

Resistive RAM based Compute-in-Memory Architecture for Content Addressable Memory

MS Research Thesis

By

RADHESHYAM MANOJKUMAR SHARMA



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
JUNE 2024

Resistive RAM based Compute-in-Memory Architecture for Content Addressable Memory

A Thesis

*Submitted in partial fulfillment of the
requirements for the award of the degree
of*

MS Research Thesis

by

RADHESHYAM MANOJKUMAR SHARMA



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
JUNE 2024



INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Resistive RAM based Compute-in-Memory Architecture for Content Addressable Memory** in the partial fulfillment of the requirements for the award of the degree of **Master of Science - Research** 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 July 2022 to June 2024 under the supervision of Dr. Santosh Kumar Vishvakarma, Professor, Indian Institute of Technology Indore, Indore, India.

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.



Signature of the Student with Date

(Radheshyam Manojkumar Sharma)

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

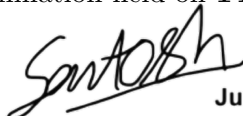

June 20, 2024

Signature of the Supervisor of MS-R Thesis with Date

(Prof. Santosh Kumar Vishvakarma)

Radheshyam Manojkumar Sharma has successfully given her MS - Research Oral

Examination held on **14/11/2024**


June 20, 2024

Signature of Supervisor of MS-R Thesis

Date: 20/06/2024

Signature of Convener, DPGC

Date:

ACKNOWLEDGEMENTS

I am immensely grateful to my MS Research thesis supervisor and mentor, Prof. Santosh Kumar Vishvakarma, for consistently encouraging and supporting me in both my research and personal growth. His unwavering belief in my abilities and his invaluable guidance have served as constant motivation, pushing me to exceed my own limits. I owe him a debt of gratitude for granting me the freedom to explore my research interests and allowing my novel ideas to flourish.

I would also like to extend my sincere appreciation to Prof. Trapti Jain, Co-ordinator of my thesis evaluation committee and evaluation committee. Their impartial evaluations and thought-provoking questions have contributed significantly to expanding my research perspective.

My family has played a major role in supporting my research work throughout the course of my master's. They have always boosted my confidence and always motivated me to push my limits. I will always be grateful to them for all their guidance, love and sacrifices. Their faith in me has brought me this far, and it will drive me further, as well, to achieve greater things. I deeply appreciate the Nanoscale Devices, VLSI Circuit System Design Lab (NSDCS) research group, especially Dr. Vishal Sharma, Mr. Narendra Dhakad, Mr. Ravi kumar, Mr. Akash Sankhe, Mr. Govindu Sathvik Reddy for their continuous support and guidance. I am also grateful to my friends and labmates, Mrs. Neha Maheshwari, Mr. Sonu Kumar, Mr. Shashank Singh Rawat, Mr. Mukul Lokhande, Mr. Vikash Vishwakarma, Mr. Ankit Tenwar, Ms. Komal Gupta, and Mr. Sagar Patel, whose camaraderie and encouragement made my time at the institute truly memorable.

Radheshyam Manojkumar Sharma

This Thesis is Dedicated

to

*My Parents, My Brother, My Grandparents
and the Almighty God*

ABSTRACT

The increasing demands of data-intensive applications necessitate high-speed, energy-efficient solutions that offer superior performance. Non-volatile memory devices, such as Resistive Random Access Memory (RRAM), have emerged as promising options for enhancing computing systems. In this thesis, we present an innovative 3T1R bitcell designed specifically for Binary Content Addressable Memory operations, implemented using 65nm CMOS technology. Our design achieves a 1.27x reduction in sensing latency and a 2.67x decrease in search energy consumption for a 64-bit word size compared to the current state-of-the-art. Additionally, the proposed bitcell demonstrates robust performance across various process corners and temperature variations, ensuring reliability in diverse operational environments.

This thesis also introduces a novel approach for designing Ternary Content Addressable Memory bitcells using a Hybrid CMOS-RRAM (4T2R) configuration, enhancing the conventional 2T2R cell by adding extra comparison transistors. This enhancement addresses signal mismatch issues and effectively maintains the precharged value for the match signal. The proposed bitcell achieves significant improvements in latency and energy consumption over existing designs, with a latency of 0.35 ns for a 256-bit word size and an energy consumption of 0.81 fJ/bit/search. It surpasses existing designs in terms of the energy-delay product by an impressive 26.85%. These performance metrics, determined using 65 nm CMOS technology, highlight the bitcell's versatility and effectiveness across various word sizes and applications. Overall, this work represents a significant advancement in TCAM design, offering enhanced speed and energy efficiency, which are critical for modern computing systems.

Contents

Abstract	i
List of Figures	iv
List of Tables	vi
List of Abbreviations	vi
1 Introduction	1
1.1 Background and Motivation	1
1.2 Emerging Nonvolatile Memory(eNVM) Devices	3
1.3 Overview of Compute-in-Memory	7
1.4 Research Objectives	8
1.5 Organization of Thesis	9
2 Literature Review	11
2.1 Overview to Content Addressable Memory	11
2.1.1 Existing CAM Implementation	12
2.1.2 Types of CAM	14
3 Resistive RAM Digital Content Addressable Memory Using Novel 3T1R Bitcell	17
3.1 Proposed Novel 3T1R digital bitcell	17
3.1.1 Structure of 3T1R nvCAM Cell	17
3.1.2 Bitcell Operation	18

3.2	BCAM row-wise search	20
3.3	Results and Discussion	21
3.3.1	Analysis of the impacts of process and temperature variation	21
3.3.2	Performance evaluation of novel 3T1R based array	22
3.4	Conclusion	23
4	HEART-CAM: Hybrid CMOS-ReRAM Based Energy Efficient And Rapid Ternary Content Addressable Memory	24
4.1	Introduction	24
4.2	Variability Model	27
4.3	Proposed Hybrid 4T2R bitcell Structure	30
4.4	Write Operation	31
4.4.1	Write for ‘0’	32
4.4.2	Write for ‘1’	32
4.4.3	Write For ‘X’ State	33
4.5	Search Operation	33
4.6	TCAM Architecture	35
4.7	Results and Discussion	36
4.7.1	Bitcell Simulation and Analysis	36
4.8	Example of 4x4 array using proposed bitcell	38
4.8.1	Precharge Stage	39
4.8.2	Evaluation Stage	40
4.9	Conclusion	43
5	Future Scope	44

List of Figures

1.1	Applications of emerging nonvolatile memories [1]	4
1.2	1T + 1R bit cell configurations of emerging NVMs	6
1.3	Processing unit and Conventional memory Vs Processing unit and Computational memory [4]	7
2.1	Typical NAND-Type Content-Addressable Memory. [5]	12
2.2	Typical NOR-Type Content-Addressable Memory. [5]	13
2.3	(a) Shows BCAM operates only on binary data (b) TCAM allows third matching state of X.	14
3.1	Proposed 3T1R bit cell (a), (b) write and read operation and their conditions, (c) match case equivalent, (d) mismatch case equivalent, and (e) conditions for XNOR functionality. [37]	18
3.2	Simulation waveforms of nvCAM operation for all four cases. using proposed 3T1R, here HRS=0 and LRS=1. [37]	19
3.3	nvCAM implementation using proposed 3T1R bitcell for 4x4 array. [37]	20
3.4	Bitcell performance for different word size (a) sense margin calculation of CAM operation for different crossbar sizes, (b) energy calculation of CAM operation for different crossbar sizes. [37]	21
3.5	Proposed bit cell performance analysis (a) match line voltage analysis of CAM for process corners, (b) match line voltage analysis of CAM for different temperatures. [37]	22
4.1	Mapping a logic function to BCAM and TCAM [34].	25

4.2	JART VCM 1b model (a) I-V characteristics (b) endurance characteristics.	28
4.3	Proposed Hybrid 4T2R bitcell Structure.	28
4.4	Signal conditions involved in storing data for proposed TCAM bitcell write operation.	30
4.5	Waveform of signal involved in storing data for proposed TCAM bitcell write operation.	31
4.6	Different cases for stored data and the search data (a-d) for search '0' case and (e-h) for search '1'.	32
4.7	Search conditions for store '0', '1', and 'X' data for proposed TCAM architecture.	33
4.8	Illustrating waveform for match and mismatch cases for stored '0', '1', and 'X'.	34
4.9	TCAM Architecture with $m \times n$ crossbar array	35
4.10	Search Latency of TCAM bitcell search 0 and search 1 delay for mismatch cases for all corners	36
4.11	Search Latency of TCAM bitcell search 0 and search 1 delay for temperature variation.	37
4.12	Histogram for latency distribution of TCAM bitcell for 1000 monte carlo simulations. The standard deviation is less than 60 ps for mismatch case.	38
4.13	Layout of the proposed hybrid-CMOS bitcell structure.	38
4.14	Detailed Analysis of a 4x4 Array Integrated with Self Reference Sensing Scheme (SRSS), Enhancing Efficiency and Performance in Memory TCAM using proposed bitcell.	39
4.15	Search waveform in a 4x4 Array configuration with exact match, 1 bit mismatch, 2 bit mismatch and 3 bit mismatch in search string.	40
4.16	Search energy and latency with different word size ranging from 32 to 256.	41

List of Tables

1.1	Device Characteristics of Mainstream & Emerging Memory Technologies. [1]	5
2.1	Qualitative Comparison of Key Properties of CAM Structures Realized With Nanoscale Devices. [2]	15
2.2	Quantitative Comparison of Different Nanoscale Memory Based CAMs. [2]	15
3.1	Performance comparison with state-of-the-art.	23
4.1	Simulation Parameters [11]	29
4.2	Performance comparison with state-of-the-art.	42

List of Abbreviations

DL	- Deep Learning
ML	- Machine Learning
AI	- Artificial Intelligence
DNN	- Deep Neural Network
VPN	- Virtual Private Networks
SRAM	- Static Random Access Memory
PCM	- Phase-Change Memory
ReRAM	- Resistive Random Access Memory
STT-RAM	- Spin-Transfer Torque RAM
FeRAM	- Ferroelectric RAM
IMC	- In-Memory Computing
CAM	- Content Addressable Memory
NVM	- Non-Volatile Memory
nvCAM	- non-volatile Content Addressable Memory
SSRS	- Self Reference Sensing Scheme
HRS	- High Resistance State
LRS	- Low Resistance State
SoC	- System-on-Chip
ASIC	- Application Specific Integrated Circuit
STA	- Static Time Analysis
IC	- Integrated Circuits

Chapter 1

Introduction

1.1 Background and Motivation

In today's era of digitization and artificial intelligence (AI), there is an ever-growing demand for efficient and high-performance memory architectures to support the processing of vast amounts of data has become a defining characteristic of modern computing. This surge in data volume poses significant challenges for traditional computing architectures, particularly in terms of processing speed, memory efficiency, and energy consumption. At the technological level, various scientific and technical factors have led to the gradual decline of Moore's Law, a principle that has driven the growth of the semiconductor industry for several decades [2]. At its core, Moore's Law states that the number of transistors in an integrated circuit (IC) doubles about every two years, resulting in increased computational power and performance. But when transistor sizes smaller dimensions, underlying physics and technology constraints become apparent. First, there are restrictions on the geometric scaling for a particular transistor structure due to the increasing prominence of quantum processes like electron tunneling and leakage currents. Second, as transistor sizes get smaller, power density rises and overheating problems arise, making heat dissipation a significant problem. Moore's Law has also slowed down as a result of the rising costs of constructing ever-more complicated fabrication facilities and the declining returns on performance advances per transistor. As the technology node gets closer

to the 1nm regime, it becomes harder to produce increases in performance and energy efficiency through transistor scaling[[3].

The von Neumann bottleneck increasingly limits system performance at the architectural level. The von Neumann bottleneck refers to a fundamental limitation in traditional computer architectures, where the speed of data transfer between the CPU and memory (the data bus) significantly constrains overall system performance. This bottleneck arises because the CPU can process data much faster than it can fetch it from memory, leading to inefficiencies and delays in computation. Named after John von Neumann, who proposed the architecture, this bottleneck highlights the mismatch between the processing power of modern CPUs and the relatively slower speed of memory access. In a von Neumann architecture, the CPU and memory are separate entities, connected by a data bus. This design requires that instructions and data be fetched from memory, executed by the CPU, and then written back to memory. As CPUs have become faster and more powerful, the bandwidth and speed of the data bus and memory have struggled to keep up, creating a bottleneck.

Key factors contributing to the von Neumann bottleneck include:

- **Limited Data Bus Bandwidth:** The data bus, which transfers information between the CPU and memory, has a finite bandwidth. As CPUs become faster, the bandwidth of the data bus often cannot keep up, leading to delays.
- **Memory Latency:** Accessing data from memory takes time (latency), which can slow down the CPU. While CPUs can execute billions of instructions per second, waiting for data from memory can cause significant delays.
- **Instruction Fetching:** In a von Neumann architecture, both program instructions and data share the same memory space and bus. This means that the CPU must fetch instructions and data sequentially, which can further slow down processing.

Efforts to mitigate the von Neumann bottleneck have led to several innovations, including:

- **Cache Memory:** Adding levels of cache memory between the CPU and main memory can help bridge the speed gap. Cache memory is much faster than main memory and stores frequently accessed data and instructions to reduce latency.
- **Parallel Processing:** Techniques such as pipelining, superscalar architecture, and multi-core processors allow CPUs to execute multiple instructions simultaneously, reducing the impact of memory latency.
- **Non-Volatile Memory:** Emerging non-volatile memory technologies, like ReRAM, can offer faster access times compared to traditional DRAM, potentially alleviating some of the bottlenecks.
- **Compute in Memory (CIM):** This innovative approach aims to bring computation closer to where the data is stored, reducing the need for data transfer between the CPU and memory. By integrating processing capabilities directly into memory devices, CIM architectures can significantly alleviate the von Neumann bottleneck.

Understanding and addressing the von Neumann bottleneck is crucial for developing more efficient and powerful computing systems, especially as the demand for processing large amounts of data continues to grow.

1.2 Emerging Nonvolatile Memory(eNVM) Devices

In modern computing systems, memory hierarchy plays a crucial role in managing data access and storage. The memory hierarchy as shown in Figure 1.1 typically consists of multiple levels, including registers, cache memory(SRAM), main memory (DRAM), secondary storage, and NAND HDD/SDD. Each level offers varying

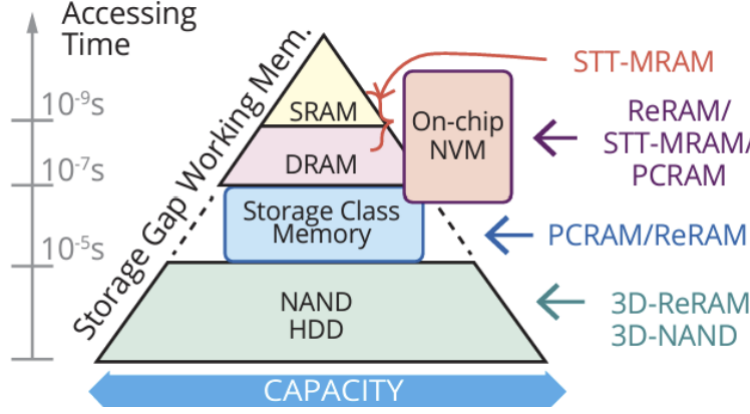


Figure 1.1: Applications of emerging nonvolatile memories [1]

degrees of speed, capacity, and cost, with the aim of optimizing overall system performance. Emerging non-volatile memory (NVM) technologies, such as ReRAM, hold great promise for revolutionizing the memory hierarchy. Unlike traditional volatile memory technologies like DRAM, NVM retains stored data even when power is removed, offering the potential for faster boot times, reduced energy consumption, and increased data persistence. Integrating ReRAM into the memory hierarchy enables new opportunities for enhancing system performance and efficiency. By leveraging the non-volatile nature of ReRAM, we can reduce the reliance on power-hungry volatile memory technologies and improve overall system reliability and resilience. Additionally, the recent development of developing NVMs toward low cost and high density has made it possible for them to be used in storage class memory and high density memory.

NAND flash memory is starting to take the place of hard disk drives as large-capacity nonvolatile storage due to its low cost per bit. The basic write mechanism based on quantum mechanical tunneling still limits the programming speed of NAND memory. To address the issues facing today's mainstream memories, emerging nonvolatile memory (eNVM) technologies like resistive memory (ReRAM), spin-torque transfer memory (STT-MRAM), and phase change memory (PCM) are being thoroughly investigated. Table 1.1 provides a summary of the figure-of-merit for several developing NVMs, NAND flash, and DRAM. When compared to traditional

Table 1.1: Device Characteristics of Mainstream & Emerging Memory Technologies.
[1]

	Mainstream Memories				Emerging Memories		
			Flash				
	SRAM	DRAM	NOR	NAND	STTMRAM	PCRAM	ReRAM
Cell Area	$>100F^2$	$6F^2$	$10F^2$	$<4F^2$ (3D)	$6 \sim 50F^2$	$4 \sim 30F^2$	$4 \sim 12F^2$
Multibit	1	1	2	3	1	2	2
Voltage	$<1V$	$<1V$	$>10V$	$>10V$	$<1.5V$	$<3V$	$<3V$
Read Time	~ 1 ns	~ 10 ns	~ 50 ns	~ 10 us	<10 ns	<10 ns	<10 ns
Write Time	~ 1 ns	~ 10 ns	10 us - 1 ms	100 us - 1 ms	<10 ns	~ 50 ns	<10 ns
Retention	N/A	~ 64 ms	>10 y	>10 y	>10 y	>10 y	>10 y
Endurance	$>1E16$	$>1E16$	$>1E5$	$>1E4$	$>1E15$	$>1E9$	$>1E6 \sim 1E12$
Write energy (J/bit)	~ 38 fJ	~ 10 fJ	~ 100 pJ	~ 10 fJ	~ 0.1 pJ	~ 10 pJ	~ 0.1 pJ

Note: F: feature size of the lithography.

NVMs, new NVMs often provide better write performance with quick speed and low power consumption. Among them, PCM and ReRAM offer benefits in terms of storage density, while STT-MRAM demonstrates exceptionally quick write operations and strong endurance.

Moreover, these new storage devices may serve as the foundation for cutting-edge applications like computing-in-memory, neuromorphic circuits, hardware security, hardware accelerator and nonvolatile logic. Emerging NVMs have been created in multiple bit cell configurations for various purposes. Figure 1.2 illustrates a typical one-transistor-one-resistor (1T1R) for PCRAM, STT-MRAM, and ReRAM. In this type of memory cell integration, the back-end-of-line (BEOL) method integrates the memory cells into the drain sides of transistors. Here, the cell access is controlled by word lines (WLs), and the write/read operations can be done by controlling bit lines (BLs) and source lines (SLs).

The 1T1R configuration is commonly adopted for embedded memory because of its;

1. Better selectivity in memory crossbar
2. High compatibility with CMOS process

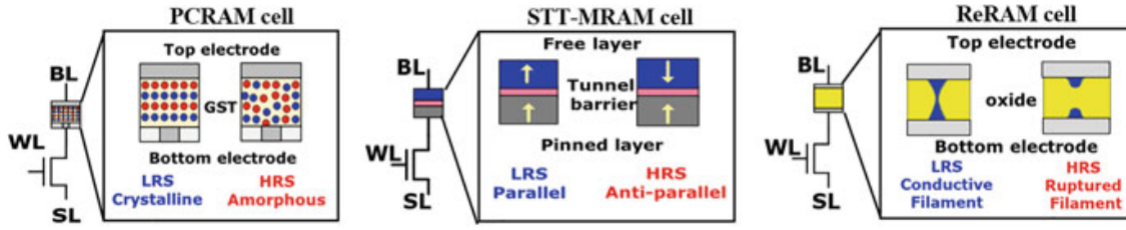


Figure 1.2: 1T + 1R bit cell configurations of emerging NVMs

3. Good immunity of write/read disturb
4. Efficient leakage current suppression

High-density storage, for instance, can help one-selector-one-resistor (1S1R), but certain diodes and/or selectors are needed. While the 1T1R array for energy-efficient systems is the primary emphasis of this chapter, it is important to note that the circuit concepts presented here can be applied to other designs as well. These technologies address the limitations of current memory systems and offer new possibilities for improving performance and efficiency. Their non-volatility ensures data retention without power.

- **Resistive Random-Access Memory (ReRAM):** ReRAM, also known as RRAM, operates by changing the resistance of a material to store data. It offers fast switching times, low power consumption, and high endurance, making it suitable for both storage and computational tasks.
- **Phase-Change Memory (PCM):** PCM stores data by changing the phase of a material between crystalline and amorphous states, which have different electrical resistances. PCM combines high speed, endurance, and non-volatility, making it a strong candidate for future memory applications.
- **Magnetoresistive Random-Access Memory (MRAM):** MRAM uses magnetic storage elements to store data. It offers high speed and endurance, along with non-volatility.

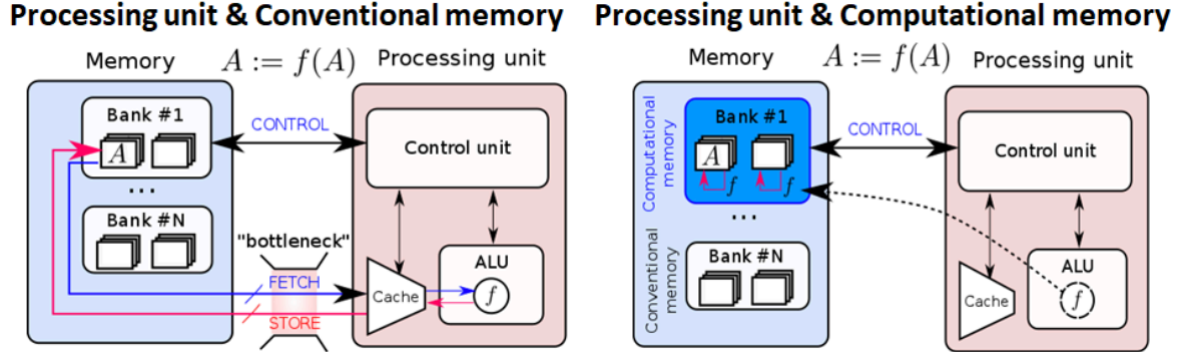


Figure 1.3: Processing unit and Conventional memory Vs Processing unit and Computational memory [4]

- **Spin-Transfer Torque RAM (STT-RAM):** A variant of MRAM, STT-RAM improves on traditional MRAM by using spin-polarized currents to alter magnetic states. This technology provides faster write speeds and lower power consumption.
- **Ferroelectric RAM (FeRAM):** FeRAM uses ferroelectric materials to store data. It combines the speed of DRAM with non-volatility and has lower power consumption. However, its density and scalability are currently lower than other emerging NVMs.

1.3 Overview of Compute-in-Memory

Compute in Memory (CIM) represents a transformative approach to computing that seeks to integrate computation directly within memory units, rather than relying solely on separate processing units.

This paradigm shift holds the potential to revolutionize memory-centric computing systems by minimizing data movement, reducing latency, and improving energy efficiency. At the heart of CIM lies the concept of performing computations in close proximity to data storage, thereby leveraging the inherent parallelism and high bandwidth of memory architectures. By executing computational tasks within memory cells themselves, CIM architectures can significantly accelerate data process-

ing tasks, particularly those involving large-scale data sets and complex algorithms. As illustrated in the Figure 1.3 it is not necessary to transfer data into a processor unit while using in-memory computing. Utilizing the physical characteristics of the memory devices, their array-level architecture, the peripheral circuitry, and the control logic, computing is accomplished. According to this paradigm, memory actively participates in the computational process. Because of the massive parallelism provided by a dense array of millions of nanoscale memory devices acting as compute units, in-memory computing offers the potential to improve the computational time complexity associated with specific tasks in addition to lowering latency and energy costs associated with data movement.

Resistive Random-Access Memory (ReRAM) emerges as a promising technology for realizing CIM architectures. ReRAM offers several advantages over traditional memory technologies, including non-volatility, low power consumption, and compatibility with existing CMOS fabrication processes. These properties make ReRAM well-suited for integrating computation with memory operations, enabling efficient CIM implementations. In the context of Content Addressable Memory (CAM), CIM based on ReRAM holds particular promise. CAM architectures, which facilitate rapid data retrieval based on content rather than explicit memory addresses, are integral to various computing applications such as database management, pattern recognition, and network routing. By incorporating CIM capabilities into CAM designs using ReRAM technology, researchers aim to enhance the speed, efficiency, and scalability of CAM operations.

1.4 Research Objectives

The primary objective of this research is to investigate and develop a novel Compute in Memory architecture based on ReRAM technology for Content Addressable Memory applications. The specific research objectives include:

- Designing a ReRAM-based CIM architecture tailored for CAM operations, with a focus on improving speed, energy efficiency, and area efficient compared

to conventional CAM designs.

- Developing efficient circuit designs and similarity search to enable computation within ReRAM-based memory cells, leveraging the unique properties of ReRAM devices.
- Implementing a architecture ReRAM-based CAM system and conducting comprehensive performance evaluations to validate the effectiveness of the proposed architecture in real-world scenarios.
- Exploring the potential applications and practical implications of ReRAM-based CIM for various computing tasks, such as pattern recognition, database searches, and AI inference.

1.5 Organization of Thesis

This thesis is organized into several chapters, each focusing on different aspects of the research conducted on Resistive RAM (ReRAM) based Compute-in-Memory architectures for Content Addressable Memory (CAM). The chapters are structured as follows:

Chapter 1: This chapter provides the background and motivation for the research, highlighting the increasing demands of data-intensive applications and the need for high-speed, energy-efficient solutions. It introduces non-volatile memory devices like ReRAM and their potential for enhancing computing systems. The chapter concludes with the research objectives and a brief overview of the thesis organization.

Chapter 2: The literature review offers a comprehensive overview of Content Addressable Memory, including existing CAM implementations and various types of CAM. This chapter sets the stage for the novel contributions of this research by discussing the limitations of current technologies and the potential improvements offered by ReRAM-based solutions.

Chapter 3: This chapter details the design and implementation of the novel 3T1R bitcell for Binary Content Addressable Memory (BCAM). It covers the structure of the 3T1R nvCAM cell, its operation, and the row-wise search mechanism. The chapter also presents results and discussions, including an analysis of the impacts of process and temperature variations and a performance evaluation of the 3T1R-based array.

Chapter 4: In this chapter, the focus shifts to the design of Ternary Content Addressable Memory (TCAM) using a Hybrid CMOS-RRAM (4T2R) configuration. It introduces the variability model and the proposed hybrid 4T2R bitcell structure. The chapter explains the write and search operations, and provides detailed results and discussions on bitcell simulation and analysis, including examples of arrays using the proposed bitcell.

Chapter 5: The final chapter summarizes the key findings and contributions of the research. It discusses the implications of the novel 3T1R and 4T2R bitcell designs for the future of CAM technology. Additionally, it outlines potential directions for future research to further enhance the performance and applicability of ReRAM-based CAM systems.

Chapter 2

Literature Review

2.1 Overview to Content Addressable Memory

Content addressable memory (CAM) draws inspiration from the biological brain, where data retrieval occurs independently of its location [2] [12]. Unlike conventional computer memory like random-access memory (RAM), which relies on indexing (addresses) for data access, the human brain stores information without needing indexes for retrieval. In the human brain, the content of the information, or a part of it, serves as the trigger for recalling information, leading to the concept known as CAM [13]. Content Addressable Memory (CAM) stands out as a specialized form of memory that allows for rapid data retrieval based on content rather than explicit memory addresses. This feature makes CAM particularly useful for applications where fast searching and pattern recognition are essential, such as database management, network routing, and image processing. However, traditional CAM designs, typically based on Static Random-Access Memory (SRAM) or ternary Content Addressable Memory (TCAM), face several limitations that hinder their effectiveness in modern computing environments. One of the primary challenges with traditional CAM architectures is their relatively high power consumption, which can limit scalability and increase operational costs, especially in large-scale computing systems. Additionally, these architectures often struggle to efficiently integrate computational tasks with memory operations, leading to sub optimal performance and resource

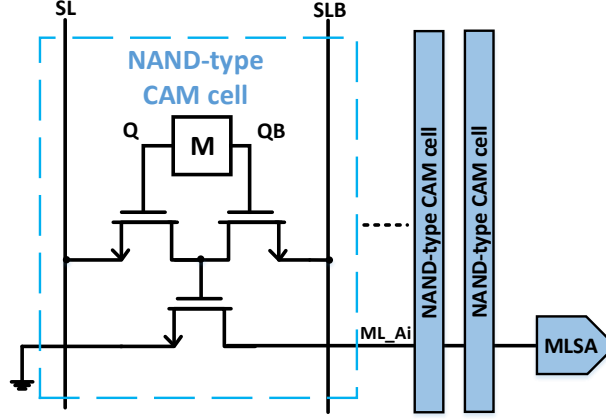


Figure 2.1: Typical NAND-Type Content-Addressable Memory. [5]

utilization. In response to these challenges, researchers have been exploring novel approaches to enhance memory performance and efficiency, with a particular focus on integrating computation directly into memory units.

2.1.1 Existing CAM Implementation

CAM operates as a fully associative memory receiving search patterns and providing the address of matched stored patterns. In the literature, it has been observed that fully parallel comparison architectures in CAM can result in significant power consumption [2] [13]. A primary contributor to power dissipation in such architectures is the charging of the match line (ML). Two prior typical CAM designs, i.e., the NAND and NOR types, are shown in Figure 2.1 and 2.2 [5], respectively. Memory Block “M” serves as a cell for storing data bits and is interchangeable with various types of memory cells. Many memory-based CAMs typically follow the NOR-type CAM design, wherein the match-line discharges when a mismatch occurs. Given that mismatches significantly surpass matches in typical CAM/TCAM applications (where only one memory row matches while the others mismatch), all match-lines must be pre-charged before each search/lookup. This process leads to considerable energy wastage. In contrast, the NAND-type CAM design ensures that only the matching row discharge, resulting in a substantial reduction in energy consumption during search/lookup by orders of magnitude. As an example, memory

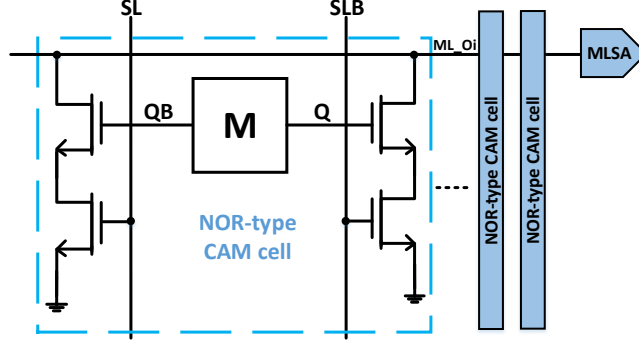


Figure 2.2: Typical NOR-Type Content-Addressable Memory. [5]

element (M) in Figure 2.1 and Figure 2.2 comprises two cross-coupled inverters, like the SRAM cell. Furthermore, regardless of the state of the match lines (ML_{Ai} and ML_{Oi}), the match line sense amplifiers (MLSAs) detect voltage changes on the MLs to determine the comparison result for each word. It's evident that the NAND-type CAM cell consumes less power but requires a longer search time, given that the ML is discharged through an extended transistor chain. Conversely, the power consumption of the NOR-type CAM cell surpasses that of the NAND-type CAM cell in exchange for higher speed. The NOR-type CAM cell features parallel discharge paths, enabling faster speed at the cost of increased power consumption.

Various types of memory have been selected for the implementation of CAM. When employing CMOS technology, CAMs designed with SRAM cells [5] [23] typically provide high speeds and reliability. However, they face challenges such as significant power dissipation and low integration density. Moreover, the non-zero standby power caused by leakage current poses issues for the adoption of SRAM CAM in battery-operated devices. To address these concerns, CAMs based on emerging non-volatile memories (NV-CAM) are being seriously considered as an alternative to SRAM based CAM. NV-CAM achieves zero standby power by reducing the number of transistors in a bitcell and effectively blocking leakage currents during the idle mode.

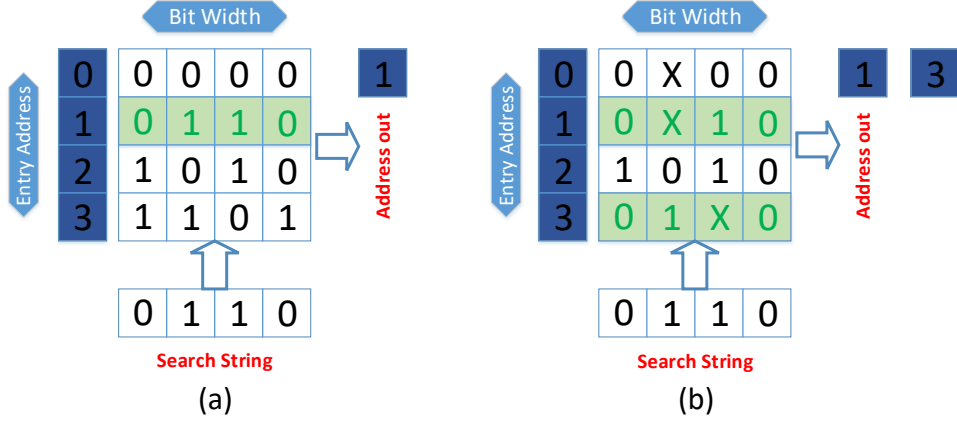


Figure 2.3: (a) Shows BCAM operates only on binary data (b) TCAM allows third matching state of X.

2.1.2 Types of CAM

As illustrated in Figure 2.3 CAM, including binary content addressable memory (BCAM) operates with binary data, necessitating exact binary matches during searches. It is well-suited for scenarios where precise binary pattern matching is essential, such as in packet classification in many network devices, such as routers and firewalls to perform services such as packet filtering, virtual private networks (VPNs), etc performing binary address lookups [15]. On the other hand, TCAM accommodates ternary data, allowing for the inclusion of don't care (X) bits in search patterns. TCAM's strategic use of "don't care" bits enhances match rates, reduces search delays, and minimizes power dissipation making it appealing for applications requiring high-speed searches such as internet packet forwarding, internet protocol routing, translation look aside buffer (TLB), tag directories in associative cache memories, database engines, image processing, pattern matching, and neural networks benefit from TCAM's high-speed search operation [17]. Additionally, increasing data generated by IoT applications poses a challenge for current computing systems [20]. General-purpose processors experience energy and performance inefficiencies when handling IoT applications like machine learning and multimedia. To address this, there is a growing need for computing systems capable of efficiently managing large volumes of streaming data.

Table 2.1: Qualitative Comparison of Key Properties of CAM Structures Realized With Nanoscale Devices. [2]

Nano-CAM Type	Non-Volatile	Area (Density)	Power		Latency	Scalability (size)
			Search/update	Leakage		
CMOS-CAM	No	Mod. (Mod.)	High	High	Moderate	Poor
STTRAM-CAM	Yes	Low (High)	Moderate	Negligible	Low	Good
ReRAM-CAM	Yes	Low (High)	Moderate	Negligible	Low	Good
DWM-CAM	Yes	Low (High)	Moderate	Negligible	Low	Good
FeRAM-CAM	Yes	Mod. (Mod.)	Moderate	Negligible	Moderate	Moderate

Table 2.2: Quantitative Comparison of Different Nanoscale Memory Based CAMs. [2]

Nano-CAM Type	Search	Area	Power		Endurance
	Delay	(Density)	Search/update	Leakage	
CMOS-CAM	1.077x	2.77x	1.33x	0.15x	1016
MTJ-CAM	2.24x	2.34x	1.49x	5.7x	1016
ReRAM-CAM	1x	1x	1x	1x	106
FeRAM-CAM	NA	NA	NA	NA	1014
PCRAM-CAM	1.81x	1.12x	1.48x	21x	109

The motivation behind this research lies in the potential of ReRAM-based CIM architectures to address the limitations of traditional CAM designs and unlock new opportunities for enhancing computing performance and efficiency. By developing innovative ReRAM-based CIM solutions tailored for CAM applications, researchers aim to overcome existing bottlenecks and enable more efficient and scalable memory systems capable of meeting the demands of emerging computing tasks, such as artificial intelligence, machine learning, and big data analytics.

Table 2.1 presents a qualitative comparison between the key characteristics of CAM structures used in CMOS and the previously stated future memory technologies. Phase-change memory (PCM), carbon nanotube field-effect transistors (CNTFETs), multigate transistor technologies like FinFET, and single electron transistors are some of the other cutting-edge technologies that have been envisaged for the realization of CAM. The quantitative comparison of the nano-CAMs with respect to area, speed,

energy, and endurance is shown in Table 2.2. It should be mentioned that ReRAM CAMs have lower durability than MTJ, PCM, and CMOS CAMs, but they can attain higher density, energy, and latency. [2]

Chapter 3

Resistive RAM Digital Content Addressable Memory Using Novel 3T1R Bitcell

The main contributions of our work are as follows:

- An in-memory matching circuit based on the proposed 3T1R cell topology for a low-latency parallel search.
- A thorough evaluation of the proposed bitcell, considering the impact of process and temperature.
- The performance evaluation for latency and search energy per bit for 4, 8, 16, 32 and 64 bit word size, using proposed 3T1R nvCAM.

3.1 Proposed Novel 3T1R digital bitcell

3.1.1 Structure of 3T1R nvCAM Cell

Figure 3.1 shows a structure of the proposed 3T1R nvCAM cell comprising an NVM device (ReRAM [11]), cell selector M1 (forming 1T1R) and an inverter using

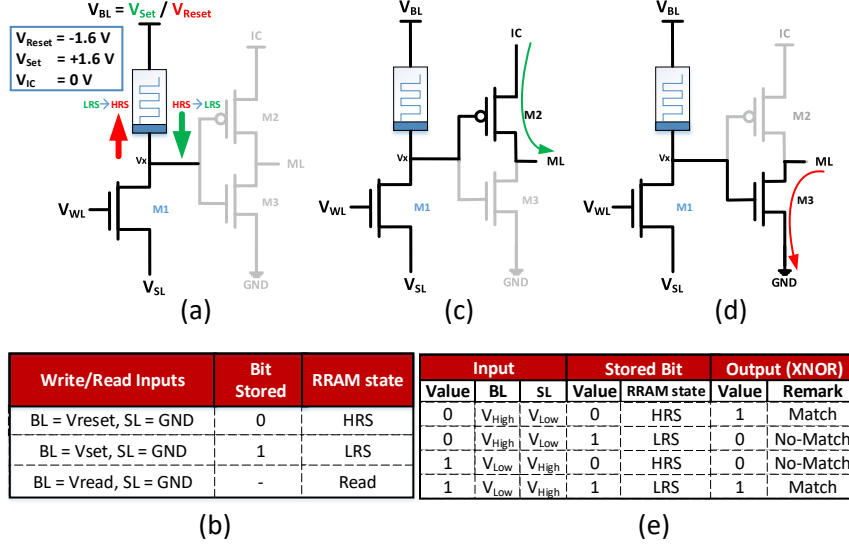


Figure 3.1: Proposed 3T1R bit cell (a), (b) write and read operation and their conditions, (c) match case equivalent, (d) mismatch case equivalent, and (e) conditions for XNOR functionality. [37]

M2 and M3 at the drain of M1. An inverter with 1T1R bitcell prevents the crosstalk between adjacent cells in a larger array. The V_x acts as the input of the inverter. The 3T1R cell can be designed using less area than necessary for 2-NVM-based nvCAM cells by utilizing a single NVM device [8]. Moreover, it lowers the 3T1R cell's NVM-write energy consumption to levels below those of 2-NVM nvCAM cells. Inverter Control (IC) signal used to disable the inverter during writing operation to reduce the overall power consumption. We propose an in-situ XNOR boolean logic functionality for content addressable memory applications.

3.1.2 Bitcell Operation

3.1.2.1 Write and Read Operation

ReRAM typically exhibits a binary behavior HRS and LRS. Switching is achieved by applying appropriate V_{Set} and V_{Reset} voltages. In the SET operation, the word line (WL) activates the bitcell, and a positive voltage is applied across the bit line (BL) and the select line (SL). This causes a conductive filament in the oxide layer,

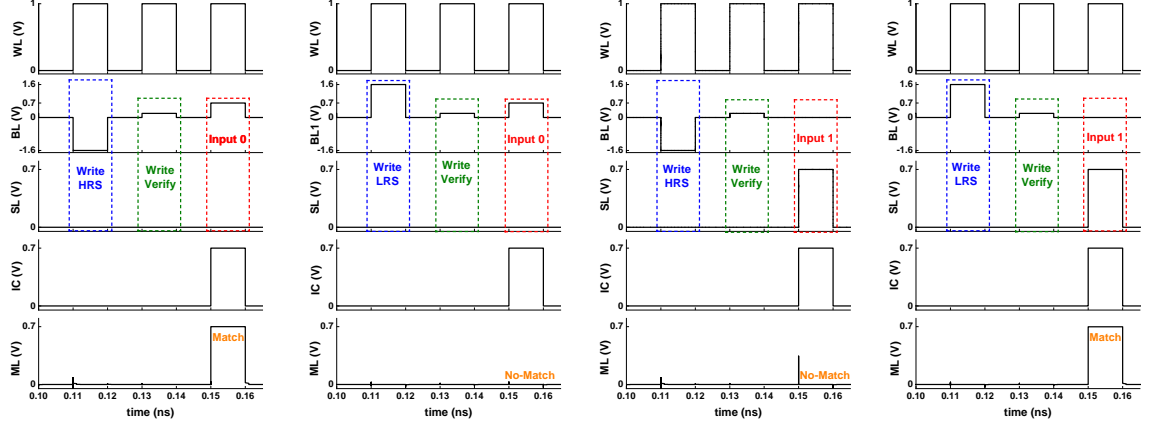


Figure 3.2: Simulation waveforms of nvCAM operation for all four cases. using proposed 3T1R, here HRS=0 and LRS=1. [37]

leading the device to switch from HRS to LRS. On the other hand, the RESET operation involves applying a negative voltage, which breaks the conductive filament and returns the device to the HRS. The proposed 3T1R nvCAM eliminates the need for program verification for medium-capacity macros by utilizing ReRAM devices with a high R-ratio. Inverter is disabled ($V_{IC} = 0V$) for write operation. The ReRAM used here can switch between HRS and LRS with R-ratio of 44 is achieved with set and reset voltages of 1.6V and -1.6V, respectively. During the operations, HRS is encoded as stored 0, and LRS encoded as stored 1 bit. After every write operation, a small read voltage of 200mV is applied to verify the stored bit. Hence writing is similar to the 1T1R ReRAM array [11].

3.1.2.2 In-situ XNOR logic for search operation

Figure 3.1 (e) shows the conditions, and Figure 3.2 illustrates waveforms associated with the input pattern search operation. The binary input 0 is encoded as $V_{BL} = \text{High}$ and $V_{SL} = \text{Low}$, 1 as $V_{BL} = \text{Low}$, and $V_{SL} = \text{High}$. For search operation, V_{WL} and V_{IC} are enabled. As illustrated in Figure 3.1 (b), when the input and stored bit are matched, V_x is below the inverter's switching threshold, M2 pulls ML up to the V_{IC} . When the input and stored bit are not matched, V_x is above the inverter's switching threshold, M3 pulling ML down as shown Figure 3.1 (c). Therefore bitcell

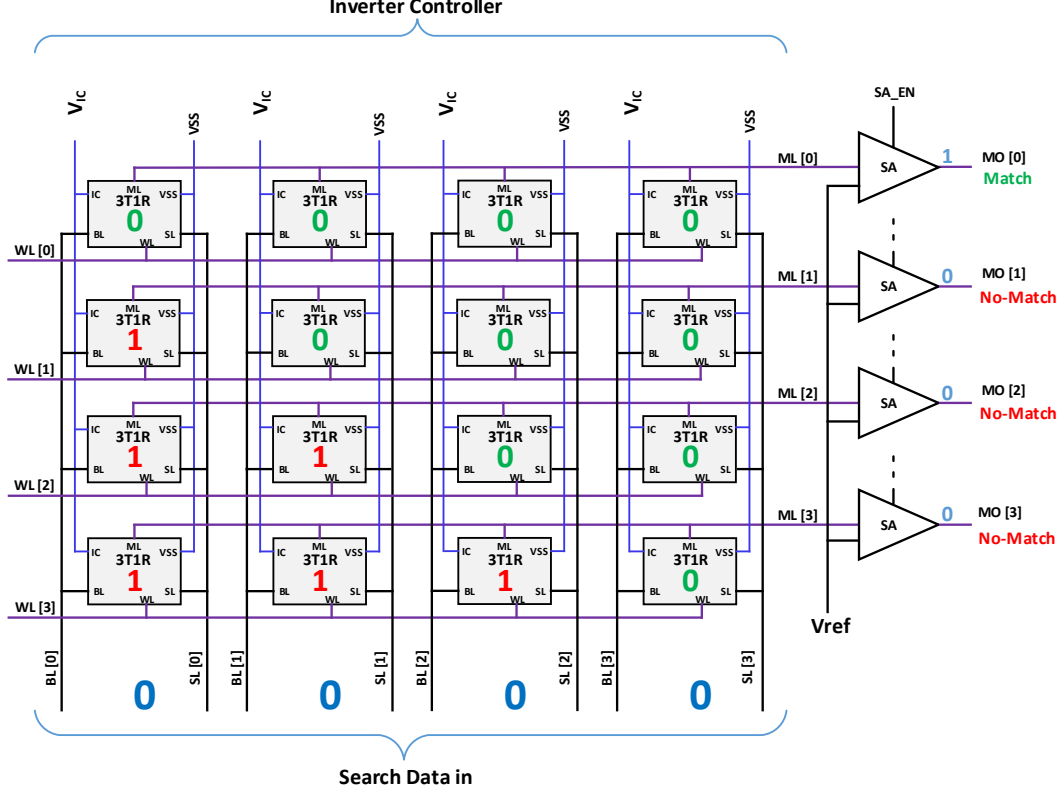


Figure 3.3: nvCAM implementation using proposed 3T1R bitcell for 4x4 array. [37]

shows in-situ XNOR behavior. We use this specific behavior to mimic the bit-level matching operation in CAM.

3.2 BCAM row-wise search

Figure 3.3 shows a 4x4 array example of pattern search. The input bits are compared with the data stored in each row of the memory array. If the input bit is matched, then ML pulls up to V_{IC} , and for unmatched, the ML is pulled down to VSS. Only the data in the first row match the search data-in. Row 2, 3, and 4 have 1, 2, and 3 bit unmatched. The operating principle is described as a competition between the pull-up network (PUN) and pull-down networks (PDN) of inverters of the bitcell. For the full match, i.e., row1, all the pull-up networks are enabled, whereas for 1-bit mismatch, i.e., row2 has a 1 PDN and 3 PUN networks are enabled, which restricts the output from reaching V_{IC} . PUN and PDN network forms a voltage divider for

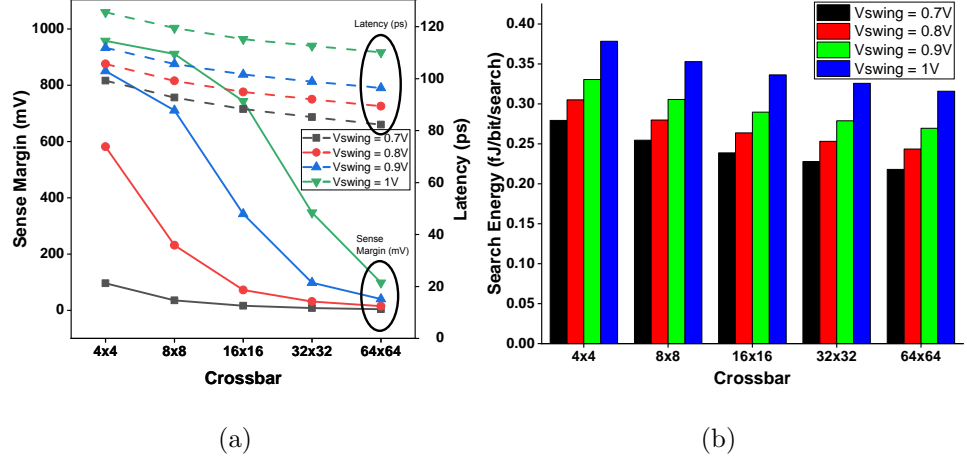


Figure 3.4: Bitcell performance for different word size (a) sense margin calculation of CAM operation for different crossbar sizes, (b) energy calculation of CAM operation for different crossbar sizes. [37]

unmatched rows. V_{ref} of the sense amplifier can be set between full match and 1 bit unmatched. This architecture can be used as a general memory storage unit as 1T1R, and CAM can be enabled by activating the inverters of each bitcell and the sense amplifiers. Then compared with the strings stored in all the rows simultaneously. All word lines and inverter control lines are enabled for the search operation by respective driver circuitry.

3.3 Results and Discussion

3.3.1 Analysis of the impacts of process and temperature variation

The 3T1R nvCAM is sensitive to the inverter behavior. The V_x input to the inverter, which forms the voltage divider by the ReRAM and M1, is a critical node for deciding the output of the bitcell. The JART VCM ReRAM model [11] includes device-to-device and cycle-to-cycle variability. Figure 3.5 shows the effects of the process corners and temperature on the behavior of bitcell performance. This proves

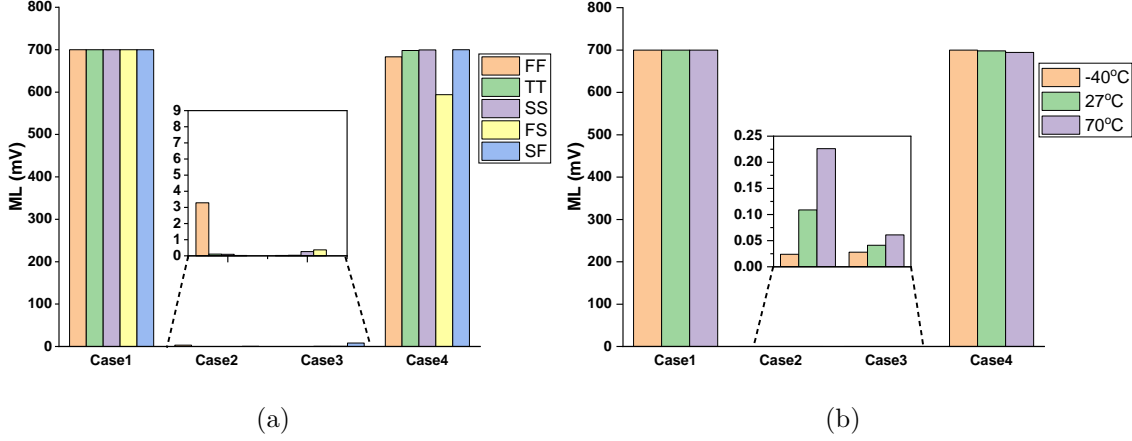


Figure 3.5: Proposed bit cell performance analysis (a) match line voltage analysis of CAM for process corners, (b) match line voltage analysis of CAM for different temperatures. [37]

that the bitcell is highly reliable for CAM applications. However, V_{ML} of a bitcell is affected by increasing word size. This lead to a small sense margin and may lead to misjudging the match.

3.3.2 Performance evaluation of novel 3T1R based array

The proposed bitcell latency decreases as the word size increases since the major contributor for V_{ML} is the PUN and PDN formed by the inverters of bitcell. Figure 3.4 illustrates the impact on the sense margin for different crossbar sizes (solid lines —) and latency (dashed lines - - -). To increase the sense margin, there can be two approaches; 1) decrease the propagation delay of the inverter such that it pulls down ML voltage faster for mismatch string; 2) increase the voltage swing of the inverter. Decreasing the propagation delay will increase the overall area, which may not improve the search latency. Considering the second approach in our proposed design. The sense margin comparison at different V_{swing} is illustrated in Figure 3.4(a). Also, the search energy gradually decreases with the crossbar but increases with the V_{swing} illustrated in Figure 3.4(b). Table 3.1 shows the comparison of prior CAM systems that are based on NVM technologies. The search latency and energy of our

Table 3.1:
Performance comparison with state-of-the-art.

Parameters	[5]	[3]	[6]	[4]	This work
NVM Devices	ReRAM	ReRAM	ReRAM	ReRAM	ReRAM
Cell Type	3T1R	5T2R	8T2R	2T2R	3T1R
Technology Node (nm)	90	45	45	65	65
Word Size	64	64	64	64	64
Latency (ns)	0.96	0.75	0.14	0.24	0.11
Search Energy (fJ/bit/search)	0.51	0.55	0.85	10.64	0.32

proposed bitcell are superior to the other CAM-based systems.

3.4 Conclusion

This paper presented a novel ReRAM-based 3T1R bitcell. The proposed bitcell gives reliable and robust performance for process and temperature variations. Further, the proposed cell performs CAM operation for the memory crossbar. For 64 bit wordlength, our proposed design shows 1.27x less sensing latency and 2.67x low search energy compared to the state-of-the-art. The proposed cell has been validated on 65nm CMOS technology node.

Chapter 4

HEART-CAM: Hybrid CMOS-ReRAM Based Energy Efficient And Rapid Ternary Content Addressable Memory

4.1 Introduction

The CAM finds the location in the stored data that corresponds to the search data that was entered. A pre-charge transistor and a match line sense amplifier (MLSA) are connected to each n -bit word in this architecture of an $n \times n$ CAM array that contains 1-bit memory elements. The ML is pre-charged to a high voltage during the search process, and the differential search line signals SL_1 and SL_0 receive the search data. The required search line signals are produced by the search line driver when the search data is decoded. The ML keeps its pre-charged voltage if the search and stored data match. Conversely, when there is no match, the ML voltage decreases or vice versa. The MLSA senses the ML voltage, feeding it to the encoder, which then provides the address of the matched data. The TCAM bitcell-based architecture offers exceptional density and resource utilization, particularly suited for

A	B	C	F
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

BCAM		
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	1	0

TCAM		
0	0	X
0	1	X
1	X	0

Figure 4.1: Mapping a logic function to BCAM and TCAM [34].

data-intensive tasks. In terms of stored bits, TCAM can store the same amount of information as BCAM while offering additional flexibility due to its ternary nature. This means that TCAM can achieve the same functionality as BCAM with potentially fewer memory cells or with the ability to represent more complex patterns using the same number of cells as shown in the Fig. 4.1 the truth table for function F is realized using BCAM, requiring 6 memory addresses, while TCAM achieves the same functionality using only half the number of addresses utilized by BCAM. Hence, it demonstrates a significant advantage over BCAM counterparts.

The proposed TCAM cell architectures in [?, ?, 30, 31] integrates two ReRAM devices to facilitate ternary logic storage, capable of representing states as 1, 0, and X. While ReRAM devices inherently operate with two distinct resistance states: LRS and HRS, the utilization of two ReRAM devices theoretically allows for four unique combinations of these states. The approach discussed in reference [8] leverages the multistate capability of ReRAM to accommodate the addition of a don't-care bit. While this approach offers flexibility, it introduces challenges related to error rates and programming complexity due to the utilization of multistate ReRAM. In contrast, the method proposed in [9] introduces an 8T2R bitcell for TCAM. However, this approach overlooks the consideration of both ReRAMs being in high-resistance states (HRS). It also adds area overhead due to 8T per bitcell. Similarly, the approach presented in [9] employs a 4T2R bitcell, which increases search energy consumption. However, this approach lacks description regarding the states of ReRAM when both are in low-

resistance states (LRS), leaving an important aspect unaddressed. Furthermore, [31] utilizes MRAM for TCAM operation, resulting in the highest latency and search energy consumption among the compared designs. In contrast, this proposed hybrid CMOS 4T2R bitcell takes into account all possible combinations while ensuring low latency and search energy consumption. By considering various scenarios and optimizing the bitcell design accordingly, this approach offers a promising solution that balances performance and efficiency in TCAM operations. However, a notable limitation arises in the other proposed TCAM design, as it fails to consider the scenario where both ReRAM devices are in the LRS [30] [32,33] whereas [9] fails to consider case for both HRS. In such a configuration, the TCAM cell encounters a failure during the search operation, rendering it unable to accurately match the stored ternary logic values against the search query. This limitation poses a significant challenge to the reliability and functionality of the TCAM cell, as it overlooks a critical scenario where the stored data cannot be effectively accessed or compared. Addressing this limitation is crucial for ensuring robust performance and accurate data retrieval in ReRAM-based TCAM architectures. Strategies such as incorporating additional logic or refining the design to accommodate all possible combinations of ReRAM states are essential to overcome this limitation and enhance the overall efficiency and reliability of the TCAM cell. A significant portion of the overall power dissipation in TCAM is attributed to the match line (ML) [34]. This is primarily because the ML is pre-charged to a high voltage initially and may discharge to a low voltage at the conclusion of the search operation. The process of pre-charging and discharging the ML is a dominant factor in the total dynamic power dissipation during the search operation. Additionally, the total dynamic power is influenced by the ML capacitance, which is determined by the length of the ML and the number of transistors connected to it [34]. Furthermore, even in an idle mode of operation, the parallel functioning of all TCAM cells results in substantial power waste because these cells never stop operating. In order to improve the power efficiency of the TCAM cell, the design integrates newly developed non-volatile memories with least standby power usage.

The primary objective of the novel proposed bitcell for TCAM architecture is to address and overcome the limitation inherent in traditional TCAM designs utilizing ReRAM devices, particularly the scenario where both ReRAM devices are in the LRS or both HRS. By acknowledging this critical scenario, the proposed bitcell aims to develop innovative strategies to effectively handle and mitigate this challenge, ensuring reliable operation and accurate data retrieval. The proposed bitcell architecture also utilizes a self-reference sense amplifier [35]. This innovative approach eliminates the need for an additional sense amplifier enable control signal, which traditionally contributes to additional control circuitry which indeed increases power consumption and complexity in TCAM architectures. The self-referenced sense amplifier ensures robust and reliable sensing operation, immune to variations across different chips. This ensures consistent performance and reliability across a range of manufacturing processes and operating conditions.

4.2 Variability Model

When designing circuits using VCM cells, it is essential to incorporate reliability effects into a suitable compact model. In a recent study, we enhanced our JART VCM 1b compact model [11, 26–28] to account for switching variability and read instability [11] [29]. The JART VCM 1b model captures changes in electronic transport through the metal/oxide interface and conduction near this electronically active interface as a function of ionic defect concentration. These defects migrate within the applied electric field, moving toward or away from the active interface. Additionally, Joule heating is considered, raising the local temperature and exponentially accelerating ion migration. Comprehensive details and equations of this model are available in [26–28].

The model’s switching variability stems from changes in parameters describing the filament geometry, such as the filament radius, the length of the “disc” region near the active interface, and the minimum and maximum defect concentration boundaries in the disc. Device-to-device variability is modeled by randomly selecting each

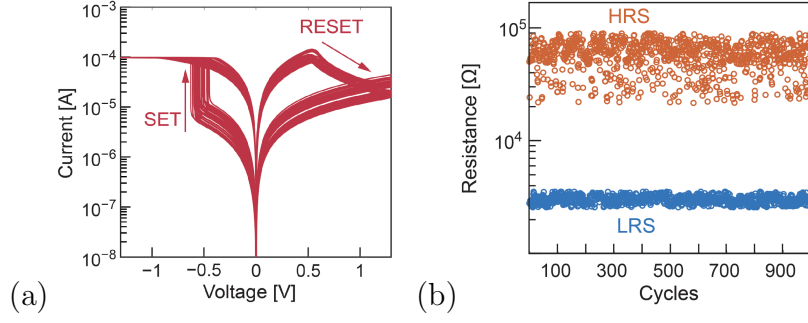


Figure 4.2: JART VCM 1b model (a) I-V characteristics (b) endurance characteristics.

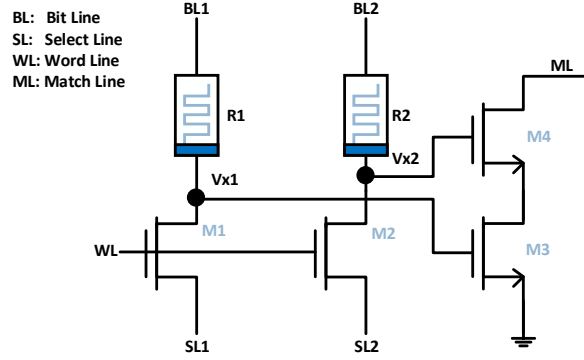


Figure 4.3: Proposed Hybrid 4T2R bitcell Structure.

parameter from a truncated Gaussian distribution, while cycle-to-cycle variability is implemented using a random walk algorithm that updates these parameters with each cycle. Detailed variability modeling and equations are reported in Table 4.1.

For read instability modeling, we assumed the defect concentration in the “disc” region could change by ± 1 or ± 2 defects, implemented via a state machine with probabilities for remaining in the current state or changing the number of defects by ± 1 , achieving a total change of ± 2 defects in two steps. The update frequency is aligned with measurement results, detailed in [29].

Fig. 4.2 displays the simulation results from the variability-aware JART VCM v1b model with parameters listed in Table 4.1 fitted to experimental data of a HfOx-based cell. The simulation results clearly demonstrate that the model can reproduce the characteristics of the experimental data.

Table 4.1: Simulation Parameters [11]

Symbol	Value	Symbol	Value
l_{cell}	3 nm	A^*	$6.01 \times 10^5 \text{ A}/(\text{m}^2 \text{K}^2)$
l_{disc}	0.4 nm	$e\phi_{BnO}$	0.18 eV
r_{fil}	45 nm	$e\phi_n$	0.1 eV
z_{vo}	2	μ_n	$4 \times 10^{-6} \text{ m}^2 / (\text{V} \cdot \text{s})$
a	0.25 nm	N_{plug}	$20 \times 10^{26} \text{ m}^{-3}$
v_o	$2 \times 10^{13} \text{ Hz}$	$N_{disc,max}$	$20 \times 10^{26} \text{ m}^{-3}$
ΔW_A	1.35 eV	$N_{disc,min}$	$0.008 \times 10^{26} \text{ m}^{-3}$
ε	$17 \varepsilon_0$	R_{series}	1370 Ω
$\varepsilon_{\phi B}$	$5.5 \varepsilon_0$	$R_{th,eff,SET}$	$15.72 \times 10^6 \text{ K/W}$
T_o	293 K	$R_{th,eff,RESET}$	$4.24 \times 10^6 \text{ K/W}$

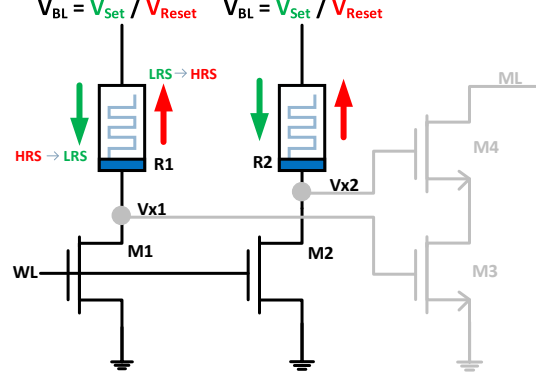


Figure 4.4: Signal conditions involved in storing data for proposed TCAM bitcell write operation.

4.3 Proposed Hybrid 4T2R bitcell Structure

Figure 4.3 shows a structure of the proposed 4T2R nvCAM cell comprising an NVM device (ReRAM [11]), the fundamental idea of the proposed bitcell is transforming the robust and silicon proven 1T1R cell [11] (M_1 and M_2 with ReRAM₁(R_1) and ReRAM₂(R_2) respectively) for BCAM and TCAM. Write operation is exactly similar to the 1T1R, as illustrated in Figure 4.4 with waveform in Figure 4.5.

The output (ML) depends on the critical node voltages V_{x1} and V_{x2} . The bit line (BL) and select line (SL) signals are used for inserting search query. Transistors M_3 and M_4 serially connected forming NAND style pull down network also form the discharging path for ML during mismatch case. At least one of these transistors runs in the cutoff zone during the match scenario. As a result, the pre-charged voltage of the ML does not change. Both of the transistors (M_3 , M_4) create an active path for ML to discharge at a mismatch. For the purpose of conducting searches at low voltages, every transistor in the bitcell is of the low threshold voltage variety.

The threshold voltage of transistors M_3 and M_4 is V_{th} , and the voltages at nodes x_1 and x_2 as V_{x1} and V_{x2} , respectively. The resistive states of memristors R_1 and R_2 , denoted as R_L and R_H , are to be determined. It is essential to set R_L and R_H in a manner that ensures critical node (V_x) is below (above) the V_{th} when ReRAM in HRS (LRS) during search 0 whereas above (below) the V_{th} during the search ‘1’.

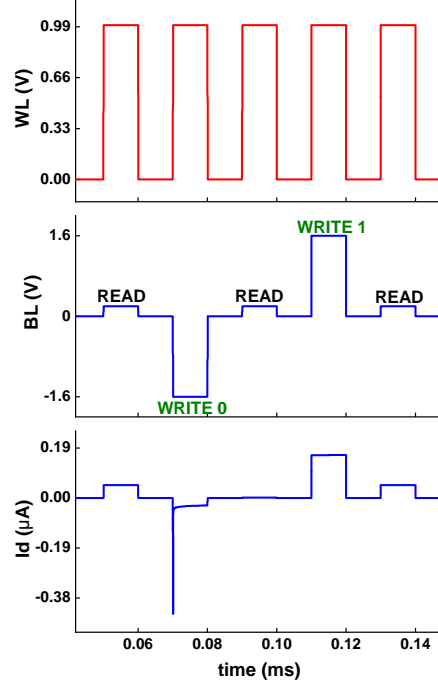


Figure 4.5: Waveform of signal involved in storing data for proposed TCAM bitcell write operation.

The effectiveness of the discharging path, particularly the transistors M_3 and M_4 , significantly influences the search delay. Increasing the width of these transistors, given a constant length, reduces the search delay but results in higher power dissipation. Importantly, even when the search data matches the stored data, ML discharges to a lower voltage if any of these transistors surpass a particular width. The misclassification causes a higher rate of search errors. Transistors M_3 and M_4 must be sized to keep ML at its pre-charged voltage during a match and to discharge more quickly following a miss in order to lessen this effect. The size of M_1 and M_2 determines the programming current and critical node voltages (V_x), thus they are sized to meet these needs.

4.4 Write Operation

During the write operation, the control signals, BL_1 (BL_2) and SL_1 (SL_2) determining the data to be stored in the memristors R_1 (R_2), match line reset (MLRST)

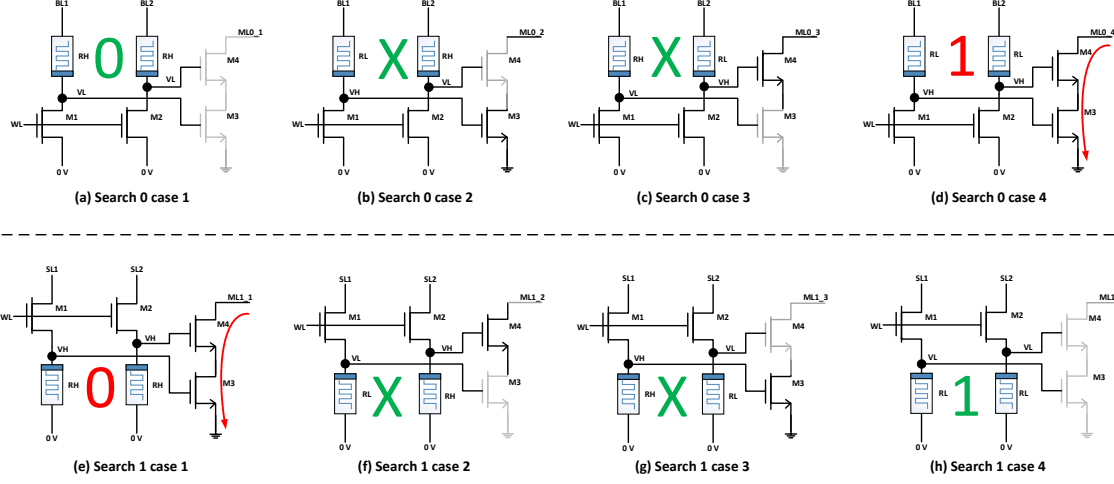


Figure 4.6: Different cases for stored data and the search data (a-d) for search ‘0’ case and (e-h) for search ‘1’.

and precharge signals (PRE) set by sense amplifier are high disabling the search mode. V_{RESET} of -1.6V and V_{SET} of +1.6V is used to switch the state of ReRAM, achieving R_{ratio} of 50. Write conditions with respect to 0, 1 and X are as described:

4.4.1 Write for ‘0’

When the stored bit value is 1 (with $R_1 = R_L$ and $R_2 = R_L$), assuming the initial resistive state of memristor R_1 is R_H and R_2 is R_H , with $V_{BL1} = V_{RESET}$ and $V_{BL2} = V_{RESET}$ switching the states of the stored bit. Consequently, the resistance of memristor R_1 and R_2 increases.

4.4.2 Write for ‘1’

When the stored bit value is 0 (with $R_1 = R_H$ and $R_2 = R_H$), assuming the initial resistive state of memristor R_1 is R_L and R_2 is R_L , with $V_{BL1} = V_{SET}$ and $V_{BL2} = V_{SET}$. Consequently, the resistance of memristor R_1 and R_2 decreases.

Search data	R1	R2	Stored data	Vx1	Vx2	VML	Remark
0 BL ₁ = 0.7 V BL ₂ = 0.7 V SL ₁ = 0V SL ₂ = 0V	RH	RH	0	VL	VL	V _{PRE}	Match
	RL	RH	X	VH	VL	V _{PRE}	Match
	RH	RL	X	VL	VH	V _{PRE}	Match
	RL	RL	1	VH	VH	0	Mismatch
1 BL ₁ = 0V BL ₂ = 0V SL ₁ = 0.7 V SL ₂ = 0.7 V	RH	RH	0	VH	VH	0	Mismatch
	RL	RH	X	VL	VH	V _{PRE}	Match
	RH	RL	X	VH	VL	V _{PRE}	Match
	RL	RL	1	VL	VL	V _{PRE}	Match

Figure 4.7: Search conditions for store ‘0’, ‘1’, and ‘X’ data for proposed TCAM architecture.

4.4.3 Write For ‘X’ State

To store bit value is X i.e., don’t care state (with $R_1 = R_L$ (R_H) and $R_2 = R_H$ (R_L)), the write X operation occurs with $V_{BL1} = V_{SET}$ ($V_{BL1} = V_{RESET}$) and $V_{BL2} = V_{RESET}$ ($V_{BL1} = V_{SET}$), mirroring the write 1 (0) operation for R_1 (R_2) and write 0 (1) operation for R_2 (R_1). Without influencing R_2 ’s resistive state, the current flowing through R_1 modifies its resistive state, and vice versa.

4.5 Search Operation

The search operation is a streamlined single-clock operation comprising two essential phases: pre-charge and evaluation. Figure 4.7 provides an overview of the logic values of control line signals during the search operation. ML are connected to the sense amplifier which also sets the precharge voltage. Output of sense amplifier is labelled as MLout. Search 0 and 1 is as described below;

4.5.0.1 Search for ‘0’

To initiate a search for 0, the control lines BL₁ and BL₂ are set to a small read voltage of 0.7V, while SL₁ and SL₂ are set to 0V. Figure 4.6(a-d) illustrate the equivalent circuit for a search 0 operation in scenarios where the search data matches

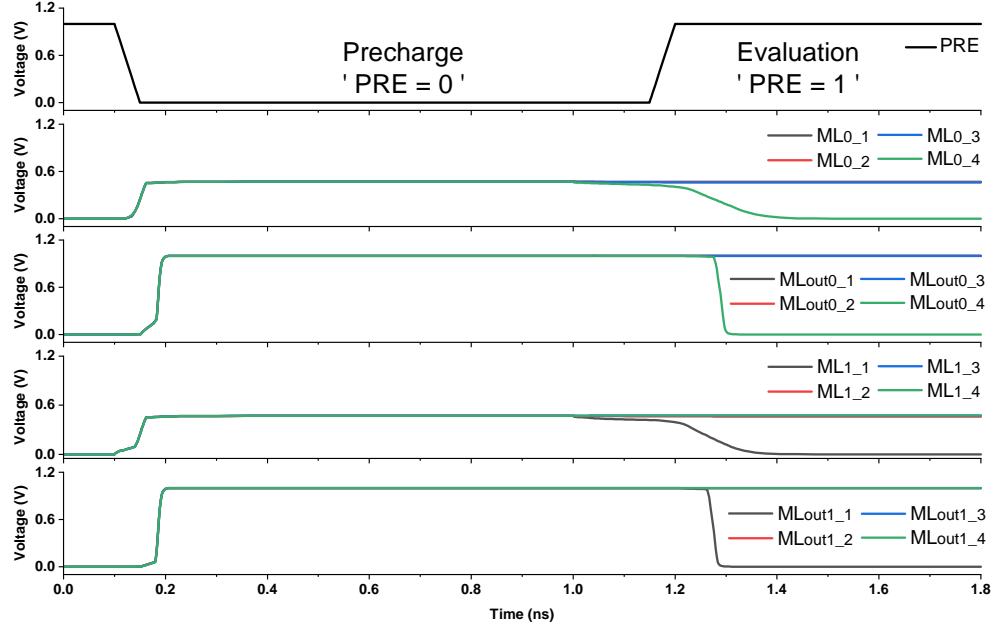


Figure 4.8: Illustrating waveform for match and mismatch cases for stored ‘0’, ‘1’, and ‘X’.

or mismatches the stored data. In the event of a match Figure (4.6(a-c)) where the stored data is 0 or X, either transistor M_3 or M_4 or both remains in cutoff due to the voltage developed at V_{x1} or V_{x2} being below the threshold voltage. Consequently, no discharging path for ML is available. Depicted by ML_{0-1} , ML_{0-2} and ML_{0-3} of the bitcell and consequently connected to the $MLout_{0-1}$, ML_{0-2} and ML_{0-3} through the sense amplifier shown in Figure 4.8. Conversely, when the stored data is 1, M_3 and M_4 form the discharging path for ML, as depicted in Figure 4.6(d) and Figure 4.8 by ML_{0-4} and $MLout_{0-4}$, leading to the detection of a mismatch.

4.5.0.2 Search for ‘1’

Search 1 begins by setting the control lines BL_1 and BL_2 are set to 0, while SL_1 and SL_2 are set to a small voltage of 0.7V. Figure 4.6 (e-h) showcase the equivalent circuit for a search 1 operation in scenarios where the search data matches or mismatches the stored data. In the case of a match (Figure 4.6 (f-h)) where the stored data is 1 or X, either transistor M_3 or M_4 or both remains in cutoff due to the voltage developed at V_{x1} or V_{x2} being below the threshold voltage. Consequently,

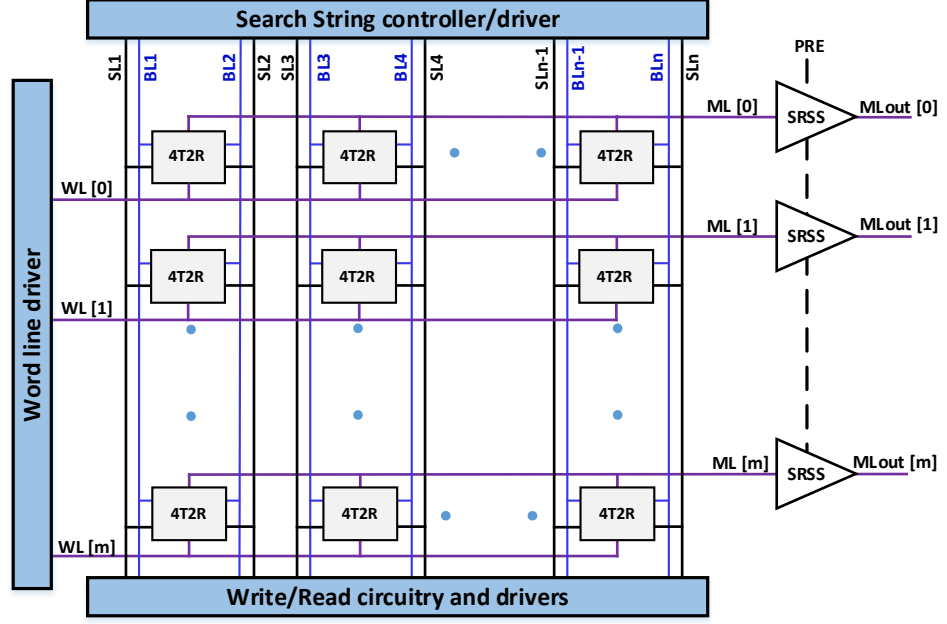


Figure 4.9: TCAM Architecture with $m \times n$ crossbar array

there is no available discharging path for ML. Depicted by ML_{1-2} , ML_{1-3} and ML_{1-4} of the bitcell and consequently connected to the $MLout_{1-2}$, ML_{1-3} and ML_{1-4} through the sense amplifier shown in Figure 4.8. Conversely, when the stored data is 0, M_3 and M_4 form the discharging path for ML, as depicted in Figure 4.6(e) and Figure 4.8 by ML_{0-4} and $MLout_{0-4}$, leading to the detection of a mismatch.

4.6 TCAM Architecture

The proposed research introduces a $m \times n$ CAM architecture using novel bitcell, aiming to contribute to the advancement of high-performance memory systems. This innovative design seeks to optimize the efficiency of content-based memory retrieval, offering potential improvements in speed, power consumption, and overall performance. The unique features of the $m \times n$ CAM architecture are expected to make a valuable contribution to the field, addressing key challenges and paving the way for enhanced memory solutions in future computing applications. Illustrated in Figure 4.9, the ML architecture highlights $m \times n$ array with the proposed bitcell. Any bit-mismatch creates a path from ML to GND through the mismatched bit-compare

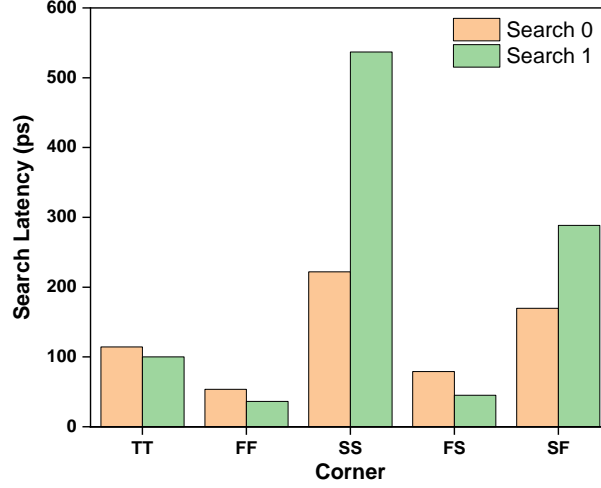


Figure 4.10: Search Latency of TCAM bitcell search 0 and search 1 delay for mismatch cases for all corners

circuits of bitcell (M_3 and M_4). The ML can either be in a pre-charged state or be pulled to GND by at least one bit-compare circuit. Before sensing, search data is provided on the SLs, and MLRST=1 resets the MLs to GND.

4.7 Results and Discussion

4.7.1 Bitcell Simulation and Analysis

The proposed 4T2R bitcell were designed in 65 nm technology. The ReRAM used for the simulation is silicon proven JART VCM model [11] achieving Rratio of 50. Circuits manufactured with technology nodes smaller than 100 nm are very vulnerable to changes in the manufacturing process. A TCAM bitcell was subjected to corner case analysis in order to predict the possible impact. The sensing amplifier also has a big impact on the delay computation. Figure 4.10 shows the impact of all five corner simulations including the effect of the sense amplifier on ML. The results of the simulation show that in the slow PMOS and slow NMOS (SS) corner case, the search latency is at its maximum, and in the fast PMOS and fast NMOS (FF) corner case, it is at its minimum. Therefore, a simulation was run to vary the temperature from -40°C to 125°C in order to analyze the impact of temperature variation on

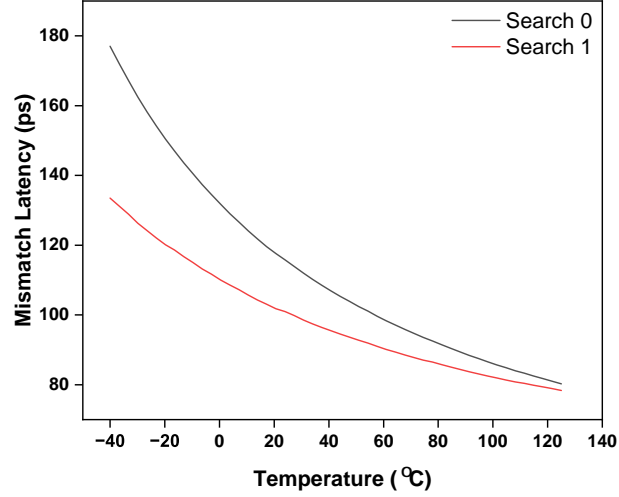


Figure 4.11: Search Latency of TCAM bitcell search 0 and search 1 delay for temperature variation.

the performance of the suggested design. The search delay of the suggested design at various temperatures is shown in Figure 4.11. The results show that when the temperature rises, the search delay goes down.

1000 Monte Carlo simulations were run on the proposed bitcell in order to examine the impact of PVT on the bitcell. In Figure 4.12, the distribution of search delay for mismatch cases is depicted. For Search 0, the mean delay is 117.387 ps with a standard deviation of 27.274 ps, while for Search 1, the mean delay is 113.702 ps with a standard deviation of 57.531 ps. Low standard deviation signifies that the simulated outcomes are more tightly distributed around the mean, indicating greater consistency resulting in the robustness of the proposed bitcell.

The proposed TCAM's layout is shown in the Fig4.13. A transistor indicated by 'X' on the M_1 and M_2 can have a memristor 3D-stacked on top of it. Nevertheless, we used a Verilog-A model [11] of the memristor for simulation reasons. With a $1.456 \text{ } \mu\text{m}^2$ size, the suggested TCAM cell offers great density and compactness.

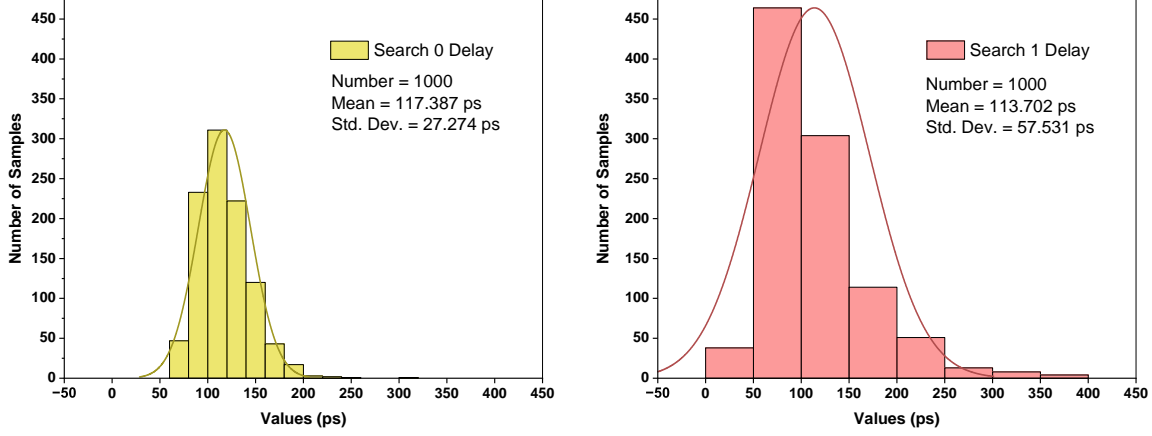


Figure 4.12: Histogram for latency distribution of TCAM bitcell for 1000 monte carlo simulations. The standard deviation is less than 60 ps for mismatch case.

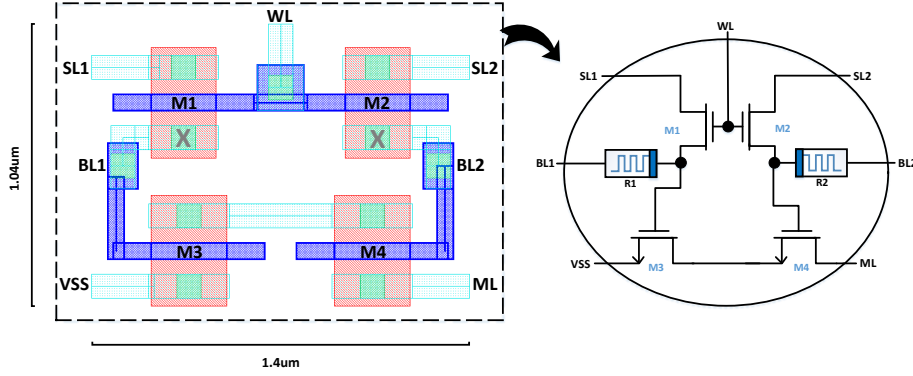


Figure 4.13: Layout of the proposed hybrid-CMOS bitcell structure.

4.8 Example of 4x4 array using proposed bitcell

The TCAM array in a 4x4 example is shown in Figure 4.14. The TCAM array allows for efficient search operations, enabling rapid retrieval of stored data based on user-defined search criteria. This example showcases the compact and dense storage capability of TCAM arrays, making them well-suited for applications requiring fast and flexible data retrieval. Before search operation CAM cells are programmed to the values shown in the Figure 4.14. Keeping $PRE=1$, and $MLRST=1$ to avoid unnecessary the power consumption by M_3 and M_4 . Search operation is done in 2 stages, precharge and the evaluation stage described as:

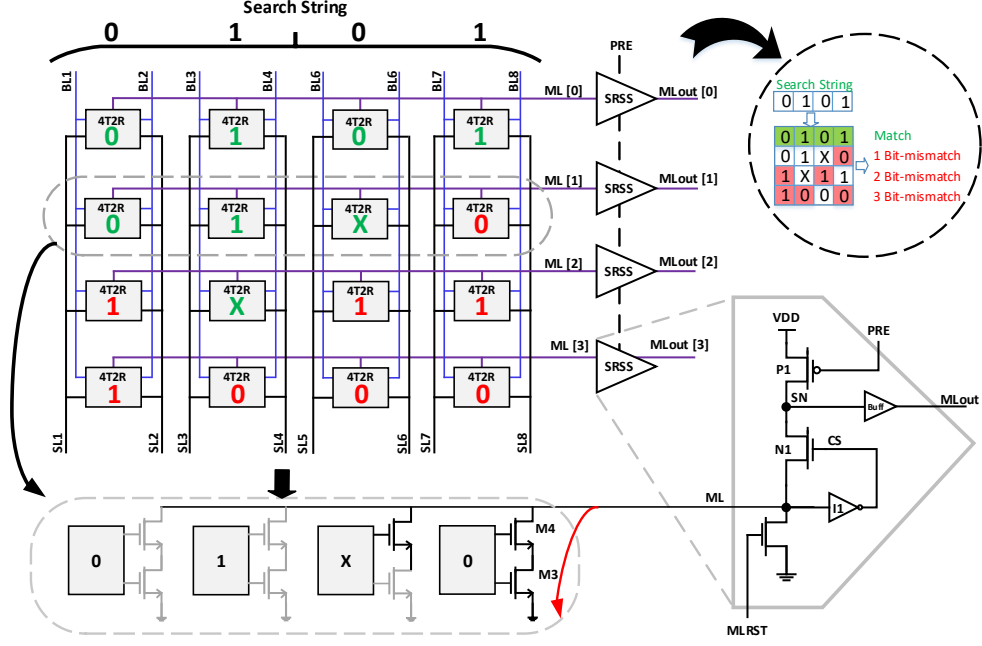


Figure 4.14: Detailed Analysis of a 4x4 Array Integrated with Self Reference Sensing Scheme (SRSS), Enhancing Efficiency and Performance in Memory TCAM using proposed bitcell.

4.8.1 Precharge Stage

During the precharge stage of the operation, indicated by $MLRST=0$ and $PRE=0$ signals, the P_1 devices within each sense amplifier (SA) begin precharging the match lines (MLs) towards the precharge voltage. At this point, any multi-bit mismatched MLs start to draw this precharge current, effectively holding the ML voltage at ground level. As the ML voltage surpasses the threshold of the I_1 device, the voltage at the compare sense (CS) node decreases, causing the N_1 device to enter the cutoff region, thus halting the precharge process for the ML.

Subsequently, each ML is precharged to a level slightly above the threshold of its respective SA, ensuring a small and consistent voltage difference (delta) between the precharge and sense voltages for each SA. As the ML voltage approaches this predetermined level, the precharge current is redirected towards the sense node (SN), rapidly charging it to the supply voltage (V_{DD}). This action, facilitated by a buffer (buff), leads to the swift transition of the ML output (MLout) to a high state,

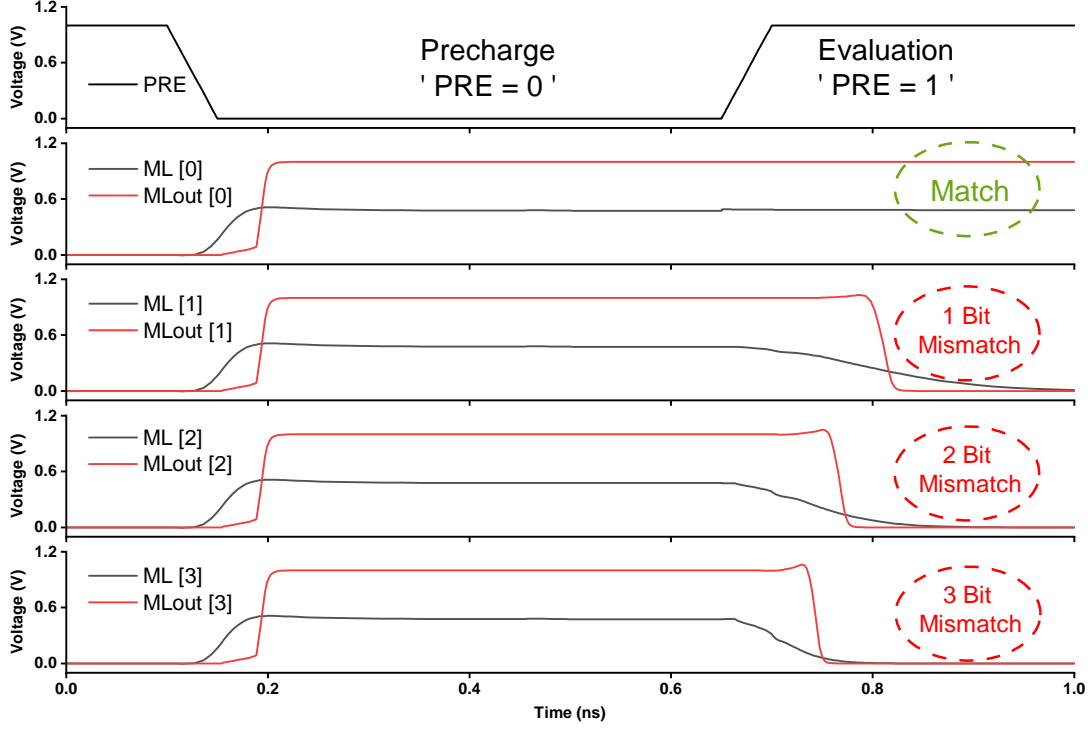


Figure 4.15: Search waveform in a 4x4 Array configuration with exact match, 1 bit mismatch, 2 bit mismatch and 3 bit mismatch in search string.

effectively completing the precharge stage.

In essence, during the precharge stage, the MLs are prepared for subsequent operations by being precharged to levels conducive to efficient sensing and data retrieval, ensuring uniformity and consistency. As shown in Figure 4.15 all the ML and MLout are precharged by the sense amplifier.

4.8.2 Evaluation Stage

The design of the proposed bitcell deployed in array features like a NOR-type configuration ensures that only the row with a mismatch discharges, leading to a faster search during search or lookup operations. This architecture also effectively minimizes energy wastage by using the self referenced sense amplifier, thus optimizing power efficiency. In the evaluation stage (where $PRE=1$), and all the word lines are activated the mismatch cases in Figure 4.15 ML[0] and ML[2] i.e., the 2nd and 4th address discharging its precharge state, while ML[0] and ML[2] starts to discharge

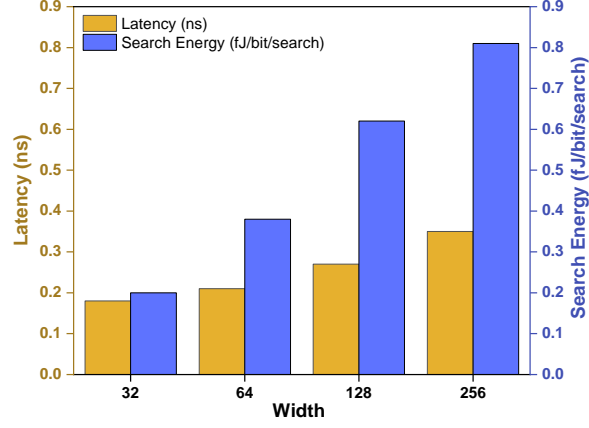


Figure 4.16: Search energy and latency with different word size ranging from 32 to 256.

through the matched bit-compare circuits. This discharges, tripping I_1 , causing SN and ML to equalize and then discharge, bringing MLout[0] and MLout[2] to discharge to 0. The PRE=1 portion of Figure 4.15 shows ML[1] and ML[3] in a high-impedance state, maintaining its precharge state and value of MLout[1] and MLout[3].

Furthermore, to ensure rapid transition of MLout, the PRE can be made high when the ML of the mismatch is discharged. This design choice facilitates swift switching of MLout[1] back to its high state, ensuring efficient operation of the evaluation stage.

In summary, during the evaluation stage with PRE=1, the mismatched bit-compare circuits facilitate the discharge of ML[1], leading to the restoration of MLout[1] to its high state. Meanwhile, ML[0] remains precharged, maintaining its high-impedance state until required for subsequent operations.

Figure 4.14 showcasing various search scenarios such as match, 1-bit mismatch, 2-bit mismatch, and 3-bit mismatch, along with their corresponding stored strings. Additionally, the previously discussed sense amplifier is depicted within this figure.

The equivalent waveform provided in Figure 4.15 represents different cases. In the case of a match, depicted by MLout[1] remaining high, this signifies that the match line (ML) remains at a high impedance due to an exact match between the search

Table 4.2: Performance comparison with state-of-the-art.

Parameters	[8] JSSC'17	[9] TETC'21	[30] TED'21	[31] JETCAS'23	This work
NVM Device	ReRAM	ReRAM	ReRAM	MRAM	ReRAM
Cell Type	3T1R	8T2R	4T2R	2T2MTJ	4T2R
Technology Node (nm)	90	45	180	28	65
Word size	64	64	64	32	64
Latency (ns)	0.96	0.14	0.6	1.2	0.21
Search Energy (fJ/bit/search)	0.51	0.85	0.18	1.73	0.38
Energy-Delay Product	0.49	0.119	0.108	2.07	0.079

query and the stored data. On the other hand, MLout[2] represents a worst-case scenario where only 1 bit is mismatched, resulting in a longer time for ML[2] to discharge below the tripping point, causing MLout to switch from high to low.

Conversely, MLout[3] and MLout[4] demonstrate quicker switching as the number of mismatched bits increases, leading to decreased latency for the stored data. Essentially, the illustration provides insights into how different search scenarios affect the behavior of the sense amplifier and the corresponding ML outputs, highlighting the impact of mismatched bits on the latency and overall performance of the system.

Figure 4.16 illustrates how increasing the word size impacts both latency and search energy. As the word size increases, there is a noticeable trend in latency, with larger word sizes generally correlating with longer search times. This is due to the increased complexity of matching larger data sets, requiring more time for the search operation to complete. This is primarily because larger word sizes typically involve more complex search operations, requiring additional energy to perform comparisons across a greater number of bits or data units. Table 4.2 provides a comparison with state-of-the-art designs in the field.

4.9 Conclusion

In conclusion, the proposed 4T2R bitcell architecture presents a notable breakthrough in Ternary Content-Addressable Memory (TCAM) design, particularly in addressing challenges associated with ReRAM devices, such as scenarios where both devices are in the same resistance state. The integration of a self-referenced sense amplifier contributes to a reduction in power consumption and system complexity while ensuring reliable sensing operations across Process, Voltage, and Temperature (PVT) variations. Compared to conventional designs and state-of-the-art alternatives, the 4T2R bitcell stands out by offering a promising solution with lower latency and search energy consumption. Remarkably, it outperforms existing designs in terms of the energy-delay product by an impressive 26.85%. These outcomes emphasize the effectiveness of the proposed architecture in achieving efficient TCAM operations. The novel approach of employing a 4T2R configuration, an improvement upon the 2T2R cell, by incorporating additional comparison transistors to address signal mismatch issues and maintain precharged values for match signals effectively. The SPICE simulations conducted at the CMOS 65nm node reveal compelling performance metrics, including a latency of 0.35 ns for a word size of 256 and an energy consumption of 0.81 fJ/bit/search. Furthermore, the evaluation of the 4T2R bitcell's performance and energy efficiency across various word sizes demonstrates its versatility and effectiveness in diverse applications and scenarios. Overall, this work signifies a promising advancement in TCAM design, offering improved speed and energy efficiency critical attributes for the demands of modern computing systems.

Chapter 5

Future Scope

Binary Content-Addressable Memory (CAM)

- **Advanced Technology Nodes:** The proposed ReRAM-based 3T1R bitcell can be further explored and optimized for more advanced technology nodes beyond the 65nm CMOS, such as 40nm, 28nm, etc. This can potentially lead to further improvements in performance, power efficiency, and integration density.
- **Scaling and Integration:** Investigating the scalability of the proposed 3T1R bitcell for larger memory arrays and its integration into system-on-chip (SoC) designs can help in realizing high-density and low-power CAMs for various applications, including artificial intelligence and machine learning accelerators.
- **Error Correction and Tolerance:** Future research can focus on incorporating error correction codes (ECC) and techniques to enhance the fault tolerance and reliability of the ReRAM-based CAMs, especially under extreme process and temperature variations.
- **Material and Device Engineering:** Exploring different materials and device structures for ReRAM elements could further improve the robustness, endurance, and retention characteristics of the 3T1R bitcell, potentially leading to commercially viable CAM solutions.

- **Power Management Strategies:** Developing advanced power management strategies and dynamic power-saving techniques for ReRAM-based CAMs could significantly reduce their overall energy consumption, making them more suitable for portable and battery-operated devices.

Ternary Content-Addressable Memory (TCAM)

- **Process Technology Scaling:** Similar to Binary CAM, further research is required to adapt and optimize the 4T2R bitcell architecture for more advanced technology nodes. This will help in achieving even lower latency and energy consumption, critical for high-speed and low-power applications.
- **Application-Specific Customization:** The versatility of the 4T2R bitcell can be leveraged to tailor TCAM designs for specific applications, such as networking hardware, security systems, and database accelerators. Customizing TCAMs for such domains can provide targeted performance and efficiency gains.
- **Hybrid Memory Architectures:** Investigating hybrid architectures that combine TCAM with other memory types, such as SRAM, DRAM, or emerging non-volatile memories, could lead to innovative solutions that balance speed, power, and capacity requirements for various computational tasks.
- **Circuit-Level Enhancements:** Further improvements in the self-referenced sense amplifier and other circuit-level innovations can enhance the robustness of the 4T2R bitcell against variations in Process, Voltage, and Temperature (PVT), ensuring consistent performance in diverse operational environments.
- **AI and Machine Learning Integration:** Exploring the use of TCAMs in AI and machine learning hardware accelerators can open up new avenues for high-speed pattern matching and data retrieval operations, thereby enhancing the overall computational efficiency of AI systems.

By addressing these future directions, the advancements presented in these works can be further refined and expanded, contributing to the development of more efficient, reliable, and high-performance memory and frequency synthesis solutions for modern and future computing systems.

Bibliography

- [1] S. Yu and P. -Y. Chen, “Emerging Memory Technologies: Recent Trends and Prospects,” in *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43-56, Apr. 2016.
- [2] R. Karam et al., “Emerging Trends in Design and Applications of Memory-Based Computing and Content-Addressable Memories,” *Proc. of the IEEE*, Aug. 2015.
- [3] Zhou et al., “The trend of emerging non-volatile TCAM for parallel search and AI applications,” *Chip* 1, no. 2, Jan. 2022.
- [4] Mehonic., et. al., “Memristors—From In-Memory Computing, Deep Learning Acceleration, and Spiking Neural Networks to the Future of Neuro-morphic and Bio-Inspired Computing”. *Advanced Intelligent Systems*. pp. 2. 10.1002/aisy.202000085, Aug. 2020.
- [5] S.-G. Ahn and K.-W. Kwon, “Local NOR and global NAND match-line architecture for high performance CAM”, *Proc. IEEE Int. Midwest Symp. Circuit Syst. (MWSCA)*, pp. 707-710, Aug. 2017.
- [6] L. Zheng et al., “ReRAM-based TCAMs for pattern search,” in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, QC, Canada, May 2016.
- [7] Quang-Kien Trinh et al., “A Novel In-memory Matching Circuit Based on Non-volatile Resistive Memory,” in *2022 International Conference on IC Design and Technology (ICICDT)*, pp. 97-100. IEEE, Sept. 2022.

- [8] Meng-Fan Chang et al., “A 3T1R Nonvolatile TCAM Using MLC ReRAM for Frequent-Off Instant-On Filters in IoT and Big-Data Processing,” *IEEE Journal of Solid-State Circuits*, vol. 52, no. 6, pp. 1664-1679, June 2017.
- [9] K. P. Gnawali and S. Tragoudas, ”High-Speed Memristive Ternary Content Addressable Memory,” in *IEEE Transactions on Emerging Topics in Computing*, pp. 1349-1360, Sept. 2022.
- [10] V. Sharma et al., “A 64 Kb Reconfigurable Full-Precision Digital ReRAM-Based Compute-In-Memory for Artificial Intelligence Applications,” in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 8, pp. 3284-3296, Aug. 2022.
- [11] Christopher Bengel et al., “Variability-Aware Modeling of Filamentary Oxide-Based Bipolar Resistive Switching Cells Using SPICE Level Compact Models,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4618-4630, Dec. 2020.
- [12] Y. V. Pershin and M. D. Ventra, “Neuromorphic digital and quantum computation with memory circuit elements”, *Proceedings of the IEEE*, vol. 100, no. 6, pp. 2071-2080, May 2012.
- [13] K. Pagiamtzis and A. Sheikholeslami, “Content-addressable memory (CAM) circuits and architectures: A tutorial and survey”, *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 712-727, March 2006.
- [14] Venkataramesh Bontupalli, Chris Yakopcic, Raqibul Hasan, and Tarek M. Taha, “Efficient Memristor-Based Architecture for Intrusion Detection and High-Speed Packet Classification”, *ACM Journal on Emerging Technologies in Computing Systems*, 14, 4, Article 41, Nov. 2018.
- [15] A. X. Liu, C. R. Meiners and E. Torng, “Packet classification using binary content addressable memory”, *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1295-1307, June 2016.

- [16] Y. Sasaki, "A survey on IoT big data analytic systems: Current and future", *IEEE Internet Things Journal*, vol. 9, no. 2, pp. 1024-1036, Jan. 2022.
- [17] J.-Y. Huang and P.-C. Wang, "TCAM-based IP address lookup using longest suffix split", *IEEE/ACM Trans. Netw.*, vol. 26, no. 2, pp. 976-989, April 2018.
- [18] João Paulo Cardoso de Lima, Marcelo Brandalero, Michael Hübner, and Luigi Carro, "STAP: An Architecture and Design Tool for Automata Processing on Memristor TCAMs". *ACM Journal on Emerging Technologies in Computing Systems*, 18, 2, Article 39, Dec. 2022.
- [19] Mohammad M. A. Taha and Christof Teuscher, "Approximate Memristive In-Memory Hamming Distance Circuit", *ACM Journal on Emerging Technologies in Computing Systems* 16, 2, Article 18, March 2020.
- [20] Rafael Fão de Moura, Joao Paulo Cardoso de Lima, and Luigi Carro., "Data and Computation Reuse in CNNs Using Memristor TCAMs", *ACM Transactions on Reconfigurable Technology and Systems*, 16, 1, Article 14, Dec. 2022.
- [21] Anteneh Gebregiorgis, Hoang Anh Du Nguyen, Jintao Yu, Rajendra Bishnoi, Mottaqiallah Taouil, Francky Catthoor, and Said Hamdioui, "A Survey on Memory-centric Computer Architectures", *ACM Journal on Emerging Technologies in Computing Systems* 18, 4, Article 79, Oct. 2022.
- [22] Hoang Anh Du Nguyen, Jintao Yu, Muath Abu Lebdeh, Mottaqiallah Taouil, Said Hamdioui, and Francky Catthoor, "A Classification of Memory-Centric Computing", *ACM Journal on Emerging Technologies in Computing Systems* 16, 2, Article 13, Jan. 2020.
- [23] N. Dhakad, E. Chittora, V. Sharma, SK. Vishvakarma, "R-inmac: 10T SRAM based reconfigurable and efficient in-memory advance computation for edge devices." *Analog Integrated Circuits and Signal Processing* vol. 116, no. 3, pp. 161-184, Sept. 2023.

- [24] Fieback, Moritz, Mottaqiallah Taouil, and Said Hamdioui. “Testing resistive memories: Where are we and what is missing?.” *IEEE International Test Conference (ITC)*, pp. 1-9. IEEE, Oct. 2018.
- [25] L. Chua, “Memristor-the missing circuit element”, *IEEE Trans. Circuit Theory*, vol. 18, no. 5, pp. 507-519, Sept. 1971.
- [26] F. Cueppers, S. Menzel, C. Bengel, A. Hardtdegen, M. von Witzleben, U. Boettger, R. Waser and S. Hoffmann-Eifert. “Exploiting the switching dynamics of HfO₂-based ReRAM devices for reliable analog memristive behavior”, *Appl Materials*, 7, 91105/1-9, Sept. 2019.
- [27] A. Hardtdegen, C. La Torre, F. Cüppers, S. Menzel, R. Waser and S. Hoffmann-Eifert. “Improved Switching Stability and the Effect of an Internal Series Resistor in HfO₂/TiO_x Bilayer ReRAM Cells”, *IEEE Transactions on Electron Devices*, 65, 3229-3236, Aug. 2018.
- [28] C. Bengel, F.Cüppers, M. Payvand, R. Dittmann, R. Waser, S.Hoffmann-Eifert and S. Menzel. “Utilizing the Switching Stochasticity of HfO₂/TiO_x-Based ReRAM Devices and the Concept of Multiple Devices for the Classification of Overlapping and Noisy Patterns”, *Frontiers in Neuroscience*,: 15:661856, June 2021.
- [29] S. Wiefels, C. Bengel, N. Kopperberg, K. Zhang, R. Waser and S. Menzel. “HRS Instability in Oxide based Bipolar Resistive Switching Cells”, *IEEE Transactions on Electron Devices*, 67, 4208-4215, Oct. 2020.
- [30] X. Wang, L. Wang, Y. Wang, J. An, C. Dou, Z. Wu, et al., “A 4T2R ReRAM bit cell for highly parallel ternary content addressable memory”, *IEEE Trans. Electron Devices*, vol. 68, no. 10, pp. 4933-4937, Dec 2021.
- [31] E. Garzón, M. Lanuzza, A. Teman and L. Yavits, “AM4: MRAM crossbar based CAM/TCAM/ACAM/AP for in-memory computing”, *ACM Journal on*

Emerging Technologies in Computing Systems, vol. 13, no. 1, pp. 408-421, March 2023.

- [32] J. Min, C. Kim, S.-Y. Kim and K.-W. Kwon, “A study of read margin enhancement for 3T2R nonvolatile TCAM using adaptive bias training”, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 8, pp. 1840-1850, Aug. 2019.
- [33] D. R. B. Ly, J-P. Noel, B. Giraud, P. Royer, E. Esmanhotto, N. Castellani, et al., “Novel 1T2R1T ReRAM-based ternary content addressable memory for large scale pattern recognition” in *IEDM Tech. Dig.*, pp. 35.5.1-35.5.4, Nov. 2019.
- [34] B. Agrawal and T. Sherwood, “Modeling TCAM power for next generation network devices”, *IEEE International Symposium on Performance Analysis of Systems and Software*, pp. 120-129, March 2006.
- [35] I. Arsovski and R. Wistort, “Self-referenced sense amplifier for across-chip-variation immune sensing in high-performance content-addressable memories”, *Proc. IEEE Custom Integr. Circuits Conf.*, pp. 453-456, Sept. 2006.
- [36] O. Tyshchenko and A. Sheikholeslami, “Match sensing using match-line stability in content-addressable memories (CAM)”, *IEEE Journal of Solid-State Circuits*, vol. 43, no. 9, pp. 1972-1981, Sept. 2008.
- [37] R. Sharma, N. S. Dhakad, G. S. Reddy, V. Sharma and S. K. Vishvakarma, “ReCAM: Resistive RAM Digital Content Addressable Memory Using Novel 3T1R Bitcell”, 8th *IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, Bangalore, India, pp. 1-3, March 2024.