A Hybrid Analog-Digital RRAM CIM System with 4T2R1C Charge-Domain ACIM and RL-Driven Mixed-Precision Deployment for AI Inference

Minghui Yin[1#], Jiukai Fang[2#], Longhao Yan[1#*], Zelun Pan[1], Bowen Wang[1], Zhe Zhan[1], Xile Wang[1], Yaoyu Tao[1], Yuchao Yang[1,2,3,4*]

[1]New Cornerstone Science Laboratory, Beijing Advanced Innovation Center for Integrated Circuits, School of Integrated Circuits, Peking University, Beijing, China. [2]New Cornerstone Science Laboratory, Guangdong Provincial Key Laboratory of In-Memory Computing Chips, School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China. [3]Institute for Artificial Intelligence, Peking University, Beijing, China.
[4]Center for Brain Inspired Intelligence, Chinese Institute for Brain Research (CIBR), Beijing, Beijing, China.
[#]Equally contributed authors. *Corresponding Author's Email: yanlonghao@pku.edu.cn, yuchaoyang@pku.edu.cn

***Abstract*—Efficient edge AI demands high-precision, low-power hardware, yet conventional RRAM-based compute-in-memory (CIM) designs are hindered by high DC power, IR-drop, and transistor mismatch. We propose a highly scalable heterogeneous A/D hybrid CIM system featuring hierarchical co-optimization. At the circuit level, a 4T2R1C charge-domain cell is introduced to eliminate steady-state DC conduction, achieving a 15.6× energy reduction over current-domain designs while ensuring MOS-mismatch-free operation and linear multi-bit support. At the system level, an RL-driven mixed-precision framework orchestrates sensitivity-aware task allocation between high-efficiency Analog CIM (INT8) and high-precision Digital CIM (FP16) clusters. By employing CoreMatch (spatial core replication) and CoreStream (temporal inter-core streaming), the framework bridges the heterogeneous throughput disparity to achieve a 3.3× system-level speedup. Ultimately, this work demonstrates that the synergy of charge-domain computing and RL-driven orchestration offers a scalable and robust pathway toward bridging the efficiency-precision gap for sophisticated embodied AI applications.**

***Keywords*—ReRAM, charge-domain, compute-in-memory, mixed-precision**

I. Introduction

The rapid evolution of Edge AI applications, such as autonomous mobile robots and Unmanned Aerial Vehicles, demands hardware accelerators that deliver low latency and high energy efficiency without compromising precision[1], [2]. While Resistive Random Access Memory (RRAM)-based Non-volatile Compute-in-Memory (nvCIM) offers a promising solution, conventional designs face critical hardware bottlenecks (**Fig. 1**) [3], [4]. Current-domain architectures suffer from substantial DC power consumption and IR-drop induced inaccuracies[5], [6]. Meanwhile, conventional charge-domain designs based on near-threshold gain are highly vulnerable to transistor mismatch and exhibit poor compatibility with high-precision multi-bit weights[7].

To address these challenges, we propose a highly scalable heterogeneous A/D hybrid CIM system featuring hierarchical co-optimization. At the circuit level, we introduce a 4T2R1C charge-domain RRAM cell based on a voltage-division mechanism. Unlike gain-based charge-domain designs, our voltage-division approach is inherently MOS-mismatch-free and supports linear multi-bit Matrix-Vector Multiplication (MVM) operations. By transitioning to this transient computing mode, the design eliminates steady-state DC conduction, achieving a 15.6× energy reduction over current-domain counterparts while ensuring high robustness against IR-drop.

At the system level, an RL-driven mixed-precision framework orchestrates the deployment[8], [9]. It performs sensitivity-aware Precision Mapping between high-efficiency ACIM (INT8) and high-precision DCIM (FP16) clusters.

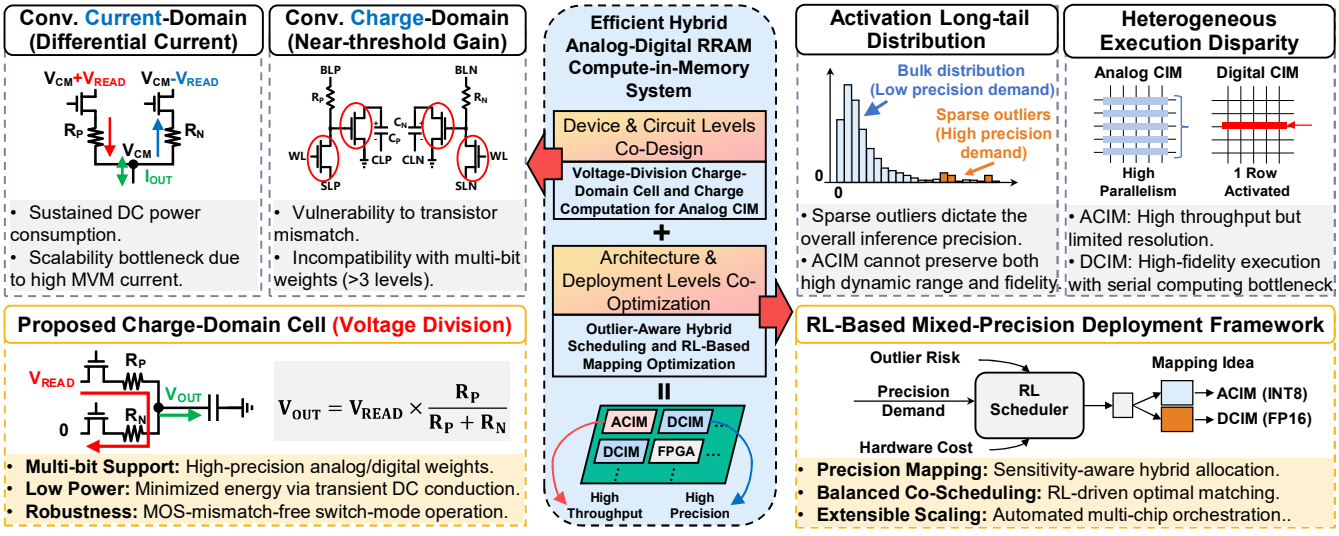

Fig. 1. Overview of the efficient hybrid analog-digital RRAM CIM system with hierarchical co-optimization from device levels to deployment frameworks.

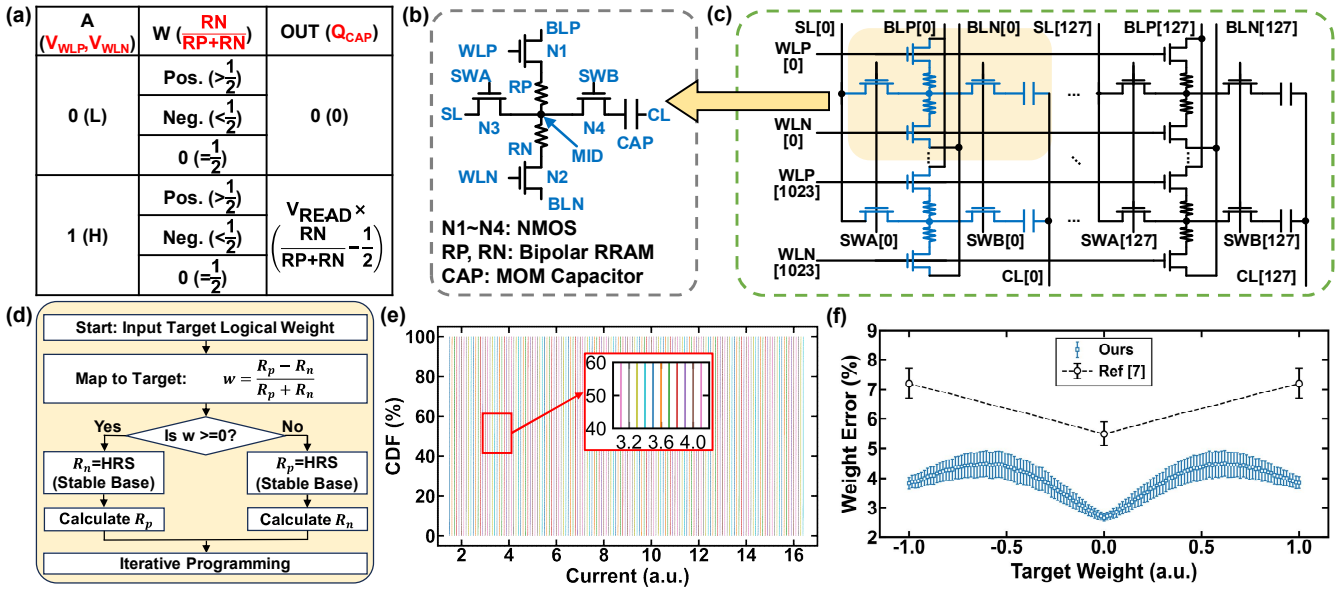

Fig. 2. Proposed 4T2R1C charge-domain cell and its array-level implementation: (a) Operation truth table; (b) 4T2R1C cell schematic; (c) 1024* 128 array architecture; (d) Iterative programming algorithm; (e) Measured CDF of 150 resistance states (>7bits); (f) Weight mapping error comparison with Ref [7].

Furthermore, it employs a synergistic scheduling strategy consisting of CoreMatch (spatial core replication) and CoreStream (temporal inter-core streaming) to bridge the throughput disparity, achieving a 3.3× system-level speedup. Experimental results show that our platform achieves 358× higher energy efficiency and 16× higher throughput than the NVIDIA H100 GPU, marking a transformative architecture for embodied intelligence.

II. Proposed Charge-Domain 4T2R1C RRAM CIM

We propose a 4T2R1C charge-domain cell, as depicted in **Fig. 2(b)**. The topology utilizes a series-connected RRAM pair (R_P, R_N) integrated with a high-density Metal-Oxide-Metal (MOM) capacitor to perform Multiply-Accumulate (MAC) operations via capacitive voltage division.

Unlike conventional current-domain designs that rely on absolute current summation, this cell operates in a transient voltage-domain mode. During the evaluation phase, the RRAM pair forms a potential divider network that charges the MOM capacitor to a specific voltage level. According to the truth table in Fig. 2(a), which defines the resistance states for digital weight mapping, the output voltage (V_{OUT}) follows the linear relationship:

$$V_{out} = V_{READ} \times \frac{R_P}{R_N + R_P} \quad (1)$$

This switch-mode operation is inherently robust against transistor mismatch and ensures high precision. To achieve high-fidelity weight representation, an iterative programming algorithm is developed based on an HRS-stable-base strategy **Fig. 2(d)** . By fixing one device at the High Resistance State (HRS), the target conductance is mapped precisely to the other device. Experimental results in **Fig. 2(e)** show over 150 distinguishable resistance states (>7 bits) . Compared to our previous work, this implementation achieves significantly lower weight mapping error across the entire target range, as illustrated in **Fig. 2(f)** .

The proposed charge-domain computing scheme operates through three synchronized phases: Reset, Multiply, Accumulate and and ADC Conversion,as illustrated in the timing and circuit diagrams of **Fig. 3(a)**. During the Reset phase, the storage MOM capacitors within the 4T2R1C cells are fully discharged to a baseline reference to eliminate residual charges from previous computational cycles. In the Multiply phase, the word-line (WL) pulses trigger the voltage-division mechanism of the RRAM pair, where the resistive ratio modulates the charging voltage, effectively mapping the weight-input product into stored charge packets. Subsequently, the Accumulate phase facilitates passive charge sharing across the array rows by interconnecting the

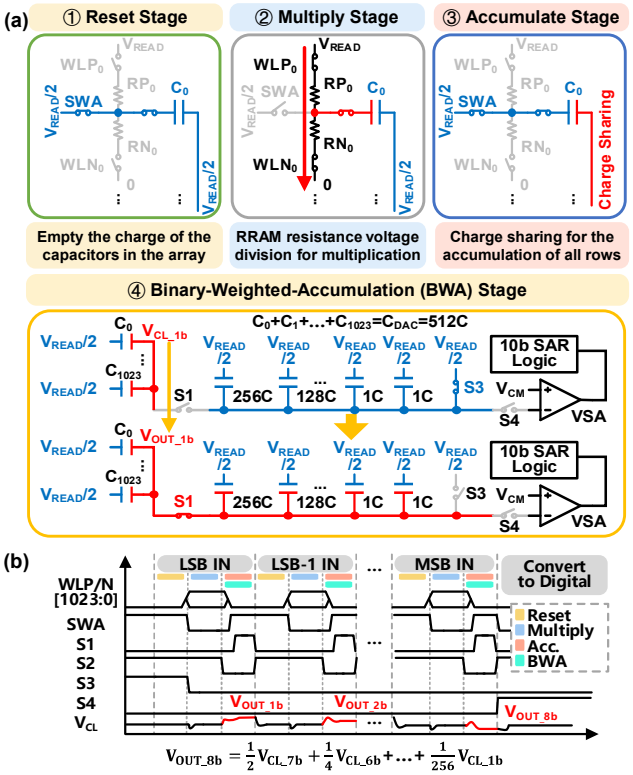

Fig. 3. Operation principles and binary-weighted accumulation: (a) Three-phase charge-domain computing scheme consisting of Reset, Multiply , and Accumulate (charge sharing) stages; (b) Circuit architecture, control timing, and recursive weighted-sum logic for multi-bit activation inputs.

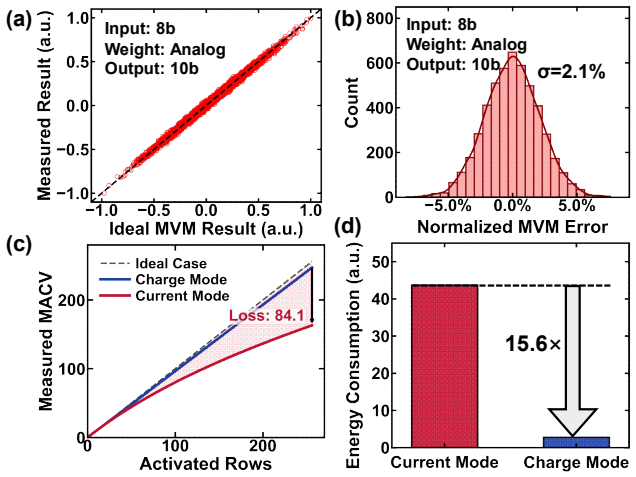

Fig. 4. Measured performance: (a) MVM results; (b) Error distribution; (c) IR Drop Resilience Comparison; (d) Energy consumption comparison.

local capacitors, allowing the discrete charges to redistribute and yield an analog partial sum. Finally, in the ADC Conversion phase, the integrated 10-bit SAR ADC digitizes the resulting analog voltage for further processing.

To support high-precision AI inference, a Binary-Weighted-Accumulation (BWA) stage is implemented using a capacitive DAC (CDAC) shown in **Fig. 3(b)** . It recursively accumulates analog partial sums of multi-bit inputs according to the timing diagram. This recursive weighted-sum logic allows the 10-bit SAR ADC to efficiently convert multi-bit results while maintaining high linearity .

The computational performance of the fabricated CIM macro was characterized through measurements presented in **Fig. 4**. Linear MVM operations with 8b-input and 10b-output precision are shown in **Fig. 4(a)**, yielding a normalized error distribution with a of sigma = 2.1% in **Fig. 4(b)** . A critical advantage of the charge-domain scheme is its superior resilience against IR-drop, as demonstrated by the comparison in **Fig. 4(c)** . Unlike current-domain baselines that suffer from significant Multiply-Accumulate Value (MACV) loss, the proposed method maintains high fidelity across the array. Furthermore, as shown in **Fig. 4(d)**, the transition from the current to the charge domain results in a 15.6× reduction in energy consumption compared to conventional current-domain implementations

III. Hybrid A/D Architecture & RL-Driven Deployment

The proposed heterogeneous A/D hybrid CIM system is validated on a scalable multi-chip PCB platform (**Fig. 5(a)**), integrating 9 ACIM and 3 DCIM chips coordinated by an FPGA-based controller. To bridge the throughput disparity between the ACIM and DCIM cores, an RL-driven mixed-precision framework (**Fig. 5(b)**) automates deployment via five stages: Stage 1 performs sensitivity-aware partitioning; Stage 2 constructs a hardware-aware Tiled Directed Acyclic Graph (Tiled DAG); and Stages 3-5 utilize an RL agent to iteratively optimize two core strategies: (i) CoreMatch, which eliminates the execution rate gap by replicating DCIM cores to match ACIM throughput, and (ii) CoreStream, which enables inter-core pipeline overlapping to allow dependent cores to initiate execution before preceding stages fully complete.

As quantified by the performance characterization in **Fig. 5(d)**, the synergy of these RL-orchestrated strategies yields a 3.3× system-level speedup by maximizing the functional utilization of heterogeneous resources. Empirical evaluations using a ResNet-18 workload on the ImageNet dataset **(Fig. 5(c))** confirm that the system preserves a top-1 accuracy of 66.3%, exhibiting negligible degradation relative to the full-precision software baseline. When benchmarked against NVIDIA H100 GPU **(Fig. 5(f))**, the proposed platform demonstrates a 358× improvement in energy efficiency and a 16× enhancement in throughput. These results substantiate that the convergence of charge-domain computing and RL-driven architectural orchestration provides a robust and scalable paradigm for sophisticated embodied AI at the edge.

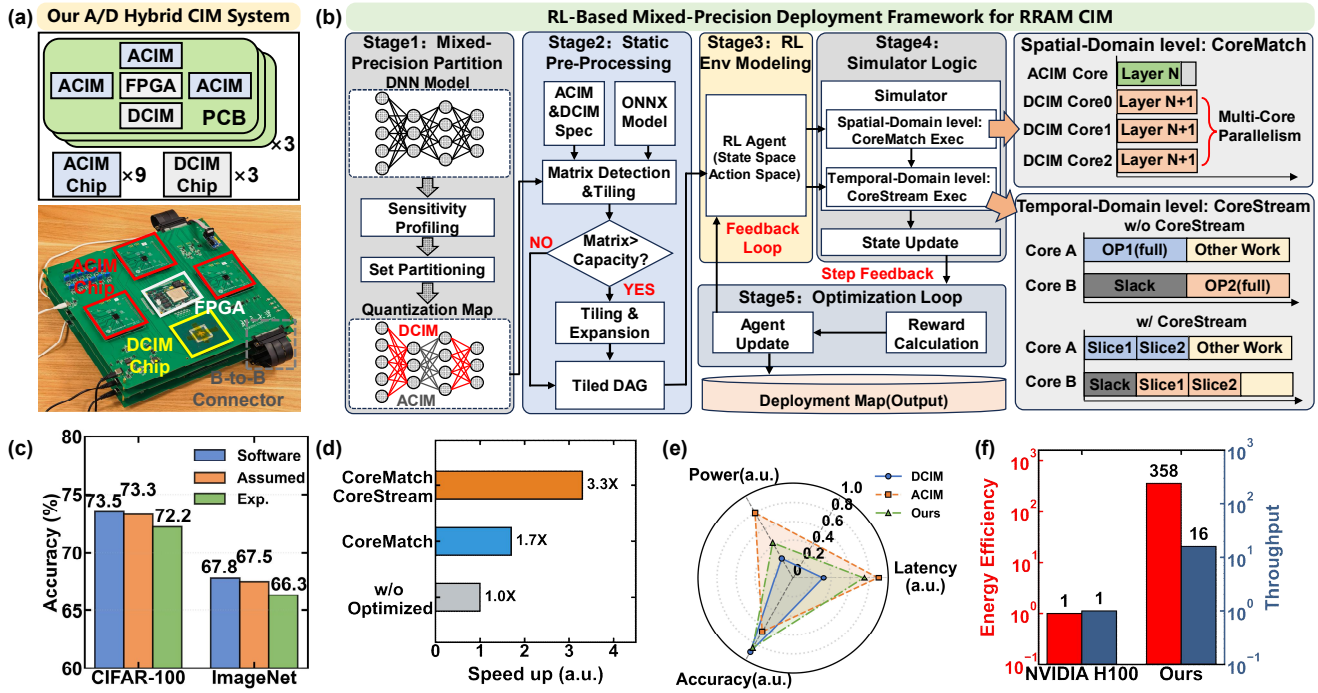

Fig. 5. System-level demonstration and evaluation: (a) Hardware platform; (b) RL-Based Mixed-Precision Deployment Framework; (c) Performance Comparison on ResNet-18; (d) Speedup with proposed optimizations; (e) Multi-Metric Performance Profile; (f) Performance comparison with NVIDIA H100.

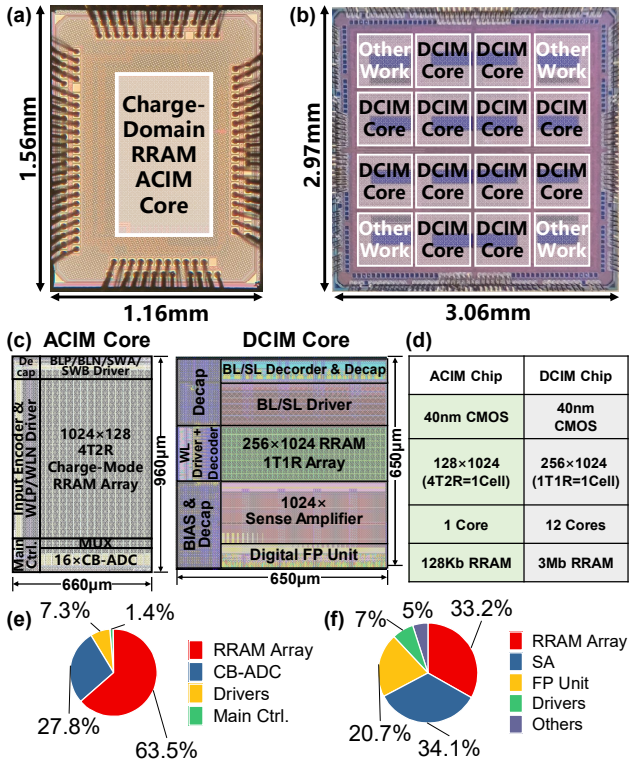

Fig. 6. Physical implementation: (a) ACIM die micrograph; (b) DCIM die micrograph; (c) Core floorplans; (d) Characteristics summary; (e) Measured ACIM power breakdown; (f) Measured DCIM power breakdown.

IV. Physical Implementation

The heterogeneous CIM platform was fabricated using a 40nm CMOS process, with die micrographs shown in **Fig. 6(a)** and **Fig. 6(b)**. The floorplans of the ACIM and DCIM cores are detailed in **Fig. 6(c)**, and the chip characteristics are summarized in Fig. 6(d) . Power breakdown analysis in **Fig. 6(e)** shows that the RRAM array and peripheral CB-ADC consume 63.5% and 27.8% of the ACIM power, respectively . For the DCIM core in **Fig. 6(f),** the RRAM array, sense amplifiers, and floating-point units are the primary contributors . This heterogeneous implementation optimizes area and energy for scalable in-memory computing.

V. Conclusion

This paper presents a 40nm heterogeneous A/D RRAM CIM system integrating a 128Kb charge-domain ACIM and a 3Mb DCIM. The 4T2R1C cell achieves a 15.6× energy reduction and 286.8 TOPS/W efficiency. An RL-driven framework optimizes inter-core task partitioning and pipeline, yielding a 3.3× system speedup. Compared to NVIDIA H100, the platform delivers 358× higher energy efficiency and 16× higher throughput, establishing a new paradigm for edge AI.

Acknowledgment

This work was supported by the National Key R&D Program of China (2023YFB4502200), the National Natural Science Foundation of China (92164302, 62404004, U25A20487), Guangdong Provincial Key Laboratory of In-Memory Computing Chips (2024B1212020002), Shenzhen Science and Technology Program (ZDCY20250901103401002, JCYJ20241202125907011), Beijing Natural Science Foundation (L234026, L257010) and the 111 Project (B18001). This work has been supported by the New Cornerstone Science Foundation and Financial Support for Outstanding scientific and technological innovation Talents Training Fund in Shenzhen.

Table 1. Comparison results with prior published works.

	Ours		ESSERC 2025 [7]	VLSI 2025 [10]	ESSERC 2024 [11]
Technology	40nm		40nm	180nm	28nm
Memory type	RRAM		RRAM	RRAM	RRAM
CIM Type	ACIM	DCIM	ACIM	ACIM	ACIM
RRAM Level	MLC/ Analog	SLC	SLC	MLC	MLC
Capacity	128Kb	3Mb	1Mb	2Kb	512Kb
Computing Mode	Charge	Current	Charge	Current	Current
Energy Efficiency (TOPS/W)	286.8 (IN:8b W:Analog O: 10b)	7.6 TFLOPS/W (FP16)	103.5 (IN:8b W:Ternary O:6b)	\	39.3 (INT 8)
Heterogeneous Computing	Hardware Implementation (PCB-Level)		No	No	No
Mixed-Precision Deployment	Yes (Reinforcement-Learning-Driven)		No	No	No
Multi-Chip Scalability	High		Low	Low	Low

References

[1] W. S. Khwa et al., "A mixed-precision memristor and SRAM compute-in-memory AI processor," Nature, vol. 639, no. 8055, pp. 617–623, 2025, doi: 10.1038/s41586-025-08639-2.

[2] X. Zheng et al., "Point-of-care testing (POCT) system based on self-recovery memoristor chip with low energy consuption (1.547 TOPS/W) and high recognition (1142 fram/s)," in Proc. IEEE Int. Electron Devices Meeting (IEDM), 2023, pp. 1–4, doi: 10.1109/IEDM45741.2023.10413719.

[3] S. Liu et al., "HARDSEA: Hybrid analog-ReRAM clustering and digital-SRAM in-memory computing accelerator for dynamic sparse self-attention in Transformer," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 32, no. 2, pp. 269–282, 2024, doi: 10.1109/TVLSI.2023.3337777.

[4] C. Mu et al., "A 28-nm RRAM/SRAM collaborative CIM accelerator supporting RRAM-endurance-latency awareness for edge fine-tuning," IEEE J. Solid-State Circuits, vol. 60, no. 10, pp. 3626–3638, 2025, doi: 10.1109/JSSC.2025.3577335.

[5] W.-H. Huang et al., "A nonvolatile AI-edge processor with 4 MB SLC-MLC hybrid-mode ReRAM compute-in-memory macro and 51.4–251 TOPS/W," in Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC), 2023, pp. 258–259, doi: 10.1109/ISSCC42615.2023.10067610.

[6] W. Ye et al., "A 28-nm RRAM computing-in-memory macro using weighted hybrid 2T1R cell array and reference subtracting sense amplifier for AI edge inference," IEEE J. Solid-State Circuits, vol. 58, no. 10, pp. 2839–2850, 2023, doi: 10.1109/JSSC.2023.3280357.

[7] L. Yan, et al., "A 1308 TOPS/W charge-mode ReRAM CIM macro with 4T2R2C differential cell and FIA-based analog accumulation for AI inference," in Proc. IEEE Eur. Solid-State Electron. Res. Conf. (ESSERC), 2025, pp. 473–476, doi: 10.1109/ESSERC66193.2025.11214056.

[8] Y. Yuan et al., "A 28 nm 72.12 TFLOPS/W hybrid-domain outer-product based floating-point SRAM computing-in-memory macro with logarithm bit-width residual ADC," in Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC), 2024, pp. 576–578, doi: 10.1109/ISSCC49657.2024.10454313.

[9] M. R. H. Rashed, S. K. Jha, and R. Ewetz, "Hybrid analog-digital in-memory computing," in Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD), 2021, pp. 1–9, doi: 10.1109/ICCAD51958.2021.9643526.

[10] M. Pallo et al., "On chip customized learning on resistive memory technology for secure edge AI," in Proc. Symp. VLSI Technol. Circuits, 2025, pp. 1–3, doi: 10.23919/VLSITechnologyandCir65189.2025.11074789.

[11] P. Yao et al., "A 28 nm RRAM-Based 81.1 TOPS/mm2/bit Compute-In-Memory Macro with Uniform and Linear 64 Read Channels under 512 4-bit Inputs," 2024 IEEE European Solid-State Electronics Research Conference (ESSERC), Bruges, Belgium, 2024, pp. 577-580.