A 28-nm 808.68 TOPS/W Resilient Charge-Domain RRAM CIM Macro with Integrated Power Supply Using Inter-Block Capacitive Reusing Scheme for Edge AI

Linfang Wang[1,3] Zhidao Zhou[1,2], Hongyang Hu[1,2], Zhi Li[1,2], Weizeng Li[1,2], Wang Ye[1,2], Jinshan Yue[1,2], Xiaoxin Xu[1,2], Jianguo Yang[1,2], Xiaoming Chen[4], Zhenghao Lu[5], Ziou Wang[5], Jing Liu[6], Chunmeng Dou[1,2]
Email: douchunmeng@ime.ac.cn
[1]Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Department of Electrical Engineering, Columbia University, New York, NY, USA
[4]Institute of Computing Technology of the Chinese Academy of Sciences, Beijing, China
[5]School of Electronic and Information Engineering, Soochow University, Jiangsu, China
[6]Beijing Memsilicon Techology Co. Ltd., Beijing, China

***Abstract*—Resistive random-access memory (RRAM)-based computing-in-memory (CIM) holds promise for meeting the growing performance demands of edge AI applications. However, achieving high performance and resilient computing with low overhead peripheral circuits remains challenging. This work presents a high efficiency and reliable RRAM based CIM macro featuring: (1) a foundry-friendly gain-cell-integrated (GCI)-2T2R array that performs resilient multiply-accumulate (MAC) over 10k program/erase (P/E) cycling and -20~100°C; (2) inter-block capacitive reuse between the ADC and charge pump, reducing area by 58.2%; and (3) an all-element-sparsity-aware ADC with fine-grained sparsity prediction based on input, weight, and output, achieving 52.6% lower readout power per task. We fabricated a 28-nm 256k cells GCI-2T2R RRAM CIM macro supporting 1~8-bit MAC operation with up to 22-bit output. It delivers a peak normalized energy efficiency of 808.68 TOPS/W and 1.25-496× higher area FoM than prior arts. This design demonstrates a variation- and drift-tolerant (VAD-tolerant) charge-domain RRAM CIM macro with small periphery-overhead (⩽ 44%).**

Keywords*—Computing-in-memory (CIM), resistive random access memory (RRAM), variation- and drift-tolerant (VAD-tolerant) design, charge domain computing

I. Introduction

Resistive random-access memory (RRAM)-based nonvolatile computing-in-memory (nvCIM) chips provide a competent solution for energy-efficient edge AI inference[1-8]. Compares to SRAM base CIM, nvCIM suits for storage-intensive AI inference with large memory capacity under constrained footprint. However, existing RRAM CIM designs [1-8] have mainly focused on optimizing energy efficiency and throughput, yet the overhead of peripheral circuits are rarely discussed. In particular, RRAM and other nonvolatile memory programming requires high voltages, yet few works integrate power management circuits. They usually rely on off-chip power supply, which is impractical for edge deployment. Furthermore, achieving high computing precision in RRAM CIM demands variation-and-drift (VAD)-tolerance. Write-and-verify scheme consumes a significant amount of time and energy.

To address these gaps, this work presents a reliable, energy- and area-efficient RRAM CIM macro featuring (1) a gain-cell-integrated (GCI)-2T2R array that performs reliable charge-domain computing over 10k write cycles and can operate at temperatures up to 100°C, (2) an inter-block capacitors reusing scheme between the analog-to-digital converter (ADC) and the charge-pump (CP), resulting in a 58.2% reduction in area, (3) an all-element-sparsity-aware (AESA)-ADC with fine-grained sparsity prediction, reducing readout power consumption by 52.6% per task. The 28-nm test-chip shows it achieve a peak normalized energy efficiency of 808.68 TOPS/W and 1.25-496× higher area FoM than prior arts.

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 92364202, 62488101, and U2441247.

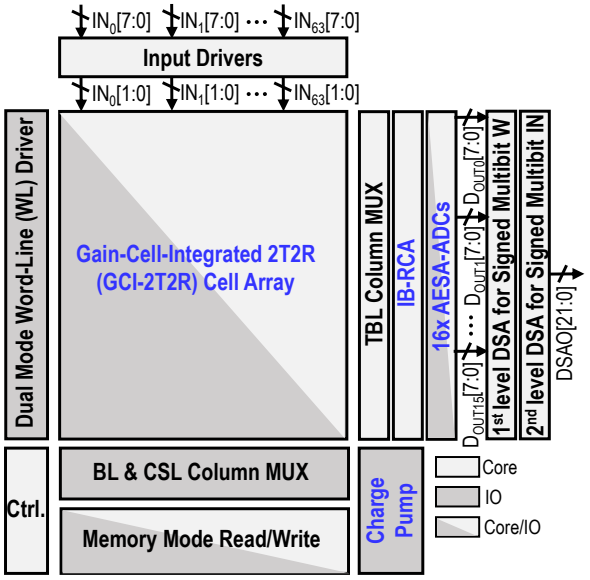


Fig. 1. Structure of the proposed 28-nm 256k cells GCI-2T2R RRAM CIM macro.

II. Proposed CIM Macro Design

A. Overall Architecture

The overall structure of the proposed GCI-2T2R RRAM CIM macro is shown in Fig. 1. The macro supports 1-8-bit MAC operations with up to 22-bit output (O). It consists of a CGI-2T2R array with 256k RRAM cells, drivers for 64 input channels (IN_0-IN_{63}), 16 8-bit AESA-ADCs for data outputs (D_{OUT0}-D_{OUT15}), a CP for high voltage generation during RRAM writing, an inter-block reusable capacitive array (IB-RCA), a two-level digital shifters-and-adders (DSAs) for combining signed multi-bit weights (W) and inputs (IN), a memory and computation dual-mode word-line (WL) driver, multiplexers (MUXs) for bit-line (BL), complementary bit-line (BLB), and common source-line (CSL), a timing and mode control (Ctrl), and other typical RRAM memory peripherals for RRAM read and write.

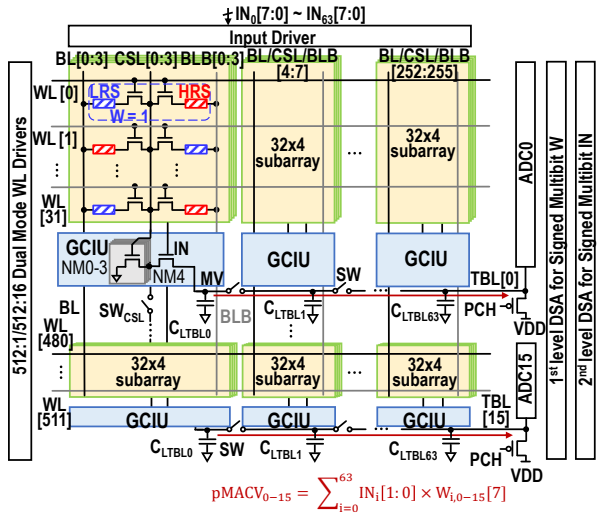


Fig. 2. Proposed GCI-2T2R array comprising 2T2R subarrays and GCIUs.

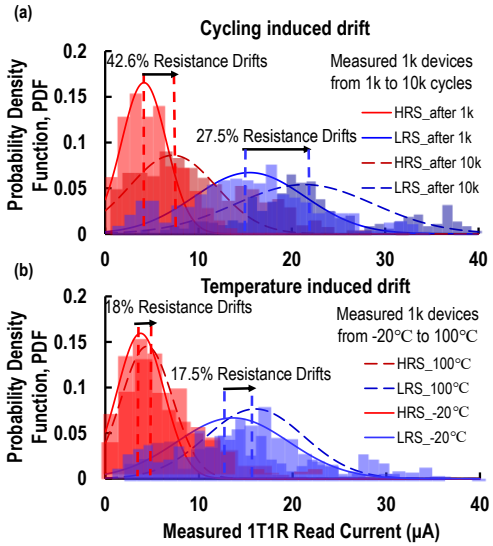


Fig. 3 Measured resistance-drift effects of 1kb RRAM cells due to (a) cycling and (b) temperature.

B. *Variation-and-Drift (VAD)-Tolerant GCI-2T2R Array*

Fig. 2 depicts the structure of proposed GCI-2T2R array, which is compatible with the foundry high-density 2T2R arrays with a common source-line (CSL). The array consists of 16×64 GCI-2T2R blocks, each containing a GCI unit (GCIU) shared across 32×4 2T2R subarray to reduce the area overhead. The transistors NM0-NM3 amplify the signal of CSL using transient discharging, determined by the voltage division of the selected one 2T2R differential pair between bit-lines (BL and BLB) in a subarray. The NM4 transistors is gated by the IN signal represented by pulse widths after the input driver. The GCI-2T2R array performs MAC operation by discharging local parasitic capacitors, followed by row-wise global charge sharing. A typical MAC cycle consists of three phases (PH0-2). During PH0, all the local transposed BLs (LTBLs) are firstly connected into a global TBL (SW = 1), and parasitic capacitors of LTBLs (C_{LTBLS}) are pre-charged to VDD (PCH = 0). In PH1, TBL is segmented (SW = 0), and the word-line (WL) and INs are activated to determine whether to discharge C_{LTBLS} or not, with multiply values stored on the MV nodes. In PH2, global charge-sharing across all C_{LTBLS} in GCIUs is activated (SW = 1) to complete the MAC operation.

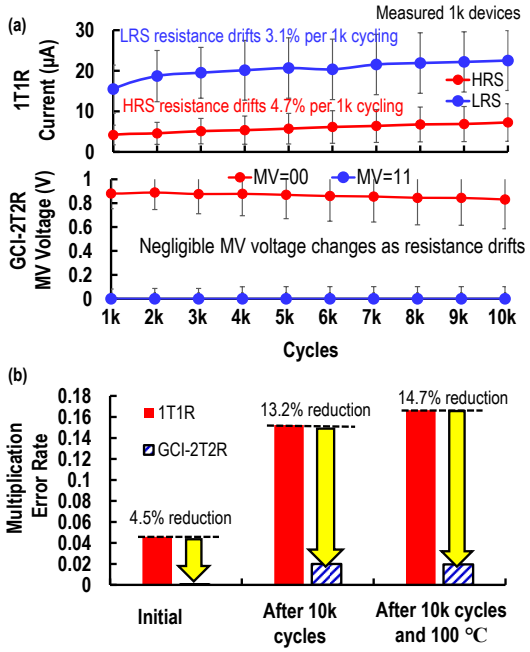


Fig. 4 (a) Measured 1T1R read current and GCI-2T2R MV voltage distributions over multiple set and reset cycles, and (b) Measured multiplication error rates at initial state, 10k cycling, as well as 100 ℃.

The accuracy of MAC operation is sensitive to RRAM VAD. Fig. 3 illustrates the measured readout currents of 1-kb cells under different conditions, which indicate RRAM resistance relaxation of ⩾27.5% after 10k set and reset cycles and ⩾17.5% after temperature changes from -20 to 100 ℃. Because the differential voltage-division processing in GCI-2T2R array can cancel out the common-mode resistance variation [3, 8, 9], as well as the gain cell further amplifies the signal, the MV voltage in GCIU can remain a stable rail-to-rail signal margin in local multiplication phase (Fig. 4(a)). Fig. 4(b) further shows that compared to traditional 1T1R array, GCI-2T2R array reduces the multiplication error rates by 14.7% after 10k set and reset cycles and at 100 ℃.

C. *Proposed IB-RCA and AESA-ADC for Reduced Hardware Overhead*

To reduce the periphery area overhead in RRAM-based CIM, we propose the IB-RCA scheme that reusing multiple configurable MOM unit capacitors to reduce the capacitive components in the ADC and CP, which typically consume significant area. Fig. 5(a) illustrates the structure of IB-RCAs, which can serve as capacitive digital-to-analog converters (DACs) for successive approximation register (SAR)-ADCs in computation mode, and form into two groups of large flying capacitors (UG and BG) for CP in write mode. The top-plates of IB-RCA connect to ADC comparators or CP branches, while the bottom plates are controlled by the SW_{RCA} signal depending on working mode.

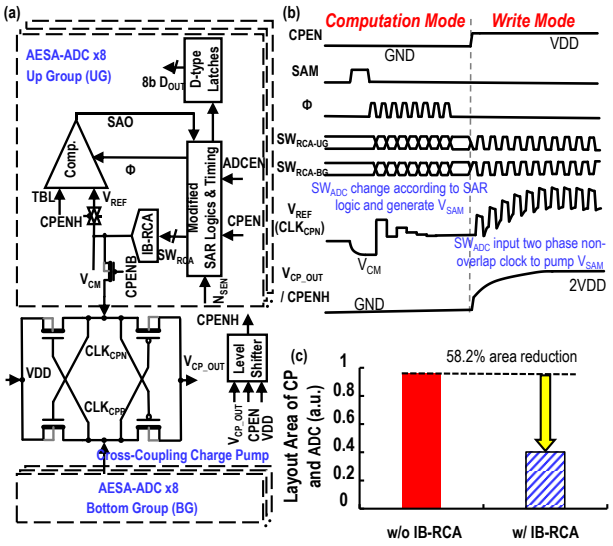


Fig 5 (a) Proposed schematic for the proposed IB-RCA and CP. The capacitive DAC arrays in ADC can be re-grouped to form the flying capacitors in CP. (b) Operational waveforms of ADC and CP at the computation mode and the memory mode. (c) Comparison on the area costs of ADCs and CP without and with the proposed IB-RCA scheme

Fig. 5(b) also depicts the operational waveforms of IB-RCA. In CIM mode, CP is disabled (CPEN = 0). After sampling the MAC result in TBL, the comparator is triggered (Φ = 1) to generate 1-bit output (SAO), and the reference voltage (V_{REF}) successively approximate the MAC result according to the SW_{RCA} controlled by SAR logic. In write mode, CP is enabled (CPEN = 1). The SW_{RCA} nodes of two IB-RCA groups receive opposite pumping clocks (SW_{RCA-UG}/SW_{RCA-BG}), causing the CP output voltage (V_{CP-OUT}) to ramp up. Notice that an additional high voltage transmission gate, which can be turned off by V_{CP-OUT} itself, is used to isolate the high voltage generated by CP from core devices in the ADC comparators. The IB-RCA reduces the total layout area of ADC and CP by 58.2% (Fig. 5(c)).

In order to further reduce the ADC power consumption, we propose an AESA control logic as shown in Fig .6(a). By counting the number of activated INs (N_{IN}) and using a priori knowledge of the weight sparsity factor (N_{WS}), which can be orthogonally superimposed on N_{IN} and learned by training, the AESA-ADC can adaptively adjust the number of sensing steps (N_{SEN} = N_{IN} - N_{WS}). Measured results show a 63.2% reduction in ADC power consumption from N_{SEN} = 8 to N_{SEN} = 1 (Fig. 6(b)). Moreover, we propose a fine-grained output sparsity prediction (FG-OSP) scheme for zero values caused by activation functions like ReLU. In a given cycle of multiple bit-serial input cycles, the theoretical maximum input data (IN_{TMAX}) is calculated based on the previously counted N_{IN}. The maximum MAC value prediction ($MACV_{PRD}$) of an output channel is obtained by adding the IN_{TMAX} to the cumulative partial MAC values (pMACVs) in the current cycle. If the $MACV_{PRD}$ < 0, meaning the final MACV of this channel is negative and will set to zero by ReLU, the AESA-ADC is disabled (ADCEN = 0), and the computation for the current output channel is terminated. The FG-OSP and AESA schemes outperform the previous coarse-grained output sparsity prediction method in [6], achieving 29% and 52.6% reductions in readout power per

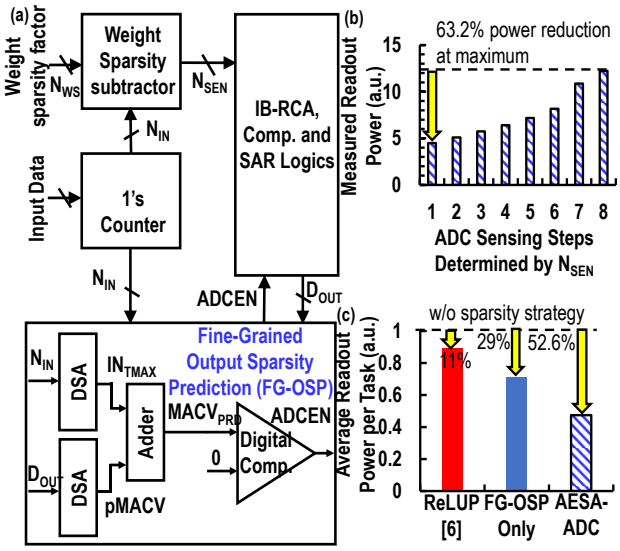
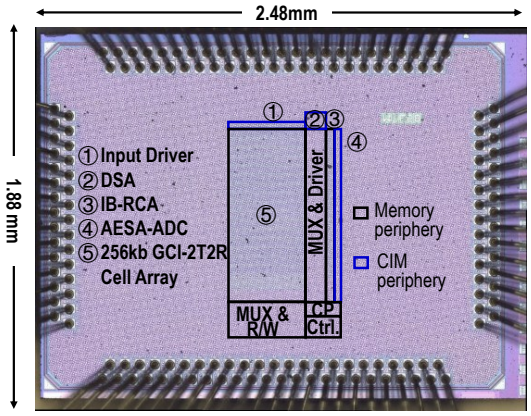


Fig. 6 (a)Proposed AESA-ADC structure, (b) Measured readout power. (c) Evaluation of average power per inference task (ResNet-18 on CIFAR-10).



Technology	28nm CMOS
Memory	SLC RRAM
Cell Type	GCI-2T2R
Macro Area (mm^2)	0.663
Resolution	1-8b-IN 1-8b-W 22b-O
Norm. TP (GOPS) *1	655.36
Norm. EF (TOPS/W) *2	808.68
Periphery Overhead *3	43.9%

*1. Measured at 160MHz, VDD = 1.4V.
*2. Measured at 1MHz, VDD = 0.6V, 90% sparse.
*3. Periphery Overhead = (Memory + CIM) periphery area / total area.

Fig. 7 Die micrograph and chip summary.

inference task, respectively, when tested on an INT-8 ResNet-18 model trained on CIFAR-10 dataset (Fig. 6(c)).

III. Measurement Results

The proposed macro is fabricated using 28-nm HKMG process with BEOL RRAM (Fig. 7). Fig. 8 shows the macro-level measurement results. The waveform of CP shows it achieves a pumping efficiency of 78% (Fig. 8(a)). The measured transfer curve, obtained by fixing the INs while sweeping different weight patterns stored in different row of

TABLE I. COMPARISON WITH PREVIOUS RRAM CIM MACROS

	ISSCC'23 [6]	VLSI'23 [5]	JSSC'24 [4]	ISSCC'25 [3]	JSSC'25 [2]	JSSC'25 [1]	This Work
Technology	22-nm	40-nm	65-nm	28-nm	28-nm	65-nm	**28-nm**
CIM Bit-Cell	1T1R	1T1R	1T1R	2T2M	1T1R	1T1R	**GCI-2T2R**
Capacity	4 MB	64 kb	4 kb	8 kb	144 kB	64 kb	**128 kb**
Resistance Drift Resilient	N.A.	Yes (ECC)	N.A.	N.A.	N.A.	N.A.	**Yes (After 10k cycling and -20-100℃)**
Computing Method	Current Summation	Current Summation	Digital Adder Tree	Current Summation	Digital Adder Tree	Current Summation	**Parasitic Capacitor Charge Sharing**
On chip CP	N.A.	No	No	N.A.	No	No	**Yes**
Inter-Blocks Circuit reusing	No	No	No	No	No	No	**Yes**
MAC precision (IN/W/O)	1-8b/ 1-8b/24b	1b/1b/6b	1b/1b/6b	1b/1b/4b	4-8b/ 4-8b/24b	8b/4b/8b	**1-8b/1-8b/22b**
ADC Precision	5b	6b	N.A.	4b	N.A.	8b	**1-8b**
Sparsity aware	Output (11% ↓)	No	No	No	No	No	**AESA w/ FG-OSP (52.6% ↓)**
Normalized EF (TOPS/W)[*1]	4601.6	350	2.07	59.8	1220	662.4	**808.68**[*2]
Output Ratio[*3]	71.4%	75%	100%	57.1%	100%	40%	**100%**
Periphery Overhead[*4]	N.A.	76%	62.2%	79%[*5]	63.5%	79.9%	**43.9% (18.3%-36% ↓)**
Energy FoM[*6]	3285.5	262.5	2.07	34.15	1220	264.96	**808.68**
Area FoM[*7]	N.A.	0.73[*5]	0.003[*5]	0.16	1.19	0.92[*5]	**1.49 (1.25x-496x ↑)**

*1. Normalized to 1-bit MAC. *2. Measured at 1MHz, VDD = 0.6V, 90% input sparsity.
*3. Output Ratio = Actual output precision / Full output precision. *4. Periphery Overhead = periphery circuits area / macro area.
*5. Estimated. *6. Energy FoM = Norm. Energy Efficiency * Output Ratio.
*7. Area FoM = Norm. Throughput * Output Ratio / Periphery Overhead.

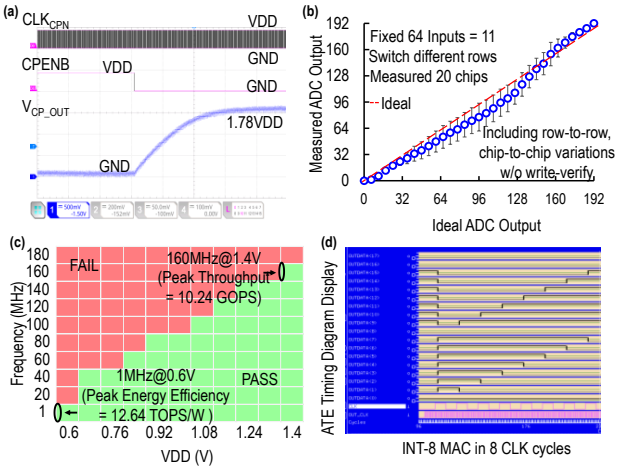


Fig. 8 (a) Measured IB-RCA-based CP waveforms. (a) Measured CIM transfer curves over 20 chips. (c) Measured shmoo plot. (d) Measured waveforms for MAC with 8b IN, 8b W, and 22b O.

20 chips (including row-to-row and chip-to-chip variations without any write verifications), demonstrates a highly linear relationship between the ideal results and measured ADC outputs (Fig. 8(b)). A shmoo plot indicates that the test-chip can perform INT-8 MAC operations at working frequency up to 160 MHz and achieve a peak throughput of 655.36 GOPS normalized to 1-bit MAC. The test-chip also achieves a peak normalized energy efficiency of 808.68 TOPS/W at 0.6 V ((Fig. 8(c, d))). Table I summarizes the chip performance and compares it with recent state-of-the-art nvCIM designs. This work achieves a peak normalized energy efficiency of 808.68 TOPS/W and delivers 1.25-496× improved area FoM with only 43.9% periphery overhead compared to previous arts.

REFERENCES

[1] J. M. Correll et al., "An 8-bit 20.7 TOPS/W Multilevel Cell ReRAM Macro With ADC-Assisted Bit-Serial Processing," in IEEE Journal of Solid-State Circuits, vol. 60, no. 8, pp. 2995-3008, Aug. 2025.

[2] C. Mu, Z. Huang et al., "A 28-nm RRAM/SRAM Collaborative CIM Accelerator Supporting RRAM-Endurance-Latency Awareness for Edge Fine-Tuning," in IEEE Journal of Solid-State Circuits, vol. 60, no. 10, pp. 3626-3638, Oct. 2025.

[3] S. Kwon et al., "37.3 Monolithic in-Memory Computing Microprocessor for End-to-End DNN Inferencing in MRAM-Embedded 28nm CMOS Technology with 1.1Mb Weight Storage,"in 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2025, pp. 1-3.

[4] F. Zhang et al., "A 65-nm RRAM Compute-in-Memory Macro for Genome Processing," in IEEE Journal of Solid-State Circuits, vol. 59, no. 7, pp. 2093-2104, July 2024.

[5] S. D. Spetalnick et al., "A 2.38 MCells/mm2 9.81 -350 TOPS/W RRAM Compute-in-Memory Macro in 40nm CMOS with Hybrid Offset/IOFF Cancellation and ICELL RBLSL Drop Mitigation,"in 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 2023, pp. 1-2.

[6] W. -H. Huang et al., "A Nonvolatile AI-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W," in 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 15-17.

[7] D. -Q. You et al., "14.1 A 22nm 104.5TOPS/W μ-NMC-Δ-IMC Heterogeneous STT-MRAM CIM Macro for Noise-Tolerant Bayesian Neural Networks," in 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2025, pp. 1-3

[8] L. Wang et al., "Efficient and Robust Nonvolatile Computing-In-Memory Based on Voltage Division in 2T2R RRAM With Input-Dependent Sensing Control," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 5, pp. 1640-1644, May 2021.

[9] K. Zhang, "1.1 Semiconductor Industry: Present & Future," in 2024 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2024, pp. 10-15.