SCIENCE CHINA
Information Sciences

• LETTER •

November 2026, Vol. 69, Iss. 11, 219402:1–219402:2
https://doi.org/10.1007/s11432-025-4864-y

A 28 nm compute-in-memory SRAM macro with adaptive quantization and replica-column-assisted computation

Zhiting LIN[1,2†], Miao LONG[1†], Yang YANG[1], Fugui JIANG[1], Juntao GE[1], Chenglong DUAN[1], Jianping XIA[1] & Xiulong WU[1,2*]

[1]*School of Integrated Circuits, Anhui University, Hefei 230601, China*
[2]*Anhui Provincial High-Performance Integrated Circuit Engineering Research Center, Hefei 230601, China*

Received 29 September 2025/Revised 27 December 2025/Accepted 2 March 2026/Published online 9 November 2026

Citation Lin Z T, Long M, Yang Y, et al. A 28 nm compute-in-memory SRAM macro with adaptive quantization and replica-column-assisted computation. Sci China Inf Sci, 2026, 69(11): 219402, https://doi.org/10.1007/s11432-025-4864-y

The present study proposes a highly efficient, linear, and robust static random-access memory (SRAM) analog computing-in-memory (CIM) macro designed to address critical bottlenecks in edge artificial intelligence (AI) applications. This macro introduces three main innovations. (1) A shared SRAM cell architecture for computation and quantization, utilizing the same discharge path for both computation and reference voltages in quantization, effectively canceling out non-linearities and variations caused by process, voltage, and temperature (PVT). (2) An adaptive quantization circuit (AQC) that replaces area-hungry analog-to-digital converters (ADCs) by utilizing a two-stage (coarse and fine) capacitance comparison approach, dramatically reducing area overhead. (3) A replica-column-assisted circuit (RC-AC) paired with an adjustable-step circuit (ASC), which dynamically modulates the wordline pulse width to adjust the least significant bit (LSB) voltage adaptively. Experimental results demonstrate that the proposed macro, fabricated in a 28 nm process, effectively suppresses variations and non-linearity. At a supply voltage of 0.85 V, it achieves a peak energy efficiency of 67.1 TOPS/W (normalized to 4-bit precision), while the quantization circuit accounts for merely 3.7% of the total macro area.

The von Neumann architecture, characterized by physically separated memory and processing units, imposes severe performance and energy bottlenecks for data-intensive neural network computations [1]. Analog-domain SRAM CIM has emerged as a transformative solution, integrating multiply-and-accumulate (MAC) operations directly within the memory arrays. While analog CIM achieves higher energy efficiency and computational density than its digital counterparts, it faces severe analog characteristic challenges (see Appendix A). Firstly, simultaneous activation of multiple wordlines causes non-linear and mismatched bitline discharge, compromising computational linearity [2]. Secondly, converting analog MAC results into digital outputs typically requires power- and area-intensive ADCs, which often dominate the macro's total cost [3]. Lastly, conventional fixed LSB voltage encoding schemes suffer from a precision-reliability trade-off: small MAC results are heavily affected by quantization errors, while large MAC results risk excessive bitline discharge, leading to precharged node disturbance and "pseudo-write" data corruption [4]. To break these bottlenecks, we propose a 28 nm SRAM-based CIM macro that fundamentally enhances linearity, discharge consistency, and resilience to PVT variations.

SRAM-CIM architecture. The proposed SRAM-CIM macro, illustrated in Figure 1(a), supports accurate and energy-efficient multi-bit MAC operations. It integrates a 64 × 64 (4 kb) core array utilizing 4096 6T SRAM cells, alongside the RC-AC and AQC peripheral modules. The RC-AC module, dynamically interacting with the pulse width modulation circuit (PWMC), resolves the precision-reliability trade-off. As illustrated in Figure 1(b), the RC-AC contains a replica column where one access transistor of the 6T cell is permanently tied to ground (V_{SS}). This asymmetric hardening guarantees a reliable "0" retention to act as a robust reference for bitline discharge dynamics. Driven by external input controls, the RC-AC dynamically adjusts the pulse width of the main wordlines (MWL), terminating discharge rapidly for large MAC results to avoid pseudo-writes, while extending the pulse width for small MAC results to increase the LSB margin and minimize relative errors. Comprehensive principles are provided in Appendix B.

To address ADC overhead and non-linear discharge, the AQC module (Figure 1(c)) converts the accumulated analog voltage on the bitline (V_{SUM}) into a digital value through a compact capacitance comparison strategy. The key innovation is sharing the exact same replica row 6T cells and bitline capacitance. The quantization is executed in coarse and fine phases. During coarse quantization, dynamic reference voltages are generated internally on a local capacitor and compared against V_{SUM} using a shared sense amplifier (SA). This process determines the upper and lower bounds of the MAC result. Fine quantization subsequently resolves the exact value. Because the reference voltage experiences the exact same non-linear discharge curve and PVT-induced jitter as the computation bitline, error cancellation is inherently achieved. To further enhance throughput, the ASC adjusts the coarse quantization granularity based on input sparsity, significantly reduc-

* Corresponding author (email: xiulong@ahu.edu.cn)
† These authors contributed equally to this work.

© Science China Press 2026

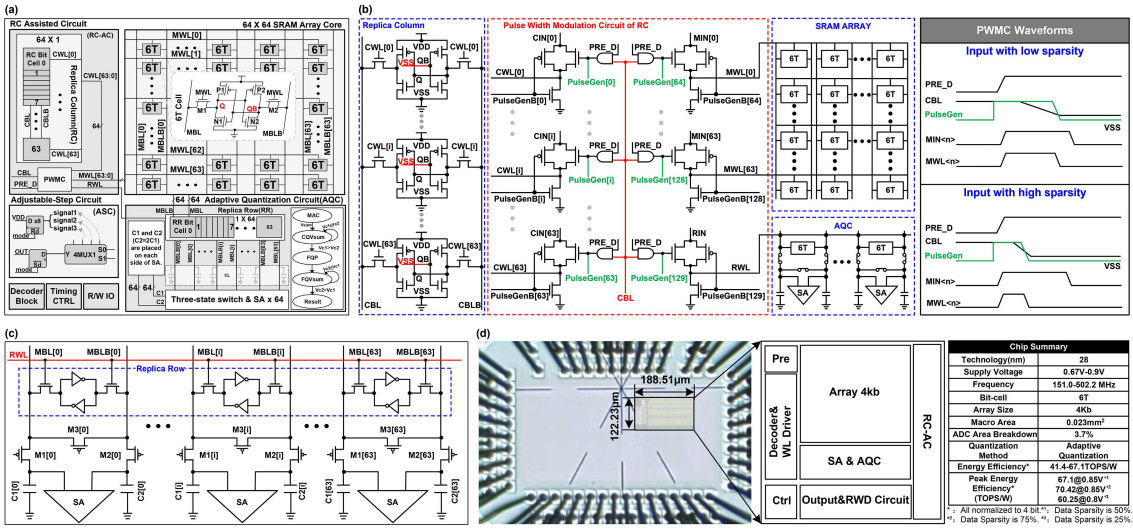

Figure 1 (Color online) (a) Structure of the proposed SRAM-CIM macro; (b) the structure of RC-AC PWMC and PWMC waveforms; (c) adaptive quantization circuit; (d) chip die photo and chip summary.

ing necessary comparison cycles. The detailed switching phases (PH1 to PH4) and corresponding equations are elaborated in Appendix C.

Experiment and result. The SRAM-CIM architecture has been rigorously validated and fabricated using the 28 nm CMOS (complementary metal-oxide-semiconductor) technology. The macro operates robustly across a wide supply voltage range of 0.67 to 0.9 V and a broad clock frequency range spanning from 151.0 to 502.2 MHz. Figure 1(d) presents the chip die micrograph and key performance summary. The total macro area is 0.0230 mm^2. Notably, the proposed AQC drastically minimizes the peripheral footprint, occupying only 3.7% of the total macro area—a massive reduction compared to traditional ADC-heavy architectures that typically consume over 15% to 20% of the silicon area. When normalized to 4-bit precision, the energy efficiency of the proposed macro achieves 41.4 to 67.1 TOPS/W, peaking at 0.85 V and 454.54 MHz. Extensive silicon measurements and PVT simulations demonstrate exceptional robustness; even under disparate temperature corners (0°C to 75°C) and severe process variations, the maximum voltage discrepancy between the computation bitline and the replica reference remains within a mere 3 mV. This minimal deviation is mathematically and physically insufficient to trigger erroneous SA flips, thereby guaranteeing high quantization pass rates across diverse operating conditions. Furthermore, system-level evaluations on standard datasets confirm its practical efficacy. When deploying ResNet20 and VGG8 neural network models, the proposed macro achieves inference accuracies of 87.38% (ResNet20) and 85.25% (VGG8) on the CIFAR-10 dataset, and up to 99.59% on the MNIST dataset, demonstrating negligible accuracy degradation compared to ideal software baselines. Overall, the proposed design achieves superior linearity and supports higher LSB voltages than state-of-the-art architectures [5].

Conclusion. This work systematically analyzes and successfully addresses the critical bottlenecks of analog SRAM-based CIM systems—namely, massive ADC area overhead, non-linear bitline discharge, and the precision-reliability trade-off. We propose an ultra-compact and robust SRAM CIM macro featuring an AQC and an RC-AC. By creatively reusing the computational SRAM cells and bitline capacitance for reference voltage generation, the macro naturally tracks and cancels out both non-linear bitline discharge characteristics and PVT-induced array inconsistencies. Furthermore, the dynamic pulse width modulation seamlessly eliminates "pseudo-write" data corruption risks for large MAC results while simultaneously optimizing LSB precision for small MAC results. Achieving a peak energy efficiency of 67.1 TOPS/W and compressing the entire quantization circuitry to a mere 3.7% of the macro area, this design vastly outperforms conventional ADC-reliant solutions in both density and linearity. Ultimately, the proposed macro delivers a highly scalable, accurate, and energy-efficient hardware foundation, unlocking new potentials for advanced, resource-constrained edge AI applications, including the Internet of Things, autonomous multisensor platforms, and ultra-low-power machine learning inference.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 92464203, 62074001), Joint Funds of the National Natural Science Foundation of China (Grant No. U19A2074), University Synergy Innovation Program of Anhui Province (Grant No. GXXT-2023-013), Key Research and Development Program of Anhui Province (Grant No. 2022a05020044), and Science Fund for Distinguished Young Scholars of Anhui Province (Grant No. 2108085J35).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Wang D, Lin C-T, Chen G K, et al. DIMC: 2219TOPS/W 2569F2/b digital in-memory computing macro in 28 nm based on approximate arithmetic hardware. In: Proceedings of the 2022 IEEE International Solid-State Circuits Conference (ISSCC), 2022. 266–268
2 Chen Z, Jin Q, Wang J, et al. MC2-RAM: an in-8T-SRAM computing macro featuring multi-bit charge-domain computing and ADC-reduction weight encoding. In: Proceedings of the 2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2021. 1–6
3 Qiao X, Guo Q, Tang X, et al. A 16.38TOPS and 4.55POPS/W SRAM computing-in-memory macro for signed operands computation and batch normalization implementation. IEEE Trans Circuits Syst I, 2024, 71: 1706–1718
4 Lin Z, Zhan H, Chen Z, et al. Cascade current mirror to improve linearity and consistency in SRAM in-memory computing. IEEE J Solid-State Circuits, 2021, 56: 2550–2562
5 Lee K, Kim J, Park J. A 28-nm 50.1-TOPS/W P-8T SRAM compute-in-memory macro design with BL charge-sharing-based in-SRAM DAC/ADC operations. IEEE J Solid-State Circuits, 2024, 59: 1926–1937