Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28449

In-memory Computing Architectures for Energy-efficient AI

Zijun Liu

International Engineering College, Xi'an University of Technology, Xi'an, China
ciyun768@gmail.com

Abstract. The exponential growth of AI—especially deep learning and generative AI—is severely constrained by the "memory wall" in von Neumann architectures, where frequent data movement between processors and memory consumes up to 90% of energy and creates critical latency bottlenecks. To address these limitations, this paper examines in-memory computing (IMC) as a transformative paradigm that co-locates computation and storage, targeting energy-efficient acceleration for AI workloads from edge inference to large-scale training. The analysis of DRAM, SRAM, and non-volatile memory (NVM) approaches reveals significant breakthroughs: capacitorless IGZO DRAM enables monolithic 3D-stacked, multibit arrays; ReRAM/PCM crossbars deliver ultra-efficient analog multiply-accumulate operations; and heterogeneous architectures (e.g., integrated analog-digital tiles with 2D mesh interconnects) achieve 22–64 TOPS/W efficiency—40–140× higher than GPUs. However, challenges persist in precision management, device variability, system programmability, and 3D integration scalability. This study concludes that IMC is pivotal for sustainable AI, potentially reducing operational carbon footprints by 10–100× through eliminated data movement. By overcoming current limitations via hybrid designs and standardized interfaces, IMC can extend beyond neural networks to graph processing and scientific computing, establishing itself as the cornerstone of future intelligent systems from edge to cloud.

Keywords: In-Memory Computing, IGZO DRAM, 3D Integration, Energy-Efficient AI, Heterogeneous Architectures

1. Introduction

The exponential growth of Artificial Intelligence (AI), particularly deep learning (DL) and generative AI (GenAI), has strained the limits of traditional computing architectures. Conventional von Neumann systems—where processors and memory are physically separated—suffer from the "memory wall": frequent data movement between units consumes up to 90% of energy and creates latency bottlenecks for data-intensive operations like matrix multiplications in neural networks [1]. In-Memory Computing (IMC) emerges as a transformative solution by co-locating computation and storage within memory arrays, drastically reducing data movement and enabling massive parallelism.

IMC leverages the physical properties of memory cells to perform computations directly where data resides. While early IMC research targeted niche applications, recent advances demonstrate its

DOI: 10.54254/2755-2721/2026.TJ28449

viability for mainstream AI workloads, from edge inference to large language model (LLM) training. This review examines: motivations driving IMC adoption in AI. Technical approaches across memory technologies (DRAM, SRAM, non-volatile memories), key breakthroughs in scalability, precision, and programmability, future directions for hardware-software co-design.

This study contextualizes breakthroughs like Li et al.'s 3D stacked IGZO DRAM array [2], Jain et al.'s heterogeneous analog-AI accelerator [3], and Ma et al.'s IMC taxonomy [4] within the broader IMC landscape.

2. Motivation: why IMC for AI

2.1. The von neumann bottleneck

AI workloads, particularly large language models (LLMs), impose stringent demands on both memory bandwidth and computational density: Matrix multiplication operations (GEMM/GEMV) emerge as the dominant contributor to Transformer inference, accounting for over 70% of latency and system energy consumption [3,4]; whereas processing high-resolution images or complex sequences necessitates transferring terabytes of weights and activations, with data movement energy exceeding computation energy by 100–1000×, thereby establishing a critical "energy disparity gap" [1,4]. These dual pressures collectively define the fundamental bottleneck in contemporary AI hardware design.

2.2. Energy and latency constraints

The energy demands of AI systems manifest differently across computing domains yet converge on memory bottlenecks: in cloud/server environments, the training of large-scale models like GPT-3 consumes massive power resources, with documented energy expenditures reaching 1,300 MWh per training cycle, equivalent to the annual electricity consumption of 500 average U.S. households [5]; meanwhile, at the edge computing frontier, devices such as autonomous drones and medical wearables require ultra-low-power inference capabilities operating within sub-watt thermal dissipation envelopes to ensure sustainable deployment, yet face fundamentally constrained energy budgets [4]. Conventional GPU and ASIC architectures struggle to address this dual-scaling challenge, as their performance and efficiency remain critically limited by energy-dominant off-chip memory accesses that can constitute over 60% of total system power consumption during inference workloads, thereby exacerbating both latency overheads and thermal management complexities across the AI hardware spectrum.

2.3. IMC’s value proposition

In-memory computing fundamentally addresses these systemic limitations through four interlocking technological vectors: First, by eliminating von Neumann-era data transfers through intrinsic colocation of computation and storage, as exemplified by Li et al.'s IGZO 2T0C DRAM array which exploits indium-gallium-zinc-oxide's ultra-low leakage currents (<10⁻¹² A/μm) to achieve >100-second retention times, thereby minimizing refresh energy by 89% compared to conventional 1T1C DRAM [2]; second, by harnessing massive spatial parallelism where crossbar architectures (ReRAM) and bank-level bitwise operations (DRAM) enable O(1)-complexity matrix computations, further amplified by Jain et al.'s dense 2D mesh interconnect that supports concurrent vector transport across 512 parallel pathways with implicit concatenation/scatter operations [3]; third, through inherent technology scalability where emerging non-volatile memories provide analog

DOI: 10.54254/2755-2721/2026.TJ28449

multiply-accumulate capabilities, multi-level cell storage (up to 4-bit/cell), and near-zero standby power – properties now extending to BEOL-compatible IGZO transistors enabling monolithic 3D integration [2,4]; and fourth, via architectural specialization manifest in heterogeneous acceleration fabrics that deploy application-optimized cores adjacent to memory tiles, such as the analog-digital partitioning in Jain et al.'s framework where dedicated units execute matrix multiplication, attention mechanisms, and activation functions within localized energy domains, collectively achieving 22-64 TOPS/W efficiency across diverse neural workloads [3].

3. Technical approaches

IMC architectures are classified by memory substrate, compute paradigm, and precision support. Key categories include:

3.1. DRAM-based IMC

DRAM technology provides critical advantages for in-memory computing implementations through its high-density storage capacity (>10GB/mm² in advanced nodes) and inherent bank-level parallelism, yet faces volatility-driven refresh constraints that complicate energy-optimized operation; beyond traditional capacitor-based implementations, two architectural pathways have emerged: bulk bitwise engines leveraging charge-sharing mechanisms within DRAM banks to perform energy-efficient Boolean operations (exemplified by the Ambit framework which accelerates bitwise AND/OR through coordinated row activation [4]), and capacitorless IGZO-based variants epitomized by Li et al.'s pioneering 8×8 3D stacked 2T0C DRAM array that achieves three transformative innovations—monolithic 3D integration vertically stacks indium-gallium-zinc-oxide transistors using BEOL-compatible processes to transcend planar scaling limits, sustaining >100-second retention with 8 distinct analog voltage levels, as further supported by recent advances in IGZO FETs for high-density 2T0C DRAM [6], thereby doubling effective density over binary implementations; and neural-optimized computing where the array functions as an int4 weight matrix for fully-connected layers, achieving 94.95% MNIST recognition accuracy through column-wise in-memory current summation that directly implements multiply-accumulate operations. While offering compelling strengths including standard-process compatibility, exceptional capacity scaling, and native multibit capability, these architectures necessitate careful co-design of asymmetrical write/read transistors (TW channel length = 180nm vs. TR = 270nm in [2]) and rigorous management of parasitic capacitances that can degrade signal margins in 3D configurations [2,7].

3.2. SRAM-based IMC

Despite offering sub-10ns access latency and full CMOS logic compatibility that enables rapid digital in-memory computation, SRAM-based architectures face intrinsic density limitations constraining practical implementations to sub-100MB capacities [4]; this technology bifurcates into two dominant implementation paradigms: analog current-domain approaches where simultaneous activation of multiple rows allows input activations—converted from digital via integrated DACs—to modulate word-line voltages, with resultant bitline currents summed to produce analog MAC outputs requiring subsequent ADC quantization (exemplified by Zhang et al.'s charge-domain computing variant that replaces current summation with precise capacitor integration to mitigate PVT variations [4]); and digital/near-digital implementations including Twin-8T SRAM cells enabling 4-bit multiply operations through dual-channel read ports, alongside bit-serial architectures

DOI: 10.54254/2755-2721/2026.TJ28449

like Wang et al.'s programmable vector processor supporting arbitrary-precision arithmetic through multi-cycle carry propagation [4]. These implementations collectively contend with three fundamental challenges: susceptibility to process-voltage-temperature variations degrading analog computation fidelity, write-disturb effects compromising storage integrity during multi-row access, and constrained signal margins limiting multilevel input precision—challenges progressively addressed through decoupled read structures (10T/12T bitcells isolating storage nodes from computing paths), charge-domain operation replacing transistor-dependent currents, and asymmetric transistor sizing [4]. While delivering exceptional strengths including industry-low <10ns latency critical for real-time systems, native integration with standard CMOS logic layers, and GHz-range operational speeds, SRAM-IMC remains ultimately bounded by its inherent capacity ceiling, substantial leakage power consumption (>30% of array energy at advanced nodes), and persistent analog non-idealities requiring complex calibration circuits that diminish area efficiency.

3.3. Non-Volatile Memory (NVM) IMC

Non-volatile memories (NVMs)—spanning resistive RAM (ReRAM), phase-change memory (PCM), spin-transfer torque MRAM (STT-MRAM), and Flash technologies—provide foundational advantages for in-memory computing through intrinsic support for analog computation and multi-level data storage, albeit with distinct material-specific tradeoffs: ReRAM crossbar arrays implement matrix-vector multiplication via Ohm's law-driven current modulation and Kirchhoff's current law summation, as demonstrated in Jain et al.'s accelerator architecture where pulse-width modulated inputs enable 8-bit activation processing while contending with non-ideal effects including >100pJ/bit write energy, >20% device variability, sneak-path currents distorting read accuracy, and substantial analog-to-digital conversion overhead consuming >35% of tile energy [3]; PCM devices achieve 4-bit/cell storage through crystallinity modulation for analog computing applications, yet suffer from resistance drift (>10%/decade at 85°C) and prohibitive >1μJ/bit SET energy that limits endurance to $\approx 10^8$ cycles [4]; NOR/NAND Flash memories leverage existing 3D integration for high-density vector-matrix multiplication, though their threshold voltage tuning requires specialized 6-20V programming regimes incompatible with logic-layer voltages, necessitating area-intensive charge pumps [4]; STT-MRAM implementations explore near-memory Boolean logic and addition operations benefiting from $>10^{15}$ write endurance, but encounter <10% read margin due to tunneling magnetoresistance ratio (TMR) variations that constrain analog precision [4]. While collectively offering critical strengths—including zero standby power non-volatility, 3D stackable density (>1Tb/in² for V-NAND), and theoretically optimal 10-100 TOPS/W analog MAC efficiency—these technologies face universal limitations encompassing stochastic device switching (σ/μ >0.3 for ReRAM/PCM), finite endurance in analog-mode operation (10^4-10^8 cycles), programming inaccuracies requiring iterative write-verify, and fundamental ADC/DAC energy costs exceeding 1pJ/conversion at >6-bit precision, establishing clear co-design priorities for next-generation NVM-IMC systems.

3.4. Heterogeneous & programmable architectures

Modern IMC systems integrate diverse components, exemplified by Jain et al.'s Analog-AI Accelerator employing a heterogeneous spatial architecture with Analog Fabric (AF) tiles: these incorporate analog CIM tiles (512x512) primarily for MAC operations using NVM (e.g., PCM/RRAM), featuring tile-level power/clock gating segmented to address poor mapping efficiency; specialized digital cores including Heavy Compute Cores (CC_H) for Self-

DOI: 10.54254/2755-2721/2026.TJ28449

Attention/Scratchpad functions, Light Compute Cores (CC_L) for LSTM-aux/ReLU/LayerNorm operations, and Memory Cores (MC) for data staging; a massively parallel 2D mesh interconnect enabling circuit-switched data transport with on-the-fly Concat/Scatter/Multicast operations controlled by Border Guards (BGs); double-buffered scratchpad SRAM for pipelined sequence storage (e.g., transformers); and fine/coarse-grained power gating critical for spatial efficiency. The architecture utilizes distributed programmable controllers executing pre-loaded instructions per component and scales via on-chip tiling or chip-to-chip I/O of multiple AFs, achieving 22-64 TOPS/W sustained efficiency across CNNs/LSTMs/Transformers representing 40-140× improvement over GPUs like NVIDIA A100 [3].

3.5. Comparative analysis of IMC architectures

Table 1 provides a comparative overview of the major IMC architectures—DRAM, SRAM, ReRAM/PCM, Flash, and heterogeneous designs—highlighting their strengths, limitations, and primary application domains. From the table, it is evident that DRAM-based IMC, particularly IGZO 2T0C variants, delivers advantages in density and retention time through 3D stacking and low-leakage oxide semiconductors, but still faces challenges such as parasitic capacitance and write uniformity. SRAM-based IMC excels in speed and CMOS compatibility, offering sub-10ns latency, though its limited capacity and high leakage power restrict large-scale deployment.

ReRAM and PCM crossbars, as shown in Table 1, stand out for their high analog MAC efficiency and non-volatility, making them attractive for energy-efficient multiply-accumulate engines [3,8]. However, device variability and endurance remain open issues. Flash-based IMC is highlighted in Table 1 as a mature, high-density option with analog potential, but voltage incompatibility and tuning precision challenges limit its current practicality. Finally, heterogeneous architectures combine multiple substrates to achieve programmability and high sustained efficiency, yet their complexity in mapping and compiler support represents a significant barrier to widespread adoption.

Overall, Table 1 emphasizes that no single architecture fully dominates across all dimensions. Instead, each memory technology offers unique tradeoffs that position it for specific use cases: DRAM for high-density arrays, SRAM for fast edge inference, ReRAM/PCM for efficient analog MAC operations, Flash for potential 3D analog CIM, and heterogeneous systems for end-to-end neural acceleration. This comparative analysis underscores the necessity of hybrid and co-designed approaches to unlock the full potential of IMC in future AI workloads.

Table 1. Showcased the characteristics, advantages, and disadvantages of different architectures

Architecture	Reference	Key Advantages	Key Limitations	Typical Applications
DRAM (IGZO 2T0C)	Li et al. [2]	High density, 3D stacking, multibit, BEOL compat	Parasitic capacitance, write speed, uniformity	High-density CIM arrays, edge-cloud
SRAM	Various [4]	High speed, CMOS logic compat, low latency	Limited capacity, leakage, variations	On-chip acceleration, inference edge
ReRAM/PCM	Jain et al. Tiles [3]	High analog MAC efficiency, non-volatile, density	Device variability, write endurance, ADC cost	High-efficiency MACengines
Flash	Lue et al. [4]	Mature, high-density (3D), analog potential	Voltage incompat, tuning precision	Potential high-density analog CIM
Heterogeneous	Jain et al. [3]	Programmability, handles Aux Ops, high sustained eff	Complexity, mapping, compiler support	Full DNN acceleration (LSTM, Transformer, CNN)

DOI: 10.54254/2755-2721/2026.TJ28449

4. Challenges and future directions

Despite significant progress, critical challenges remain for the practical deployment of in-memory computing (IMC). These challenges mainly concern four aspects: device precision and variability, system integration and programmability, technology scaling with 3D integration, and sustainable application in broader AI workloads. The following subsections discuss these issues in detail.

4.1. Precision, robustness & variability

Analog IMC implementations exhibit susceptibility to device variations (ReRAM/PCM), noise, parasitic effects (DRAM), and PVT variations (SRAM), with Li et al. [2] addressing array-level uniformity challenges specifically in IGZO DRAM to enable reliable 3-bit operation; future development requirements consequently include advanced calibration circuits supporting either in-situ or background operation, robust data representations employing ternary or binary formats, error compensation techniques exemplified by Extra Ops methodologies as implemented in Jain et al. [3], and hybrid analog-digital approaches combining analog MAC operations with digital refinement stages [3,9].

4.2. System integration & programmability

Critical system integration challenges persist across three domains: programming models and compilers require efficient mapping of complex DNN graphs onto heterogeneous spatial architectures exemplified by Jain et al. [3], necessitating standardized ISAs such as RISC-V PIM extensions and mature compiler toolchains as essential foundations [1,3]; data mapping and movement demand optimization of data tiling/sharing strategies with minimized inter-tile/core communication, leveraging architectural features like Jain et al.'s 2D mesh concat/multicast capabilities for performance-critical operations [3]; while integration with digital logic necessitates seamless co-design of IMC macros alongside conventional processors/accelerators, coupled with efficient handling of non-MAC operations that remain unresolved across the hardware stack [1,3,4].

4.3. Technology scaling & 3D integration

Monolithic 3D integration demonstrates significant density advantages as validated by Li et al.'s IGZO-based implementation [2], with future advancement priorities encompassing scaled layer counts, enhanced thermal management solutions, and heterogeneous integration paradigms such as logic-under-memory configurations or memristor-DRAM hybrid architectures [1,4]; concurrently, emerging device exploration targets novel material systems including IGZO semiconductors alongside innovative structures like vertical transistors and ferroelectric FETs (FeFETs) to advance performance metrics, density scaling, and non-volatile functionality [2,4].

4.4. Sustainable AI & broader applications

In-memory computing delivers transformative energy efficiency advantages promising 10–100× reductions in AI's operational carbon footprint through elimination of data movement bottlenecks [1,3,5]; extends applicability beyond deep neural networks to domains including graph processing, scientific computing (particularly sparse solvers), and database operations [1]; and enables neuromorphic synergy by integrating its architectural efficiency with event-driven spiking neural networks (SNNs) to achieve ultra-low-power sensory processing capabilities [1,4].

DOI: 10.54254/2755-2721/2026.TJ28449

5. Conclusion

In-memory computing (IMC) has evolved from a promising concept into a critical enabler for energy-efficient AI across the entire computing spectrum, ranging from edge devices to large-scale cloud infrastructures. By co-locating storage and computation, IMC fundamentally addresses the von Neumann bottleneck and unlocks new pathways for performance and efficiency. Key developments demonstrate the breadth of innovation: (1) DRAM Innovations: Capacitorless IGZO-based DRAM [2] leverages ultra-low leakage currents and monolithic 3D integration to achieve high-density, multibit arrays with long retention times, providing a viable platform for low-energy inference. (2) NVM Maturity: ReRAM/PCM crossbars [3] showcase ultra-efficient analog multiply-accumulate (MAC) operations and non-volatility, positioning them as strong candidates for large-scale AI accelerators, though variability and endurance remain active research challenges. (3) SRAM Refinement: Advanced bitcells and mixed-signal approaches improve robustness and computation precision, making SRAM-based IMC highly attractive for high-speed, real-time inference at the edge [4]. (4) System-Level Heterogeneity: Heterogeneous architectures integrate CIM tiles with digital cores, optimized interconnects, and fine-grained power management, delivering sustained efficiency improvements across diverse DNN workloads (CNNs, RNNs, and Transformers) [3].

Looking ahead, the future of IMC lies in overcoming device-level non-idealities through hybrid analog-digital co-design and calibration, scaling density and bandwidth via 3D integration, and establishing standardized programming interfaces to improve accessibility. Moreover, Table 1 highlights that each IMC technology offers unique trade-offs, underscoring the necessity of hybrid solutions tailored to specific workloads. As AI models continue to grow in complexity and energy demand, IMC architectures are poised to become the cornerstone of sustainable, real-time intelligent systems [5,10], reducing carbon footprints while broadening applicability to domains such as graph analytics, neuromorphic processing, and scientific computing.

References

[1] Horowitz, M. (2014) Computing's energy problem (and what we can do about it). 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, pp. 10–14.
[2] Li, Q., et al. (2025) 3D stacked IGZO 2T0C DRAM array with multibit capability for computing in memory applications. Science Advances, 11, eadu4323.
[3] Jain, S., Tsai, H., Chen, C.T., Muralidhar, R., et al. (2023) A heterogeneous and programmable compute-in-memory accelerator architecture for analog-AI using dense 2-D mesh. IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, vol. 31, no. 1, pp. 114–127.
[4] Ma, Y., Du, Y., Du, L., Lin, J., & Wang, Z. (2020) In-Memory Computing: The Next-Generation AI Computing Paradigm. In Proceedings of the 2020 on Great Lakes Symposium on VLSI (GLSVLSI '20), pp. 265–270.
[5] Patterson, D., Gonzalez, J., et al. (2021) Carbon Emissions and Large Neural Network Training. arXiv preprint arXiv: 2104.10350.
[6] Yan, S.Z., Cong, Z.R., Lu, N.D., Yue, J., Luo, Q. (2023) Recent progress in InGaZnO FETs for high-density 2T0C DRAM applications. Science China Information Sciences, 66(10): 200404.
[7] Belmonte, A., Oh, H., Subhechha, S., et al. (2021) Tailoring IGZO-TFT architecture for capacitor-less DRAM: achieving >10^3 s retention and Lg scalability down to 14 nm. In IEEE International Electron Devices Meeting (IEDM), pp. 226–228.
[8] Wan, W., Kubendran, R., Schaefer, C., Eryilmaz, S. B., et al. (2021) Edge AI without compromise: Efficient, versatile and accurate neurocomputing in resistive RAM. arXiv preprint arXiv: 2108.07879.
[9] Ambrogio, S., Narayanan, P., Tsai, H., et al. (2018) Equivalent-accuracy accelerated neural-network training using analogue memory. Nature, 558, 60–67.
[10] Joshi, V., Le Gallo, M., Haensch, W., & De Micheli, G. (2020) Analog AI hardware: Challenges and opportunities. Nature Electronics, 3, 292–300.