WEP Warwick Evans Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 10
ISET 2025

Research and Application of Compute-in-Memory Architectures: RRAM, MRAM, and FeRAM

Yichang Chen*

Department of Electronic Science and Technology, Jinan University, Guangzhou, 511400, China

*Corresponding author: qs0018@stu2022.jnu.edu.cn

Abstract. With the rise of data-intensive applications such as artificial intelligence, big data analytics, and edge computing, the performance constraints of traditional memory technologies have become noticeable. Next-generation non-volatile memory devices, including resistive random-access memory (RRAM), magneto resistive random-access memory (MRAM), and ferroelectric random-access memory (FeRAM), are being actively explored to address demands for higher speed, lower power consumption, and greater integration density, especially within compute-in-memory (CIM) architectures. This study offers an overview of the operating principles, material mechanisms, performance characteristics, and reliability difficulties of RRAM, MRAM, and FeRAM. Furthermore, it covers their recent technological developments and illustrative applications in neuromorphic computing, embedded memory, and edge artificial intelligence devices. A comparative examination emphasizes each technology's advantages: RRAM has high density and multilevel cell potential with cell dimensions down to $4F^2$ and multilevel storage achieving over 7 bits per cell in some prototypes; MRAM features outstanding endurance exceeding 10^{12} white cycles and write/read energies as low as a few picojoules per bit; FeRAM is famous for its excellent reliability about 50-100ns and low power consumption. The research additionally examines development patterns and future potential for developing memory in advanced information processing systems.

Keywords: Non-volatile Memory; Compute-in-Memory; Neuromorphic Computing; Edge Artificial Intelligence.

1. Introduction

With the growing expansion of data-driven applications such as artificial intelligence, big data analytics, and edge computing, the limitations of conventional memory technologies have become increasingly evident. Traditional memory chips, such as static random-access memory (SRAM) and dynamic random-access memory (DRAM), fail to meet the needs for higher speed, lower power consumption, and greater integration density demanded by modern CIM systems. As computing operations become more complicated and data-intensive, there exists a pressing demand for new types of non-volatile memory devices that may offer quick switching speed, high durability, low energy consumption, and compatibility with logic processes [1-3]. In the context, developing memory technologies, such as RRAM, MRAM, and FeRAM, have gained substantial attention. These innovative memory devices not only provide the non-volatility and scalability required for CIM but also enable in-situ computation, paving the way for more efficient and powerful computing systems.

The traditional von Neumann architecture separates memory and the processor, but the traditional design causes the well-known "von Neumann bottleneck." Specifically, the shared bus for instructions and data limits system performance and creates a significant bandwidth issue between modern high-performance processors and slower memory access speeds. Research shows that the separated architecture is inefficient for data-intensive tasks, particularly in artificial intelligence and machine learning, where delays and energy consumption from data movement become major challenges. For example, von Neumann processors operate sequentially, executing one operation at a time with a single processing unit and a sequential control unit [4,5]. The phenomenon restricts speed improvements and requires parallel architectures for fifth-generation computers to achieve higher speeds. Additionally, the general nature of von Neumann architecture leads to energy dissipation that

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/). Published by Warwick Evans Publishing.

far exceeds that of specialized array processors, especially in image processing tasks, highlighting its limitations in energy efficiency.

The fundamental principle of the compute-in-memory architecture is to integrate computational capabilities within memory, thereby reducing the overhead of data movement between the processor and memory, which alleviates the "memory wall" issue [5]. Research indicates that this architecture is particularly well-suited for data-intensive tasks, such as training neural networks and real-time data analysis. Evidence supports its advantages, including reduced latency, lower energy consumption, and enhanced computational efficiency. For instance, CIM significantly decreases the energy and time costs associated with data movement by merging logic and memory units, achieving up to a 100-fold improvement in PageRank throughput, especially in machine learning applications [6]. Furthermore, CIM facilitates high parallelism, exemplified by RRAM-based ternary content-addressable memory (TCAM), achieving a 25-fold improvement in energy efficiency in genomic sequencing, with a computational complexity of O(1).

This paper investigates the key mechanisms, material characteristics, and application prospects of emerging non-volatile memory devices in CIM systems. We systematically review the physical mechanisms and structural advantages of mainstream memories such as RRAM, MRAM, and FeRAM, as well as their compatibility with logic processes. Furthermore, we analyze their potential and recent progress in enabling high-performance, low-power, and highly integrated CIM architectures. This paper also summarizes the latest research advances in the application of non-volatile memory devices for artificial intelligence, edge computing, and other cutting-edge fields, and discusses the future trends and challenges in this rapidly evolving area.

2. Principles and Characteristics of New Memory Devices

2.1. RRAM Technology

RRAM is a sort of non-volatile memory that preserves data by altering the resistance of the material within a metal-insulator-metal arrangement. Research reveals that its operation is reliant on the production and disruption of oxygen vacancies, which move under an electric field's influence, permitting the creation or dissolution of conductive filaments. The switching mechanism of RRAM largely includes the creation and severance of these conductive channels, with the SET/RESET process driven by the migration of oxygen ions and vacancies. The performance of RRAM is strongly influenced by materials such as HfO_x and TiO_x [7]. The SET/RESET stages switching processes of RRAM are presented in Fig. 1.

The main advantages of RRAM include excellent scalability with sizes less than 10 nanometers and a simple structure, nanosecond-level high-speed switching that is faster than flash memory, low power consumption making it suitable for mobile devices, long data retention time exceeding 10 years, and compatibility with complementary metal-oxide-semiconductor (CMOS) processes which facilitates integration.

However, RRAM additionally presents several difficulties, primarily concerning reliability issues related to durability, data retention, and device consistency [8]. Although the endurance of the device can achieve up to 10^{12} cycles, which remains inadequate for applications demanding ultrahigh write-cycle requirements, such as DRAM replacement.

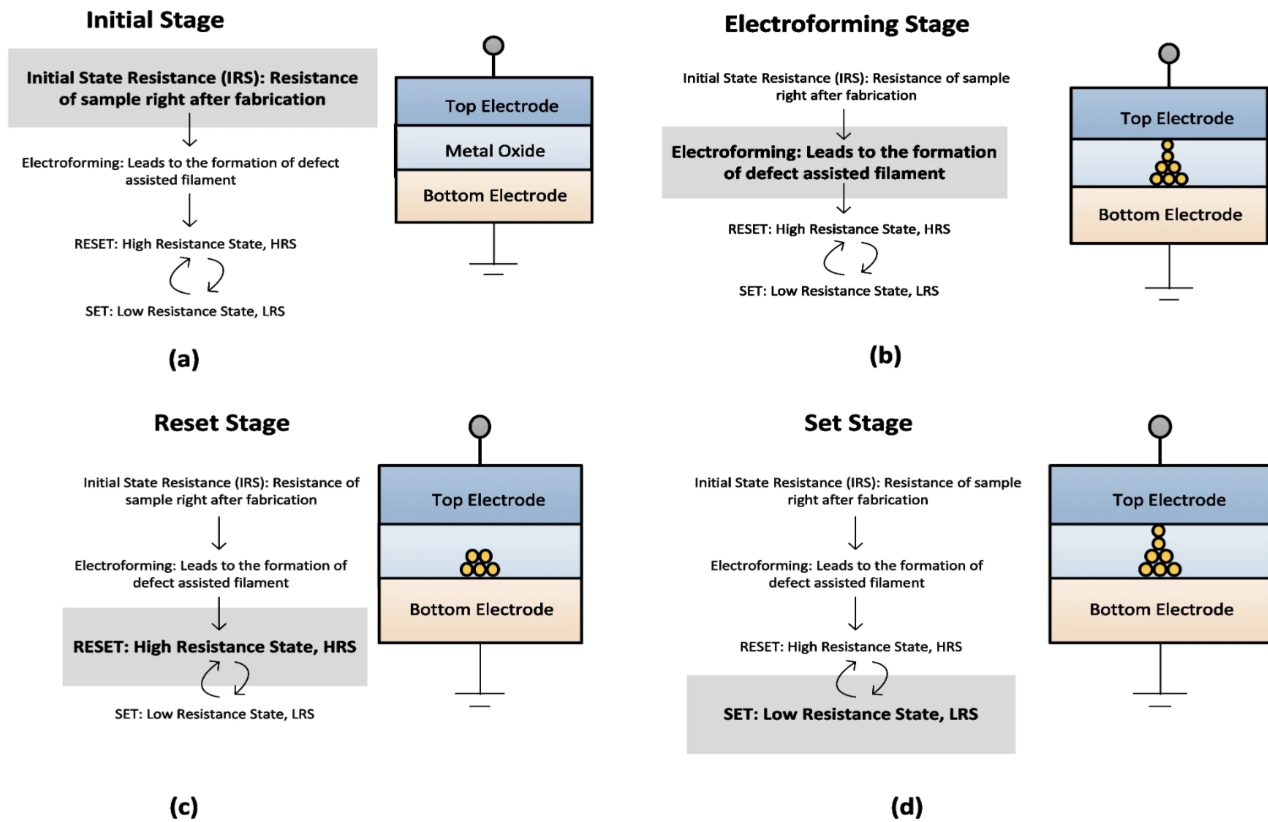

Fig. 1 Description of the switching stages of RRAM [1]

2.2. MRAM Technology

MRAM is a non-volatile storage device that exploits the spin of electrons to store data. At its heart is the Magnetic Tunnel Junction (MTJ), which contains a fixed reference layer and a variable free layer separated by an insulating layer. Parallel and anti-parallel magnetic alignments between the two layers are represented by low-resistance (logic '0') and high-resistance (logic '1') states in the device, respectively. Read operations are accomplished by sensing resistance [9]. Write operations commonly exploit the Spin-Transfer Torque (STT) phenomenon, where the current affects the magnetization direction of the free layer.

MRAM has major advantages such as non-volatility, high-speed read and write capabilities (equivalent to SRAM), low power consumption, and great endurance. For instance, STT-MRAM has been employed in the automotive sector, greatly cutting code update times; spin-orbit torque MRAM (SOT-MRAM), with its sub-nanosecond speed and good durability, is suited for cache applications.

However, MRAM also tackles challenges: high production costs, complexity, inferior storage density compared to DRAM and flash memory, slower read speeds due to tunneling magneto resistance (TMR), and difficulties relating to data retention, consistency, and write window imposed by the shrinkage of MTJ [10]. These hurdles hinder its commercialization and scalability, necessitating optimization of materials and processes.

2.3. FeRAM Technology

Based on the reversible polarization of ferroelectric capacitors, FeRAM is a technology that makes use of the polarization state of ferroelectric materials to accomplish non-volatile storage. Traditional FeRAM includes ferroelectric materials such as lead zirconate titanate (PZT), however, in recent years, hafnium oxide (HfO_2)-based ferroelectric films have drawn interest due to their outstanding scalability and compatibility with CMOS technology [11]. HfO_2-based FeRAM eliminates the size effects and compatibility concerns associated with conventional materials in high-density integration, presenting itself as a significant direction for storage technology in the post-Moore era.

FeRAM offers several advantages, including non-volatility, rapid read and write rates of less than 50 nanoseconds, low power consumption ideal for mobile devices, and high endurance exceeding 10^{13} switches. HfO_2-based FeRAM has demonstrated tremendous gains in density and performance, making it suited for applications in high-performance computing, the internet of things (IoT), and smart devices.

Nevertheless, FeRAM also confronts other issues. Specifically, in HfO2-based materials, the wake-up effect and fatigue mechanisms are major concerns determining device reliability and lifespan. Additionally, the requirement for low-temperature processing during manufacture below 650 degrees for ferroelectric behavior and reliability factors such as fatigue and the need for improvement in initial polarization values limit its continued expansion. Addressing these problems will require material optimization and process advances. MRAM is a non-volatile storage device that exploits the spin of electrons to store data. At its heart is the MTJ, which contains a fixed reference layer and a variable free layer separated by an insulating layer. When the magnetization orientations of the two layers are parallel, the resistance is low (logic 0), and when they are anti-parallel, the resistance is high (logic 1).

3. Application of RRAM, MRAM, and FeRAM

3.1. Neuromorphic RRAM (NeuRRAM) chip based on RRAM

The application of RRAM in in-memory computing architectures is predominantly observed in neural network computations. An exemplary case is the NeuRRAM chip, which amalgamates 3 million RRAM devices with CMOS technology to execute in-memory computing for artificial intelligence applications [9]. The chip structure utilizes a 1T1R configuration, facilitating bi-directional transpose neural synapse arrays (TNSA), with each core consisting of 256×256 RRAM cells and 256 CMOS neuron circuits. Typical application scenarios encompass image classification, such as ResNet-20 attaining a 14.34% error rate on the Canadian Institute for advanced research 10-class dataset (CIFAR-10), along with voice command recognition and image recovery [12]. NeuRRAM demonstrates exceptional energy efficiency and adaptability, achieving 96.6% accuracy in handwritten digit recognition and 91.5% in image classification without necessitating supplementary training.

The performance of RRAM is defined by rapid switching speeds less than 10ns, minimal writing energy between 0.1 and 1 pJ per bit, and moderate endurance ranging from 10^5 to 10^8 cycles, alongside an on/off ratio of 10 to 100, enabling 2-bit multi-level cell operation. Nonetheless, unpredictability and reliability persist as concerns. Optimization strategies encompass the enhancement of the switching window through material doping, such as NH4OH doping in Ta_2O_5 elevating the on/off ratio to 9.12×10^2, nanostructure embedding, for example MoS_2-Pd NPs augmenting the window by 32 times, and annealing processes; the improvement of endurance via bilayer configurations, such as TaO_x bilayers increasing endurance to 10^{12} cycles; and the enhancement of synaptic linearity through pulse modulation and interlayer insertion, for instance Al doping in HfO_2.

3.2. Application of MRAM

MRAM, particularly STT-MRAM and SOT-MRAM, is highly suitable for in-memory computing, neuromorphic computing, edge computing, and other applications due to its high endurance, low latency, and energy efficiency. Fast write and read speeds below 10ns, astonishingly low write energy approximately 100 fJ/bit for STT-MRAM and less than 100 fJ/bit for SOT-MRAM, and excellent endurance of more than 10^{15} cycles define MRAM's performance. In in-memory computing, SOT-MRAM supports high-precision floating-point operations, achieving up to 26.9 GOPS/W energy efficiency. In neuromorphic computing, MRAM's magnetic tunnel junctions emulate neurons and synapses, enabling low-power AI hardware. For edge computing, MRAM is used in IoT nodes,

medical devices, and RFID tags, offering low latency and power savings. Additionally, MRAM serves as a standalone memory, replacing SRAM/DRAM, and as embedded memory in SoCs, reducing power in IoT devices. Nevertheless, the structure suffers from a very low on/off ratio (1.5-2), and multi-level cell operation is usually limited to 1 bit [13]. Increasing the Tunnel Magnetoresistance (TMR) ratio will increase state distinguishability, produce dependable multi-level cell operating approaches, and lower write current through circuit design optimization, so enhancing energy efficiency.

3.3. Neural network applications based on FeRAM

FeRAM's low power consumption and great dependability are mostly driving its integration into CIM systems, especially in applications like multi-level ferroelectric field-effect transistors (FeFETs) for multi-bit MAC operations [14]. FeFET devices made of 28 nm high-K metal gate (HKMG) technology in a 1FeFET-1R configuration help to enable 2-bit weight storage for deep neural network inference. With a loss of less than 2% against the 99.11% accuracy of the floating-point model, the LeNet model achieves 96.64% accuracy on the MNIST dataset; the VGG-19 model attains 91.55% accuracy on the CIFAR-10 dataset, with a loss of less than 2% [14]. The findings show that the models are fit for wearables and sensors, among edge AI devices.

Although multi-level operation is normally confined to 1 bit, FeRAM indicates a write time of roughly 30 ns, a read time of less than 10 ns, a write energy of around 100 fJ/bit, endurance of up to cycles, and a significant on/off ratio [15]. Optimization efforts are focused on three areas: attaining more density by technology scaling, enhancing write speed by reducing latency through enhanced ferroelectric materials, and improving the dependability of ferroelectric materials by extending retention time through annealing processes.

3.4. Comparative Analysis and Development Trends

DRAM, SRAM, and RRAM all support random access, allowing data to be retrieved directly from any memory address without the need for sequential dependency. This characteristic is crucial for efficient data processing in modern computing systems. Despite their distinct operational mechanisms, such as DRAM requiring periodic refresh, SRAM offering stable storage without refresh, and RRAM providing non-volatility, their fundamental purpose remains consistent: to store data for system use. Each type serves as a critical component in electronic devices for managing and maintaining data. Moreover, DRAM, SRAM, and RRAM are integral to computer systems, providing the necessary memory infrastructure to support a wide range of applications. Whether facilitating high-speed processing or enabling persistent data storage, these memory types are essential to the functionality and performance of modern computing architectures.

4. Analysis of RRAM, MRAM, and FeRAM

Table 1 indicates that MRAM has superior endurance and minimal write energy consumption, making it well-suited for high-performance computation. RRAM has a greater on/off ratio and multi-level cell capability, ideal for high-density storage. FeRAM achieves a balance in endurance and on/off ratio, but with slower write speeds, rendering it suited for low-power embedded applications.

Table 1. Comparison of three RRAM, MRAM, and FeRAM devices

Parameters	RRAM	MRAM	FeRAM
Write time (ns)	<10	<10	~30
Read Time(ns)	<10	<10	<10
Write Energy(fJ)	100-1000	~100	~100
Endurance	10^5-10^8	$>10^{15}$	10^{10}
On/Off Ratio	10-100	1.5-2	100-1000
Multilevel(bit)	2	1	1

RRAM is still being developed. Several enterprises have started developing embedded RRAM chips in substantial quantities for use in mini-LED, LCD, and TV driver integrated circuits. Variability and dependability difficulties, however, continue to be substantial impediments to commercialization. While RRAM faces these hurdles, MRAM has made notable strides in its technological maturity. MRAM technology is comparatively sophisticated. Everspin expects to mass-produce 64Mb and 128Mb chips in 2025, particularly for artificial intelligence gear, and STT-MRAM has entered the production stage. The purpose of SOT-MRAM's continued development is to further boost performance. in contrast to MRAM's rapid progress, FeRAM has carved out a niche in specific markets. FeRAM has already found use in some specialized industries, including IoT and automotive systems. With a compound annual growth rate of around 5%. Nevertheless, there are still numerous barriers to overcome before FeRAM can be scaled to higher-density structures, such as 3D structures.

By exploring phase separation methods and building higher-density memory arrays, for instance, future RRAM development will concentrate on boosting dependability and minimizing variability. While additional study is necessary, it reveals significant potential in the field of neuromorphic computing. The primary objectives for MRAM include enhancing the TMR ratio, advancing reliable multi-level storage units via circuit optimization, and reducing write current to improve energy efficiency. MRAM presents numerous potential applications in industrial IoT and edge AI. Future FeRAM initiatives will focus on reducing process nodes to 7 nm, improving write speed using novel ferroelectric materials, and combining with sophisticated logic processes to promote its development in embedded applications. FeRAM is increasingly sought after, particularly for low-power applications.

5. Summary

This paper investigates emerging non-volatile memory technologies, RRAM, MRAM, and FeRAM have shown considerable advantages over traditional memory for data-intensive and compute-in-memory applications. RRAM stands out for its potential in high-density and neuromorphic computing, albeit with persistent issues relating to device variability and endurance. MRAM's maturity, outstanding endurance, and energy efficiency make it suited for high-performance and industrial applications, with further breakthroughs in multi-level storage and TMR improvements projected to expand its position in edge AI and IoT. FeRAM, benefiting from high durability and ultralow power operation, is increasingly utilized in embedded and low-power situations, however, process scaling and device optimization remain necessary for broader deployment. Overall, these memory technologies are positioned to promote breakthroughs in future computer systems, supporting more efficient, scalable, and intelligent architectures. Continued research into material science, device engineering, and system-level integration will be needed to fully exploit their potential across a wide spectrum of unique applications.

References

[1] Furqan Z, Azmadi H F, Bature I U, et al. Resistive random access memory: introduction to device mechanism, materials and application to neuromorphic computing. Discover Nano, 2023, 18(1).

[2] Van-Duong N, Sreeramulu R, Koen W, et al. Recent progress in spin-orbit torque magnetic random-access memory. npj Spintronics, 2024, 2:48.

[3] Jeong-Yoon L, Jaeho S, Seonghyeon C, Jae-Min S, Young-Hee S. A Novel 3D 2TnC FeRAM Architecture and Operation Scheme with Improved Disturbance for High-Bit-Density Dynamic Random-Access Memory. Electronics, 2024, 13(22): 4474.

[4] Alan L. Davis. Computer architecture. IEEE Spectrum, 1983, 20(11): 94-99.

[5] Onur M, Saugata G, Juan G-L, Rachata A. Processing data where it makes sense: Enabling in-memory computation. Microprocessors and Microsystems, 2019, 67: 28-41.

[6] Pietro M, Matteo F, Niccolò L, et al. In-memory computing with emerging memory devices: Status and outlook. APL Machine Learning, 2023, 1(1): 010902.

[7] Deepak P, Pradipta K S, Tseung-Yuen T. A Collective Study on Modeling and Simulation of Resistive Random Access Memory. Nanoscale Research Letters, 2018, 13:8.

[8] Furqan Z, Tengku Zulfikri Azni Z, Fawnizu Azmadi K. Resistive Random Access Memory (RRAM): an Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications. Nanoscale Research Letters, 2020, 15:90.

[9] Shoko I, Frank B M, Jim J, Sumit A. Magnetoresistive Random Access Memory: Present and Future. IEEE Transactions on Electron Devices, 2020, 67(4): 1407-1419.

[10] Cai Kaiming, Jin Tianli, Lew Wen Siang. Spin-based magnetic random-access memory for high-performance computing. National Science Review, 2024, 11(3): nwad272.

[11] Liao Jiajia, Dai Siwei, Peng Ren-Ci, et al. HfO2-based ferroelectric thin film and memory device applications in the post-Moore era: A review. Fundamental Research, 2023, 3(3): 332-345.

[12] Wen Wan, Ranjith K, Carsten S, et al. A compute-in-memory chip based on resistive random-access memory. Nature, 2022, 608: 504-512.

[13] Jeonghyun Y, Young-Woong S, Wooho H, et al. A review on device requirements of resistive random access memory (RRAM)-based neuromorphic computing. APL Materials, 2023, 11(9): 090701.

[14] Pietro M, Matteo F, Niccolò L, et al. In-memory computing with emerging memory devices: Status and outlook. APL Machine Learning, 2023, 1(1): 010902.

[15] Tarek S, Sourav C, Nicholas L, et al. First demonstration of in-memory computing crossbar using multi-level Cell FeFET. Nature Communications, 2023, 14:6348.