A Review on Device-Based Processing-In-Memory

Zimo Li[1,*], Longxuan Li[2], Boyi Wang[3]

[1]College of Electronic Science and Engineering, Jilin University, Changchun, 130012, China
[2]Computer School, Beijing Institute of Technology Zhuhai College, Zhuhai, 519088, China.
[3]College of Artificial Intelligence , Shenyang Normal University, Shenyang, 110034, China
* lizm1923@ mails.jlu.edu.cn

Abstract. To address the "memory wall" problem that limits the computing power of processors due to the traditional von Neumann architecture's separation of storage and computation, the Processing-in-Memory architecture, which integrates storage and computation, has become a research hotspot in the current new computing architectures. The core idea of the memory-computation integration architecture is to break the traditional von Neumann bottleneck by directly processing data within the memory, thereby significantly reducing data transmission overhead and enhancing computing speed and energy efficiency. This architecture not only can solve the performance bottleneck faced by current computing systems but also provides a new technical path for emerging fields such as intelligent computing and edge computing. This paper first introduces the concept of memory-computation integration and focuses on elaborating different volatile and non-volatile storage devices. These include the principles and applications of SRAM, DRAM, PCM, MRAM, FLASH, and RRAM in the Processing-in-Memory field. Finally, it looks forward to the future development direction of Processing-in-Memory devices. It is expected that through the summary of the research progress of memory-in-operation-related devices, this paper can provide a reference for the development of memory-computation integration

Keywords: Memory Wall, Processing-In-Memory, Devices

1 Introduction

With the rapid development of information technology, particularly in the domains of big data, artificial intelligence, and machine learning, traditional computing architectures have increasingly revealed limitations in handling large-scale data and complex tasks. Although the von Neumann architecture has long underpinned advancements in computer technology, its design separating storage from computation no longer meets the demands of modern applications requiring efficient computation. The fundamental bottleneck of this architecture stems from the frequent data transfers between computing and storage units. Such high-frequency data exchange not only escalates energy consumption but also constrains overall system performance. This bottleneck becomes increasingly pronounced with growing data volumes, resulting in computational inefficiency and energy waste.Consequently,optimizing the integration

© The Author(s) 2025
J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_102

of storage and computation through reduced data transmission has emerged as a pivotal challenge in contemporary computing architecture innovation.

To address this limitation, Compute-in-Memory(CIM) technology has gained prominence.By integrating computational and storage units, CIM mitigates frequent data transfers, thereby enhancing computational efficiency and substantially lowering energy consumption.In recent years, driven by surging demands in artificial intelligence, big data processing, and high-performance computing, storage-compute convergence technologies have attracted growing attention.Within this domain, approaches such as near-memory computing (PNM), processing-in-memory (PIM), and computation-in-memory (CIM) have become focal research areas.These technologies incorporate computational capabilities into storage units through distinct methodologies,enabling parallel execution of data access and computation, which markedly elevates system processing capacity. Nevertheless, practical implementations face persistent challenges, including hardware integration, energy efficiency optimization, and computational accuracy control.

The realization of storage-compute convergence technologies critically depends on the selection of memory media. Diverse memory technologies exhibit unique characteristics in performance, cost, and energy efficiency, profoundly influencing the implementation and optimization of integrated architectures. Prominent memory technologies include static random-access memory (SRAM), dynamic random-access memory (DRAM), NOR flash memory, resistive random-access memory (RRAM/ReRAM), magnetoresistive random-access memory (MRAM), and phase-change memory(PCM).Each technology assumes distinct roles in storage-compute integrated architectures, presenting both advantages and challenges.

In this paper, we systematically explore the applications and evolution of these memory technologies in storage-compute integrated architectures.We focus on analyzin the characteristics, merits, and limitations of SRAM, DRAM, NOR Flash, RRAM/ReRAM, MRAM, and PCM, alongside their potential applications in storage-compute convergence.Additionally, we evaluate their suitability for diverse scenarios.It is anticipated that this research will provide theoretical foundations and practical insights for advancing memory-compute integration technology.

2 Overview of Integrated Memory and Computing Technology

Integrated memory and computing technology is a new type of computing architecture, which combines storage units and computing units. It aims to solve the "storage wall" problem faced by traditional von Neumann architecture. In traditional computing architecture, the storage unit and computing unit are separated from each other. Data needs to be repeatedly transferred between storage devices and processors, which not only consumes a lot of time but also energy [1]. Integrated memory and computing technology can effectively solve these problems. Its basic working principle is to integrate the computing function into the storage device and use the high parallelism of the storage unit to directly execute the computing task. This technology can satisfy demand in real time and reduce data transmission, thereby improving computing efficiency and reducing energy consumption. Integrated memory and computing architecture is suitable for data-intensive tasks such as big data processing, deep

learning and other data intensive applications. In recent years, with the rapid development of artificial intelligence and the Internet of Things, integrated memory and computing technology has been widely concerned and has become a hot topic in today's computer field. In the field of memory computing, for the device level, the main development direction is to study the device that can realize "Integrated storage and computing". It means to implement the time-division multiplexing of memory and calculation within the memory unit itself. Storage media is mainly divided into two categories based on its storage medium characteristics: "volatile memory" and "non-volatile memory". Volatile memory devices (such as SRAM and DRAM) have a long history. Although with the continuous progress of technology and the emergence of new storage devices, volatile memory still faces some competition and challenges. However, volatile memory not only has great advantages in terms of industrial scale and cost-effectiveness but also with the emergence of 3D stacking technology and through-silicon technology it has great potential for new development in memory computing. Non-volatile memory devices (such as PCM, ReRAM, MRAM and Flash memory) can maintain data without loss after power failure while providing significant improvements in performance,power consumption, integration density. Moreover, it has some performance advantages that traditional memory does not have, so it has great advantages in the development of memory computing [2]. At the same time, with the progress of technology integrated memory and computing technology is expected to play a greater role in the future and promote the innovation of computing technology [3].

2.1 DRAM

Dynamic Random Access Memory(DRAM) each storage unit usually consists of a transistor and a capacitor. When the capacitor is charged, it represents the stored data "1", and when it is discharged, it represents the data "0".DRAM uses dynamic storage units to represent data through capacitive storage charges, which makes it have a very high storage density and low cost. It is also due to this structural feature that DRAM has limited computing accuracy and weak computing capabilities. DRAM near-memory computing places computing units in close proximity to DRAM chips, which aims to reduce the transmission delay and power consumption of data between the processor and memory. With the development of artificial intelligence, big data analytics and other technologies, the traditional von Neumann architecture limits the improvement of system performance. DRAM near-memory computing makes data processing more convenient and efficient by combining computing unit with memory. It has become an important research direction in the field of computing architecture in recent years. The development history of DRAM can be traced back to the 1960s when the United States IBM researchers invented the first DRAM chip, which not only marked the birth of DRAM technology but also laid the foundation for modern DRAM technology. With the advancement of technology, DRAM near memory computing has also gone through multiple stages of development. First, the storage capacity of DRAM has been continuously increased from a few thousand bits to tens of gigabytes. Second, reducing DRAM access time improves the efficiency of data processing. Besides, the

integration of DRAM chips has also been improved, which increases storage capacity and performance. Nowadays, some new DRAM architecture designs such as Hybrid Memory Cube (HMC) and High Bandwidth Memory (HBM) provide higher bandwidth and better performance for near-memory computing. It further promotes the development of DRAM near memory computing. The 3D stacking technology vertically integrates the computing unit and DRAM to shorten the data transmission distance and improve the computing efficiency. DRAM near memory computing also has a wide range of application prospects in deep learning, big data analysis and other applications requiring high bandwidth and low latency. With the continuous progress of technology, DRAM near-memory computing will further improve the computing speed and performance.

2.2 SRAM

With the rapid development of information technology, the amount of data is increasing explosively. The delay and energy consumption of data processing have become a major bottleneck in reducing the efficiency of data processing. In this setting, SRAM came into being. Static Random Access Memory(SRAM) is a type of semiconductor memory usually composed of six metal oxide field effect transistors [4]. The characteristic of SRAM is that in the power state, the data stored in the storage unit can be constantly maintained without periodic updates like DRAM. SRAM is fully compatible with CMOS (complementary metal oxide semiconductor) logic processes with high integration, fast speed of read and write, low operating voltage, no durability limitations, and high process maturity. SRAM also has many other advantages. First of all, the access speed of SRAM is very fast, which can quickly respond to CPU requests in order to improve the computer running speed. Secondly, the power consumption of SRAM is low. It does not need to be updated regularly, so it helps to extend the battery life of the device in the standby or idle phase. Finally, SRAM has strong anti-interference ability, which can ensure the accuracy and integrity of the stored information. In addition, currently SRAM technology is widely used in the design of various chips. The emergence of SRAM can be traced back to the 1980s. As computer architectures evolved, researchers realized problems with traditional von Neumann architectures and the concept of SRAM was proposed. By the 21st century, integrated memory and computing technology had gradually matured and had achieved certain results in processing. After 2010, with the rise of artificial intelligence, the demand for computing and storage has increased dramatically.SRAM integrated memory and computing has become a research hotspot. For the past few years, SRAM integrated memory and computing technology has gradually attracted industrial attention. A number of companies have launched commercial integrated memory and computing chip products, which are applied to AI acceleration, intelligent driving and other fields. At present, although the SRAM integrated memory and computing technology has the advantages of high maturity and high robustness. It is still necessary to pay attention to issues such as architecture design, reliability, area optimization, and computing power improvement.

2.3 PCM

Phase-change memory (PCM) is a memory technology that uses phase-change materials as the storage medium [5]. The basic mechanism of PCM was invented by Robert Ovshinsky of Stanford University in the 1960s. It involves the transformation of phase-change materials between a solid phase and a liquid phase. By controlling the phase transition process between the solid and liquid phases of the phase-change material, data can be stored and retrieved. The core material of PCM is typically a germanium-antimony-tellurium (GST) alloy, which has a lower resistance in the crystalline state and a higher resistance in the amorphous state. Compared to traditional hard disks and flash memory, PCM has a much higher read and write speed than flash memory, which is essential for tasks that require large amounts of data. The write speed of flash memory is limited by the speed of charge movement. PCM stores data through the solid and liquid phase transitions of phase-change materials and can switch states in real time. Due to its non-volatility, PCM retains information even when the power is disconnected, making it suitable for tasks such as storing important data backups. PCM can store relatively larger amounts of data in a smaller space, and its chip size is smaller than that of flash memory chips, naturally reducing manufacturing costs and contributing to larger data capacities. When the memory is not performing read or write operations, the power consumption is minimized. This power-saving feature makes PCM perform better in battery-powered devices.

In-memory optical computing, which leverages the high speed, low energy consumption, and high parallelism of optical signals to achieve non-volatile storage and computing by altering the optical properties (such as refractive index or absorption rate) of PCM (phase change materials). It is one of the mainstream directions in the development of in-memory computing with PCM today. The typical process involves fabricating PCM into thin films for information storage, using laser heating to change its crystalline or amorphous state for writing, inputting optical signals, and representing the data to be processed by controlling the intensity, phase, or wavelength of the light. The interaction between the optical signals and the PCM results in interference, diffraction, or transmission changes to complete the computing operation. Finally, the changes in the optical signals are read by an optical detector to obtain the computing results. In 2024, the Oxford team made a new breakthrough in the development of in-memory optical computing. Their team put forward [6] a photonic convolution processing system that utilizes reduced temporal coherence (a partially coherent system) to enhance processing parallelism without significantly sacrificing accuracy. And potentially enabling large-scale photonic tensor cores. This was demonstrated by testing the classification of the MNIST handwritten digit dataset, achieving a test accuracy of 92.4%. Thus, it is evident that PCM holds broad prospects in in-memory optical computing.

However, as an emerging storage medium, PCM-related technologies still have some shortcomings at present. The storage density of PCM is relatively low, making it difficult to compete with traditional storage media such as NAND Flash. Although it has advantages in random read and write speed and lifespan, the limited number of three-dimensional integration layers restricts the rapid increase of its capacity. Material fatigue and the decline in durability are also key challenges that PCM needs to overcome. At the same time, the switching time and energy consumption issues of PCM

limit its application in in-memory computing fields such as optical neural networks.

2.4 MRAM

MRAM (Magnetic Random Access Memory, magnetic random access memory) is also a non-volatile storage technology. Its core is to use the magnetoresistance effect of the magnetic tunnel junction (Magnetic Tunnel Junction, MTJ) to store information. The MTJ is composed of two layers of ferromagnetic materials and an insulating layer. The magnetization direction of one layer of ferromagnetic material is fixed (referred to as the fixed layer).The magnetization direction of the other layer can be changed (referred to as the free layer). When the magnetic field direction of the free layer is parallel to that of the fixed layer, the storage unit presents a low resistance. Otherwise, it presents a high resistance. By detecting the high or low resistance of the storage unit, it can be determined whether the stored data is 0 or 1.

Due to the high-speed reading and writing, high durability, and low power consumption of MRAM, it has application potential in areas such as weight storage and matrix multiplication and addition operations for neural networks, and edge computing [7]. For instance, in 2022, Donhee Ham's team designed a neural network computing architecture based on MRAM cross arrays [8]. This architecture utilizes the magnetic resistance characteristics of MRAM to store and compute the weights and input data of neural networks. Experimental results show that the neural network computing architecture based on MRAM cross arrays performs well in terms of accuracy and power consumption. Compared with traditional CPU or GPU computing, this architecture can significantly reduce computing power consumption and increase computing speed.

Although MRAM has limitations such as relatively low resistance, a low tunneling magnetoresistance (TMR) ratio, and less favorable write energy consumption and latency compared to CMOS computing units, which restrict its application in analog in-memory computing. With the increasing maturity of MRAM technology, MRAM is bound to become one of the key points in the development of in-memory computing.

2.5 Flash memory

In the architecture of Computing-in-Memory (CIM), the choice of storage medium is critical. Flash memory technology has become an ideal option for CIM due to its non-volatile nature, low power consumption, and high density. Flash memory is primarily categorized into NOR Flash and NAND Flash, each offering distinct advantages and application scenarios.

NOR Flash is characterized by high read speeds and low access latency. Supporting random read operations, it is suitable for applications requiring high speed and low latency. Particularly in CIM tasks, NOR Flash provides high parallel processing capabilities, making it ideal for embedded systems, real-time processing, and similar scenarios.

NAND Flash, on the other hand, boasts higher storage density and lower cost, making it suitable for large-scale storage applications. While its random access performance is relatively poor, it excels in sequential read/write operations. In CIM,

NAND Flash leverages its large storage capacity and cost efficiency to handle big data storage and processing, especially excelling in high-throughput tasks.

In recent years, the emergence of 3D NAND technology has significantly enhanced the storage density of NAND Flash. By vertically stacking multiple storage layers, 3D NAND delivers greater storage space and faster access speeds. This technology optimizes spatial efficiency and reduces costs, further strengthening its competitiveness in CIM.

Durability was once a limitation for flash memory, particularly in CIM scenarios with frequent read/write operations. Today, advancements in wear leveling and error-correcting code (ECC) technologies have extended flash memory's lifespan. In fields like data centers and edge computing, the endurance of flash memory has been substantially improved.

Computing-in-Memory enhances efficiency by integrating computation with storage, reducing data transfer requirements. Both NOR Flash and NAND Flash hold immense potential in CIM applications. NOR Flash's high performance and low latency make it suitable for embedded systems and real-time computing tasks, while NAND Flash's high density and cost-effectiveness solidify its role as a critical technology for big data processing.

2.6 Memristors

A memristor (memory resistor) is a non-volatile memory device based on resistance changes, capable of recording current history and altering its resistance value [9]. Its unique properties enable memristors not only to store information but also to perform computational tasks, making them a critical technology in Computing-in-Memory (CIM) architectures.

2.6.1 Resistive RAM (RRAM).

RRAM stores data by modulating the resistance of memory cells through current adjustments. Compared to traditional memory technologies (e.g., DRAM and Flash), RRAM offers the following advantages in CIM:

RRAM can perform calculations directly within memory cells, reducing data transfer latency and enhancing parallel computing efficiency [10]. This makes it ideal for highly parallel tasks like deep learning and neural network processing.

RRAM consumes significantly less power during computation and storage compared to conventional memory, making it suitable for power-constrained scenarios such as edge computing and IoT devices [11].

RRAM provides greater storage density (enabling more data storage in compact spaces) and improved durability, supporting high-frequency read/write operations in demanding applications [12].

2.6.2 Applications of RRAM in CIM.

By enabling computation within memory, RRAM drastically reduces data transfer delays, accelerating AI inference processes.

Its low power consumption and high density allow RRAM to meet the needs of

embedded systems and connected devices, extending operational lifetimes.

Non-volatile storage: RRAM retains data during power loss, making it suitable for high-speed caching and long-term data retention scenarios.

With technological advancements, RRAM is poised to play an increasingly vital role in CIM. Integrating RRAM with Flash, DRAM, and other memory technologies [13], future CIM systems will achieve enhanced computational power, improved energy efficiency, and breakthroughs in storage capacity. However, challenges remain, including optimizing memory cell stability and advancing manufacturing feasibility.

In summary, RRAM demonstrates immense potential in CIM applications, particularly in deep learning, artificial intelligence, and big data processing. Its adoption could drive revolutionary advancements and propel the development of next-generation computing architectures.

3 Conclusion

This paper explores the principles and applications of in-memory computing in six different categories of memory. Among them, DRAM integrates computing units with DRAM chips closely through 3D stacking technology and through via holes in silicon, reducing the distance between the processor and the memory in space, thereby lowering transmission delay and energy consumption. SRAM holds an important position in the field of memory-acceleration integration due to its reliable stability and mature process. PCM has become a research hotspot thanks to the particularity of its phase-change materials. MRAM for memory-acceleration integration has achieved certain results in neural network computing, logical operations, and matrix multiplication and addition operations, and is worthy of further in-depth exploration. Flash memory technology is also highly concerned due to its low power consumption and high storage density. Meanwhile, the current research on RRAM still focuses on small-scale arrays, and further exploration is still needed for large-scale arrays and storage reliability in the future.

Although volatile memory has mature manufacturing processes and reliable stability, it still needs to be further explored to achieve key technologies such as low power consumption and high integration. Analog storage, non-volatility, and low power consumption are all significant advantages of non-volatile memory. In the future, the migration from relatively mature DRAM and other in-memory computing architectures to RRAM and other new non-volatile device architectures can be carried out, thereby enabling technological iteration and innovation as well as product continuity.

Overall, the development of in-memory computing is rapid. However, a series of practical technical issues such as hardware resource reuse, unit design, and optimization of analog computing still need to be addressed.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Guo, X., Wang, G., Wang, S.: 'Research progress and application of memory computing chip', J. Electron. Inf. Technol., 2023, 45, (5), pp. 1888–1898
2. Ye, L., Jia, T., Chen, P., et al.: 'Research on SRAM integrated memory and computing chip: Development and challenges', Sci. China Inf. Sci., 2019, 54, (1), pp. 25–33
3. Kang, W., Kou, J., Zhao, W.: 'Development status, trend and challenge of integrated memory and computing chip', Sci. China Inf. Sci., 2019, 54, (1), pp. 16–24
4. Li, J., Yao, P., Jie, L., et al.: 'Research status of integrated storage and computing technology', Acta Electron. Sin., 2024, 52, (4), pp. 1103 – 1117
5. Jeyasingh, R., Ahn, E. C., Eryilmaz, S. B., et al.: 'Phase change memory', in 'Emerging Nanoelectronic Devices', Wiley, 2015, pp. 78–109
6. Dong, B., Brückerhoff-Plückelmann, F., Meyer, L., et al.: 'Partial coherence enhances parallelized photonic computing', Nature, 2024, 632, pp. 55–62
7. Li, Y., et al.: 'A survey of MRAM-centric computing: From near memory to in memory', IEEE Trans. Emerg. Top. Comput., 2023, 11, (2), pp. 318–330
8. Jung, S., Lee, H., Myung, S., et al.: 'A crossbar array of magnetoresistive memory devices for in-memory computing', Nature, 2022, 601, pp. 211–216
9. Jiang, Z. X., Xi, Y., Tang, J. S., et al.: 'Research progress on memristors and their applications in computing-in-memory', Ph.D. Dissertation, Sch. Integr. Circuits, Tsinghua Univ., Beijing, China, 2023
10. Zhang, Z., Li, C., Han, T. T., et al.: 'A review of memristor-based sensing-memory-computing integration technology', J. Shanghai Jiao Tong Univ. (Electron. Inf. Electr. Eng. Ed.), 2021, (6), pp. 1498–1512
11. Yin, Z. Q.: 'Integrated circuit design for RRAM-based computing-in-memory multipliers', Master's Dissertation, Anhui Univ., Anhui, China, 2019
12. Wang, W.: 'Design of memristive devices and their applications in computing-in-memory processors', Ph.D. Dissertation, Grad. Sch., Natl. Univ. Def. Technol., Changsha, China, 2019
13. Zhu, Z. H., Zhu, Y., Wang, Y., et al.: 'A survey and prospects for simulator design targeting memristor-based computing-in-memory architectures', J. Tsinghua Univ. (Sci. Technol.), 2021, 61, (2), pp. 123–135

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.