage and sent to reference generation stage which is designed to minimize the variation caused by LE.

digital intermediate data processing in analog computation. With increased feature image size, the crossbar array size supplements. Therefore, the loading effect in voltage reference generations for ADC in RRAM array increases with increasing array size. In ADC, a reference voltage is required for the decision and conversion. In order to design the reference voltage generation circuit depicted in Figure 5.1, a resistive ladder is not an excellent choice due to its high power consumption and more on-chip area utilization.

High-throughput computation is the primary necessity in edge-AI computing. The RRAM-based architecture performs the computation in the analog domain and needs converters such as ADC and DAC for intermediate data conversion [17]. In this reference, flash-type ADC is preferred in the conversion. The flash ADC requires ladder of reference voltages where N-bit ADC requires $2^N - 1$ node voltages (i.e $V_1, V_2, \ldots V_N$) for the conversion. Typically reference node voltages are generated

using a resistive ladder which is very power-hungry, not an efficient way for on-chip implementation. Further, higher precision conversion circuitry acquires most of the area in any memory-based chip, leading to enhanced power requirements. The higher precision ADC/DAC is more sensitive to the design parameters due to its small voltage comparison [18]. Besides, the offset voltage impact in the comparator have at lower voltage has the main challenge. Therefore, efficient design is essential mainly in applications where large converters are required.

In the RRAM crossbar array, the peripheral circuit requires around 90% area and 95% power consumption [19]. The design requires many ADCs for the conversion and, therefore, a voltage reference ladder. Therefore, it occupies a large area and consumes a large amount of power when its resolution becomes high. Therefore, low precision circuitry is desirable for low-power applications. An application-specific efficient flash ADCs have been design in state-of-the-art [20–22]. Design specific efficient architecture have been implemented in [23–26]. The designs have used a restive ladder and discussed the loading effect at reference voltage nodes. A reliable reference voltages ladder is paramount in circuit design.

In RRAM-based analog computing, each column accumulated current in voltage (V) form compared reference voltage. In this work, we proposed efficient reference generation circuitry for ADC using PMOS transistors. The parallel transistor is recommended for generation of reference node voltage. Moreover, Pareto studies evolved the significant parallel MOSs for the design of reference ladder. The overall power consumption using resistive ladder is 1.26mW whereas the proposed design consumes very small power which is o.98uW. Furthermore, power gating technique is also applicable in the proposed design and it saved 43% of total power using sleep signal in the ladder. The rest of the paper is organized as follows. Section 5.2 details the related work and introduces preliminaries required for this work. Section 5.3 shows the design and analysis for an efficient reference generation circuit. The detailed small-signal analysis for the Pareto studies of requires parallel PMOS is discussed in Section 5.3.3. The simulation results and discussion is given in Section 5.4. Finally, the summary is given in Section 5.5.

5.2 Related Work and Motivation

RRAM-based In-Memory computing techniques have proven to be an effective solution among several stages, as shown in Figure5.1. However, RRAM-based architecture is susceptible to the design process parameters [15]. Since computing is analog in nature and storage is digital, ADC and DAC must be added to complete the operation. Moreover, with the increasing size of the crossbar array, the requirement of the number of ADCs increases. An extensive crossbar array has a more loading effect on reference voltage generation in ADCs. Therefore, we describe the related works that seek to achieve a reference generation circuit with a more reliable and minor variant to design parameters and the motivation behind its implementation.

The vast energy required to tune RRAM during training neural networks and performing any operation within the network. The two primary reasons for high tuning energy are enormous training iterations and frequent weight up-gradation while the network operates. Thus, there exists a large number of tuning behavior, which changes the resistance of RRAM cells into the target resistance. Hence, overall the loading at the reference node voltage is variable in the RRAM crossbar array [17]. Furthermore, in the case of In-memory computing, the data has to be stored temporarily in some buffer memory other than the computing memory after performing any operation within the array. Then, the next iteration must be sent back to the computing memory to complete the process.

The high precision computation circuitry increases the resolution of ADC/DAC, which comes with exponential increment in the number of hardware units [17, 27]. The simple and faster flash converters have been investigated based on the TIQ (Threshold Inverter Quantization) [28], and other techniques [29, 30]. Although they reduce the converter's complexity, they are sensitive to process variation and consume significant power. The CMOS inverter is used as an analog voltage comparator. By keeping the transistor length constant and varying its width, one can adjust the comparator's threshold (as a reference to the comparator). It's a long and

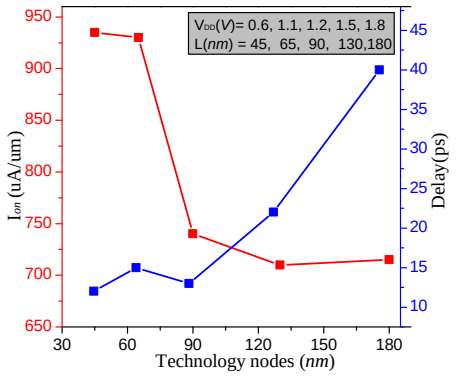


Figure 5.2: With respect to technology nodes (180,130,90,65,45), variations in ON current and Delay was obtained from an Inverter circuit simulated in the cadence environment. [12]

repetitive task, and the reference point is highly dependent on the load, which may generate erroneous output. In order to reduce the design process time and improve the circuit reliability, it is required to investigate ways to mitigate these designs into a more digital form by using stander-cells.

This work proposed an efficient voltage reference generation ladder for flash ADC in the RRAM crossbar array. We have used parallel PMOS transistors for precise generation of the reference voltage; in the case of 4-bit resolution, ADC with 1.8V supply requires step size 0.1125V (16 equal steps). To make the design efficient and significant parallel Pareto analyses have been performed. The study helped to select minimum similar MOSs requirements with no performance loss. The current in the MOSFET is proportional to the ratio W/L, where W is the width of the gate and L is its length. So, the MOS device allows more current at lower technology nodes, which comes with higher power dissipation and minimum delay as depicted in Figure 5.2. Therefore, the power-gating technique introduces in our design allows keeping the ladder in sleep mode. The proposed ladder uses voltage division property within MOS to obtain varying voltage drops across itself with the variation in width and length of the transistor.

The transistor's sizing (width and length) depends upon the drop required across it. To select the transistor sizing for the required voltage drop, the need to generalize

equation to extend the reference ladder for different configurations of node voltages generation. We used PMOS for circuit design with better stability and power consumption of PMOS-based resistive ladder circuitry and better temperature stability having a temp coefficient of Ambient (atmosphere) temperature. Therefore we checked circuit behavior for PVT variations and device mismatch. The total power has been reported for with and without power-gated ladder circuits. The leakage power has been reduced through power Gating (power shut off).

5.3 Robust Voltage Reference Generation Circuit in RRAM

Arithmetic computing requires a higher precision design architecture to obtain better output accuracy. ADC/DAC is an important circuit in any analog In-memory computing array used in the intermediate layer interface. Considering flash-type ADC, with an increased single bit resolution, the number of resources involved gets approximately doubled, and voltage comparison step size gets halved. With the reduced step size, the voltage reference node in the comparator of ADC gets more sensitive, and offset error occurs due to the loading effect presented, i.e., change in overall impedance at the comparator input side. For an N-bit flash-type ADC, the required number of comparators and no. of bits is given as

$$\mathbf{n} = 2^{\mathbf{N}} - 1$$

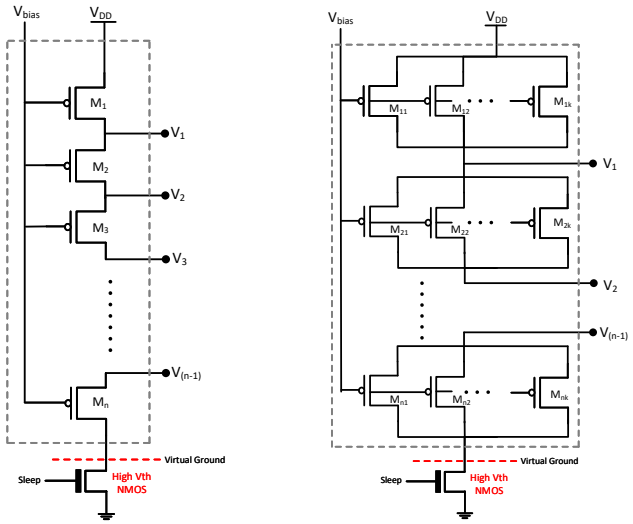
Here `n` represents the required number of a comparator in `N`-bit flash ADC and one can observe that number gets doubled for increasing 1-bit resolution. The analog to digital conversion accuracy depends on the reference voltage input at the ADC which needs to be more precise and must be error tolerant.

In order to address the reference circuit accuracy and robustness, the proposed work focuses majorly on energy efficiency and minimum variability due to loading effect. This section is divided into three sub-sections. In the first section, we have explained the proposed single and multiple parallel MOS based reference voltage

generation ladder circuit. The proposed circuit have entirely used PMOS transistors as it has shown better stability towards temperature variation and process & mismatch than NMOS based circuitry. Moreover, a single-level ladder circuit is efficient in terms of area and power consumption as depicted in 5.3(a). Still, it has certain snag, which are explained in the later stages, followed by its compensated multi-level reference generation circuit, which has shown excellent stability and it is least immune to variability. However, circuit comes with some constraint in terms of area utilization and power consumption. Subsequently Pareto analysis have been done for selection of number of parallel stages. In which compares both the previous designs. The insignificant loss in variability has achieved a significant amount of efficiency in terms of both area and overall power consumption of the entire ladder circuitry.

5.3.1 Parallel PMOS transistors Analysis in reference ladder circuit

We have replaced the resistors of a conventional resistive ladder with the PMOS device. Exploiting the basic feature of resistance of a MOS transistor, we can say that the effective resistance of any transistor depends on the type of the device, size and external biases of the transistor and also on the load of the overall circuit. The implemented design for voltage reference generations using PMOS have shown in Figure 5.3(a). Here, the $V_1, V_2, ..., V_{n-1}$ are the reference node voltages used for the comparison in the N-bit ADC. If the transistor is in the deep triode region then we can represent the transistor as a voltage-controlled resistor, that is, with $V_{DS} << 2\ (V_{GS} - V_{th})$. An equivalent resistance `R` in the case can be given by the ratio between the drain to source voltage and the drain current shown in Eq 5.1. We have not shown the term $0.5V_{DS}^2$ in the above equation due to its insignificant impact. However, the equation is valid in the triode region for $V_{DS} < 0.2(V_{GS} - V_{th})$ which is typically smaller than $0.15V$ for deep submicron CMOS technologies. Whereas, in the saturation region, the resistance can be estimated as the ratio between V_{DS}



(a) Single-level reference generator MOS ladder

(b) Multi-level voltage reference generator ladder.

Figure 5.3: Output voltage is taken out as V1, V2, ..., V(n-1). Output voltage can be varied by changing the (W/L) ratio of the respective parallel transistors of a single reference level.

and I_D having both V_{DS} and V_{GS} equal to the power supply i.e. V_{DD} and therefore an equivalent resistance is shown in Eq. 5.2.

$$R = \frac{1}{K'_p(\frac{W}{L})_p(V_{GS} - V_{thp})} \tag{5.1}$$

$$R = \frac{V_{DD}}{\frac{1}{2}K'_p(\frac{W}{L})_p(V_{GS} - V_{thp})^2} \tag{5.2}$$

An empirical estimation was adopted for the resistances of the NMOS and PMOS devices in estimating the charging and discharging delay time with the dependence between equivalent resistance and (W/L)n and (W/L)p being inversely proportional which is observed in [31]. According to the approximation, the equivalent resistances of the NMOS and PMOS devices have represented using Eq. 5.3(a) and 5.3(b)

respectively and are given by:

$$R_N = \frac{12.5}{(\frac{W}{L})_n} K\Omega \tag{5.3a}$$

$$R_P = \frac{30}{(\frac{W}{L})_p} K\Omega \tag{5.3b}$$

The factors 12.5 and 30 depend on the technology. These values works well for most of the CMOS technology including 0.25um, 0.18um, and 0.13um [32]. Several parameters such as threshold voltage and the short channel effects were not taken into account but although estimating the performance of design using these equivalent resistance is simple. So, depending upon the size and the control voltage applied between source and drain, the overall resistance of the PMOS device and of the ladder gets changed.

We have observed MOS ladder having a single PMOS for each reference generation has more variability with the process parameters. Keeping in mind the importance of variability of any reference generation circuit and especially in the RRAM based circuit, we increased the number of MOSs in the reference levels of the ladder. With the increase in number of levels the overall stability of the circuit was improving. So more the number of levels that were increased, more number of control parameters we have gained such as sizing of each transistor in a single row of multiple levels. The circuit with less variability is shown in Figure 5.3 (b). Here we have used more than one PMOS transistor in Parallel for each reference node (i.e $1, 2, \ldots k$). It is observed that with increasing parallel transistors circuit show more endurance with variation in process parameters. However, more transistor comes with area overhead. In order to make the design efficient in all aspect, we have analysed the Pareto points.

5.3.2 Pareto analysis for finding parallel MOSs in reference circuit

The high level RRAM system model is shown in Figure 5.4. On can observed that the accumulate input at the ADC is depends on the size of Resistive RAM array.

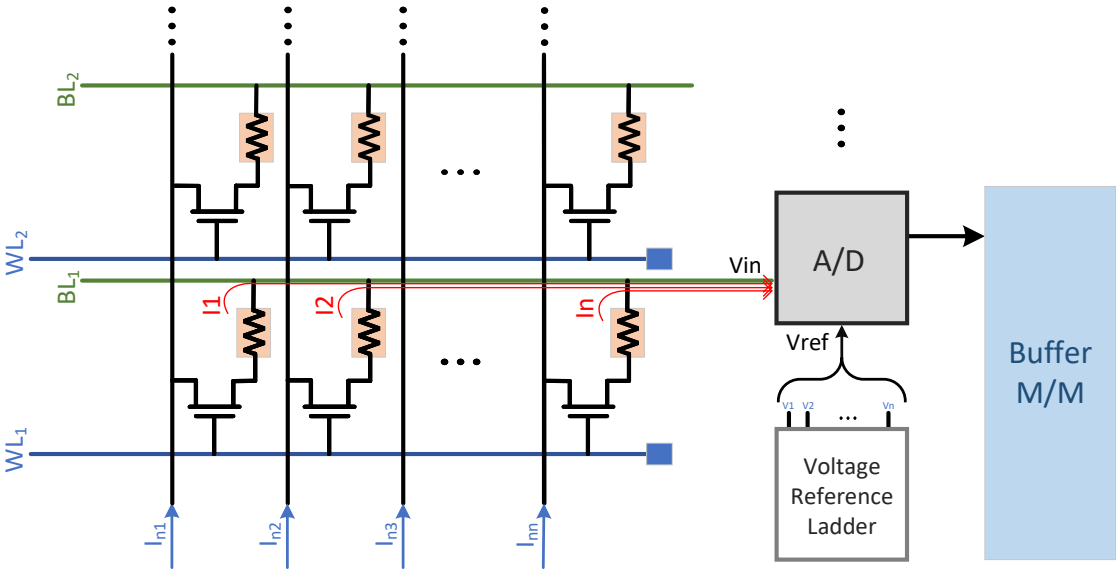


Figure 5.4: Output generated from RRAM array goes to ADC and gets compared with the reference voltage generated from the proposed Voltage reference ladder block.

Therefore with increasing size of array and impedance variability at RRAM shows the loading effect at reference voltage of ADC as depicted in Figure 5.4. In order to analyse parallel MOS versus reference voltage variability at desirable load, we perform a Pareto analysis. The single transistor based ladder is unable generate all the required reference voltages due to constraint of W/L aspect ratio. Furthermore, it has more variability with respect to change in output load. The limitations are overcome by implementing parallel transistor. There is a trade-off between Area, Power, and variability. Though Pareto analysis shows with insignificant compromise in the variability, a significant amount of area and power have been reduced. The two parallel transistors have used to generate the all reference node voltages that required at analog to digital conversion. The detail analog analysis and design parameters has explained in Section 5.3.3. We have operated all the MOS in the triode region. Simplified circuit architecture have shown in Figure 5.5.

5.3.3 Small signal analysis of the two parallel MOS in node voltage

For an N-bit flash-type ADC, the required comparators increases exponentially i.e, n=2^{N-1} where n is the number of voltage comparators. The proposed work focuses majorly on energy efficiency along with minimum variability. We have consider the both the MOSFET's are in operated in linear region. The equation of current (I_D) in linear region at each reference node is shown in Figure 5.3 (a) is elaborated bellow.

$$I_D = \mu_p C_{ox}\Big(\frac{W}{L}\Big)\Big[(V_{GS} - Vth).V_{DS} - \frac{1}{2}V_{DS}^2\Big] \tag{5.4}$$

Furthermore, the transconductance of each transistor depends on the gate input voltage (V_G) and source node voltage (V_x). The transconductance of each transistor can be evaluated using below Eq 5.5:

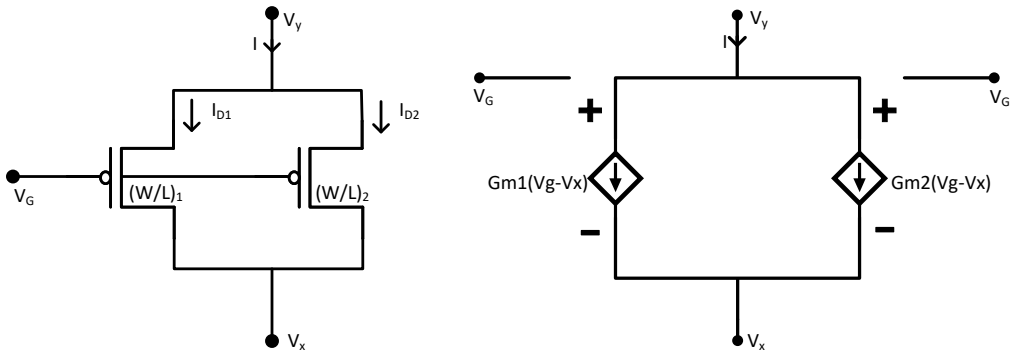
$$G_m = \frac{I_D}{V_G - V_x} \tag{5.5}$$

The resultant current (I) passing through the PMOS-based reference circuit shown in Figure 5.3 (b) have been calculated using bellow equation as:

$$\begin{aligned} I = {} & \mu_p C_{ox}\Big(\frac{W}{L}\Big)_1\Big[(V_{GS} - V_x - Vth).(V_y - V_x) - \frac{1}{2}(V_y - V_x)^2\Big] \\ & + \ \mu_p C_{ox}\Big(\frac{W}{L}\Big)_2\Big[(V_{GS} - V_x - Vth).(V_y - V_x) - \frac{1}{2}(V_y - V_x)^2\Big] \\ & + \ \ldots \end{aligned}$$

At lower technology, the circuit is more sensitive for process, mismatch, and variation in environmental parameters. Further, it is dominating at higher resolution. Therefore, the reliability of transconductance is variable with applied voltage variation and current through the MOS transistor. The equation for transconductance is given bellow as

$$G_m = \frac{\partial I}{\partial(V_G - V_x)}$$



(a) Reference generator parallel MOS ladder. Considering V_y applied as input and V_x is taken as output voltage.

(b) Equivalent small-signal model where both the MOS transistors are replaced by their GM equivalents.

Figure 5.5: Single section of the proposed 2-level MOS based circuit. VG is the voltage applied at the Gate terminals of both the PMOS transistors. Here, GM1 and GM2 are the transconductance of the M1 & M2 transistors respectively.

We have discussed about the edge of parallel MOSs in ladder returns less variability , and the overall conductance for the K^{th} parallel MOS is evaluated using Eq. 5.6. We can observe that overall transconductance depends on the number of MOS in the parallel. Therefore, we can achieve as minimum reference voltage by implicating many transistors unlike the case of reference generation using single transistor.

$$G_m = \mu_p C_{ox}\Big(\frac{W}{L}\Big)_1\Big(V_y - V_x\Big) + \mu_n C_{ox}\Big(\frac{W}{L}\Big)_2(V_y - V_x) \;+\; \ldots \tag{5.6}$$

From the Pareto points we have analysed that using two transistors we can achieve required minimum reference node voltage by changing the W/L of both the transistors. In this respect two parallel MOS is shown in Figure 5.5 (a) and the small signal model of efficient proposed circuit model in Figure 5.5 (b). Here, V_y and V_x are the upper and lower node voltage respectively for each reference. Therefore, for Figure 5.5, we can evaluate the total conductance at the each reference using Eq. 5.7. Systematically, we analysed W/L for both the PMOS and generated sixteen node voltages (i.e $V_1, V_2, \ldots V_{16}$). The reference voltages have used as a input to two-stage comparator considering as a load at reference point and output reference

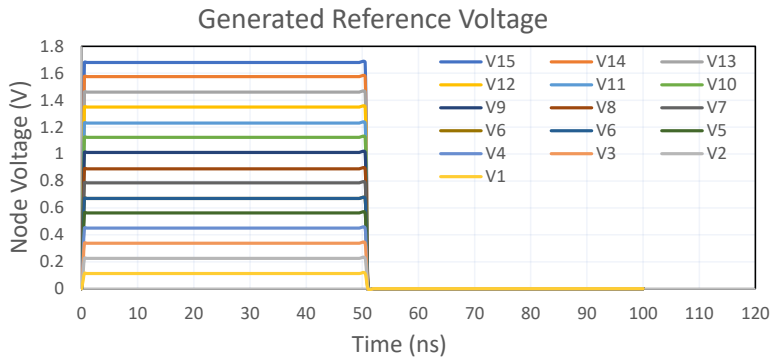


Figure 5.6: Parallel generated reference voltage output from V1-V15. 1.8V is divided into 16 equal steps having minimum step size as $112.5mV$.

voltages observed as shown in Figure 5.6.

$$G_m = \mu_p C_{ox}(V_y - V_x)\Big[\Big(\frac{W}{L}\Big)_1 + \Big(\frac{W}{L}\Big)_2\Big] \tag{5.7}$$

5.3.4 Low power and robust voltage reference ladder circuit

A power supply's primary objective is to regulate the output voltage at a desired fixed value even when variations in input voltage or load current. Therefore, load and line regulation are performance parameters of a power supply. Making the circuit robust for the input variation current mirror circuit has efficiently used and evaluated the performance analysis. The current mirror circuit copies or mirrors the current flowing in one active device in another, keeping the output current constant regardless of loading.

The proposed efficient design have shown in Figure 5.7. The design should be less power-consuming concerning static and dynamic power. Static power accounted for many things concerning a device (CMOS or process technology). The major thing a designer can do to reduce it is leakage power reduction through power gating (power shut off) with the help of Common Power Format (CPF) and Unified Power Format (UPF). Power gating is a technique in which we shut off power to a domain reference ladder V_{DD} to ground connection when that is not in use; thus overall saving power on a chip. The power gated transistor with high V_{th} is shown in Figure 5.7.

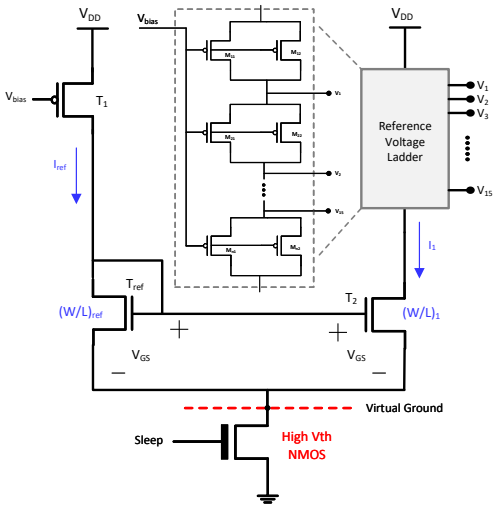


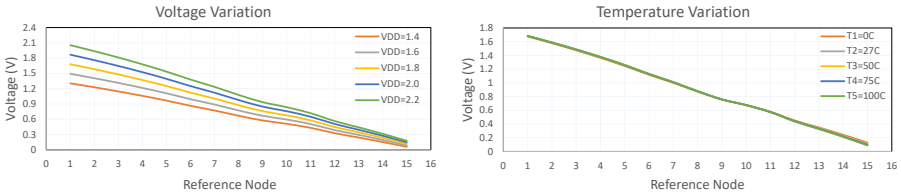
Figure 5.7: Proposed 2-level reference generator MOS ladder having optimized (W/L) ratio in order to minimize variations caused by LE. Power gating technique is also applied for the Ladder block. The CLK signal is used to isolate the circuits from power supply.

5.4 Simulation Results and Discussion

To validate the design, we perform the simulation at different abstraction levels. The results evaluation and observation are discussed in the following subsections.

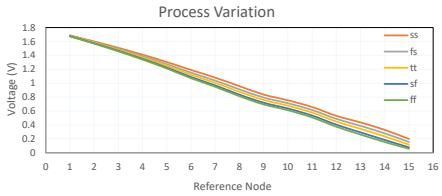
5.4.1 Process-Voltage-Temperature (PVT) variations impact

On-Chips circuit may suffer from these variations due to process, voltage, or temperature change, making the transistor have differing performances. The overall design has been evaluated for (Process, Voltage, Temperature) PVT variations. From Figure 5.8 (a), it can be observed that for supply voltages ranging from 1.4V to 2.2V, the maximum variation detected was approximately 7% from its expected value. To achieve the same performance from a circuit at a higher temperature, it consumes more power than a lower temperature in order to evaluate the circuit stability for temperature variation. In Figure 5.8 (b), temperature variation



(a) Node voltages variation due to $\pm 10\% V_{DD}$(V) supply.

(b) Node voltages variation due to on-chip Temperature (°C).



(c) Node voltages variation at different Process corners (ss, fs, tt, sf, ff).

Figure 5.8: On-chip generated reference node voltage variation at various Voltage, Temperature, and Process corners. At lower node, voltages variation is less than the high potential reference nodes. Whereas, Variation is large at lower node voltages as compared to high potential reference nodes.

can be seen, which was least detected at higher potential and maximum at lower potential ($< 1\%$). The design has also been tested to run on different process corners, and it can be observed from Figure 5.9 similar to temperature variation. The maximum variation detected was 3.47% from its expected reference value, decreasing towards a higher potential. Further, during the chip manufacturing process at the foundry, there are minute attribute variations in transistors considering oxide thickness, length, and wealth. Therefore, the circuit has been simulated at different process corners. We have depicted in Figure 5.8 (c), the performance at both slow(s) and fast(f) process corners, i.e., tt, ss, sf, fs, and ff. One can observe that lower voltage nodes are more sensitive to process variation than the higher voltage node.

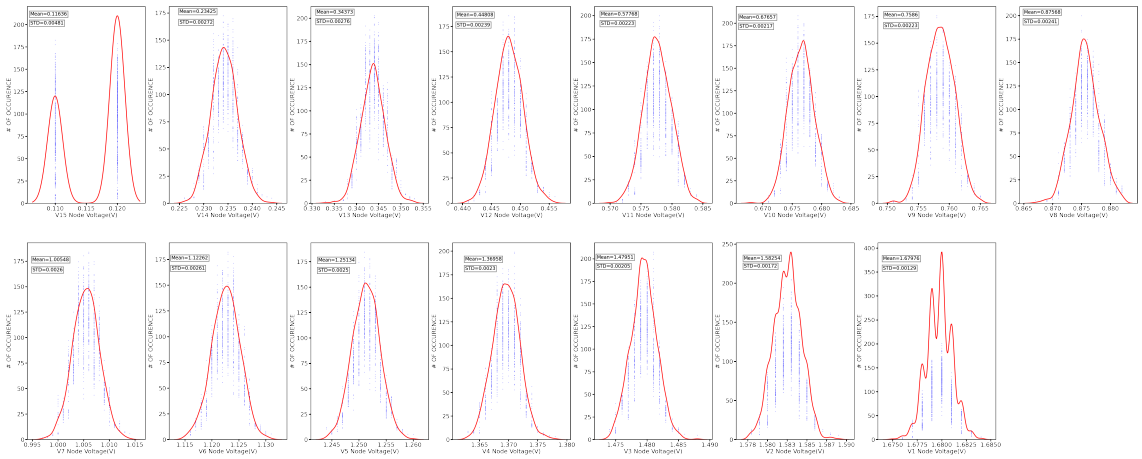


Figure 5.9: Monte-Carlo Simulation for Process variation and Mismatch. Voltage variation from Node V_1 to Node V_{15} is shown, and mean along with standard deviation at each node is also calculated and shown in the graph.

5.4.2 Monte-Carlo for Process-Variation and Mismatch Analysis

Process variation is the device and peripherals characteristics such as length, widths, oxide thickness when the integrated circuits are fabricated. The amount of process variation becomes particularly pronounced at smaller process nodes (below $180nm$) as the variation becomes a more significant percentage of the total length or width of the device and as feature sizes approach the fundamental dimensions during the lithography masks. Therefore, process variation and device mismatch have a significant role in circuit physical parameters' stability and reliability at lower technology. There is a substantial variation in static current due to process variation and mismatch. Monte-Carlo simulation calculates the probabilistic distribution of device conductance variation due to process variation and device mismatch in the characteristics of similar design devices, which occur during the manufacturing of IC's. At lower technology, MOS triode region characteristics are more sensitive for naturally occurring variation. Therefore, the Monte-Carlo simulation is carried out for 1,000 samples to validate the power variation due to process and mismatch.

The Mont-Carlo simulation is performed in *Virtuoso-Cadence* for 1000 samples

Table 5.1: Performance parameters of proposed MOS based resistive ladder and state-of-the-art designs for reference generation in 4 bit resolution ADC applications.

Physical Parameter	**[33]**	**[34]**	**[35]**	**[36]**	**Proposed**
Technology (μm)	0.18	0.18	0.18	0.25	0.18
Supply Voltage (V)	1.25 - 1.8	$\geqslant$1.2	2.3	1.2-3	1.4 - 2.2
Power Consumption (μW)	0.67	0.23	1.4-32.7*	1.1	0.7
Temperature ($^{\circ}C$)	-40 - 85 (Room temp.)	0 - 80 (Room temp.)	0 - 100 (Room temp.)	-20 - 100 (Room temp.)	0 - 100 (Room temp.)
Line Regulation (%/V)	7.5	0.58	0.13**	$\pm$0.60	0.02
Load Regulation (%/V)	-	0.25	NA	-	0.01

for all 15 reference voltages generated. The simulation result is extracted at *180nm* technology node. The proposed design has less dynamic power variation and standard deviation. The mean dynamic power and σ deviation at 180nm technology for reference V_{15} is 115.4mV and 2.98mV respectively and for all other 14 nodes deviation and mean value is shown in the individual graphs. To draw conclusion on individual basis, all the graphs from node V_1 to node V_{15} is shown in Figure 5.9.

5.4.3 Physical performance parameters evaluation of proposed MOS-based resistive ladder and comparison with the state-of-the-art

The proposed circuit was designed using 180nm standard CMOS process. We have designed the circuit to generate voltages of equal with maximum accuracy. For 4-bit reference voltage generation, we require 16 quantization levels and thus, complete voltage range needs to be divided into equal step size of 112.5mV for 1.8V as input. The overall power consumption came out to be 0.7uW at 1.8V power supply. Almost a constant voltage reference was obtained when analysed over varying factors such as temperature, supply voltage and process corners. Supply

voltage was varied for $V_{DD} \pm 10\%$. The line regulation observed was 0.02%/V while the load regulation was 0.01%/V which is compared with the state-of-the-art designs as shown in Table 5.1. Our design was able to achieve an appropriate trade-off between area, power consumption and overall stability.

5.5 Summary

RRAM-based architectures require high-speed detection of generated current, which is very low in magnitude(uA) and with minimum variations. In this paper, a low-power, high-speed reference generation circuit is designed using SCL 180nm technology at the supply voltage of 1.8V. The proposed design provides better results and low power consumption for Flash-type ADC and addresses the Loading effect for RRAM-based circuits. Hence, the proposed reference generation circuit can be used in any RRAM array to provide accurate results with Flash-type ADC. The power gating technique has also been used to avoid static power loss and minimize the power requirements. The temperature-compensated voltage reference generation circuit has been presented. The proposed circuit is useful as a voltage reference circuit for low-power LSIs. In future work, low-power startup circuitry can be designed, which can be used along with a voltage reference generator in several RRAM-based circuit applications to further improve efficiency.

Chapter 6

Conclusion

Towards the efficient design and implementation of next-generation ALUs in RRAM-based computational memories, this work highlighted some promising design concepts to consider, introducing a segmented NV-SRAM (array) that uses an augmented peripheral circuitry to improve logic latency of non-stateful logic schemes, where computations are performed via modified memory read operations. Alternative designs for the in-memory circuits were proposed that were proved to be robust in the presence of device-to-device variability in memristors. We identified the set of all supported primitive operations/instructions of the proposed computational memory system and addressed system-level design issues towards the design of a ReRAM-based general-purpose computational memory with ALU functionality. Circuit simulation results validated the functionality of the designed system, which demonstrated important performance improvements over other state-of-the-art in-memory computing approaches both for elementary logic operations and for other combinational operations.

6.1 Future scope of work

Logic-in-memory computing designs using RRAM has been first proposed in this work. Later, we moved on to the designing of Non-volatile SRAM cell and array design. However, there are certain steps that may be taken in the future to increase the quality of the planned work. The following are some of the most important points in this regard:

1. In this thesis work we focused on the CMOS custom design approach for memory architecture design. The proposed $NV - SRAM$ can be further investigated based on semi custom VLSI design flow and hardware implementation.

2. NV-SRAM memory architecture can be applied to a variety of applications, including sensor applications in distant places, smart energy metres, and other high-speed applications since it provides instant non-volatility.

3. For the study, we employed the $130nm$ and $45nm$ technology nodes. For further research, recent technological nodes like as $22nm$, $12nm$, and beyond can be employed.

Finally, we conclude that the goal of building an efficient logic-in-memory cell using RRAM has been accomplished.

Bibliography

[1] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device scaling limits of si mosfets and their application dependencies," *Proceedings of the IEEE*, vol. 89, no. 3, pp. 259–288, 2001.

[2] S. A. McKee, "Reflections on the memory wall," in *Proceedings of the 1st conference on Computing frontiers*, 2004, p. 162.

[3] M. B. Kamble and K. Ghose, "Analytical energy dissipation models for low power caches," in *Proceedings of 1997 International symposium on low power electronics and design*. IEEE, 1997, pp. 143–148.

[4] H. A. Du Nguyen, J. Yu, L. Xie, M. Taouil, S. Hamdioui, and D. Fey, "Memristive devices for computing: Beyond cmos and beyond von neumann," in *2017 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*. IEEE, 2017, pp. 1–10.

[5] P. A. Packan, "Pushing the limits," *Science*, vol. 285, no. 5436, pp. 2079–2081, 1999.

[6] K. Kim, "Future memory technology: challenges and opportunities," in *2008 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*. IEEE, 2008, pp. 5–9.

[7] Y. H. Do, J. S. Kwak, J. P. Hong, H. Im, and B. H. Park, "Nonvolatile unipolar and bipolar resistive switching characteristics in co-doped tio2 thin films with

different compliance currents," *J. Korean Phys. Soc*, vol. 55, no. 2009, pp. 1009–1012, 2009.

[8] K. Ishibashi and K. Osada, *Low power and reliable SRAM memory cell and array design*. Springer Science & Business Media, 2011, vol. 31.

[9] G. Gielen and W. Dehaene, "Analog and digital circuit design in 65 nm cmos: End of the road?" in *Design, Automation and Test in Europe*. IEEE, 2005, pp. 37–42.

[10] I. Chakraborty, A. Jaiswal, A. Saha, S. Gupta, and K. Roy, "Pathways to efficient neuromorphic computing with non-volatile memory technologies," *Applied Physics Reviews*, vol. 7, no. 2, p. 021308, 2020.

[11] D. Shin and H.-J. Yoo, "The heterogeneous deep neural network processor with a non-von neumann architecture," *Proceedings of the IEEE*, vol. 108, no. 8, pp. 1245–1260, 2019.

[12] G. Raut, S. Rai, S. K. Vishvakarma, and A. Kumar, "Recon: Resource-efficient cordic-based neuron architecture," *IEEE Open Journal of Circuits and Systems*, vol. 2, pp. 170–181, 2021.

[13] H. H. Li, Y. Chen, C. Liu, J. P. Strachan, and N. Davila, "Looking ahead for resistive memory technology: A broad perspective on reram technology for future storage and computing," *IEEE Consumer Electronics Magazine*, vol. 6, no. 1, pp. 94–103, 2016.

[14] M. Cheng, L. Xia, Z. Zhu, Y. Cai, Y. Xie, Y. Wang, and H. Yang, "Time: A training-in-memory architecture for rram-based deep neural networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 5, pp. 834–847, 2018.

[15] Z. Zhu, H. Sun, Y. Lin, G. Dai, L. Xia, S. Han, Y. Wang, and H. Yang, "A configurable multi-precision cnn computing framework based on single bit rram,"

in *2019 56th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2019, pp. 1–6.

[16] Y. Osaki, T. Hirose, N. Kuroki, and M. Numa, "A 95-na, 523ppm/° c, 0.6-μw cmos current reference circuit with subthreshold mos resistor ladder," in *16th Asia and South Pacific Design Automation Conference (ASP-DAC 2011)*. IEEE, 2011, pp. 113–114.

[17] T. Chou, W. Tang, J. Botimer, and Z. Zhang, "Cascade: Connecting rrams to extend analog dataflow in an end-to-end in-memory processing paradigm," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 114–125.

[18] B. Wu, D. Feng, W. Tong, J. Liu, C. Wang, W. Zhao, and M. Peng, "Reram crossbar-based analog computing architecture for naive bayesian engine," in *2019 IEEE 37th International Conference on Computer Design (ICCD)*. IEEE, 2019, pp. 147–155.

[19] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams, "Dot-product engine for neuromorphic computing: Programming 1t1m crossbar to accelerate matrix-vector multiplication," in *2016 53nd acm/edac/ieee design automation conference (dac)*. IEEE, 2016, pp. 1–6.

[20] S. Naraghi, "A 4-bit analog-to-digital converter for high-speed serial links." 1984.

[21] S.-i. Gotoh, T. Takahashi, K. Irie, K. Ohshima, N. Mimura, K. Aida, T. Maeda, T. Yamamoto, K. Sushihara, Y. Okamoto *et al.*, "A mixed-signal 0.18/spl mu/m cmos soc for dvd systems with 432 ms/s prml read channel and 16 mb embedded dram," in *2001 IEEE International Solid-State Circuits Conference. Digest of Technical Papers. ISSCC (Cat. No. 01CH37177)*. IEEE, 2001, pp. 182–183.

[22] D. J. Foley and M. P. Flynn, "A low-power 8-pam serial-transceiver in 0.5 mum digital cmos," in *PROCEEDINGS OF THE IEEE CUSTOM INTEGRATED CIRCUITS CONFERENCE.* IEEE; 1999, 2001, pp. 123–126.

[23] S. Banik, D. Gangopadhyay, and T. Bhattacharyya, "A low power 1.8 v 4-bit 400-mhz flash adc in 0.18/spl mu/digital cmos," in *19th International Conference on VLSI Design held jointly with 5th International Conference on Embedded Systems Design (VLSID'06).* IEEE, 2006, pp. 6–pp.

[24] S. S. Chauhan, S. Manabala, S. Bose, and R. Chandel, "A new approach to design low power cmos flash a/d converter," *International Journal of VLSI design & Communication Systems (VLSICS)*, vol. 2, no. 2, p. 100, 2011.

[25] J. Vandenbussche, K. Uyttenhove, E. Lauwers, M. Steyaert, and G. Gielen, "A 8-bit 200 ms/s interpolating/averaging cmos a/d converter," in *Proceedings of the IEEE 2002 Custom Integrated Circuits Conference (Cat. No. 02CH37285).* IEEE, 2002, pp. 445–448.

[26] Y. Li and E. Sanchez-Sinencio, "A wide input bandwidth 7-bit 300-msample/s folding and current-mode interpolating adc," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 8, pp. 1405–1410, 2003.

[27] G. Raut, A. P. Shah, V. Sharma, G. Rajput, and S. K. Vishvakarma, "A 2.4-gs/s power-efficient, high-resolution reconfigurable dynamic comparator for adc architecture." *Circuits, Systems & Signal Processing*, vol. 39, no. 9, 2020.

[28] J. Yoo, *A TIQ-based CMOS flash A/D converter for system-on-chip applications.* The Pennsylvania State University, 2003.

[29] P. Iyappan, P. Jamuna, and Y. S. Vijayasamundiswary, "Design of analog to digital converter using cmos logic," in *2009 International Conference on Advances in Recent Technologies in Communication and Computing.* IEEE, 2009, pp. 74–76.

[30] M. S. Njinowa, H. T. Bui, and F.-R. Boyer, "Novel threshold-based standard-cell flash adc," 2012.

[31] X. Li, J. Qin, and J. B. Bernstein, "Compact modeling of mosfet wearout mechanisms for circuit-reliability simulation," *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 98–121, 2008.

[32] K. Suzuki, S. Matsui, and Y. Ochiai, *Sub-half-micron lithography for ULSIs*. Cambridge University Press, 2000.

[33] S. S. Chouhan and K. Halonen, "A 0.67-μw 177-*ppm*/°c all-mos current reference circuit in a 0.18-μm cmos technology," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 8, pp. 723–727, 2016.

[34] M. Choi, I. Lee, T.-K. Jang, D. Blaauw, and D. Sylvester, "A 23pw, 780ppm/° c resistor-less current reference using subthreshold mosfets," in *ESSCIRC 2014-40th European Solid State Circuits Conference (ESSCIRC)*. IEEE, 2014, pp. 119–122.

[35] J. Lee and S. Cho, "A 1.4-μw 24.9-ppm/° c current reference with process-insensitive temperature compensation in 0.18-μm cmos," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 10, pp. 2527–2533, 2012.

[36] T. Hirose, Y. Asai, Y. Amemiya, T. Matsuoka, and K. Taniguchi, "Ultralow-power temperature-insensitive current reference circuit," in *SENSORS, 2005 IEEE*. IEEE, 2005, pp. 4–pp.

Publications

1. **Varun Bhatnagar**, Gopal Raut, and Santosh Kumar Vishvakarma. "Loading Effect Free MOS-only Voltage Reference Ladder for ADC in RRAM-crossbar Array." *In Proceedings of the Great Lakes Symposium on VLSI 2022*, pp. 199-202. 2022. DOI: 10.1145/3526241.3530354