A Literature Review of Analog In-Memory Computing for AI and Machine Learning

Executive Summary See: https://circularastronomy.com/2025/09/29/breaking-the-memory-wall-a-structured-literature-review-of-analog-in-memory-computing-aimc-for-next-generation-ai-and-deep-learning/

1. Introduction to Analog In-Memory Computing

1.1 The Von Neumann Bottleneck and the Rise of In-Memory Computing

The architecture of modern computers, which has been in place for more than 70 years, is based on the foundational principles of the von Neumann model. This model fundamentally separates the central processing unit (CPU) from its memory, with a data bus serving as the channel for communication between the two components [1, 2]. This design, while robust, has given rise to a critical limitation known as the von Neumann bottleneck, or the "memory wall" [1]. This bottleneck is not a theoretical curiosity but a practical and increasingly severe constraint on performance and energy efficiency, particularly for data-intensive applications like large language models (LLMs) and other complex AI algorithms [1].

In these applications, the constant, repeated transfer of vast amounts of data, such as neural network weights, activations, and key-value caches, back and forth between memory (DRAM) and the processor (CPU or GPU) becomes the primary limiting factor [1]. This data movement is highly power-consuming and time-intensive [1, 4]. The escalating demands of modern artificial intelligence (AI), machine learning (ML), and high-performance computing (HPC) for greater computational capacity and tighter system integration necessitate a new architectural paradigm [5, 6]. The impact of this bottleneck extends beyond raw performance, creating a cascade of challenges for resource-constrained edge devices. These include high power consumption that drastically reduces battery life, increased latency that hinders real-time decision-making, and the need for larger physical footprints and complex thermal management solutions [3, 6]. A direct consequence of these issues is a paradigm where much of the heavy AI workload is offloaded to the cloud, which inherently undermines the benefits of edge computing, such as privacy and low latency [6].

1.2 Defining Analog In-Memory Computing (AIMC)

Analog In-Memory Computing (AIMC), also referred to as Compute-in-Memory (CIM) or Processing-in-Memory (PIM), is an architectural paradigm designed to overcome the von Neumann bottleneck by physically integrating computation directly within the memory itself [3, 6]. The core principle of AIMC is the execution of fundamental neural network operations, most notably matrix-vector multiplication (MVM), by leveraging the intrinsic physical laws of electronics [1, 7]. This is accomplished by storing neural network weights directly within a dense, non-volatile memory (NVM) array, such as phase-change memory (PCM) or resistive random-access memory (RRAM), where individual memory elements function as tunable resistors [2, 7, 8].

The computational process unfolds by applying input data as a voltage on the memory array’s word lines. According to Ohm's law (V=IR), the current that flows through each memory element is proportional to the product of the input voltage and the stored resistance (weight) [1, 7]. The currents from a given column are then summed on the bit line, yielding the result of the MVM operation in accordance with Kirchhoff’s Current Law [1, 7]. This inherent parallelism, which allows hundreds of thousands of multiply-accumulate operations to occur simultaneously, dramatically reduces or eliminates the need for data movement, leading to significant improvements in energy efficiency and computational throughput [7, 9]. It is important to note that the term "AIMC" is used in other contexts, such as by the Acupuncture and Integrative Medicine College [10] and the Associazione Internazionale Mosaicisti Contemporanei [11], which are unrelated to the field of computer engineering. Similarly, academic papers on "AI-Generated Content" (AIGC) [12, 13] should not be confused with this topic.

2. Historical Context and Foundational Principles

2.1 Early Analog Computing and its Precursors

The concept of using physical properties for computation is not a recent development but an ancient one, predating the digital era by centuries. Historical devices such as the Antikythera mechanism (c. 150-100 BC), an analog computer for astronomical calculations, and later inventions like the slide rule (c. 1620-1630) and the Differential Analyzer (1930s) developed by Vannevar Bush, served as important historical precursors to modern analog computing [14, 15]. These early analog machines were often faster and more efficient for solving specific problems, such as differential equations, than their early digital counterparts [16]. However, the trajectory of computing changed with the rise of digital computers, which began to dominate the landscape from the 1950s onward. The reasons for this shift were not simply a matter of raw speed but were rooted in the superior precision, programmability, and versatility inherent in digital systems [16]. This historical progression, where digital approximation eventually overtook continuous analog calculation, is essential for understanding the ongoing debates and challenges of modern AIMC, particularly regarding the trade-offs between precision and efficiency.

2.2 The Evolution of Memory Technologies for In-Memory Computing

The modern resurgence of analog computing for AI is inextricably linked to the development and maturation of specific memory technologies [17]. The history of computer memory, from early drum memory [18, 19] and magnetic core memory [18, 19] to modern semiconductor memory like DRAM and SRAM, demonstrates a continuous drive toward higher density, lower cost, and faster access [18, 19]. However, for AIMC to be effective, a new class of memory was required: non-volatile memories (NVMs) that could not only store data without power but also retain multiple, continuously tunable states to represent the analog-like values of neural network weights [2, 17, 20].

Key memory technologies that have enabled the AIMC paradigm include:

- **Phase-Change Memory (PCM):** At the heart of several IBM analog AI chips, PCM stores data by reversibly changing the physical state of a chalcogenide glass between amorphous (high resistance) and crystalline (low resistance) states [2, 4, 20]. This continuous range of resistance can be used to represent the synaptic weights of a neural network [2].

- **Resistive RAM (RRAM/ReRAM):** This technology works by generating conductive filaments or "defects" in an insulating material, which in turn changes its resistance [17, 20]. RRAM is recognized for its scalability, fast read/write speeds, and low write energy, making it a promising candidate for neuromorphic and in-memory computing applications [17].
- **Memristors:** First conceptualized as the fourth fundamental circuit element by Leon Chua in 1971, memristors are a class of resistive memory that retains a state of resistance based on the history of voltage and current passed through them [21, 22]. Their physical realization by HP Labs in 2008 [22] has been at the core of new breakthroughs in AIMC [8, 23].

The journey from the theoretical memristor to a commercially viable system-on-a-chip (SoC) demonstrates a critical narrative of engineering problem-solving [8, 21]. Early research had to overcome practical challenges such as "sneak path" currents by adopting a one-transistor-one-resistor (1T1R) structure [8]. Subsequently, the challenge shifted to eliminating off-chip bottlenecks by integrating all components, including peripheral circuits, onto a single SoC [8]. This continuous process of tackling cascading engineering problems has paved the way for the practical realization of this technology.

2.3 Core Principles of Analog Matrix-Vector Multiplication

The central operation in many AI and ML models is matrix-vector multiplication (MVM), which is a key strength of AIMC [1, 7, 9]. The process within an analog memory array can be broken down into a series of steps:

1. **Weight Storage:** The synaptic weights of a neural network are programmed and stored as the conductive states (resistances) of a crossbar array of NVM devices [7].
2. **Input Modulation:** The input vector of activations is converted from the digital domain into a series of analog voltage signals. This is typically achieved using digital-to-analog converters (DACs) that modulate the voltage on the word lines of the array [9].
3. **Parallel Computation:** For each row in the array, the voltage on the word line passes through the NVM devices (representing the weights), producing a current proportional to their product, as defined by Ohm’s law [1, 7].
4. **Current Summation:** The currents from each column of the array are summed on the bit line, yielding the result of the MVM operation, as defined by Kirchhoff’s law [1, 7].
5. **Output Conversion:** The resulting analog current on the bit line is then converted back to a digital value using an analog-to-digital converter (ADC) [9].

This process allows for a massive number of multiply-accumulate operations to occur in parallel directly at the location of the data [7, 9]. The elimination of data movement for weights results in extremely high energy efficiency, outperforming state-of-the-art digital compute arrays by a wide margin [9].

3. Key Research Themes and Breakthroughs

3.1 AIMC for AI/ML Inference and Training

A primary theme in the literature is the application of AIMC to AI and ML workloads, specifically for both inference and training [2]. Inference, which involves using a pre-trained model to make predictions, is particularly well-suited for AIMC due to the static nature of the model weights [2, 4]. The IBM Research group, for example, has explored using resistive random-access memory (RRAM) and electrochemical random-access memory (ECRAM) for training purposes, while leveraging phase-change memory (PCM) for inference [2]. While the complexities of continuously updating weights in analog devices present challenges for on-device training, most research has focused on accelerating inference at the edge, where extreme energy efficiency is a paramount requirement [6].

3.2 Advancements in Memristor and Non-Volatile Memory (NVM) Architectures

Recent breakthroughs have extended the capabilities of AIMC far beyond its foundational role of accelerating linear operations. A landmark study from a team at Peking University, published in *Nature Electronics*, demonstrated a memristor-based hardware system capable of tackling complex, nonlinear sorting tasks without the need for traditional comparators [23]. This "comparator-free" design, which employs a novel Digit Read mechanism and a Tree Node Skipping (TNS) algorithm, achieved remarkable performance gains, including a 7.7x increase in speed and a 160.4x improvement in energy efficiency compared to leading ASIC-based sorters [23]. This finding is significant because it validates AIMC's versatility for a wider range of high-complexity, nonlinear tasks, moving it beyond its traditional focus on linear MVM.

Another significant advancement is the development of fully integrated systems-on-a-chip (SoCs) [8]. The startup TetraMem Inc., building on over a decade of academic research, has released a memristive SoC (the MX100) that integrates multiple computing cores with a RISC-V CPU [8]. This product represents a major milestone in the field, as it bridges the gap between cutting-edge academic research and the practical, scalable deployment of AI hardware. This development addresses the engineering challenges of integrating on-chip peripheral circuits and eliminating off-chip bottlenecks, which had previously limited overall system latency and energy efficiency [8].

3.3 Pushing the Boundaries: AIMC for Transformer and MoE Models

A central and active theme in modern AIMC research is its application to large, state-of-the-art models that have come to define modern AI. The work from IBM Research has been pivotal in this area [1, 4].

A paper featured on the cover of *Nature Computational Science*, with lead author Julian Büchel, demonstrated a unique 3D AIMC architecture specifically designed for Mixture of Experts (MoE) models [4]. This research showed how each "expert" in an MoE network layer can be mapped onto a distinct physical tier of a 3D non-volatile memory [4]. Through numerical simulations, this 3D AIMC architecture was shown to achieve higher throughput, higher area efficiency, and higher energy efficiency when running MoE models compared to commercially available GPUs [4].

The same research team also outlined the "first deployment of a transformer architecture on an analog in-memory computing chip" [4]. This work, which appeared in *Nature Machine Intelligence*, performed within 2% accuracy of a floating-point computation on a benchmark called the Long Range Arena [4]. This was a major breakthrough as it addressed the challenge of accelerating the attention mechanism, a key bottleneck for transformers [4]. Since the values in the attention mechanism are dynamically changing, they would typically require constant, energy-intensive reprogramming of the NVM devices [4]. To overcome this practical barrier, the researchers used a mathematical technique called "kernel approximation" to perform the necessary nonlinear functions on their experimental analog chip [4]. This solution demonstrates a co-design approach, where algorithmic innovations are used to circumvent the physical limitations of the hardware.

Table 3: Key AIMC Breakthroughs and Publications

Publication/Project	Authors/Team	Core Findings	Significance	Relevant Sources

Nature Computational Science MoE Paper	Julian Büchel (Lead), IBM Research	3D AIMC architecture outperforms GPUs for MoE models in throughput, area, and energy efficiency.	Demonstrates viability of AIMC for large-scale, modern AI architectures.	4
Nature Machine Intelligence Transformer Paper	IBM Research Team	First deployment of a full transformer architecture on an AIMC chip, achieving ~2% accuracy of FP operations.	Overcomes the key challenge of accelerating dynamic transformer attention mechanisms.	4
Nature Electronics Memristor Sorting Paper	Prof. Yang Yuchao's Team, Peking University	First comparator-free, memristor-based system for complex, nonlinear sorting. Achieved 7.7x speed and 160.4x energy efficiency gains.	Proves AIMC can go beyond linear MVM to handle complex, nonlinear tasks.	23
"Analog Foundation Models" Paper	IBM Research, ETH Zurich	LLMs can be made robust to AIMC noise through hardware-aware training	Addresses the fundamental concern of accuracy and non-determinism, proving	1

		(HAT), competing with 4-bit quantized LLMs.	AIMC's viability for high-stakes models.	
TetraMem MX100 SoC	TetraMem Inc.	First commercially available memristive SoC integrating multiple cores with a RISC-V CPU.	A significant milestone bridging the gap from academic research to practical commercial deployment.	[8]

4. Critical Analysis of Conflicting Viewpoints and Debates

4.1 The Great Divide: Analog vs. Digital In-Memory Computing

The debate between analog and digital in-memory computing represents a central point of contention and a key architectural divide in the field [24]. Digital IMC employs all-digital circuits, typically using SRAM as the memory device, to perform bit-by-bit product-accumulation operations directly within the memory array [24]. This approach provides high performance, strong noise immunity, and high robustness and reliability [24]. However, these benefits come at the cost of relatively large area and high power consumption overheads [24].

In contrast, AIMC, by leveraging non-volatile memory technologies and the physical laws of electronics, achieves unparalleled energy efficiency, high storage density, and a smaller physical footprint [24]. The trade-off is a fundamental one: analog computations are inherently noisy and non-deterministic, resulting in lower precision [1, 6, 24]. This core trade-off between precision and efficiency has led to the emergence of two distinct architectural paths. Digital IMC is better suited for high-precision, power-insensitive scenarios, such as large-scale cloud-based AI, while AIMC is more viable for energy-efficient, low-power applications at the edge where high accuracy is not a strict requirement [24]. This practical engineering choice mirrors a broader philosophical debate about whether digital systems can ever perfectly replicate the continuous nature of the analog world [16], a debate which finds a new, practical expression in the design of next-generation AI accelerators.

Table 2: Comparison of Analog vs. Digital In-Memory Computing

Feature	**Analog In-Memory Computing**	**Digital In-Memory Computing**
Primary Advantage	High energy efficiency, high parallelism	High precision, strong noise immunity
Primary Disadvantage	Lower precision, non-idealities	Higher area and power overheads
Typical Memory Medium	NVM (e.g., PCM, RRAM)	SRAM
Ideal Application	Low-power edge AI, inference	High-precision cloud AI, high-performance computing
Relevant Sources	[1]	[24]

4.2 The Precision-Efficiency Trade-off and Solutions

The inherent lack of precision and accuracy in analog computation due to noise, non-ideal device characteristics, and device-to-device variability is arguably the biggest challenge for the widespread adoption of AIMC [1, 6]. However, the literature indicates that this is not an insurmountable problem but rather a design challenge that can be mitigated through innovative co-design of hardware and software. A key solution is the use of "hardware-aware training" (HAT) [1, 6]. This technique involves training models on synthetic data while simultaneously using methods to enhance the model's robustness to the noise present in AIMC hardware [1]. A paper by IBM Research and ETH Zurich, titled "Analog Foundation Models," demonstrated that through this process, they could achieve accuracy comparable to 4-bit quantized LLMs [1]. This finding proves that a symbiotic relationship between software algorithms and hardware design can effectively mitigate a primary hardware limitation.

Furthermore, algorithmic workarounds can address other hardware constraints. For instance, the use of "kernel approximation" to perform the nonlinear functions required for transformer attention mechanisms is an example of a creative software solution to a hardware limitation [4]. The ability to address these fundamental precision concerns through both architectural innovation and algorithmic co-design indicates that the problem is not being solved by brute-force hardware perfection alone but through an integrated, holistic approach to system design.

5. Gaps in the Current Literature and Unresolved Challenges

Despite the significant breakthroughs in academic research, several critical challenges remain, forming a substantial gap between proof-of-concept prototypes and widespread commercial adoption [3, 6].

5.1 Hardware Non-Idealities and Variability

While hardware-aware training provides a viable path to mitigating noise, the underlying physical issues of non-ideal device characteristics and device-to-device variability still require significant research [6]. Mitigating these effects through circuit design and on-chip calibration techniques is an active area of research, but these solutions inevitably add to the complexity, area, and cost of the final hardware [6].

5.2 The Software and Toolchain Ecosystem Gap

The software ecosystem for AIMC is critically underdeveloped [3, 6]. Traditional programming paradigms, compilers, and frameworks are meticulously optimized for the von Neumann architecture, creating a significant lack of standardized abstractions for novel AIMC hardware [3, 6]. This forces developers to use vendor-specific tools and frameworks, which increases development costs and creates "vendor lock-in," thereby hindering broader industry adoption

[3].

5.3 Manufacturing Scalability and Cost Hurdles

The fabrication of hybrid memory-processing elements requires specialized manufacturing processes that differ from conventional CMOS manufacturing [3]. This leads to lower yields and higher production costs, resulting in AIMC solutions commanding a premium of 40-60% over traditional computing architectures, which is a significant barrier to entry for many potential customers [3].

5.4 Lack of Standardization and Industry Fragmentation

The AIMC industry is currently fragmented, with each solution operating within a proprietary ecosystem [3]. There is no clear consensus on hardware interfaces, programming models, or software development frameworks [3]. This fragmentation complicates integration into existing technology stacks and discourages significant investment from potential customers who are wary of committing to a single, non-standardized solution [3]. The challenges of manufacturing, standardization, and the software ecosystem are not isolated problems; rather, they form a circular, interdependent barrier to adoption. High costs limit the market size, which in turn stifles the development of open-source software and industry-wide standardization efforts. This lack of a robust ecosystem then further limits the market, creating a reinforcing negative feedback loop that is difficult to break without coordinated effort from both academic research and industry consortia.

Table 1: Evolution of Computing Paradigms

Paradigm	**Core Principle**	**Time Period**	**Key Limitation**	**Relevant Sources**
Von Neumann	Separate processing and memory units	~1940s-Present	Von Neumann Bottleneck (data movement)	[1]

Early Analog Computing	Physical properties used for continuous calculation	Pre-1950s	Precision, scalability, and versatility	15
Digital In-Memory Computing	Digital logic integrated with memory	~2010s-Present	High area/power overhead	24
Analog In-Memory Computing	Analog computation within non-volatile memory	~2010s-Present	Precision and noise sensitivity	1

6. Future Research Directions and Outlook

6.1 Suggestions for Future Research

The literature review reveals several promising avenues for future research to address the identified gaps and move the technology toward broader commercialization.

- **Hybrid Analog-Digital Architectures:** A central direction is the development of heterogeneous computing systems that combine the energy efficiency of analog cores with the precision and flexibility of traditional digital compute units [3, 4, 6]. This approach would allow different computational tasks to be dynamically assigned to the most appropriate processing substrate within a single system [3].
- **Standardized Software Ecosystems:** Research is urgently needed to develop open-source compilers, programming models, and frameworks that can abstract hardware complexities and enable seamless integration of AIMC with existing technology stacks [3].
- **On-Device Training and Adaptability:** While AIMC excels at inference, efficient on-device training or continuous learning remains a challenge due to the complexity of updating analog weights [6]. Future work should explore hybrid solutions where

fine-tuning occurs on the device, or entirely new methods for on-chip learning [6].

6.2 The Long-Term Vision: Hybrid and Heterogeneous Architectures

The long-term vision for in-memory computing is not to replace digital computing but to create a new computational paradigm where different tasks are dynamically assigned to the most suitable computing substrate [3]. This fully integrated, heterogeneous system, where traditional CPUs and GPUs work in concert with various forms of in-memory processors, promises to deliver unprecedented performance and energy efficiency [3]. Such an architectural evolution would be foundational to the next generation of computing applications, from resource-constrained edge devices to massive data centers, as it moves beyond the physical and economic limits of traditional hardware scaling [3].

7. Conclusion

Analog In-Memory Computing represents a paradigm shift poised to address the critical limitations of the von Neumann architecture in the era of data-intensive AI. While the core principles draw from the long history of analog computing and the evolution of non-volatile memory technologies, recent breakthroughs have validated its viability for complex workloads. These include the development of novel architectures for transformers and MoE models [4], the use of clever algorithmic workarounds to address dynamic tasks [4], and the application of hardware-aware training to mitigate the fundamental challenge of noise and precision [1]. Furthermore, breakthroughs in memristor-based hardware have proven that AIMC can handle nonlinear and high-complexity tasks beyond basic MVM [23].

Despite these advancements, significant challenges persist. The lack of standardized programming models and a mature software ecosystem, coupled with manufacturing hurdles and a fragmented industry, form a critical gap between academic proof-of-concept and widespread commercial adoption [3]. Future research is steering toward the development of hybrid architectures that combine the strengths of both analog and digital computing [4, 6], the creation of open-source software toolchains, and the exploration of new methods for on-device learning [6]. The ultimate success of AIMC will depend on its ability to transcend its current role as a specialized accelerator and evolve into a foundational component of a new, heterogeneous computing landscape that can deliver the performance and energy efficiency required for the next generation of AI applications.

Works cited

1. Overcoming accuracy limitations of Analog In-Memory Computing hardware - Reddit, accessed on September 26, 2025, https://www.reddit.com/r/MachineLearning/comments/1nkxejt/overcoming_accuracy_limitations_of_analog/
2. Analog AI - IBM Research, accessed on September 26, 2025, https://research.ibm.com/projects/analog-ai
3. What Challenges Limit Large-Scale Adoption Of In-Memory ..., accessed on September 26, 2025, https://eureka.patsnap.com/report-what-challenges-limit-large-scale-adoption-of-in-memory-computing-processors
4. Analog in-memory computing could power tomorrow's AI models ..., accessed on September 26, 2025, https://research.ibm.com/blog/how-can-analog-in-memory-computing-power-transformer-models
5. The Promise of Analog AI: Could In-Memory Computing ..., accessed on September 26, 2025, https://runtimerec.com/the-promise-of-analog-ai-could-in-memory-computing-revolutionize-edge-devices/
6. Analog Computing - Mythic, accessed on September 26, 2025, https://mythic.ai/technology/analog-computing/
7. Memristive system-on-a-chip for intelligent wireless communications, accessed on September 26, 2025, https://communities.springernature.com/posts/memristive-system-on-a-chip-for-intelligent-wireless-communications
8. Analog computer - Wikipedia, accessed on September 26, 2025, https://en.wikipedia.org/wiki/Analog_computer
9. Dyson claims: the brain is analog, and analog can't be simulated by digital : r/IsaacArthur, accessed on September 26, 2025, https://www.reddit.com/r/IsaacArthur/comments/j4m8cs/dyson_claims_the_brain_is_analog_and_analog_cant/
10. Memristor-Based Hardware Achieves Breakthrough in Nonlinear ..., accessed on September 26, 2025, https://neurosciencenews.com/memristor-nonlinear-sorting-29505/
11. Classification of Computing in Memory Principles - Digital ..., accessed on September 26, 2025, https://hackernoon.com/classification-of-computing-in-memory-principles-digital-computing-in-memory-vs-analog-computing