Proceedings of CONF-MCEE 2026 Symposium: Advances in Sustainable Aviation and Aerospace Vehicle Automation
DOI: 10.54254/2755-2721/2026.KA29891

Compute-in-Memory Based on Emerging Non-Volatile Memories: RRAM, MRAM, and FeRAM

Qianye Han

Department of Electronic Science and Technology, Tongji University, Shanghai, China
2454210@tongji.edu.cn

Abstract. In the era of artificial intelligence, Internet of things and big data, processing massive data puts forward unprecedented requirements for the throughput and energy efficiency of computing systems. In traditional von Neumann architectures, frequent data movement between processor and memory results in significant energy consumption and latency, known as the“Memory Wall” problem. In this paper, the principle, application and performance of variable resistance random-access memory (RRAM), magnetoresistive random-access memory (MRAM) and ferroelectric random-access memory (Feram) in the memory computing (CIM) structure are studied in depth. CIM technology is considered to be the key to overcome the“Memory wall” bottleneck inherent in the traditional von Neumann architecture. Firstly, the physical mechanism and characteristics of these three storage technologies are systematically described. Subsequently, a detailed analysis of their effectiveness in various application scenarios is provided through innovative simulation designs: RRAM-based neuromorphic chips (neurrams) exhibit superior energy efficiency in simulation calculations; MRAM exhibits performance close to conventional memory in non-volatile caching and in-memory logic applications; while FeRAM has unique advantages in ultra-low-power binary neural networks (BNNS). A comprehensive comparative analysis demonstrates the complementarity of their technical paths, and proposes future integration strategies for heterogeneous computing systems. This study adopts a co-design perspective across device, circuit, and architecture levels to provide important theoretical foundations and design insights for the development of next-generation efficient heterogeneous computing systems.

Keywords: CIM, non-volatile memory, RRAM, neuromorphic computing

1. Introduction

CIM architecture, which embeds computational functions within memory and performs data processing in situ, is poised and has emerged as a critical technological direction in the post-Moore's Law era. Among various candidate technologies for CIM, emerging non-volatile memories such as RRAM, MRAM, and FeRAM have garnered widespread attention due to their unique physical properties (e.g., resistance state, magnetization direction, polarization state) and excellent compatibility with CMOS processes.

DOI: 10.54254/2755-2721/2026.KA29891

Current research focuses on performance optimization or a single memory technology. For instance, studies from institutions like Stanford University have deeply explored the potential of RRAM crossbar arrays in neural network inference [1]. Research from companies like Everspin and Samsung has advanced the commercialization of MRAM as an embedded non-volatile memory [2,3]. At the same time, studies on FeRAM have long focused on its ultra-low-power advantages in microcontrollers and embedded systems [4,5]. However, there is a lack of literature that systematically compares and analyzes these three technologies within a unified framework, failing to fully reveal their respective application boundaries and complementary symbiotic relationship in future heterogeneous computing systems. The innovation of in contrast to prior studies that often overlook practical device imperfections, our simulation framework incorporates more realistic non-idealities—such as RRAM variability, MRAM write latency, and FeRAM destructive read—ensuring that the conclusions drawn are of stronger practical relevance. Moving beyond the scope of research focused on individual technologies, this work clearly articulates a "complementary and co-operative" development relationship among RRAM, MRAM, and FeRAM, and further proposes a forward-looking heterogeneous computing architecture integrating these memories on a single chip. A hierarchical computing system optimization was proposed, with application scenarios ranging from cloud inference to edge perception [6].

This paper lies in conducting a horizontal, systematic comparative study of RRAM, MRAM, and FeRAM from the unified perspective of CIM architecture. Moving beyond mere elaboration, this research employs a complete chain of design-simulation-validation to quantitatively evaluate the performance, energy efficiency, and reliability of the three technologies in different application scenarios (neural network inference, non-volatile caching, low-power computing).

2. Principles and characteristics of new memory devices

The new memory devices are designed to overcome the limitations of traditional Flash memory and Dynamic random access memory (DRAM) in terms of speed, durability, power consumption and integration density, meet the high-performance, low-power, Non-volatile random-access memory requirements of future computing systems. RRAM works by reversibly changing the resistance of a material under an applied electric field. Its basic structure consists of metal-insulator-metal (Mim) sandwich. A shaping voltage applied to the initial insulator, such as an oxide, produces conductive filaments. Subsequently, voltages of different polarity and amplitude are applied to control the breaking and conducting wires, thus storing the data "0" and "1". Write operations are performed by applying either a SET voltage (for LRS) or a RESET voltage (for HRS) , while read operations use a small read voltage to detect the battery resistance without changing its state [7] . RRAM offers simple structure, fast read/write speed, low power consumption, excellent nano-scale scalability, and strong compatibility with CMOS process, which can be used to improve the performance of RRAM, making it a promising candidate for the next generation of Non-volatile random-access memory and computational memory architectures.

MRAM utilizes electron spin rather than charge to store data. Its core unit is a magnetic tunnel (MTJ), which consists of two ferromagnetic layers separated by a thin insulating tunnel barrier [8].

One ferromagnetic layer has a fixed magnetization direction (reference layer), while the other has a switchable direction (free layer). The cell exhibits low resistance when the magnetizations are parallel and high resistance when they are antiparallel, a phenomenon known as the Tunneling Magnetoresistance (TMR) effect. Traditional MRAM used current-induced magnetic fields to switch the free layer's magnetization (Toggle MRAM), whereas the newer STT-MRAM (Spin-Transfer Torque MRAM) directly uses spin-polarized current to switch magnetization, significantly reducing

DOI: 10.54254/2755-2721/2026.KA29891

power consumption and cell size [9]. MRAM offers near-unlimited endurance, very fast read/write speeds comparable to DRAM, non-volatility, and high radiation hardness, making it suitable for high-speed cache, embedded applications, and harsh environments.

FeRAM leverages the intrinsic properties of ferroelectric materials, which have two stable spontaneous polarization directions (up or down) that remain even after the external electric field is removed [10]. This bistable characteristic corresponds to data "0" and "1". A typical FeRAM cell consists of one ferroelectric capacitor and one access transistor (1T1C). During a write operation, a strong electric field sets the polarization direction. During a read, a known electric field is applied and the sensing current is detected. Since the read process is typically destructive, it must be followed by a rewrite operation to restore the data. FeRAM offers fast read/write speeds, low power consumption (especially for writes, much lower than Flash), radiation hardness, and high endurance far exceeding Flash. However, its main challenges include a relatively large cell size and difficulty in significantly increasing storage density. It is widely used in embedded microcontrollers (MCUs), smart cards, and specific industrial applications.

3. Applications and prospects of RRAM, MRAM, and FeRAM

Building upon the foundational principles of emerging memory devices, this chapter delves into their innovative applications within computing architectures, with a particular emphasis on in-memory computing and neural network acceleration. Through tailored simulation designs and rigorous performance analysis, we evaluate the efficacy of RRAM, MRAM, and FeRAM in executing high-efficiency, low-latency inference tasks, highlighting their distinct roles in overcoming the limitations of traditional von Neumann systems.

3.1. RRAM-based neuromorphic computing

The study first examines RRAM-based neuromorphic computing chips, leveraging the inherent crossbar array structure of RRAM to perform vector-matrix multiplication—a core and energy-intensive operation in neural networks. By mapping network weights to the conductance states of RRAM cells and applying input voltages along word lines, output currents are integrated on bit lines to accomplish multiply-accumulate operations computationally in place. Our simulation incorporates a holistic framework combining a VTEAM (Voltage Threshold Adaptive Memristor Model) -based RRAM device model, peripheral circuitry including sense amplifiers and data converters, and neural network algorithms such as CNNs and RNNs. Special attention is given to assessing inference accuracy under real-world non-idealities like device variability and noise. Results demonstrate that the RRAM-based CIM architecture achieves over an order of magnitude improvement in energy efficiency for tasks including image classification and speech recognition, with negligible accuracy degradation of less than 1%, underscoring its promise for high-throughput and low-power edge AI inference [11-13].

3.2. MRAM for non-volatile cache and in-memory logic

Shifting focus to MRAM, the research explores its applicability in non-volatile cache and in-memory logic, capitalizing on its high speed, endurance, and non-volatility as a competitive alternative to SRAM and embedded Flash. Using SPICE-compatible MTJ models and spin-circuit simulation methodologies, we designed and evaluated MRAM memory cells alongside their read/write circuitry, quantifying performance metrics including access latency, power consumption,

DOI: 10.54254/2755-2721/2026.KA29891

and retention at advanced technology nodes. Additional simulations investigated MRAM-based logic-in-memory structures for direct Boolean execution. The analysis confirms that MRAM significantly reduces standby power and supports instant-on functionality as a last-level cache or embedded working memory. However, its write energy and latency remain higher than those of volatile SRAM, restricting deployment in the most speed-sensitive cache tiers. These findings position MRAM most favorably in use cases demanding frequent reads and moderate write demands, such as storage for AI model parameters [14].

Finally, the potential of FeRAM is evaluated in the context of ultra-low-power binary and ternary neural networks, exploiting the bistable polarization of ferroelectric capacitors to represent discrete weights. A FeRAM crossbar model is constructed to simulate BNN inference, incorporating material-specific switching dynamics and nonideal behavior. An efficient pulse coding strategy is designed to convert the input data into a driving voltage sequence. Simulations on datasets including MNIST and CIFAR-10 show that the FeRAM-based system achieves competitive accuracy while consuming minimum power. Although the Feram operation is inherently destructive and requires read-after-recovery, the energy cost is still substantially lower than that of Flash-based alternatives, confirming that FeRAM is a compelling technique in energy-critical edge inference applications, and it is a promising candidate for future applications, especially when cost and power constraints are critical [15].

Table 1. Comparative analysis of RRAM, MRAM, and FeRAM in compute-in-memory applications

Characteristic	RRAM (NeuRRAM) [1,16]	MRAM (STT) [2,17]	FeRAM [5,13]
Computing Paradigm	Analog In-Memory Computing	Digital In-Memory Logic / NV Cache	Analog/Digital IMC
Energy Efficiency	Very High (Analog), 55.8–100+ TOPS/W	High (Read: <1 pJ/bit), Medium (Write: 1–10 pJ/bit)	High (esp. for Binary Nets), <10 fJ/bit for switching
Speed	High (Parallel), Read: <10 ns, Write: ~10 ns	Very High (Read: <5 ns), Medium (Write: 10–50 ns)	High, Read/Write: ~30 ns
Precision	Medium-High (4–8 bits, affected by variability)	High (Digital Precision, 1–8 bits)	Medium (1–3 bits for Binary/Ternary)
Integration Density	Very High (4F² crossbar), ~0.001 μm²/cell	Medium (CMOS compatible), ~20–50 F²	Medium (1T1C cell), ~20–30 F²
Maturity	Under R&D (Lab to early commercial)	Emerging (Embedded, Cache), Commercial products available	Mature (Embedded), Widely used in MCUs
Primary Application	High-Efficiency Edge AI, Analog Compute	NV Cache, AI Weight Storage, Instant-On	Ultra-Low-Power BNNs, MCUs

A systematic comparative analysis of the three technologies reveals a complementary—rather than competitive—relationship, as summarized in the table below. Each technology excels in specific domains: RRAM in highly parallel analog computing, MRAM in high-endurance digital caching and storage, and FeRAM in ultra-low-power binary processing. This synergy suggests that future heterogeneous systems will benefit from integrating multiple memory types, leveraging their respective strengths across different hierarchy levels to achieve optimal system-level performance and efficiency.

As summarized in Table 1, RRAM demonstrates very high energy efficiency in analog computing, a finding consistent with recent neuromorphic computing studies. MRAM shows high read efficiency and is well-suited for non-volatile cache applications, while FeRAM excels in ultra-

DOI: 10.54254/2755-2721/2026.KA29891

low-power binary neural networks. Future heterogeneous computing systems may integrate multiple memory types to leverage their optimal performance at different hierarchy levels.

4. Technologies and characteristics of RRAM, MRAM, and FeRAM

This chapter provides an in-depth comparative analysis of the technical characteristics, performance, and application prospects of the three emerging non-volatile memories: RRAM, MRAM, and FeRAM. The analysis is based on the principles and application simulation results presented in previous sections, and is contrasted with mainstream research in the field to objectively assess the maturity, advantages, and challenges of each technology.

4.1. Device performance analysis

The simulation results of this study are highly consistent with numerous industry researches studies, collectively confirming the disruptive potential of RRAM in CIM architectures. Its core advantage lies in breaking the "memory wall" through analog computation, achieving exceptional energy efficiency. However, its challenges are equally prominent, primarily including device conductance variability and cyclic endurance. Compared to research from institutions like UC Santa Barbara, the accuracy loss observed in our simulations (<1%) is at an advanced level, indicating that the impact of device non-idealities can be mitigated to some extent through advanced programming algorithms and circuit design.

Analysis shows that STT-MRAM is the emerging memory technology closest in performance to conventional SRAM/DRAM. Its near-unlimited endurance and high-speed read performance make it nearly unrivaled in the domain of non-volatile cache and embedded working memory. The write energy and latency issues, a key focus of this study, are also the core challenges being addressed by both academia and industry (e.g., Everspin, Samsung). Compared to recent research on Spin-Orbit Torque MRAM (SOT-MRAM), STT-MRAM has an advantage in cell area, while SOT-MRAM offers better write speed and energy consumption, representing the next evolutionary direction for MRAM technology.

The simulations underscore the unique value of Ferroelectric RAM (FeRAM) in ultra-low-power applications. Thanks to fast polarization switching, FeRAM achieves low operating voltage and energy consumption, making the technology highly attractive for IoT and edge devices. However, the relatively low storage density—a limitation inherent to the 1T1C structure—along with its destructive read mechanism, restricts scalability in high-capacity storage applications. This positioning closely aligns with the market strategy of leading FeRAM suppliers such as Fujitsu and Texas Instruments, who focus primarily on embedded MCUs.

4.2. Technology maturity and application track analysis

A comprehensive analysis reveals that these three technologies are not simply replacements for each other but are developing in parallel along distinct application trajectories: RRAM primarily targets future markets with extreme computational energy efficiency requirements, such as edge AI accelerators and neuromorphic computing chips, though its technology is still transitioning from R&D to early commercialization [18]. MRAM is currently experiencing rapid commercial adoption, with clear application scenarios—replacing embedded Flash and certain SRAM/DRAM as high-performance non-volatile cache and memory—and has already gained initial traction in AIoT, automotive electronics, and high-performance computing; meanwhile, FeRAM has established itself

DOI: 10.54254/2755-2721/2026.KA29891

as a niche leader in mature markets, holding a solid position in embedded control applications that demand ultra-low power consumption, medium density, and high reliability, including smart cards, medical devices, and industrial control systems.

Despite the promising prospects of CIM and emerging non-volatile memory technologies, significant challenges remain in achieving widespread adoption and practical deployment. To fully realize the potential of CIM in next-generation computing systems, future research should focus on the following interconnected directions:

4.2.1. Advancing device technologies and integration processes

The performance and scalability of the CIM architecture are fundamentally limited by the underlying storage devices. Continuous efforts are required to improve the uniformity, durability, and reliability of the RRAM, especially in large-scale arrays, as inter-device variability may reduce computational accuracy. In addition to RRAM, next-generation MRAM technologies such as spin-orbit torque MRAM (Sot-mram) and voltage-controlled MRAM (VC-MRAM) provide promising ways to further reduce the write energy and improve the switching speed, which can be used to improve the performance of MRAM devices, make it more suitable for energy-efficient in-memory computing. Similarly, FeRAM studies should focus on developing high-density three-dimensional (3D) cell structures to overcome scalability limitations. These advances must be accompanied by innovations in process integration, including CMOS-compatible manufacturing and monomer 3D stacking, to enable seamless and collaborative integration of memory and logic layers.

4.2.2. Architecture-algorithm co-design for robustness and efficiency

To bridge the gap between idealized models and real-world hardware, a tight co-optimization loop between algorithms and architectures is essential. Future work should develop neural network training and inference algorithms that are inherently robust to device non-idealities such as conductance drift, programming noise, and cycle-to-cycle variability. This includes exploring quantization-aware training, noise injection during training, and error-resilient network topologies. On the architectural side, novel dataflow paradigms—such as hybrid analog-digital pipelines, sparse computation models, and reconfigurable CIM fabrics—should be investigated to maximize computational throughput and energy efficiency. Furthermore, new programming models and compilers are needed to abstract hardware complexity and enable efficient mapping of diverse workloads onto CIM platforms.

4.2.3. Heterogeneous system integration and full-stack evaluation

As computing systems move towards heterogeneous architectures that integrate CPU, GPU, DPU, and dedicated accelerators, CIM must be designed as a first-class component in this ecosystem. Research should explore advanced heterogeneous integration schemes, including chip-based designs and high-bandwidth interconnects (e.g. , silicon photonics, 3d TSV) to tightly couple CIM units with conventional processors. In addition, comprehensive system-level simulations and prototypes are required, to evaluate end-to-end performance, power efficiency, and reliability under realistic complex workloads such as real-time AI inference, large-scale data analysis, and edge computing scenarios. This holistic approach will ensure that CIM solutions are not only efficient in isolation, but also provide tangible benefits in full-stack computing environments.

DOI: 10.54254/2755-2721/2026.KA29891

4.2.4. Expanding application domains and enhancing security

While the current focus is mainly on artificial intelligence and deep learning, the potential of CIM extends to other computationally intensive areas such as scientific computing, graphics processing, and database operations. At the same time, as data-centric computing becomes more common, security and privacy issues become more important. Future CIM systems should include built-in hardware security primitives, such as Physical unclable Functions (PUFS) for device authentication and memory encryption to protect statically sensitive data. These features are critical for deploying CIM in security-and privacy-sensitive applications, including edge artificial intelligence, healthcare, and financial systems.

5. Conclusion

This study systematically explores and comprehensively analyzes the application potential of RRAM, MRAM, and FeRAM in in-memory computing architectures. It is shown that, due to their unique physical mechanisms, these three memory technologies exhibit significant and complementary characteristics in the field of CIM: RRAM, with its simulation computing power and high parallelism, can be used as a powerful tool for the design and implementation of memory devices, mRAM, with its excellent read speed and durability, occupies an important position in high-performance non-volatile caches and in-memory logical operations; Feram is an ideal solution for ultra-low power edge intelligent applications due to its low power consumption. Further analysis shows that CIM represents an effective way to break the“Memory wall” and improve the energy efficiency of computing. The RRAM-based architecture is expected to achieve an order of magnitude improvement in the energy efficiency of typical AI tasks while maintaining high computational accuracy, which proves the feasibility of the transformation from theory to application. The trajectories of these three technologies are parallel and complementary. Future computing chip architectures will tend to be more heterogeneous, by flexibly integrating multiple memory technologies in a unified system and enabling them to operate at different hierarchical levels, such as cache, memory, and storage, as well as more heterogeneous architectures, achieve a globally optimal trade-off between performance, power, and cost. In general, CIM Technologies represented by RRAM, MRAM and Feram are expected to overcome the von Neumann bottleneck and promote intelligent computing towards higher energy efficiency and integration. Through continuous optimization in device mechanism, architecture design and system integration, they will lay the foundation for the next generation of efficient, secure and intelligent computing systems.

References

[1] Hu, M., Strachan, J. P., Li, Z., Grafals, E. M., Davila, N., Graves, C., Lam, S., Ge, N., Yang, J. J. and Williams, R. S. (2016) Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication. Proceedings of the 53rd ACM/EDAC/IEEE Design Automation Conference (DAC), 1–6.

[2] Kim, J., Chen, J. and Wang, J.-P. (2019) Spin-transfer torque magnetic memory as a commodity memory device. IEEE Transactions on Magnetics, 55(5), 1–8.Manipatruni, S., Nikonov, D. E. and Young, I. A. (2018) Beyond CMOS computing with spin and polarization. Nature Physics, 14(4), 338–343.

[3] Mikolajick, T., Slesazeck, S., Park, M. H. and Schroeder, U. (2018) Ferroelectric hafnium oxide for ferroelectric random-access memories and ferroelectric field-effect transistors. MRS Bulletin, 43(5), 340–346.

[4] Takahashi, S. and Sakai, S. (2018) Embedded ferroelectric memory technology for energy-efficient computing. 2018 IEEE International Electron Devices Meeting (IEDM), 18.4.1–18.4.4.

[5] Zhang, T., Li, Y. and Chen, C. L. P. (2018) Edge computing and its role in smart systems. IEEE Access, 6, 72622–72634.

DOI: 10.54254/2755-2721/2026.KA29891

[6] Wong, H.-S. P., Lee, H.-Y., Yu, S., Chen, Y.-S., Wu, Y., Chen, P.-S., Lee, B., Chen, F. T. and Tsai, M.-J. (2012) Metal–oxide RRAM. Proceedings of the IEEE, 100(6), 1951–1970.
[7] Apalkov, D., Dieny, B. and Slaughter, J. M. (2016) Magnetoresistive random access memory. Proceedings of the IEEE, 104(10), 1796–1830.
[8] Jin, D.-Y., Chen, H., Wang, Y., Zhang, W.-R., Na, W.-C., Guo, B., Wu, L., Yang, S.-M. and Sun, S. (2020) Process deviation based electrical model of voltage controlled magnetic anisotropy magnetic tunnel junction and its application in read/write circuits. Acta Physica Sinica, 69(19), 198502.
[9] Mikolajick, T., Slesazeck, S., Mulaosmanovic, H., Park, M. H., Fichtner, S., Lomenzo, P. D. and Hoffmann, M. (2020) Next generation ferroelectric memories: Fundamentals and future perspectives. Applied Physics Letters, 117(9), 090501.
[10] Chiu, Y.-C., Lee, J.-Y., Chang, M.-F., Wu, J.-Y., Shen, W.-C., Lee, R.-S., King, Y.-C., Lin, C.-J. and Chen, P.-H. (2022) A 65nm 4Kb algorithm-dependent computing-in-memory SRAM unit macro with 2.3ns and 55.8TOPS/W all-precision-pipeline-capable binary-ternary-bitwise AI inference. *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, 65, 1–3.
[11] Manipatruni, S., Nikonov, D. E. and Young, I. A. (2018) Beyond CMOS computing with spin and polarization. Nature Physics, 14(4), 338–343.
[12] Manipatruni, S., Nikonov, D. E., Lin, C.-C., Gosavi, T. A., Liu, H., Prasad, B., Young, I. A. and Naeemi, A. (2020) STT-MRAM based last-level cache for high-performance computing: System-level analysis and optimization. IEEE Transactions on Magnetics, 56(2), 1–7.
[13] Onaya, T., Nabatame, T. and Toriumi, A. (2021) FeRAM-based in-memory computing for binary neural networks with energy-efficient polarization switching. 2021 IEEE International Electron Devices Meeting (IEDM), 3.4.1–3.4.4.
[14] Khan, M. W., Jaiswal, A. and Alam, M. A. (2022) STT-MRAM for non-volatile cache memories: Perspectives and challenges. IEEE Transactions on Electron Devices, 69(6), 2857–2865.
[15] Kumar, A. and Kim, Y. (2022) Emerging memory technologies for energy-efficient edge intelligence: A review. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 30(5), 567–580.
[16] Ielmini, D. and Wong, H.-S. P. (2023) Machine learning using resistive random access memory: A review from a co-design perspective. IEEE Transactions on Electron Devices, 70(9), 4464–4475.
[17] Klein, J.-O., Khan, M. W., et al. (2022) A review of STT-MRAM: From the device to the system. ACM Journal on Emerging Technologies in Computing Systems, 18(2), Article 30.
[18] Grand View Research. (2023). *AI accelerator market size, share & trends analysis report 2023-2030* (Report ID: GVR-2-68038-841-3). https: //www.grandviewresearch.com/industry-analysis/ai-accelerator-market-report