A 1T2R1C ReRAM CIM Accelerator With Energy-Efficient Voltage Division and Capacitive Coupling for CNN Acceleration in AI Edge Applications

Deyang Chen, Zhiwang Guo, Jinbei Fang, Chenyang Zhao, Jingwen Jiang, Keji Zhou, Haidong Tian, Xiankui Xiong, Xiaoyong Xue, *Member, IEEE*, and Xiaoyang Zeng

***Abstract*—Resistive random-access memory (ReRAM) is widely studied in computing-in-memory (CIM) for neural network acceleration in edge devices. However, the static current of conventional 2T2R array becomes prominent with increasing computing parallelism, leading to remarkable power and area costs. To resolve this issue, a voltage-style one-transistor-two-resistor-one-capacitor ReRAM (1T2R1C) CIM accelerator using energy-efficient voltage division and capacitive coupling is proposed for convolutional neural network (CNN) acceleration in AI edge applications. The 1T2R1C cell, which is 15% smaller than the previous 2T2R cell, comprises one selection transistor, two ReRAM resistors for 1-bit weight storage, and a MOS-based capacitor. The multiply-and-accumulation (MAC) operation is realized by voltage division within a cell and capacitive coupling across different cells on a plate line. The static current during computation can be effectively reduced by the series connection of low-resistance state (LRS) and high-resistance state (HRS) ReRAM resistors within a cell and the elimination of low resistance paths in the array. A corresponding weight mapping method is also proposed to transform the multi-bit weight based on 0/1 into that on −1/1, adapting the proposed accelerator for high-precision CNN applications. By evaluation in the 28nm technology, the proposed accelerator with a 384Kb array achieves a peak energy efficiency of 400.2 TOPS/W, ∼8.4X higher than previous work. The inference accuracy of CNN reaches 96.2% on the MNIST dataset.**

***Index Terms*—Computing-in-memory, ReRAM, 1T2R1C, capacitive coupling, weight mapping.**

I. Introduction

With the exponential growth of Internet of Things (IoT) devices, equipping the edge devices with certain computational capabilities especially the artificial intelligence (AI) acceleration becomes a trend [1], [2]. Computing-in-memory (CIM) provides an energy-efficient solution for the edge device to realize neural network acceleration compared with von Neumann architectures [3], [4], [5], [6]. Especially, resistive random-access memory (ReRAM) has been under extensive investigations in CIM architecture thanks to its non-volatility, high density, and process compatibility [7], [8], [9], [10], [11], [12], [13], [14].

Conventionally, ReRAM CIM accelerators often use a one-transistor-one-resistor (1T1R) cell to store an unsigned weight bit or two 1T1R cells to store a signed one. The multiply-and-accumulation (MAC) operation of the input vector and the weight matrix in the memory macro is performed in a current accumulation (CA) mechanism according to Ohm's Law and Kirchhoff's Current Law. However, with the increase in neural network scale and computing parallelism, the static current which is mainly caused by the low resistance state (LRS) ReRAM read paths tends to rise significantly, causing large power consumption at the array level and greatly offsetting the energy-efficient benefits of CIM [9]. What's more, the large static current accumulated on the bit line (BL) calls for wide interconnection for routing and results in considerable area costs [8]. Moreover, the large IR drop of BL may also result in an accuracy loss for the neural network implemented on the CIM accelerator [7]. Therefore, it is of vital importance to reduce the static current of the CIM array for the efficient implementation of low-power edge devices.

Although increasing the LRS and high-resistance state (HRS) resistances of ReRAM helps to lower the current, it is closely related to the limitations of technology capability. A sign-weighted 2T2R (SW-2T2R) array can reduce the current and IR drop on the source line by connecting the positive weight and the negative one on the same column. But it is still based on CA and the static currents consumed by the cells are almost the same [7]. A voltage-mode sensing scheme was proposed with two differential-row weight mapping to save the current duration. However, the computing accuracy can be easily affected by the nonidealities like variation and nonlinearity of ReRAM

Manuscript received 18 April 2022; revised 1 August 2022; accepted 21 August 2022. Date of publication 24 August 2022; date of current version 22 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61874028 and Grant 61834009; in part by the Science and Technology Commission of Shanghai Municipality under Grant 21TS1401200 and Grant 22ZR1407100; and in part by the Opening Project of Zhejiang Laboratory under Grant 2022PF0AB01. This brief was recommended by Associate Editor X. Miao. *(Corresponding author: Xiaoyong Xue.)*

Deyang Chen, Zhiwang Guo, Jinbei Fang, Chenyang Zhao, Jingwen Jiang, Keji Zhou, Xiaoyong Xue, and Xiaoyang Zeng are with the State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 201203, China (e-mail: xuexiaoyong@fudan.edu.cn).

Haidong Tian and Xiankui Xiong are with the State Key Laboratory of Mobile Network and Mobile Communication Multimedia Technology, ZTE, Shenzhen 518057, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2022.3201367.

Digital Object Identifier 10.1109/TCSII.2022.3201367

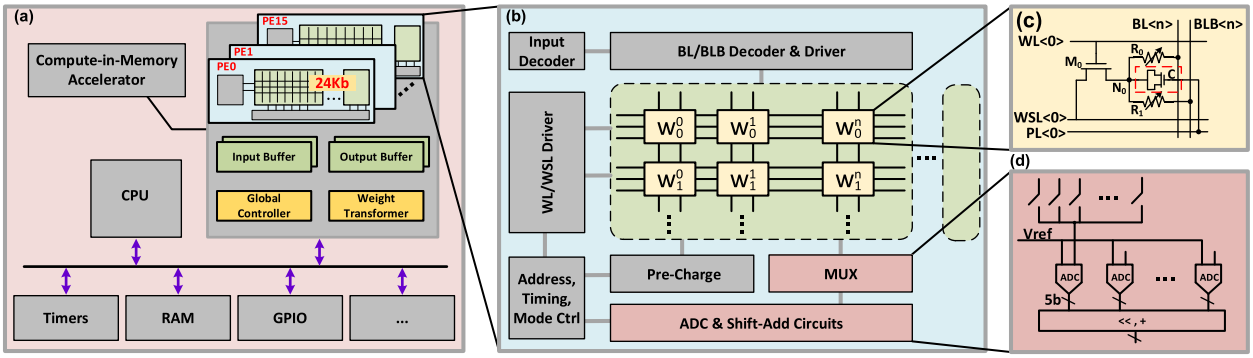

Fig. 1. The hierarchical architecture of the proposed 1T2R1C voltage-division-based accelerator. (a) The overall structure with 16 PEs, (b) the structure of one PE macro, (c) the structure of proposed 1T2R1C cell, (d) the output processing circuit.

as well as the difficulty to realize accurate intermediate resistance states for stable storage [8]. A voltage-division (VD) based circuit was proposed to reduce the large current flowing through the BL during MAC computation when multiple parallel inputs are activated. However, the current is only suppressed at the cell level and low resistance (low-R) path still exists at the array level, which causes a large current, as will be illustrated in Section II [9]. Other techniques are also helpful to reduce the current at the architecture or algorithm level [15], [16].

This brief proposes a voltage-style one-transistor-two-resistor-one-capacitor (1T2R1C) ReRAM CIM accelerator for energy-efficient convolutional neural network (CNN) acceleration in AI edge applications. The 1T2R1C structure features more area-efficient than the conventional 2T2R cell and enables voltage-style computing to realize MAC. Thus, the static current during computation can be effectively reduced at the cell level as well as the array level. Apart from the binary neural network (BNN), a weight transforming method is also proposed to accommodate the accelerator for high-precision CNN. The evaluation verifies the benefit of the proposed 1T2R1C CIM accelerator in energy efficiency.

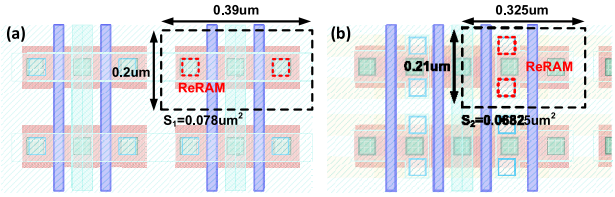

Fig. 2. The layout of (a) 2T2R cell, (b) proposed 1T2R1C cell.

TABLE I
OPERATING CONDITIONS FOR THE ARRAY

Operation	WL	BL	BLB	PL	WSL
Forming	VDD	Vform	Vform	-	GND
Set R0	VDD	Vset	Vb1	-	GND
Set R1	VDD	Vb1	Vset	-	GND
Reset R0	VDD	GND	Vb2	-	Vreset
Reset R1	VDD	Vb2	GND	-	Vreset
Precharge	VDD	Vpre	Vpre	Vpre	Vpre
Evaluation	GND	Vread	GND	Vout	-

**Note*: Vform = 1.8V, Vreset = 1.5V, Vset = Vb2 = 1.1V, Vb1 = 0.6V. Vpre = 0.15V, Vread = 0.3V.

II. PROPOSED 1T2R1C CIM ACCELERATOR

A. 1T2R1C Cell for Voltage-Style Computing

The proposed 1T2R1C accelerator is targeted for CNN acceleration in edge System-on-Chip (SoC) devices. Fig. 1(a) shows the overall architecture of the proposed accelerator which is composed of 16 processing elements (PEs), input buffers, output buffers, and other control logics. Each PE further consists of a small array of 24Kb to obtain a small leakage current and parasitic capacitance. The peripheral circuits of the PE, such as drivers, decoders, write circuits, and ADCs for MAC, are shown in Fig. 1(b). The 1T2R1C cell comprises one selection transistor M_0, two ReRAM resistors R_0 and R_1 for 1-bit weight storage, and a capacitor C for capacitive coupling with small size, as shown in Fig. 1(c). The gate and source of M_0 are connected to the word line (WL) and the write source line (WSL), respectively. The drain of M_0 is connected to the bottom electrodes of two ReRAM resistors and C at node N_0, the top electrodes of which are connected to the complementary bit lines (BL/BLB) and the plate line (PL), respectively. For the interconnections at the array level, WL/WSL/PL will run horizontally while BL/BLB will be vertically configured. Two bit-cells are studied with capacitors implemented by MOSCAP and MOMCAP. The results show that the MOMCAP with M2-M5 layers achieves a capacitance density of 1.33 fF/um^2 whereas MOSCAP achieves 13.83 fF/um^2 at 100MHz. Considering the area overhead of the bit-cell, we use the MOS capacitor for implementation to save area cost, as shown in Fig. 2.

Table I lists the conditions of write and CIM operations. For the write operation, the two ReRAM resistors are separately operated [17]. Taking R_0 as an example, the selection transistor M_0 is firstly activated by WL, and the set/reset voltage *Vset*/*Vreset* is then applied on BL/WSL while WSL/BL is grounded to apply a sufficient voltage at both ends of R_0 for a HRS (LRS) to LRS (HRS) transition. To avoid write disturbance to R_1, a biasing voltage *Vb1*/*Vb2* is applied on BLB. After the write operation of R_0, a similar operation is performed on R_1 successively. The combination of R_0 and R_1 resistance states in a 1T2R1C cell represents one bit of weight in the neural network. The 1-bit binary weight, i.e., "+1/−1", is represented by a LRS-HRS/HRS-LRS cell. Note that PLs are floating and unselected WLs are grounded to avoid disturbance during the write operation.

(a)

(R0,R1)	Weight	BL	BLB	Input	ΔVPL
(LRS,HRS)	1	Vread	GND	1	+ΔV
		Vpre	Vpre	0	-
		GND	Vread	-1	-ΔV
(HRS,LRS)	-1	Vread	GND	1	-ΔV
		Vpre	Vpre	0	-
		GND	Vread	-1	+ΔV

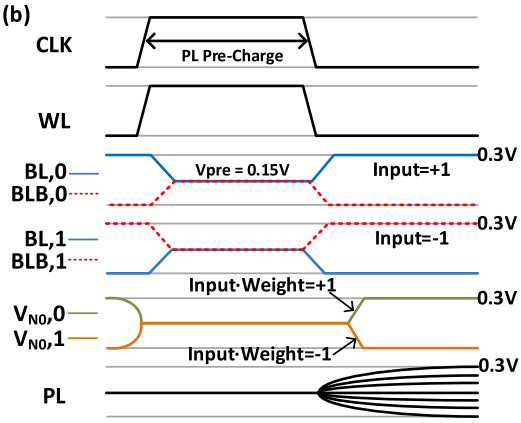

Fig. 3. (a) The truth table in CIM mode, (b) the function waveform in the readout phase.

Two phases, i.e., pre-charge and evaluation, are required for CIM operations to obtain the MAC results. In the pre-charge phase, WL is activated and the voltage *Vpre* is applied on WSL to pre-charge the node N_0. Meanwhile, BL/BLB are also kept at *Vpre* level to suppress any static current in the 1T2R1C cell. PL is also pre-charged to *Vpre* level as the initial voltage. In the evaluation phase, BL/BLB is applied to certain voltage levels depending on the input. Therefore, the local result is developed on node N_0 based on the voltage division of two ReRAM resistors in the 1T2R1C cell. The voltage of node N_0 can vary from ground to *Vread*. The truth table for CIM results at the cell level is illustrated in Fig. 3(a). The CIM computation supports ternary input (−1/0/1) and binary weight (−1/1) multiplications. Fig. 3(b) shows the waveform of MAC operation at the array level during the pre-charge and evaluation phases. To readout the global CIM results from PL, the MOSCAP in each 1T2R1C cell is used to couple the local computing results to PL. Assuming that n cells are selected in one CIM operation, the voltage of PL after charge sharing is proportional to the MAC results as follows:

$$V_{PL} = V_{pre} + \frac{\sum_{i=0}^{n-1} MAC(i) * C_c}{2(nC_c + C_p)} * V_{read} \tag{1}$$

where the MAC(i) denotes the local MAC results in the i^{th} cell, and C_p is the sum of PL parasitic capacitance and the input capacitance of ADC.

The benefits of proposed voltage-style computing enabled by 1T2R1C cell on static current reduction and readout margin are analyzed as follows. Previous 2T2R VD-based architecture [9] uses differential cells to restrain the readout current within a cell, but a low-R path across two LRS cells in different selected rows may still arise through common SL during computing, as shown in Fig. 4(a). The low-R path may

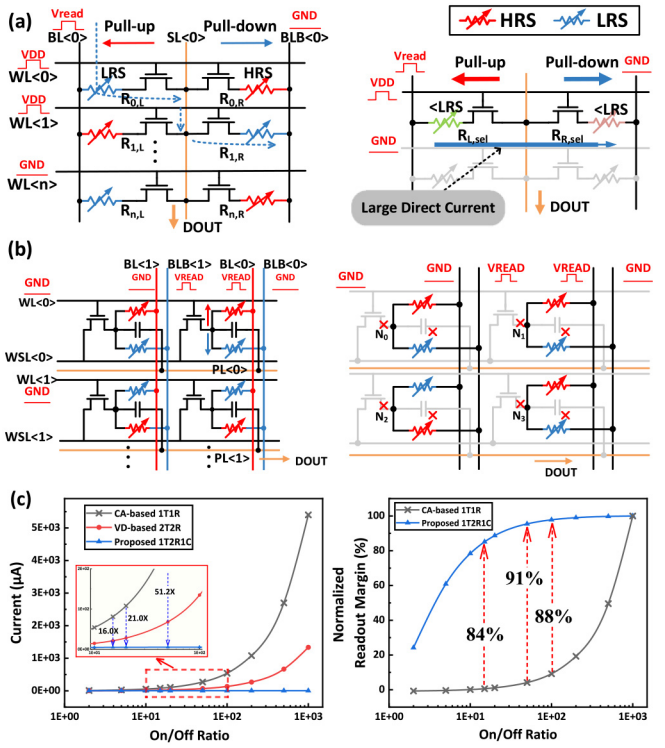

Fig. 4. The schematic and the equivalent circuit of (a) the VD-based 2T2R array, (b) the proposed 1T2R1C array, (c) the relationship of static current and readout margin with the on/off ratio for 9 parallel inputs activated.

bring considerable current, as calculated below:

$$I_{total} = \frac{V_{read}}{R_{L,set} + R_{R,sel}} \tag{2}$$

where $R_{L,sel}$ and $R_{R,sel}$ represent the parallel resistance of the left and right resistors of cells in the selected rows. As the parallelism increases, the values of $R_{L,sel}$ and $R_{R,sel}$ can be much smaller than that of LRS resistance. Thus, the large direct current path between BL and BLB will arise accordingly, aggravating the power consumption and IR-drop issues. By contrast, the proposed 1T2R1C array can reduce the currents from two aspects. Firstly, the series connection of LRS and HRS RRAM resistors in a 1T2R1C cell helps to ensure sufficient resistance in the current path within the cell, effectively suppressing the static current, as shown in Fig. 4(b). Secondly, the current paths flowing across different cells are also isolated by the unselected NMOS selectors and MOS capacitors, thus eliminating low-R current paths at the array level. Moreover, the readout results in 1T2R1C array are obtained by the voltage division of two differential ReRAMs, which is only related to the relative value of HRS and LRS resistance in the same cell, hence showing a better readout margin than the CA scheme. The simulation results in Fig. 4(c) show that the 1T2R1C array suppresses the static current better than the CA-based and the VD-based ones. When the on/off ratio of ReRAM device is 15, the current can be reduced by ∼16X compared to conventional CA-based ones, whereas the readout margin increases by 84%.

B. Multi-Bit Weight Mapping for CNN

According to the application requirements, different neural networks with different precisions are usually implemented on

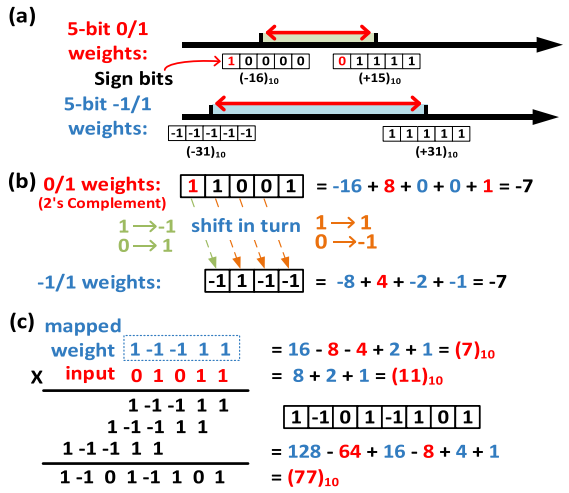

Fig. 5. (a) The representation range, (b) the transformation method, (c) the multiplications with mapped weight.

the CIM accelerator. The proposed 1T2R1C cell can represent 1-bit weight precision of ±1, which makes it suitable for binary neural network (BNN) based on XOR operations in simple AI applications. However, for most practical networks such as CNN, multi-bit weights are usually needed for high-precision MAC operations. It is necessary to use the proposed 1T2R1C cell storing a 1-bit weight of ±1 to represent the multi-bit weights based on 0/1.

A new mapping method is proposed to transform the multi-bit weight based on 0/1 to that on ±1. Since the 1T2R1C cell can inherently represent ±1, the multi-bit weight will not need a sign bit and thus can be treated as an unsigned number. Therefore, the binary number based on −1/1 can represent a 2X range than the same count of bits based on 0/1, as shown in Fig. 5(a). For the transformation method, the weight represented by 2's complement binary number based on 0/1 will be converted to that based on ±1 for storage and computation in the proposed 1T2R1C array. The sign bit and the other bits are dealt with separately, as shown in Fig. 5(b). For the sign bit based on 0/1, the "1" bit is mapped to the 1T2R1C cell of "−1" while "0" to "1". For the other bits, the "1" bit is mapped to the 1T2R1C cell of "1" whereas the "0" bit is to that of "−1". Once the weights of neural network are transformed, the subsequent calculations are the same as that in conventional CNN accelerations. Fig. 5(c) gives an example process of a 5-bit mapped weight multiplied by a 5-bit input. Note that the transformation of 2's complement weight based on 0/1 to that on −1/1 is mainly performed by shift-by-bit operation during the write operation, and it will not incur any additional delay or power consumption during the CIM operation.

III. Results and Analysis

A. Read Disturbance and Analysis

Due to the non-idealities of ReRAM devices and the effects of process fluctuations, the read voltage of ReRAM devices has to be carefully selected considering both the cell reliability and the computing performance. A large read voltage is expected to enlarge the dynamic range of output, thus allowing for high-precision output or simple ADC with low cost. However, continuous read operations with high read voltages may induce the resistance drift effect in ReRAM devices,

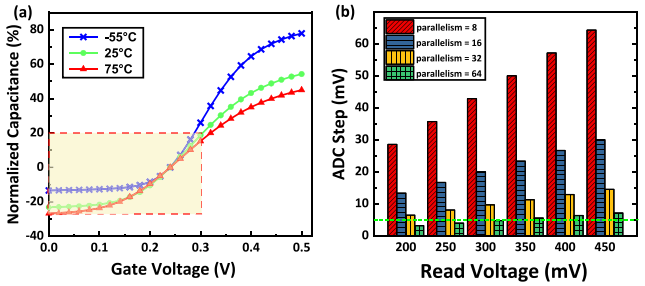

Fig. 6. (a) The MOSCAP capacitance under heterogeneous temperature and offsets, (b) the required ADC steps under different parallelisms and read voltages.

which may affect the accuracy of MAC results. Also, a high read voltage will lead to a larger voltage swing at node N_0, resulting in the fluctuation of MOSCAP capacitance. The MOSCAP capacitance is dependent on the frequency, temperature, and bias voltage. The frequency and temperature will have little effect on the function of output since all the capacitors work in a similar condition. However, the voltage may affect the readout results, thus it is expected that the MOSCAP work with small capacitance variation in the range of less than 0.3V, as shown in Fig. 6(a). A Monte Carlo simulation is performed to study the influences of the capacitance voltage and the mismatch effect on the computation results. The simulation result shows that a 6.6% deviation is brought to PL at most, which will only influence the LSB of outputs.

The value of read voltage will also be affected by the output precision and the resolution of ADC. A smaller read voltage will narrow the voltage difference between two adjacent states on PL, making it difficult for ADC to resolve. As shown in Fig. 6(b), when the same count of parallel rows is selected, the ADC resolution is reduced as the read voltage decreases. Similarly, the increase in parallelism will call for a high ADC resolution as well. Considering the area cost, the noise, the parasitic resistance in the circuit, and the variation of the ReRAM devices, the resolution of ADC is supposed to be ≥5mV to ensure output accuracy. Thus, the parallelism of ≤32 rows which can support 5-bit output is recommended. To sum up, the selection of read voltage and parallelism of 300mV and 32 are selected, respectively.

B. System-Level Evaluation

When AI algorithms are implemented on edge devices, the floating-point data are usually quantified into fixed-point numbers for computation. To investigate the influence of the proposed weight mapping method on the computing accuracy of fixed-point numbers, a CNN network is implemented on the accelerator to perform the inference for MNIST dataset, as shown in Fig. 7(a). The binary bits of weights are mapped to −1/1 with the proposed mapping method while the inputs are still based on 0/1. The outputs of ADCs from different PLs are cumulated using shift-and-add operation. The results after ReLU activation serve as the input of the next layer. The 28nm technology is used for the circuits for evaluation. The HRS and LRS of ReRAM devices are 300KΩ and 20KΩ with a 30% (3σ) variation, respectively.

Fig. 7(b) shows the impact of the proposed mapping method on the computing accuracy with different weight precisions. The customized CNN and VGG16 models are mapped on

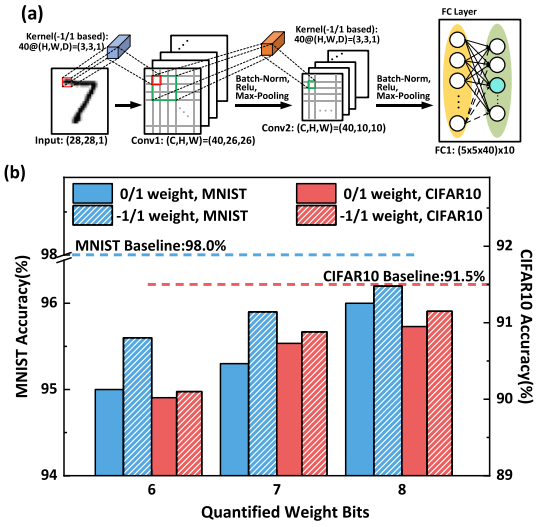

Fig. 7. (a) The convolutional neural network model, (b) the computation accuracy with different weight bits.

TABLE II
COMPARISON WITH PREVIOUS RERAM CIM ACCELERATORS

Work	ISSCC [7]	VLSI [8]	TCAS-II [9]	TCAS-II [12]	**This Work**
Process	130nm	130nm	180nm	28nm	28nm
Cell Structure	2T2R	2T2R	2T2R	1T2R	1T2R1C
Capacity	158.8Kb	64Kb	400Kb	384Kb	384Kb
Output Precision	1-8 bit	1bit	3bit	analog	1-8 bit
MNIST Accuracy	94.4%	91.4%	-	-	96.2%
CIFAR10 Accuracy	-	-	-	62.3%	91.1%*
Energy Efficiency (TOPS/W)	78.4	851.1	42.6	99 - 197.8	400.2 (BNN) 6.3-16.0 (CNN)

* Single layer of VGG16 in hardware

hardware and tested on MNIST and CIFAR10 datasets respectively. The fixed-point calculation before and after mapping as well as the full precision calculation are compared. The computing accuracy after weight mapping is higher than that before mapping. The reason for lower accuracy loss is that the $-1/1$ based weight mapping has a wider representation scale than that based on 0/1 with the same count of memory cells.

Table II compares this brief with previous ones. Since the proposed accelerator can effectively reduce the static power during MAC computation, the energy efficiency is improved by ~8.4X compared to previous work with a similar readout scheme [9]. The macro has 15.42mW power consumption, with the array accounting for 2.8% of that. The proposed structure has a limited power dissipation, which is suitable to apply in low-power devices.

IV. CONCLUSION

In this brief, a voltage-style ReRAM CIM accelerator is proposed by energy-efficient voltage division and capacitive coupling for CNN accelerations. The proposed 1T2R1C cell structure with area-efficient MOSCAP can effectively suppress the large current at the cell level and the array level during MAC computation, hence achieving high energy efficiency. The corresponding mapping method for multi-bit weight helps to accommodate the accelerator for high-precision CNN rather than low-precision BNN. The evaluation demonstrates that the CIM accelerator with our proposed techniques is promising in low-power AI edge applications.

REFERENCES

[1] N. Lotze and Y. Manoli, "Ultra-sub-threshold operation of always-on digital circuits for IoT applications by use of Schmitt trigger gates," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 11, pp. 2920–2933, Nov. 2017.

[2] L. Ye *et al.*, "The challenges and emerging technologies for low-power artificial intelligence IoT systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 12, pp. 4821–4834, Dec. 2021.

[3] R. Han *et al.*, "A novel convolution computing paradigm based on NOR flash array with high computing speed and energy efficiency," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 5, pp. 1692–1703, May 2019.

[4] X. Si *et al.*, "A twin-8T SRAM computation-in-memory unit-macro for multibit CNN-based AI edge processors," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 189–202, Jan. 2020.

[5] H. Zhang, Y. Shu, W. Jiang, Z. Yin, W. Zhao, and Y. Ha, "A 55nm, 0.4V 5526-TOPS/W compute-in-memory binarized CNN accelerator for AIoT applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 5, pp. 1695–1699, May 2021.

[6] S. Jung *et al.*, "A crossbar array of magnetoresistive memory devices for in-memory computing," *Nature*, vol. 601, no. 7892, pp. 211–216, 2022.

[7] Q. Liu *et al.*, "33.2 A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2020, pp. 500–501.

[8] W. Wan *et al.*, "A voltage-mode sensing scheme with differential-row weight mapping for energy-efficient RRAM-based in-memory computing," in *Proc. IEEE Symp. VLSI Technol.*, 2020, pp. 1–2.

[9] L. Wang *et al.*, "Efficient and robust nonvolatile computing-in-memory based on voltage division in 2T2R RRAM with input-dependent sensing control," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 5, pp. 1640–1644, May 2021.

[10] C.-X. Xue *et al.*, "16.1 A 22nm 4Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7TOPS/W for tiny AI edge devices," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2021, pp. 245–246.

[11] Y. Feng *et al.*, "Improvement of state stability in multi-level resistive random-access memory (RRAM) array for neuromorphic computing," *IEEE Electron Device Lett.*, vol. 42, no. 8, pp. 1168–1171, Aug. 2021.

[12] K. Zhou *et al.*, "An energy efficient computing-in-memory accelerator with 1T2R cell and fully analog processing for edge AI applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 8, pp. 2932–2936, Aug. 2021.

[13] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40-nm, 64-Kb, 56.67 TOPS/W voltage-sensing computing-in-memory/digital RRAM macro supporting iterative write with verification and online read-disturb detection," *IEEE J. Solid-State Circuits*, vol. 57, no. 1, pp. 68–79, Jan. 2022.

[14] W. Li, X. Sun, H. Jiang, S. Huang, and S. Yu, "A 40nm RRAM compute-in-memory macro featuring on-chip write-verify and offset-cancelling ADC references," in *Proc. IEEE Eur. Solid-State Device Res. Conf. (ESSDERC)*, 2021, pp. 79–82.

[15] J. Yue *et al.*, "14.3 A 65nm computing-in-memory-based CNN processor with 2.9-to-35.8TOPS/W system energy efficiency using dynamic-sparsity performance-scaling architecture and energy-efficient inter/intra-macro data reuse," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2020, pp. 234–235.

[16] L. Wang *et al.*, "Sparsity-aware clamping readout scheme for high parallelism and low power nonvolatile computing-in-memory based on resistive memory," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2021, pp. 1–4.

[17] J. Yang *et al.*, "A 28nm 1.5Mb embedded 1T2R RRAM with 14.8 Mb/mm^2 using sneaking current suppression and compensation techniques," in *Proc. IEEE Symp. VLSI Circuits (VLSI-C)*, 2020, pp. 1–2.