An 8Mb 16-Core Wear-Resilient RRAM Compute-in-Memory Processor with Overshoot-Optimized MLC Programming and Compute-Aware Wear Leveling for Aerial Embodied Learning

Ruiheng Wang[2#], Longhao Yan[1#*], Daijing Shi[1], Zelun Pan[1], Zhe Zhan[1], Xinjie Wen[2], Yaoyu Tao[1], Yuchao Yang[1,2,3,4*]
[1]School of Integrated Circuits, Peking University, Beijing, China. [2]School of Electronic and Computer Engineering, 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, China.
[#]Equally contributed authors. *Corresponding Author's Email: yanlonghao@pku.edu.cn, yuchaoyang@pku.edu.cn

***Abstract*—Aerial embodied intelligence requires autonomous drones to adapt to dynamic environments via edge on-chip learning (OCL). Since stringent power and latency constraints make conventional hardware inefficient for flight-critical edge OCL, RRAM-based compute-in-memory (CIM) offers a promising substrate. However, continuous weight updates introduce critical reliability bottlenecks. At the device level, stochastic conductance overshoots during conventional write-and-verify (W&V) programming severely degrade device endurance. At the system level, non-uniform workloads across multi-core arrays create localized wear hotspots, which compute-agnostic wear-leveling schemes fail to address without massive throughput degradation. This paper presents a 40-nm, 8-Mb, 16-core RRAM CIM processor that jointly mitigates these device- and system-level challenges. An Overshoot-Optimized Programming (O²P) scheme employs a three-phase adaptive pulse strategy, reducing programming latency by 2.1×, write stress by up to 5×, and correction overhead by 5×. Furthermore, a Lightweight Compute-Aware Wear Leveling (LCWL) mechanism utilizes event-driven rebalancing to extend block-level lifetime by 15.3× with minimal throughput overhead (< 6.3%). Jointly, these techniques extend overall chip lifetime by 87× over baseline RRAM CIM designs. Validated on an autonomous drone performing adaptive cruise control, the proposed chip demonstrates a 12× energy reduction and a 140× latency speedup compared to an NVIDIA Jetson AGX Orin.**

***Keywords*—RRAM, compute-in-memory, on-chip learning, wear leveling, embodied intelligence**

I. Introduction

The proliferation of aerial embodied intelligence requires autonomous drones to perceive and adapt to dynamic environments in real time across applications such as search-and-rescue and precision agriculture [1]. Subjected to abrupt illumination changes, complex terrains, and incomplete environmental information, these platforms require continuous in-situ model adaptation to maintain robust decision-making. Conventional approaches rely on cloud offloading or GPU-class edge processors such as the NVIDIA Jetson series, but cloud offloading incurs prohibitive communication latency for flight-critical tasks, while GPU-class edge platforms impose significant power budgets and form factors incompatible with payload-constrained drones [2-3]. These limitations create a strong demand for edge OCL, where models are adapted directly on ultralow-power embedded hardware.

Resistive RAM (RRAM)-based CIM provides an ideal substrate for OCL [4-5]. By executing analog multiply-accumulate (MAC) operations in situ, they circumvent the von Neumann memory wall and deliver high energy efficiency and parallelism. However, transitioning from inference-only to OCL introduces critical endurance and reliability challenges (**Fig. 1**). At the device level, the stochasticity of RRAM conductance updates frequently causes programming overshoot [6]. In conventional W&V schemes, this triggers oscillatory programming cycles that severely degrade device endurance. This vulnerability is compounded at the system level: non-uniform weight updates across multi-core CIM architectures create localized wear hotspots. Furthermore, conventional compute-agnostic wear-leveling schemes rely on fixed-interval block migrations that stall parallel execution, severely degrading system throughput [7].

To address these device- and system-level bottlenecks, this work presents a 40-nm, 8-Mb, 16-core RRAM CIM processor tailored for efficient OCL. At the device level, an O²P scheme employs a three-phase adaptive strategy: coarse-grained convergence, stochastic overshoot detection, and fine-grained residual correction. This achieves near-W&V programming precision while reducing latency by 2.1× and write stress by up to 5×. At the system level, a LCWL mechanism utilizes multi-tiered condition evaluation, triggering block swaps only upon detecting actual wear

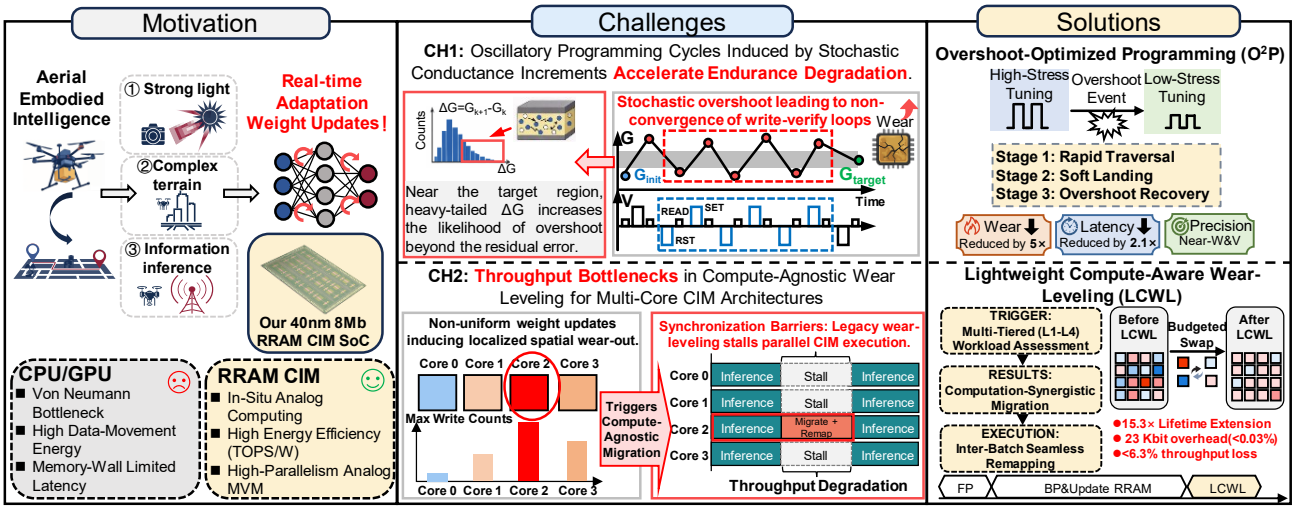

Figure 1: Motivation, reliability challenges, and proposed O²P and LCWL solutions for RRAM CIM on-chip learning in aerial embodied intelligence.

imbalance. By leveraging SRAM-buffered execution and logical-to-physical address remapping, LCWL extends block-level lifetime by 15.3× with minimal throughput overhead (< 6.3%). Jointly, O²P and LCWL improve overall chip lifetime by 87× compared to conventional RRAM CIM designs. The proposed processor is validated on an autonomous drone performing adaptive cruise control via aerial embodied learning, demonstrating a 12× energy reduction and a 140× latency speedup over an NVIDIA Jetson AGX Orin.

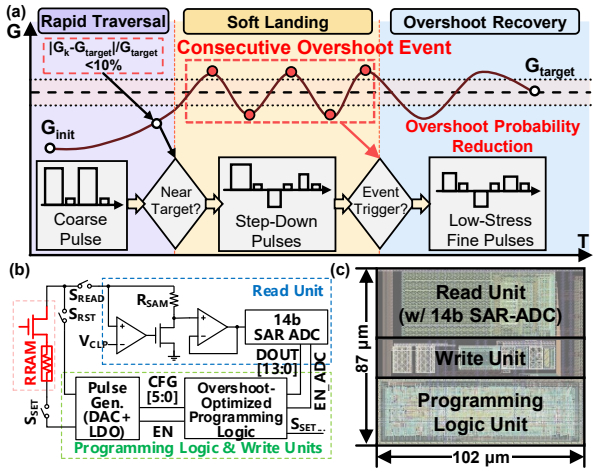

Figure 2: (a) Principle of O²P programming strategy. (b) Schematic of the read/write circuits with O²P control logic. (c) Layout of the O²P circuits.

II. Proposed Wear-Resilient RRAM CIM Design

A. Overshoot-Optimized Programming

Conventional W&V programming applies fixed-amplitude pulses iteratively until the conductance reaches the target, which guarantees precision but incurs significant latency and excessive device stress. One-Shot programming, on the other hand, applies a single calibrated pulse, offering minimal latency but suffering from poor accuracy due to the inherent stochasticity of RRAM. The proposed O² P scheme bridges this gap by employing a three-phase adaptive pulse strategy that dynamically adjusts pulse parameters according to the real-time programming state, as shown in **Fig. 2(a).**

In the rapid traversal phase, coarse high-amplitude pulses drive the conductance toward the target at maximum speed, prioritizing convergence over precision. Once the relative conductance error falls below 10%, the algorithm transitions to the soft-landing phase, where step-down pulses with progressively reduced amplitude suppress the overshoot probability. If consecutive overshoot events are detected, indicating an oscillatory regime, the algorithm enters the overshoot recovery phase, applying dedicated low-stress fine pulses to steer the conductance back to the target with minimal additional device degradation. Unlike W&V, which corrects overshoot with the same fixed-amplitude pulses, O²P confines high-energy stress to the early traversal phase and employs low-energy recovery, significantly reducing cumulative wear. The O²P control logic is implemented on-chip, as shown in the partial schematic in **Fig. 2(b)**. The corresponding layout, shown in **Fig. 2(c)**, integrates a read unit with a 14-bit Successive Approximation Register Analog-to-Digital Converter, a write unit, and a programming logic unit within a compact area of 87×102 µm².

To provide a quantitative measure of the cumulative electrical wear imposed on the RRAM device during programming, we define a write stress metric as:

$$S = \sum_i V_i^2 t_{pulse\_i} \quad (1)$$

Where V_i and t_{pulse_i} denote the voltage amplitude and width of the i-th programming pulse, respectively. The V^2 term reflects the Joule heating power dissipated across the switching layer, while t_{pulse} captures the duration of thermal and electrical stress. Together, S serves as an engineering-oriented proxy for the total energy deposited into the oxide, which correlates with filament over-growth, oxygen vacancy redistribution, and endurance degradation. **Fig. 3** presents comprehensive experimental validation. **Fig. 3(a)** confirms excellent programming linearity across the full conductance range with a tight Gaussian-like error distribution, and **Fig. 3(b)** further shows that the conductance distribution of O²P closely approaches W&V while significantly outperforming One-Shot, demonstrating near-W&V precision without sacrificing efficiency. Beyond precision, **Fig. 3(c)** reveals a 2.1× reduction in mean programming latency compared to W&V, as the coarse-to-fine pulse transition eliminates redundant verify cycles far from the target and avoids repeated overshoot-correction loops near the target. Meanwhile, **Fig. 3(d)** shows that the write stress S is consistently and substantially lower for O²P across different conductance ranges, since high-energy pulses are confined to the early traversal phase rather than applied throughout the entire process. When overshoot events are triggered, **Fig. 3(e)** shows that O²P requires 5× fewer corrective pulses to recover, owing to the dedicated low-energy recovery phase that achieves monotonic convergence instead of the oscillatory correction inherent to W&V. As a combined result, **Fig. 3(f)** confirms a 2× reduction in total programming pulses with a narrower distribution, reflecting the cumulative efficiency gains across all three phases.

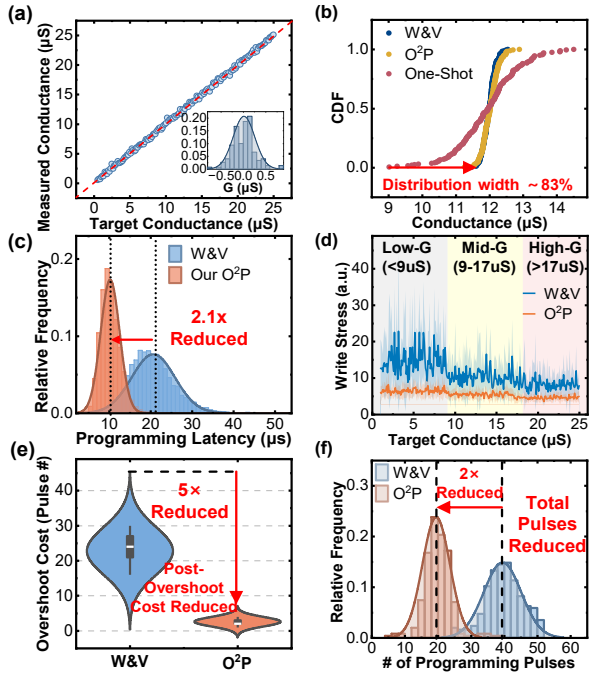

Figure 3: (a) Measured and target conductance with error distribution. (b) Programmed conductance CDF comparing three methods. (c) Programming latency distribution. (d) Write stress across different conductance ranges. (e) Post-overshoot correction cost. (f) Total programming pulse distribution.

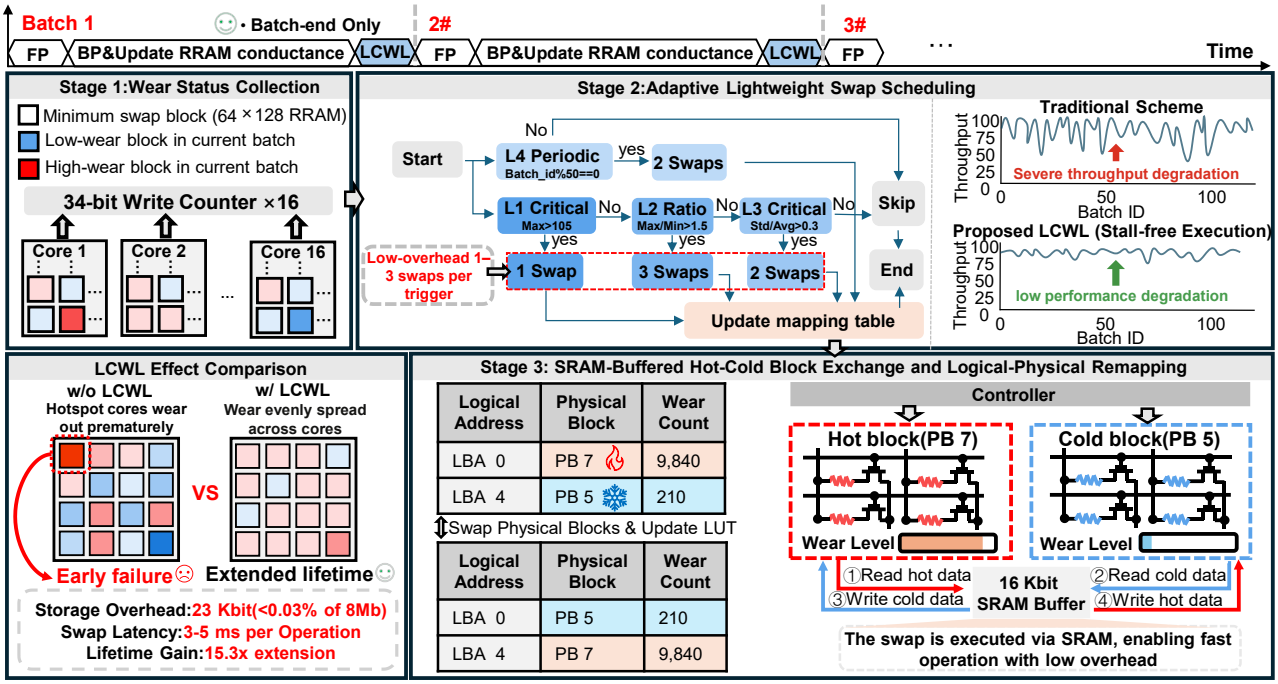

Figure. 4: Overview of the proposed three-stage LCWL scheme with hardware implementation details and wear distribution comparison.

B. *Lightweight Computer-Aware Wear Leveling*

In multi-core CIM architectures, OCL induces inherently non-uniform wear distributions across cores, as different layers and partitions of the neural network impose varying update frequencies on their mapped physical blocks. Conventional wear-leveling approaches perform fixed-interval rebalancing regardless of actual wear state, which introduces frequent and unnecessary data migrations that severely degrade training throughput. To address this, we propose LCWL, a compute-aware wear management scheme that monitors real-time wear statistics and triggers block swaps only when actual imbalance is detected, thereby achieving effective wear redistribution while preserving high training throughput.

As illustrated in **Fig. 4**, LCWL is invoked exclusively at batch boundaries to avoid interference with the training data path. The control logic is implemented entirely on-chip with minimal overhead, comprising 34-bit per-core write counters, a 512-entry logical-to-physical mapping table (9 bits per entry), and a 16 Kbit SRAM buffer for ping-pong block swapping, totaling approximately 23 Kbit (<0.03% of the 8 Mb array). The scheme operates in three stages. In the first stage, the controller collects the cumulative write count of each core at the granularity of 64 × 128 RRAM blocks via the per-core counters, establishing a real-time wear profile across all 16 cores. In the second stage, an adaptive swap scheduling algorithm evaluates four hierarchical conditions to determine the necessity and intensity of rebalancing. The algorithm first checks whether a periodic interval of 50 batches has been reached, in which case two swaps are issued. If not, it sequentially examines three event-driven conditions: whether the maximum write count exceeds 10^5 whether the max-to-min write count ratio exceeds 1.5, and whether the ratio of standard deviation to mean exceeds 0.3. The cycle is bypassed entirely if no condition is satisfied. This condition-driven hierarchy ensures that aggressive migrations are reserved for severe imbalance scenarios, while mild or absent imbalance incurs zero overhead. In the third stage, the selected hot block and cold block are physically exchanged via an on-chip SRAM buffer, and the logical-to-physical mapping table is updated upon completion.

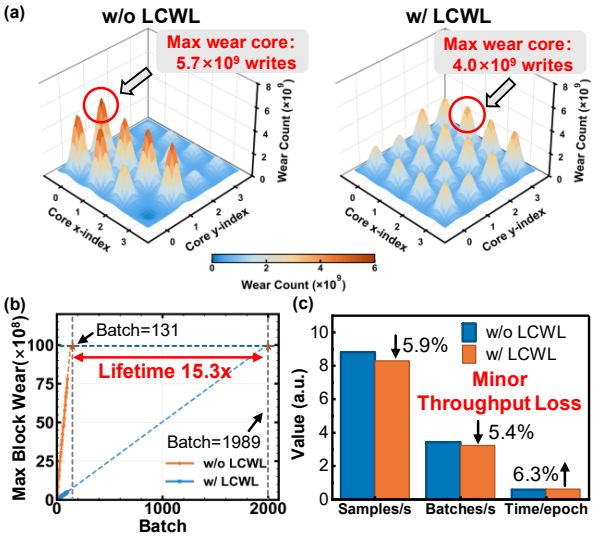

Figure. 5: Experimental validation of LCWL on ResNet-18/CIFAR-100 deployed across 16 cores. (a) Per-core cumulative wear count distribution. (b) Block-level lifetime comparison. (c) Throughput overhead comparison.

Fig. 5 quantitatively validates LCWL on a ResNet-18 network trained on the CIFAR-100 dataset deployed across 16 cores. **Fig. 5(a)** compares the cumulative write count per core with and without LCWL. Without LCWL, write counts concentrate in specific cores with a 57× peak-to-minimum imbalance, whereas LCWL produces a substantially uniform distribution with the peak reduced from 5.7 × 10^9 to 4.0 × 10^9. **Fig. 5(b)** plots the maximum block wear as a function of training batches. Without LCWL, the most heavily worn block reaches the endurance limit of 1 × 10^{10} at batch 131, whereas LCWL extends this to batch 1989, yielding a 15.3× lifetime

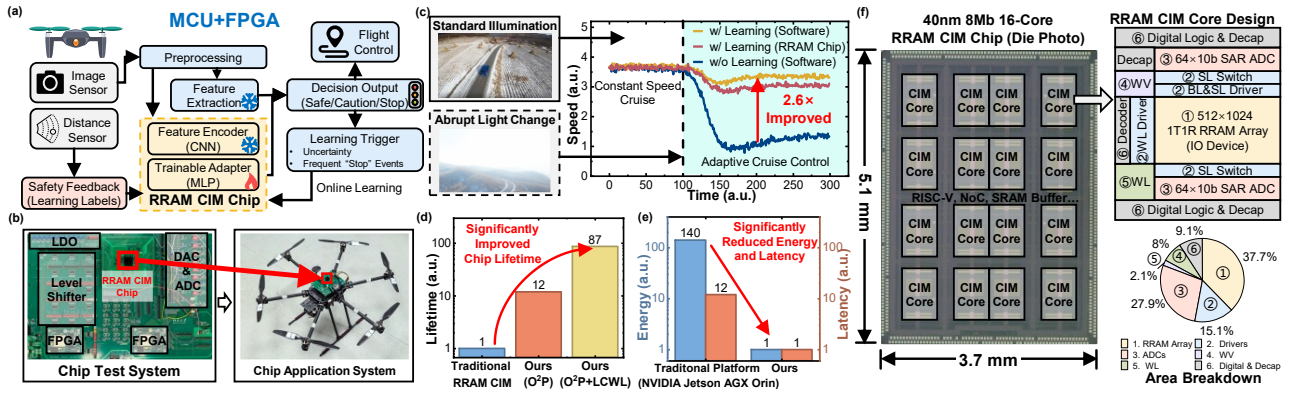

Figure. 6: (a) Aerial embodied learning system architecture. (b) Chip test system and drone platform. (c) Adaptive cruise control performance comparison. (d) Chip lifetime comparison. (e) Energy and latency comparison with NVIDIA Jetson AGX Orin. (f) Die photo of the chip, with core-level design details and a comprehensive area breakdown.

improvement at the block level. **Fig. 5(c)** shows that LCWL incurs only a 5.9% reduction in samples per second, 5.4% reduction in batches per second, and 6.3% increase in wall-clock time per epoch.

III. System Validation on Aerial Embodied Learning

To validate the proposed techniques in a real-world scenario, we deploy the RRAM CIM chip on an autonomous drone platform for adaptive cruise control, as shown in Fig. **6(a)(b)**. The system integrates an image sensor and a distance sensor, with a CNN feature encoder and a trainable MLP adapter mapped onto the RRAM chip. OCL is triggered by uncertainty or frequent stop events, enabling real-time adaptation to environmental changes such as abrupt illumination shifts. As shown in **Fig. 6(c)**, OCL achieves a 2.6× improvement in adaptive cruise control performance compared to the non-learning baseline. **Fig. 6(d)** demonstrates that O²P alone extends chip lifetime by 12× over traditional RRAM CIM, and the combination of O²P and LCWL further improves it to 87×. **Fig. 6(e)** shows that the RRAM CIM solution achieves 12× energy reduction and 100× latency reduction compared to an NVIDIA Jetson AGX Orin platform. Our proposed chip architecture and its internal functional area breakdown in **Fig. 6(f)**.

TABLE I. Comparison results with prior published works

	Ours	ESSERC 2025 [8]	ESSERC 2025 [9]	VLSI 2025 [5]	VLSI 2025 [10]	IEDM 2025 [11]
Technology	40nm	40nm	40nm	180nm	12nm	40nm
Memory type	RRAM	RRAM	RRAM	RRAM	RRAM	RRAM
Implementation Level	Processor	Macro	Macro	Macro	Macro	Macro
Array/Core #	16	4	1	1	1	1
Capacity	8Mb	1Mb	64Kb	2Kb	8Mb	1Mb
Wear-Resilient	✓	×	×	×	×	×
Low-Wear Programming	✓ (On-Chip)	×	×	×	×	×
CIM-Aware Wear Leveling	✓ (On-Chip)	×	×	×	×	×
Weight Precision	8-bit/ Analog[1]	Ternary[1]	2-bit	3-bit	2-bit	4-bit
Peak MVM Throughput (TOPS)	32.6 (INT8)	0.842[2] (INT8)	\	\	\	\
Application	Aerial Embodied Learning	AI Inference	Vector-Symbolic Architecture	Omniglot Classification	MNIST Classification	Secure Point Cloud Processing

1. Two RRAM devices represent one cell; 2. Normalized to 8-bit.

IV. Conclusion

This paper presents a 40-nm, 8-Mb, 16-core RRAM CIM processor enabling efficient on-chip learning for aerial embodied intelligence. To overcome training-induced endurance bottlenecks, we propose a device-system co-design. O²P mitigates stochastic write stress, while LCWL resolves wear imbalances without stalling throughput. Synergistically, they extend overall chip lifetime by 87×. Validated on a drone performing adaptive cruise control, the chip achieves 12× energy reduction and 140× latency speedup over an NVIDIA Jetson AGX Orin, establishing a resilient hardware foundation for next-generation self-adaptive 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.

References

[1] S. J. Chung et al., "A survey on aerial swarm robotics," IEEE Trans. Robot., vol. 34, no. 4, pp. 837–855, Aug. 2018.

[2] J. V. Anchitaalagammai et al., "Edge artificial intelligence for real-time decision making using NVIDIA Jetson Orin, Google Coral Edge TPU and 6G for privacy and scalability," ICVADV, 2025, pp. 150–155.

[3] M. Bala et al., "Democratizing drone autonomy via edge computing," SEC, 2023, pp. 40–52.

[4] W. Wan et al., "A compute-in-memory chip based on resistive random-access memory," Nature, vol. 608, pp. 504–512, 2022.

[5] M. Pallo et al., "On chip customized learning on resistive memory technology for secure edge AI," VLSI, 2025, pp. 1–3.

[6] D. Ielmini and G. Pedretti, "Resistive switching random-access memory (RRAM): applications and requirements for memory and computing," Chem. Rev., vol. 125, no. 12, pp. 5584–5625, 2025.

[7] W. Wen et al., "Wear leveling for crossbar resistive memory," DAC, 2018, pp. 1–6.

[8] 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," ESSERC, 2025, pp. 473–476.

[9] W. Yue et al., "A 40nm 48.7F²/bit 1T2R resistive memory bitcell array for adaptive vector-symbolic in-memory computing," ESSERC, 2025, pp. 241–244.

[10] C. Y. Tsai et al., "A CMOS-compatible 12nm 8Mb MLC RRAM enabling producible 2-bit per cell for high energy efficiency compute-in-memory in edge AI applications," VLSI, 2025, pp. 1–3.

[11] Y. Gao et al., "First Demonstration of In-MVM XOR-Decryption based on 40nm RRAM Chip for Secure Point Cloud Processing," IEDM, 2025, pp. 1-4.