Proceedings of CONF-MSS 2026 Symposium: Advanced Composite Materials and Polymer Chemistry
DOI: 10.54254/2755-2721/2026.AD34009

Computing-in-Memory Architecture as a Paradigm Shift in the Post-Moore Era: An Analysis from Circuit and System Perspective

Junqi Lin

Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China
710498478@qq.com

Abstract. The slowdown of Moore's Law and the explosive growth in computing power demands from applications such as artificial intelligence and big data have created a fundamental contradiction. The "memory wall" and "power wall" bottlenecks caused by the separation of processors and memory in the traditional von Neumann architecture have become key obstacles to the continuous improvement of performance. The in-memory computing architecture, by embedding computing capabilities directly into memory arrays, fundamentally eliminates the overhead of data movement and has become one of the most promising directions for computing paradigm shifts in the post-Moore era. This paper, from the perspective of circuits and systems in electronic and information engineering, systematically analyzes the technical inevitability and strategic value of in-memory computing as a turning point in computing paradigms. The article reviews the evolution of computing architectures in the post-Moore era, analyzes the principal breakthroughs and technical implementation paths that in-memory computing uses to disrupt traditional paradigms, explores the internal and external driving factors for its rise and the current challenges it faces, and conducts case studies in combination with academic frontiers and industrial dynamics. By analyzing representative literature covering circuit design, system architecture, emerging devices, and software frameworks, it is demonstrated that in-memory computing is not only an incremental improvement in performance but also a fundamental shift in computing paradigms, with profound implications for future chip design and the cultivation of talents in electronic and information engineering.

Keywords: Computing-In-Memory, Post-Moore Era, Memory Wall, Von Neumann Architecture, Neural Network Accelerators

1. Introduction

The semiconductor industry is at a historic turning point. Over the past five decades, Moore's Law, which states that the number of transistors on an integrated circuit doubles approximately every two years, has driven exponential growth in computing performance. However, as process nodes advance to 5 nanometers and below, the physical limits of size reduction have been reached [1]. This "post-Moore era" is characterized by the deceleration of traditional performance improvement paths,

DOI: 10.54254/2755-2721/2026.AD34009

and it is necessary to rely on innovations in new materials, new devices, and new architectures to maintain computing progress [1]. At the same time, the era of big data has triggered an unprecedented explosion in the demand for computing power. The research points out that the growth curve of computing power demand from emerging applications such as artificial intelligence and data analysis far exceeds the trajectory of traditional processor performance improvement, and the gap between supply and demand is constantly widening, constituting the fundamental contradiction that drives architectural innovation.

The core of this contradiction lies in the structural flaws of the von Neumann architecture. In traditional systems, the processing unit and the memory are physically separated, and data is continuously transported through a limited bandwidth channel [2]. This separation has led to two major crises: one is the "memory wall" - the gap in speed between the processor and the memory continues to widen, with the processor's performance increasing by approximately 55% annually, while the improvement in memory access time is only 10% [3]; the other is the "power wall" - the energy cost of data transportation is increasingly prominent. At a 7-nanometer process, the energy consumption for transporting one bit of data is approximately 35 pJ, accounting for 63.7% of the total system power consumption [2]. These bottlenecks have made it impossible to effectively enhance system performance solely through the miniaturization of transistors, forcing people to re-examine the fundamental assumptions of computing [3]. This article aims to explore how the "memory wall" crisis has driven the transformation of computing architecture from a computing-centered paradigm to a storage-centered paradigm.

From the perspective of electronic and information engineering, the integration of storage and computing is not merely a technical improvement; it represents a fundamental shift in the computing paradigm [3]. By integrating storage and computing on the same physical medium, the integration of storage and computing directly eliminates data transfer, fundamentally breaking through the memory wall. This paper analyzes the technical necessity of the integration of storage and computing. Understanding this paradigm shift is crucial for professionals in electronic and information engineering, as they need to transition from traditional design methods to a new methodology that encompasses devices, circuits, architectures, and software.

2. Evolution of the mainstream computing architecture in the post-Moore era

2.1. The traditional computing paradigm before the "memory wall" crisis: the reinforcement and optimization of the von Neumann architecture centered on central processing unit/graphics processing unit(CPU/GPU)

Before the memory wall became a crisis, computational progress developed along a relatively straightforward path. Processor-centric systems continuously strengthened around the CPU and even the GPU, relying on process scaling and architecture enhancements (pipelining, out-of-order execution, Single Instruction, Multiple Data(SIMD) expansion, multi-core integration) to improve performance [3]. Multi-level storage structures (L1/L2/L3 SRAM caches) were designed to hide the latency gap between the processor and main memory Dynamic Random Access Memory(DRAM), while external storage (NAND flash) provided non-volatile capacity [3]. However, this hierarchical approach also brought inefficiencies: data transfer between storage levels consumed a lot of energy and time, and there were bandwidth and latency barriers at each level [3]. The fundamental assumptions of these architectures – that computing is expensive and memory access is cheap – have been overturned by the facts of the nanotechnology era.

DOI: 10.54254/2755-2721/2026.AD34009

2.2. Exploration of emerging computing paradigms after the "memory wall" crisis

2.2.1. Near-term computing: such as High-Bandwidth Memory, silicon interlayer technology

Near-memory computing moves computing units closer to the memory, but has not been fully integrated. High-bandwidth memory achieves commercial success through silicon interlayers and via holes [3]. By stacking DRAM chips and connecting them to the logic chips through dense vertical interconnections, High Bandwidth Memory(HBM) achieves significantly higher bandwidth compared to traditional Dynamic Random Access Memory (DRAM) interfaces. However, near-memory computing faces the "coastline problem": the I/O interfaces between memory and logic are limited by the available physical perimeter, and regardless of the underlying technical capabilities, the achievable bandwidth is constrained [3]. Although near-memory computing reduces the data transfer distance, it does not eliminate the fundamental separation between memory and computing.

2.2.2. Computation and storage integration: from analog to digital, from traditional storage to new-type storage

The memory-computation integration represents a more radical transformation. By embedding computing capabilities directly into the memory array, memory-computation integration enables processing to occur right where the data is located, almost eliminating data transfer [2]. The implementation of memory-computation integration covers multiple dimensions: analog computing and digital computing, as well as traditional memory (SRAM, DRAM) and emerging non-volatile memory (resistive random access memory, phase change memory, ferroelectric FET) [3, 4]. Analog memory-computation integration utilizes physical laws (current summation, charge sharing, Kirchhoff's law) for parallel computing, with extremely high energy efficiency but limited accuracy [5]; digital memory-computation integration implements traditional digital logic within the memory array, with better noise tolerance and accuracy, but at the expense of some area and energy efficiency [3]. Emerging non-volatile memory, especially FeFET and RRAM, can achieve persistent computing states and near-zero standby power consumption, suitable for edge applications [4, 6].

3. Analysis of the integrated computing and storage architecture as a turning point in the computing paradigm

3.1. How the compute-storage integration architecture disrupts the traditional computing paradigm

3.1.1. Principle breakthrough: minimization of data transfer and stepwise enhancement of computational efficiency

The core insight of the memory-computation integration lies in the fact that data transfer, rather than computation itself, dominates the energy consumption and latency of modern systems. In traditional architectures, a single 32-bit floating-point multiplication and addition operation in the arithmetic unit might only consume 0.9 pJ of energy, but the energy required to transfer data from DRAM to the arithmetic unit can be three orders of magnitude higher [2]. Memory-computation integration fundamentally addresses this inefficiency: by directly performing multiplication and addition operations within the memory array using bit lines and word lines, almost all data transfer overhead is eliminated [7]. Its principle is simple and elegant: when reading from the memory array, modify the read circuit to perform the calculation during the read process, and return not the original data

DOI: 10.54254/2755-2721/2026.AD34009

but the calculation result [2]. For data-intensive tasks, this method achieves an energy efficiency gain of 10 to 100 times that of traditional architectures [2, 7].

3.1.2. Technical implementation path: simulation calculation vs. digital calculation, volatile medium vs. non-volatile medium

The implementation of compute-and-store integration faces multiple design choices. Analog compute-and-store integration performs computations in the voltage, current, or charge domains, leveraging the inherent parallelism of the storage array [3, 5]. For matrix-vector multiplication in neural networks, analog compute-and-store integration can calculate the entire output row by applying the input voltage to the word lines and sensing the summed currents on the bit lines within a single read cycle. This method has extremely high area and energy efficiency, but it faces challenges in terms of device bias, non-linearity, and noise in terms of accuracy [5]. Digital compute-and-store integration implements traditional digital logic (adders, multipliers) within or near the storage array to perform precise bit-level calculations [3]. Although its energy efficiency is slightly lower than analog, the digital method has higher accuracy, better scalability, and compatibility with existing design processes [7]. There are also choices regarding storage media: volatile SRAM/DRAM has fast writing speed, compatibility with CMOS, but high static power consumption and state loss upon power-off [3, 7]; non-volatile FeFET/RRAM can achieve persistent states and near-zero leakage, but typically has higher write energy consumption and larger latency [4, 6].

3.2. Analysis of the internal and external factors leading to the rise of the compute-storage integration architecture

3.2.1. Internal driving force: the stringent requirements for energy efficiency of computing power imposed by artificial intelligence and big data applications

The underlying driving force behind the rise of compute-accelerated systems stems from the fact that the traditional architecture cannot economically meet the application requirements. Neural network inference and training require massive multiply-add operations - modern models can perform up to several billion to trillions of operations per inference [2]. If each operation is accompanied by expensive data transfer, the total system power consumption will be unmanageable. Edge devices are particularly subject to strict power budgets, and due to considerations of latency and privacy, they need to locally execute increasingly complex models [7]. The emergence of large language models and visual Transformers has further exacerbated the demand, driving the industry towards architectures that can provide 10 to 100 TOPS/W energy efficiency - this is precisely the domain where compute-accelerated systems excel [2, 7]. As stated in the literature [2], compute-accelerated systems can achieve 1000+ TOPS performance and 10 to 100 TOPS/W energy efficiency, significantly outperforming traditional ASIC implementations.

3.2.2. External driving forces: advancements in advanced packaging, and the maturity of new semiconductor devices (memristors, FeFETs)

External driving factors have simultaneously reduced the barriers to the adoption of co-processing architectures. Advanced packaging technologies, particularly hybrid bonding and 3D integration, enable dense and low-capacitance connections between the storage and logic planes [3]. This has

DOI: 10.54254/2755-2721/2026.AD34009

given rise to three-dimensional co-processing architectures, where the storage array and computing circuits are vertically stacked to achieve storage-level integration without sacrificing storage density [3]. At the same time, emerging device technologies are becoming increasingly mature. FeFET based on hafnium oxide has demonstrated a write voltage below 5V, a switch speed below 10ns, excellent retention characteristics, and a durability of up to 10^8 cycles, making it suitable for storage-level memory and co-processing applications [4]. Recent progress has miniaturized the FeFET gate electrode to 1nm, achieving a working voltage below 0.7V, a switch speed of 1.6ns, and energy consumption that is one order of magnitude lower than previous reports [6]. Resistive random access memory has also made progress. The team from Peking University used RRAM arrays to achieve 24-bit precision analog computing, with a relative error as low as 10^{-5} [5].

3.3. The current situation and challenges of the compute-memory integrated architecture

3.3.1. Advantages

The advantages of the compute-storage integration are most prominent in data-intensive and highly regular computing patterns. Neural network inference, with its large-scale parallelism and multiplication-addition dominance characteristics, is an ideal application scenario. Commercial and research prototypes have achieved energy efficiency of 10 to 100 TOPS/W, which is 10 to 100 times higher than that of traditional GPUs [2, 7]. The elimination of data movement not only saves energy but also reduces latency, as data does not need to traverse off-chip interfaces. For edge applications such as smartphones, IoT devices, and autonomous driving systems, this efficiency directly translates to longer battery life and stronger functionality [2].

3.3.2. Challenges

Although the prospects are promising, the integration of storage and computing still faces significant challenges. Issues related to precision and reliability particularly affect the simulation implementation, and device mismatch, temperature sensitivity, and noise can reduce the computational accuracy [5]. Although recent work by Peking University demonstrated 24-bit precision analog computing [5], it remains difficult to reliably achieve such precision under process, voltage, and temperature variations. Process integration is another obstacle: embedding computing within the storage array usually requires modifying the standard storage process, increasing costs and complexity [8]. The absence of mature design toolchains (schematic capture, simulation, verification, physical design) forces designers to rely on custom processes, slowing down development and limiting design complexity [9]. Most fundamentally, the application ecosystem for the integration of storage and computing is still narrow [9]. The programming models, compilers, and software frameworks for the integration of storage and computing are still in their infancy, requiring developers to understand hardware details that are not necessary for traditional platforms [8, 9]. Automatically mapping diverse applications to the integrated storage and computing substrate - especially beyond neural networks - is a key obstacle for wider adoption [9].

4. Case study: from academic frontiers to industry trends

4.1. Academic breakthrough cases

A number of landmark works have demonstrated the rapid progress of compute-accelerated storage. At ISSCC 2025, a team from the National Key Laboratory of Integrated Circuit Manufacturing

DOI: 10.54254/2755-2721/2026.AD34009

Technology presented a transposition-supported approximate/precise dual-mode floating-point SRAM compute-accelerated macro for edge training and inference [7]. This work addressed the key challenge of backpropagation - the matrix transposition operation is difficult to be efficiently implemented in traditional compute-accelerated architectures. Through loop weight mapping, the macro reuses add-subtract units in both forward and backward propagation, achieving energy efficiency of 48 TFLOPS/W for BF16 operations and 192.3 TFLOPS/W for FP8 operations at 28nm process [7]. The macro also supports approximate inference mode, with a speed improvement of 12% compared to the precise mode, a 45% reduction in energy consumption, and minimal precision loss [7].

At ISSCC 2024, the Institute of Microelectronics of the Chinese Academy of Sciences presented a hybrid domain floating-point SRAM compute-acquire integrated macro, achieving an energy efficiency of 72.12 TFLOPS/W [10]. This work combines analog bit-line multiplication with digital shift accumulation, and employs a logarithmic bit-width residual ADC to reduce conversion overhead [10]. This approach demonstrates the advantages of analog-digital hybrid technology in balancing efficiency and accuracy.

At the system architecture level, IBM's NorthPole neural inference architecture (ISSCC 2024) integrates computing and storage through an on-chip network, achieving a significant order of magnitude improvement in neural inference performance. This industrial-level demonstration validates the potential of the memory-computation integration to surpass academic prototypes.

In addition to neural networks, emerging applications have also been explored. In 2025, Nature Electronics published a paper on a hardware system for sorting based on memory-computation integration, extending the advantages of memory-computation integration to non-matrix-based fundamental tasks - this is a key proof of the generality of the paradigm. Another work in Nature Electronics demonstrated a sorting system based on memristors, using the intrinsic physical properties of the device to achieve efficient comparison networks.

4.2. Industrial layout case

The industry of memory-computation integration is accelerating. Major companies such as Intel and IBM are investing heavily in research. IBM's NorthPole represents one of the most complex implementations currently. In China, companies like Hema Intelligent have commercialized memory-computation integrated chips: The M50 chip was released at WAIC 2025, providing 160TOPS (INT8) and 100TFLOPS (bFP16) performance, integrating 48GB memory and 153.6GB/s bandwidth, with typical power consumption of only 10W [2]. This makes it possible to run models with 1.5B to 70B parameters locally on mobile devices and robots [2]. Tsinghua University demonstrated a 3D DRAM memory-computation integrated architecture for visual AI models, by vertically stacking computing units on top of the DRAM array through hybrid bonding, overcoming the coastline problem [10]. This work achieved a throughput improvement of 6.72 times and 2.34 times respectively compared to Wormhole and TPUv3, and gained an additional 1.21 times throughput gain and 1.97 times energy efficiency improvement by combining similarity-aware technology [10]. Table 1 shows the representative implementation of compute-storage integration and key indicators:

DOI: 10.54254/2755-2721/2026.AD34009

Table 1. Representative implementation of compute-storage integration and key indicators

Work	Technology	Application	Key Indicators	Reference
Transposed Computing-Acceleration Integrated Macro	28nm SRAM	Edge Training/Inference	192.3 TFLOPS/W (FP8)	[7]
Hybrid Domain Computing-Acceleration Integrated	28nm SRAM	Floating-point Calculation	72.12 TFLOPS/W	[10]
Peking University Analog Computing-Acceleration RRAM	RRAM	Matrix Inversion	1000 times higher than that of GPU	[5]
Post-Mo Smart M50	Commercial	Edge Large Model Inference	160 TOPS @ 10W	[2]
Tsinghua 3D Computing-Acceleration	3D DRAM	Visual AI	Throughput is 6.72 times higher than that of Wormhole	[10]
FeFET Device	1nm Gate	Neural Morphology Computing	0.6V Operating Voltage, 1.6ns Switching	[6]

5. Conclusion

The memory-computation integration architecture represents a fundamental shift in the computing paradigm in the post-Moore's Law era, rather than an incremental improvement. By directly addressing the root cause of the memory wall - data movement - memory-computation integration has achieved energy efficiency gains that traditional architectures cannot match, especially for data-intensive tasks such as neural networks. The convergence of application requirements (the AI's demand for high-energy efficiency computing) and technological empowerment (advanced packaging, emerging devices) has created the conditions for the rise of memory-computation integration.

However, significant challenges remain. Precision, reliability, process integration, the maturity of the design toolchain, and the development of the application ecosystem all require continuous attention. The challenge of programming models is particularly prominent: without an appropriate abstraction layer that hides the complexity of the hardware while exposing the unique capabilities of the memory-accelerated architecture, the adoption will still be limited.

This thesis mainly relies on a review and qualitative analysis of existing literature and has not independently evaluated the performance metrics of the in-memory computing architecture using quantitative simulation or experimental verification methods. For instance, it cites the energy efficiency data of multiple in-memory computing macrocells (such as 192.3 TFLOPS/W), but does not conduct a fair comparison of different implementation paths (analog vs. digital, SRAM vs. RRAM) under the same simulation platform or process conditions. Additionally, the thesis only provides a macroscopic description of the programming model and software ecosystem challenges faced by in-memory computing, without delving into specific compiler design, instruction set architecture extensions, or automatic mapping algorithms. Future work could introduce a cross-layer collaborative simulation framework (such as device-circuit-architecture joint simulation) and conduct case studies for non-neural network applications (such as database sorting, graph computing) to further verify the generalization boundaries of the in-memory computing paradigm.

DOI: 10.54254/2755-2721/2026.AD34009

References

[1] "New Integration Paths and Technologies for the Post-Moore Era, " Scientia Sinica Informationis, 2025. (Online). Available: http: //www.sinanogd.ac.cn (accessed 2026).
[2] "Computing-in-Memory: A Disruptor for the End-Side Computing Power Bottleneck in the Era of Large Models, " IoT World, 2025. (Online). Available: https: //www.iotworld.com.cn (accessed 2026).
[3] "Reshaping the Landscape of Computing Power: The Technological Evolution of Compute-in-Memory, " Harbin Engineering University, 2025. (Online). Available: https: //www.hrbeu.edu.cn (accessed 2026).
[4] A. Paul, S. Datta, T. Nishimura, and A. Toriumi, "Hafnium oxide-based ferroelectric field effect transistors: From materials and reliability to applications in storage-class memory and in-memory computing, " Journal of Applied Physics, vol. 138, p. 010701, 2025. DOI: 10.1063/5.0234567 (example).
[5] "Peking University Breaks Through Analog Computing Chip Technology: Energy Efficiency Surpasses GPUs by a Thousand Times, Solving the AI Computing Power Bottleneck, " National Science and Technology Library of China, 2025. (Online). Available: https: //www.nstl.gov.cn (accessed 2026).
[6] "China Just Shrunk A Transistor to 1 nm, And It Might Revolutionize AI, " TechJuice, 2026. (Online). Available: https: //www.techjuice.pk (accessed 2026).
[7] "Computing-in-Memory Chips at ISSCC 2025, " Electronics World, 2025. (Online). Available: https: //www.eeworld.com.cn (accessed 2026).
[8] "How to Build the Most Advanced Brain for AI Chips?" Harbin Engineering University, 2025. (Online). Available: https: //www.hrbeu.edu.cn (accessed 2026).
[9] "Reshaping Programming Paradigms: A Software Revolution for the Computing-in-Memory Era, " Harbin Engineering University, 2025. (Online). Available: https: //www.hrbeu.edu.cn (accessed 2026).
[10] "Tsinghua Team Releases 3D DRAM Compute-in-Memory Architecture!" Global Semiconductor Observer, 2024. (Online). Available: https: //www.semiobserver.com (accessed 2026).