A 1308 TOPS/W Charge-Mode ReRAM CIM Macro with 4T2R2C Differential Cell and FIA-Based Analog Accumulation for AI Inference

Longhao Yan[1#], Zelun Pan[1#], Zhe Zhan[1], Daijing Shi[1], Yihang Zhu[1], Yaoyu Tao[1,3], and Yuchao Yang[1,2,3,4*]
[1]Beijing Advanced Innovation Center for Integrated Circuits, School of Integrated Circuits, Peking University, Beijing, China.
[2]Center for Brain Inspired Intelligence, Chinese Institute for Brain Research (CIBR), Beijing, China.
[3]Center for Brain Inspired Chips, Institute for Artificial Intelligence, Peking University, Beijing, China.
[4]Guangdong Provincial Key Laboratory of In-Memory Computing Chips, School of Electronic and Computer Engineering, Peking University, Shenzhen, China. [#]Equally contributed authors. *Corresponding Author's Email: yuchaoyang@pku.edu.cn

***Abstract*—ReRAM-based Compute-in-Memory (CIM) chips have shown great potential in the field of AI inference. However, traditional ReRAM CIMs use current-mode or voltage-mode computation methods. On one hand, the low on/off ratio of devices limits the enhancement of array row parallelism. On the other hand, significant direct currents within arrays and peripheral computation circuits lead to substantial energy overhead. Furthermore, due to issues like IR-drop, these methods also exhibit computational nonlinearity problems. In this work, we fabricate a 40nm charge-mode compute-in-memory chip that employs a novel charge-mode ReRAM cell structure and the FIA-based charge-domain analog accumulation method. This design significantly enhances the on/off signal radio, reduces direct currents within arrays and improves computational linearity (0.9992 of R^2), achieving peak throughput and energy efficiency of 53.47 TOPS and 1308.24 TOS/W (normalized to 1b), respectively.**

Keywords*—ReRAM, differential cell, compute-in-memory, charge-mode, inference

I. Introduction

Traditional Resistive Random Access Memory (ReRAM)-based compute-in-memory technologies typically employ current-mode and voltage-mode computation methods [1-5], which face numerous challenges in throughput, linearity, and energy efficiency in **Fig. 1**. Firstly, the limited on/off ratio of the devices restricts the improvement of computational parallelism. Secondly, traditional computation methods are affected by issues like IR-drop, leading to non-linear computation results. Lastly, these methods exhibit large direct currents within arrays and require substantial analog and digital circuits outside arrays, resulting in considerable area and energy consumption overheads.

In this work, we fabricated a 40 nm charge-mode ReRAM-based compute-in-memory chip, introducing innovations in the charge-mode ReRAM cell structure, the charge-domain accumulation scheme and the analog-to-digital converter (ADC) circuits. Measurement results demonstrate that the chip achieves great linearity (0.9992 of R^2) under large row parallelism with a throughput and energy efficiency of 53.47 TOPS and 1308.24 TOS/W (normalized to 1b), respectively. Furthermore, we deploy components of Transformer networks on the chip for testing and validation, achieving metrics close to the software. The specific contributions are listed as follows:

We propose a 4T2R2C charge-mode differential cell structure that leverages the transconductance gain of transistors to significantly enhance the ReRAM device's on/off ratio, thereby supporting higher row-level parallelism for accurate Vector Matrix Multiplication (VMM) computation.

We propose a charge-domain fully-parallel accumulation method to reduce the power consumption of the large current induced by the multiple opened rows in one column, and to suppress the non-linearity problem caused by IR drop due to the non-negligible resistance of Bit Line (BL) and Source Line (SL).

We design a FIA-based [6] binary-weighted successive-approximation analog-to-digital converter (FBS-ADC) enabling efficient binary weighting of the computation results in analog

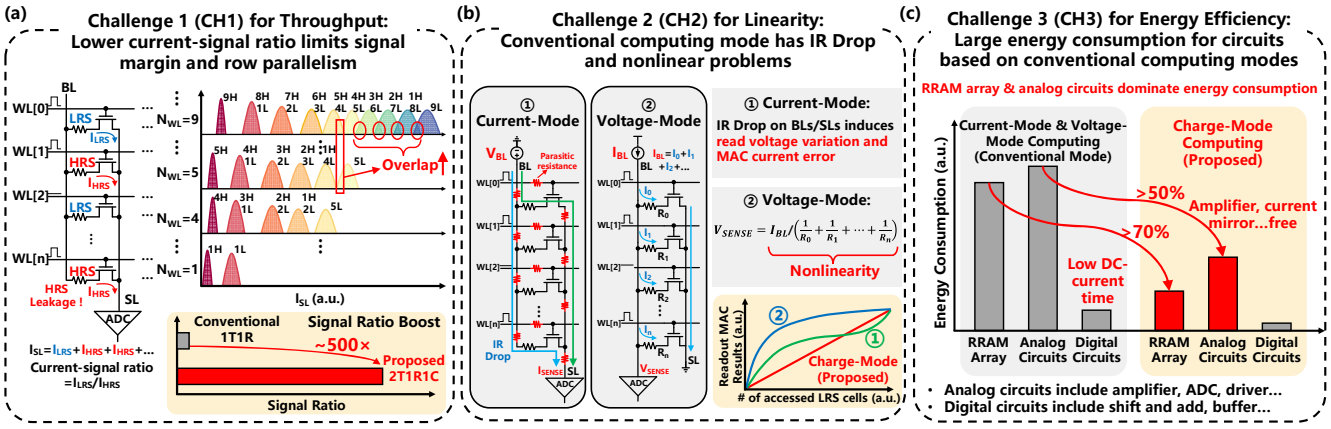


Figure 1: (a)(b)(c) Challenges faced by traditional ReRAM-based current-mode and voltage-mode compute-in-memory method.

domain, mitigating the impact of common-mode voltage caused by weight differencing, and eliminating the redundant area and power budget of peripheral digital computing circuits.

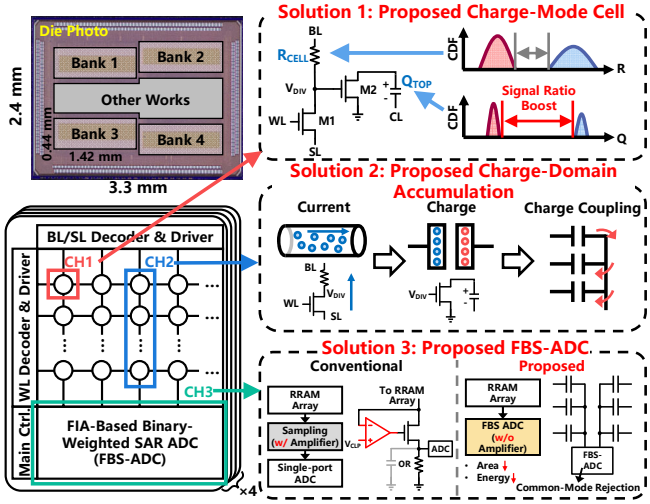


Figure 2. Our proposed charge-mode ReRAM-based CIM macro.

II. PROPOSED CHARGE-MODE RRAM CIM MACRO

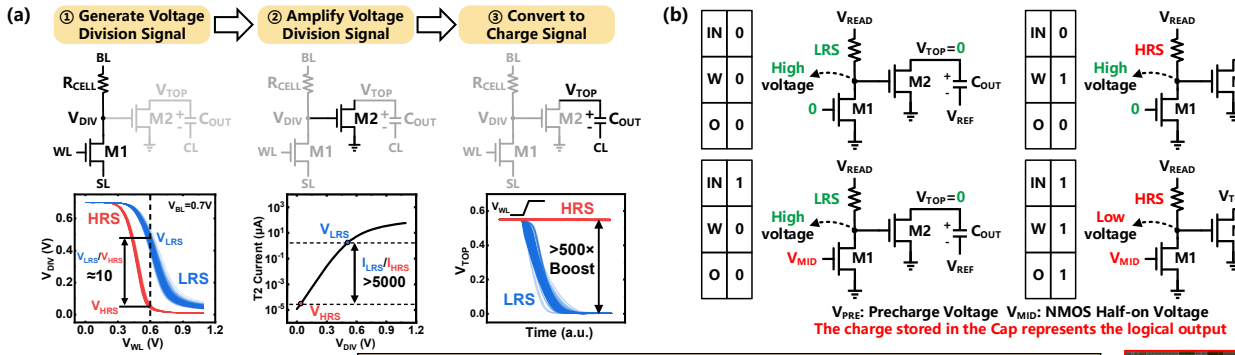
In this work, we propose a ReRAM-based compute-in-memory chip capable of charge-mode computing to address the above challenges. The architecture of the chip is shown in **Fig. 2**. The chip consists of four identical banks, each having an array of 256K ReRAM cells with independent peripheral circuits and a control module. To mitigate the limited on/off ratio of the ReRAM device, the bank utilizes a charge-mode cell structure to boost the on/off signal ratio (**Fig. 2-S1**). Leveraging the inherent advantages of charge-mode ReRAM cells, we propose a charge-sharing scheme to enable multi-row accumulation operations (**Fig. 2-S2**). We propose the binary-weighted circuit based on the Floating-Inverter-Amplifier (FIA) comparator to enable multi-bit activation inputs (8-bit activation), weight differencing (ternary weight), and analog-to-digital conversion. This scheme not only significantly reduces the area and energy overhead of peripheral circuits (e.g., amplifiers) compared to conventional approaches but also demonstrates superior common-mode rejection capability, thereby effectively supporting analog-domain differential operations (**Fig. 2-S3**).

A. Proposed 4T2R2C Charge-Mode Differential Cell

The proposed 4T2R2C differential cell consists of two identical 2T1R1C cells. **Fig 3 (a)** shows the structure of the 2T1R1C cell structure, using the characteristics of the transistor to boost the on-off ratio and to reduce the influence of the variation of ReRAM cell resistance. This structure also allows charge-mode computing to alleviate the problem of large currents in the array and to obtain increased linearity. The cell contains two NMOS devices (M1 and M2), one RRAM device (R_{CELL}), and one capacitor (C_{OUT}). Wherein the M1 which is in series with the R_{CELL} divides the voltage of BL (V_{BL}), resulting in a varying mid-point voltage V_{DIV}. And the ratio in V_{DIV} (V_{LRS}/V_{HRS}) can be amplified by the transconductance of M2, generating a larger ratio of discharging current (I_{LRS}/I_{HRS}) from the pre-charged (V_{PRE}) capacitor C_{OUT} to amplify signal margin (>500× boost) and to convert into the charge domain. As shown in **Fig. 3(b),** the top electrode potential (V_{TOP}) of C_{OUT} will be either 0 or V_{REF} according to the WL input (0 or V_{MID}) and the resistance of R_{CELL} (HRS, high resistance state or LRS, low resistance state). Two sets of 2T1R1C cells (positive cell and negative cell) from two different columns form a 4T2R2C differential pair, giving our chip the ability to represent ternary (1.58b) weight. The detailed truth table as well as the cell layout design (0.83692 μm² for 2T1R cell area) are given in **Fig. 3(c)**.

B. Charge-Domain Fully-Parallel Accumulation Scheme

Based on the charge-mode cell structure, the charge-domain fully-parallel accumulation scheme is proposed to accumulate the multiplication result to enable in-memory VMM. Our



4T2R2C Differential Pair Truth Table

Input (V_{WL})	Weight	Positive R (HRS/LRS)	Negative R (HRS/LRS)	Positive Cap (Q_{CAP_P})	Negative Cap (Q_{CAP_N})	Logic Output (Q_{CAP_P}-Q_{CAP_N})
0 (0V)	+1	1 (HRS)	0 (LRS)	0 ($-V_{REF}C_{OUT}$)	0 ($-V_{REF}C_{OUT}$)	0 (0)
	0	1 (HRS)	1 (HRS)	0 ($-V_{REF}C_{OUT}$)	0 ($-V_{REF}C_{OUT}$)	0 (0)
	-1	0 (LRS)	1 (HRS)	0 ($-V_{REF}C_{OUT}$)	0 ($-V_{REF}C_{OUT}$)	0 (0)
1 (V_{MID})	+1	0 (LRS)	1 (HRS)	1 (0)	0 ($-V_{REF}C_{OUT}$)	1 ($V_{REF}C_{OUT}$)
	0	1 (HRS)	1 (HRS)	1 (0)	1 (0)	0 (0)
	-1	0 (LRS)	1 (HRS)	0 ($-V_{REF}C_{OUT}$)	1 (0)	-1 ($-V_{REF}C_{OUT}$)

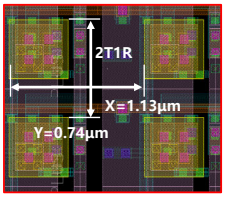


Figure 3: (a) Proposed 2T1R1C charge-mode ReRAM cell structure, (b) the multiplication operation method of the charge-mode cell, (c) the 4T2R2C structure for ternary weights, its truth table and layout design.

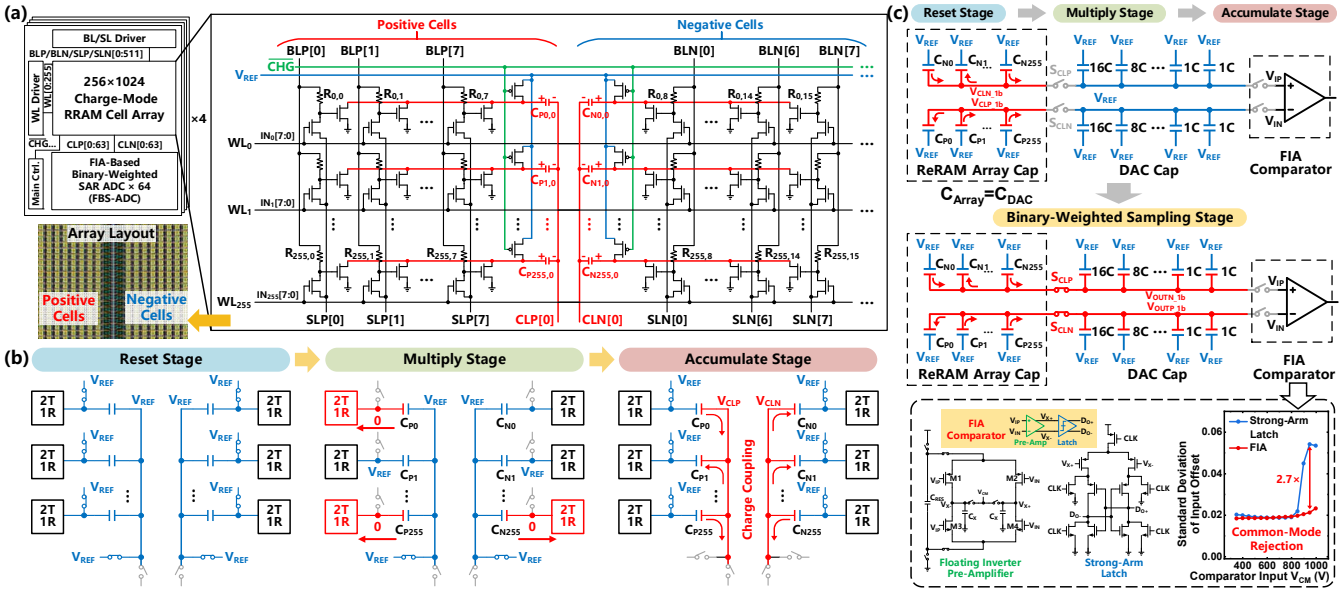


Figure 4: (a) The array design of our proposed charge-mode ReRAM CIM macro. (b) The operation principle of the charge-domain accumulation scheme. It takes 3 stages to complete the accumulation. (c) The schematic of the FBS-ADC and the advantage of FIA comparator.

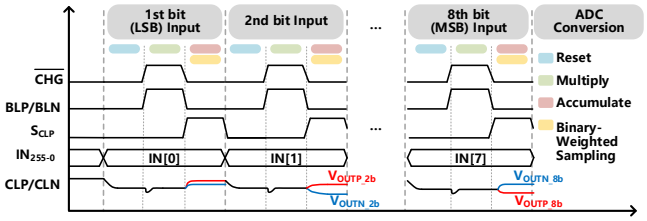


Figure 5: The wave diagram of the 8-bit binary weighting scheme.

proposed macro is comprised of 4 banks as shown in **Fig. 4(a)**. Each bank has a 256 x 1024 charge-mode RRAM cell array, as well as 64 ADCs and peripheral circuits. In the array, every eight cells from one row share a capacitor and a PMOS which provides voltage V_{REF}. As shown in **Fig. 4(b)**, there are three stages to implement the MAC operation: (a) Reset stage: Selected columns are driven to V_{READ}, and unselected columns are set to 0. The voltage of both the top and the bottom electrode of C_{OUT} is set to V_{REF}, resetting the charge to 0 from the previous outcome. (b) Multiply stage: The PMOS that connects the top electrode of C_{OUT} is turned off, leaving the capacitor connected to the 2T1R cell. The same bit of the input activation from 255 to 0 is applied to WL (fully parallel), and the result of activation multiplies weight is stored in C_{OUT} in the form of charge. (c) Accumulation stage: The top electrode is connected to V_{REF} again, and the bottom electrode is disconnected from VREF instead. As a result, the charge from the bottom electrodes of all C_{OUT} in the same column is evenly distributed to all the capacitors, representing the multiplication result stored in each capacitor is accumulated. The final MAC output is represented in the form of the voltage of CL, yielding V_{CLP} and V_{CLN} for positive cells and negative cells respectively. Such the charge-mode computing scheme can alleviate the non-linearity problem that occurs in current or voltage-mode computing and the shorter reading time can decrease the computing power. The variation of the RRAM device resistance is also suppressed by the enlarged signal margin.

C. *FBS-ADC for Multi-bit Activation Input Accumulation*

After the MAC result has been computed and stored in the capacitors in the array, the switches (S_{CLP} and S_{CLN}) between the CL (CLP/CLN) and the ADC input (IP/IN) are opened as in **Fig. 4(c)**, entering the binary-weighted sampling (BWS) stage. The total capacitance of the Capacitive DAC (CDAC) is equal to the sum of the capacitors in the array ($C_{DAC} = C_{Array}$), so that the voltage of CLP and CLN will be evenly shared with the IP and IN. Therefore, the input voltage of the ADC is VIP–VIN=1/2(V_{CLP_1b}–V_{CLN_1b}).

Moreover, the MAC result of 8b activation input (IN[7:0]) could be achieved by consecutively repeating the BWS stage as shown in **Fig. 5**. As shown in the waveform, the LSB of IN255-0 ($IN_{255\text{-}0}[0]$) is firstly input into the 256 WLs of the array, generating the first partial MAC result. After the first BWS stage, CDAC stores the MAC results of the 1st input bit, and the input of the comparator will remain floating (thus keeping the charge). Then the next bit $IN_{255\text{-}0}[1]$ is input to the array. After the second BWS stage, the charge of the CDAC and the charge of the second partial MAC result are shared among the CDAC and the capacitors of the array, which results in the weighted sum of the MAC result of the two input bits ($IN_{255\text{-}0}[1:0]$). By analogy, 8-bit activation input VMM can be implemented through this scheme, with the sampled voltage following the formula:

$$V_{OUTP_{1b}} = \frac{\sum Q_{DAC_P} + \sum Q_{array_P}}{\sum C_{DAC_P} + \sum C_{array_P}} = \frac{1}{2}\left(V_{CLP_{1b}} + V_{CM}\right) \quad (1)$$

$$V_{OUTP_{ib}} = \frac{\sum Q_{DAC_P} + \sum Q_{array_P}}{\sum C_{DAC_P} + \sum C_{array_P}} = \frac{1}{2}\left(V_{CLP_{1b}} + V_{OUTP_{(i-1)b}}\right) \quad (2)$$

$$V_{OUTP_{8b}} = \frac{1}{2^8}\left(2^7 V_{CLP_{8b}} + 2^6 V_{CLP_{7b}} + \cdots + V_{CLP_{1b}} + V_{CM}\right) \quad (3)$$

Then the ADC starts converting the result. Under this scheme, the input common-mode voltage of the comparator will be affected by the MAC result, therefore the ADC utilizes the FIA to suppress the error induced by changes in the input

common-mode voltage. The simulation result shows that FIA can achieve 2.7× better performance than traditional SA latch.

III. Measurements and Evaluation

The ReRAM CIM chip is fabricated under the 40nm foundry process, and we develop a chip testing system for chip performance testing and application. **Fig. 6(a)** illustrates the measured resistance distribution of ReRAM devices on the chip using the test board. As we can see the devices yield an on-off ratio of only 6.1x, which is inadequate for traditional current-mode and voltage-mode non-volatile in-memory computing with multiple rows activated at the same time. By adopting the 4T2R2C cell structure, the signal ratio is significantly enhanced, enabling the charge-mode in-memory computing to maintain a substantial signal margin even under large row parallelism, as shown in **Fig. 6(b)**. We then evaluate the VMM results of our chip under both 1-bit (1b) and 8-bit (8b) activation precision and statistically analyze the computational errors, as illustrated in Figure **Fig. 6(c & d)**.

The photo of our test system can be seen in **Fig. 7(a)**. We deploy a part of a ternary Transformer-based model [7] on our chip to utilize the VMM linearity and large parallelism of our chip, and the Rogue-{1, 2, L} score is shown in **Fig. 7(b)**. The energy and area breakdown of our chip is shown in **Fig. 7(c)**.

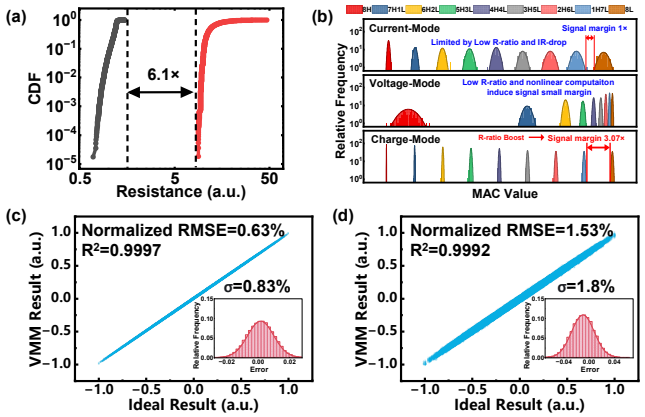


Figure 6: (a) Measured CDF of the RRAM device resistance. (b) The simulated signal margin comparison between current, voltage and charge-mode computing. (c) VMM results for 1bit activation and (d) 8bit activation.

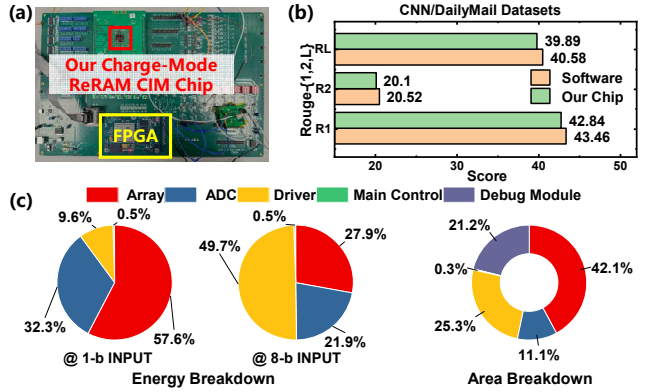


Figure 7: (a) Experiment setup. (b) Rogue-{1, 2, L} score on CNN/DailyMail dataset. (c) Energy and area breakdown of our chip.

Table 1 Comparison with other ReRAM CIM macro works

	This work	ESSERC 2024, THU [1]	VLSI 2023, GaTech [2]	ISSCC 2023 NTHU [3]	JSSC 2023 IMECAS [4]	ISSCC 2021 GaTech [5]
Technology (nm)	40	28	40	22	28	40
ReRAM Capacity	1Mb	512Kb	64Kb	32Mb	8Kb	64Kb
Calculation Precision	8b-IN/T*-W/ 6-bO	4b-IN/4b-W/ 8b-O	1b-IN/1b-W/ 6b-O	INT 8	1b-IN/3b-W/ 4b-O	1b-IN/1b-W/ 4b-O
Computing Domain	Charge	Current	Current	Current	Current	Voltage
Cell Structure	Charge-Mode 4T2R2C	2T2R	1T1R	1T1R	2T1R	1T1R
Energy Efficiency (TOPS/W)	103.5	39.3	75.2	76.5	30.34	26.56
Peak Throughput (TOPS)	4.23	2.84	0.18*	8.21	N/A	NA
1bit Normalized EE (TOPS/W)	1308.24	628.8	75.2	4896	91.02	26.56
1bit Normalized Throughput (TOPS)	53.47	45.44	0.18*	525.44	N/A	N/A
Linearity R^2	0.9992	0.9985	N/A	N/A	N/A	N/A

*T: Ternary=1.585bit

IV. Conclusion

In this paper, we propose a charge-mode ReRAM-based CIM chip. With the proposed 4T2R2C charge-mode cell structure, the signal ratio is enlarged by 500x compared with the conventional computation scheme. Moreover, our proposed charge-domain accumulation scheme and FBS-ADC implement analog weighting of 8-bit input VMM computation, significantly suppressing the impact of nonlinearity problem induced by large column current, enhancing energy efficiency and throughput. We deploy components of Transformer networks on the chip for testing and validation, achieving metrics close to the software. Our chip achieves the peak energy efficiency of 1308.24 TOPS/W, the throughput of 53.47 TOPS (normalized to 1b) and great linearity with 0.9992 of R^2.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (92164302, 8206100486 and 62404004), Guangdong Provincial Key Laboratory of In-Memory Computing Chips (2024B1212020002), Shenzhen Science and Technology Program (JCYJ20241202125907011), Beijing Natural Science Foundation (L234026, 25D40029, 25FY3314) and the 111 Project (B18001).

References

[1] 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,

[2] 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," 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 2023, pp.

[3] 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," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 15-17.

[4] 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 Journal of Solid-State Circuits, vol. 58, no. 10, pp. 2839-2850, 2023.

[5] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "29.1 A 40nm 64Kb 56.67 TOPS/W read-disturb-tolerant compute-in-memory/digital RRAM macro with active-feedback-based read and in-situ write verification," in 2021 IEEE International Solid-State Circuits Conference (ISSCC), 2021, vol. 64: IEEE, pp. 404-406.

[6] X. Tang et al., "An energy-efficient comparator with dynamic floating inverter amplifier," IEEE Journal of Solid-State Circuits, vol. 55, no. 4, pp. 1011-1022, 2020.

[7] Z. Liu, B. Oguz, A. Pappu, Y. Shi, and R. Krishnamoorthi, "Binary and Ternary Natural Language Generation," in *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.