SCIENCE CHINA
Information Sciences

• LETTER •

January 2026, Vol. 69, Iss. 1, 119406:1–119406:2
https://doi.org/10.1007/s11432-025-4615-6

An area/energy-efficient RRAM computing-in-memory macro with fully-charge-domain multi-bit computation

Zhi LI[1,2], Shengzhe YAN[1,2], Jun WANG[1,2], Zhuoyu DAI[1,2], Zeyu GUO[1,2], Zhaori CONG[1,2], Zhihang QIAN[1,2], Xiangqu FU[1,2], Xu ZHENG[1], Chunmeng DOU[1,2], Dashan SHANG[1,2*] & Jinshan YUE[1*]

[1]*Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*

Received 9 January 2025/Revised 19 May 2025/Accepted 29 September 2025/Published online 18 November 2025

Citation Li Z, Yan S Z, Wang J, et al. An area/energy-efficient RRAM computing-in-memory macro with fully-charge-domain multi-bit computation. Sci China Inf Sci, 2026, 69(1): 119406, https://doi.org/10.1007/s11432-025-4615-6

The present study proposes a compact, accurate, and efficient RRAM computing-in-memory (CIM) architecture that supports fully-charge-domain multi-bit computation with three main innovations. (1) A novel charge-domain RRAM readout and multiply-accumulate (MAC) circuit is proposed, demonstrating superior linearity and robustness against device variations. (2) A capacitive multi-bit accumulation module (MBA) has been developed to carry out 8-bit MAC operations. (3) A stacked reverse charge quantization readout (SRCQR) module that facilitates the accumulation of multiple MACs in the analog charge domain while ensuring high accuracy and reducing the analog-digital conversion to a single iteration has been designed. The experimental results demonstrate the efficacy of the proposed CIM macro in suppressing RRAM variations. The macro attains a 645 GOPS/mm^2 area efficiency of a 47.62 TOPS/W energy efficiency, exhibiting a mere 1.11% accuracy loss on the ImageNet dataset.

The utilization of CIM has been put forth as a potential solution to address the memory wall bottleneck that is prevalent in the conventional von Neumann architecture. Numerous studies have been conducted about CIM, encompassing both digital CIM [1] and analog CIM [2] approaches. The former achieves high computational accuracy but exhibits significant area and energy consumption, while the latter offers superior area/energy efficiency but experiences accuracy degradation due to device non-linearity and imprecise analog computation. In practice, the design of analog CIM architectures necessitates meticulous deliberation on the trade-off between computational accuracy and area/energy efficiency. Furthermore, the performance metrics are influenced by the computational domains (e.g., current-domain, voltage-domain, time-domain, and charge-domain) within the analog CIM (see Appendix A). To address the imprecise issues encountered in analog CIM and to further improve computational precision, together with area and energy efficiency, a proposal is hereby made for an RRAM-based 8-bit fully-charge-domain CIM architecture, named RQ-CIM. RQ-CIM can execute robust and expeditious 8-bit MAC operations with precision in the charge domain and enable efficient readout of the computed results.

RQ-CIM architecture. The proposed RQ-CIM is presented in Figure 1(a), which supports accurate and area/energy-efficient multi-bit MAC operations. It is composed of three primary components: the charge-domain grouped 1T1R RRAM bank, which facilitates readout and MAC operations; the capacitive MBA module, which processes the accumulation of multi-bit weights and input activations (IAs). Furthermore, the SRCQR module facilitates the accumulation of partial results derived from multiple CIM operations prior to the execution of the final ADC readout. The charge-domain RRAM bank contains eight RRAM groups, with each RRAM group storing 1 bit of the 8-bit weight data. Each group consists of 4×512 1T1R cells and shared peripheral circuits for charge-domain readout and MAC. One of the four rows in each group is activated to share the charge-domain peripheral circuits within each MAC phase.

Figure 1(b) illustrates the detailed structure of the charge-domain grouped 1T1R read and MAC circuit. Each column contains an MAC module consisting of a reference resistor (R_{REF}), which is connected to the SL, a local column capacitor, and a read and multiply module composed of a buffer and MUX. The resistance state of the RRAM (i.e., weight) is read out through a fine-tuned buffer based on the voltage division between the RRAM and the R_{REF}. The multiplication operation between weight and IA is implemented on MUX, and the accumulation operation is achieved by charge sharing across multiple column capacitors. A thorough examination of the circuit design considerations is provided in Appendix B.1. As illustrated in Figure 1(c), the capacitive MBA module consists of eight sample capacitors (C_S) and two pipelined hybrid C-2C-exponential capacitor networks [3]. The accumulation is achieved through three stages: (i) precharge; (ii) accumulation for 8-bit IA by successive charge sharing between C_S and the corresponding capacitor in the hybrid capacitor network; (iii) accumulation for 8-bit weight through charge sharing in the hybrid capacitor network. The SRCQR module has been proposed as a method to reduce ADC power, as shown in Figure 1(d). It comprises four stacked reverse-charge capacitor circuits, a comparator, a counter, and a 6-bit SAR ADC. The stacked capacitor

* Corresponding author (email: shangdashan@ime.ac.cn, yuejinshan@ime.ac.cn)

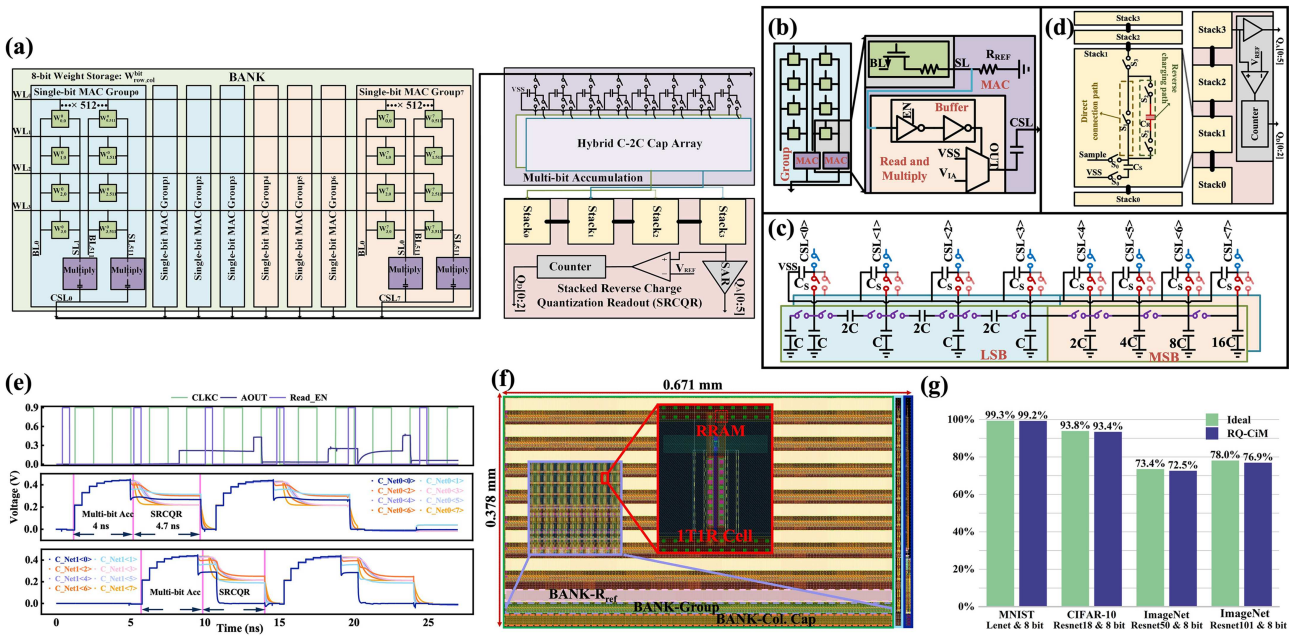

Figure 1 (Color online) (a) Overview of the proposed 8-bit charge-domain RRAM CIM architecture, RQ-CIM; (b) charge-domain grouped 1T1R read and MAC circuit; (c) MBA module; (d) SRCQR module; (e) simulated waveform of pipelined 8-bit MAC operation in the RQ-CIM architecture; (f) layout of the RQ-CIM architecture; (g) simulated inference accuracy of RQ-CIM.

charge (MAC results) is first coarsely quantized using comparators and counters, and subsequently finely quantized through a single ADC operation, which greatly reduces the overall ADC power consumption. Details of the operations and the principles of the MBA and the SRCQR are provided in Appendixes B.2 and B.3.

Experiment and result. The RQ-CIM architecture has been validated and evaluated using the SMIC 28 nm technology, with the LRS/HRS values set to 30/240 kΩ (10% variation) and a supply voltage of 0.9 V. Figure 1(e) illustrates the overall pipelined 8-bit MAC operation of the proposed RQ-CIM. The overall operation is completed in 25 ns, encompassing four 8-bit MAC computations and readouts, thereby attaining a throughput of 163.8 GOPS. The layout of the proposed RQ-CIM is shown in Figure 1(f). The total area of the proposed RQ-CIM (with a 16 Kb RRAM bank) is 0.254 mm^2. This achieves a memory density of 63 kb/mm^2 and an area efficiency of 645 $GOPS/mm^2$. The power consumption of the overall operation is 86 pJ across all computation and readout processes of the RQ-CIM, achieving an energy efficiency of 47.62 TOPS/W. This represents a 6.3× and 2.7× gain in energy efficiency compared to C-RRAM [4] and SAMBA [3], respectively. In addition, the RQ-CIM is implemented on the simulated framework proposed in [5] to verify the inference accuracy, as shown in Figure 1(g). As the scale of neural network models and the complexity of datasets increase, RQ-CIM shows a moderate rise in inference accuracy loss. Notwithstanding, the loss remains minimal, with only a 1.11% degradation observed on the ImageNet classification task using the 8-bit ResNet101 model. Appendix C contains detailed results of the experiment, as well as a comparison with other studies.

Conclusion. This work proposes a novel RRAM CIM architecture, designated as RQ-CIM, with the objective of achieving high area/energy efficiency. The RQ-CIM supports fully-charge-domain multi-bit accumulation operations. It enhances the CIM parallelism while minimizing energy and area consumption, and performs intermediate charge-domain accumulation, thereby eliminating the need for power-hungry and area-intensive digital adder trees, and consequently reducing the number of ADC conversions. The total energy consumption is 86 pJ, with a throughput of 163.8 GOPS and an area of 0.254 mm^2. It achieves 645 $GOPS/mm^2$ and 47.62 TOPS/W area and energy efficiency, respectively, exhibiting a mere 1.11% accuracy loss on the ImageNet dataset.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 92464203, 62204256, 62488101, 92364202).

Supporting information Appendixes A–C. 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 Si X, Tu Y N, Huang W H, et al. 15.5 A 28 nm 64 Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2020. 246–248
2 Wen T H, Hsu H H, Khwa W S, et al. 34.8 A 22 nm 16 Mb floating-point ReRAM compute-in-memory macro with 31.2 TFLOPS/W for AI edge devices. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2024. 67: 580–582
3 Chen Y, Yin G, Zhou M, et al. SAMBA: single-ADC multi-bit accumulation compute-in-memory using nonlinearity-compensated fully parallel analog adder tree. IEEE Trans Circuits Syst I, 2023, 70: 2762–2773
4 He Y, Huang Y, Yue J, et al. C-RRAM: a fully input parallel charge-domain RRAM-based computing-in-memory design with high tolerance for RRAM variations. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2022. 3279–3283
5 Wang Z, Yue J, He C, et al. A user-friendly fast and accurate simulation framework for non-ideal factors in computing-in-memory architecture. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2023. 1–5