SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2024, Vol. 67, Iss. 5, 152402:1–152402:8
https://doi.org/10.1007/s11432-023-3887-0

An isolated symmetrical 2T2R cell enabling high precision and high density for RRAM-based in-memory computing

Yaotian LING[1†], Zongwei WANG[1*†], Yuhang YANG[1], Lin BAO[2], Shengyu BAO[1], Qishen WANG[1], Yimao CAI[1*] & Ru HUANG[1]

[1]*School of Integrated Circuits, Beijing Advanced Innovation Center for Integrated Circuits, Peking University, Beijing* 100871*, China;*
[2]*State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing* 100876*, China*

Received 19 May 2023/Revised 7 August 2023/Accepted 2 November 2023/Published online 23 April 2024

Abstract In-memory computing (IMC), leveraging emerging memories, holds significant promise in overcoming memory limitations and improving energy efficiency. However, the prevailing IMC array structure based on serially connected transistors and memory cells (1T1R/2T2R), along with the signed weight mapping scheme, can lead to asymmetrical weight sensing issues (AWS) due to electrical asymmetry within the 1T1R/2T2R structure, particularly in highly scaled cells where the transistor's resistance becomes significant. In this paper, we propose and fabricate an electrically symmetric memory cell based on a physically isolated 2T2R structure for IMC. This design aims to enhance the precision and density of RRAM-based IMC arrays. The 2T2R cells are manufactured using the back-end-of-line (BEOL) process of a commercial 40 nm technology platform. The feasibility of this design is verified through measured and simulated results, showcasing its capability to address the issue of AWS. Compared to conventional 2T2R cells, this design achieves a considerably smaller transistor footprint without compromising accuracy, while also improving integration density by 42.2%. These innovative memory cell advancements have the potential to further advance high-energy-efficient IMC technology.

Keywords RRAM, 2T2R, multi-level storage, weight asymmetry, in-memory computing

1 Introduction

The artificial intelligence (AI) market has seen remarkable growth in recent years, particularly in voice recognition and machine vision [1–4]. As smart terminals like smartphones, watches, and bracelets continue to proliferate, the demand for energy-efficient hardware has skyrocketed. However, traditional computing architectures rely on von Neumann architecture, which separates the processor and memory. In this architecture, the processor retrieves data from memory, processes it, and stores it back to memory. This process necessitates extensive data transfers between the processor and memory, resulting in significant energy consumption and time delays. Additionally, when dealing with AI algorithms, the data size and computation required become increasingly substantial, which exacerbates energy and time consumption issues [5–7].

In-memory computing (IMC), which uses emerging devices like resistive random-access memory (RRAM), phase change memory (PCM), magnetic random-access memory (MRAM), and ferroelectric memory, has emerged as a promising solution to overcome this bottleneck [8–12]. RRAM is one of the most promising device candidates for the IMC paradigm due to its non-volatile characteristics, potential for multi-level storage, high integration density, and compatibility with existing semiconductor processes [13]. RRAM uses a resistive switching mechanism to store data, allowing multiple bits to be stored in a single cell. In an array, RRAM can perform multiply-accumulate (MAC) operations in the analog domain using

* Corresponding author (email: wangzongwei@pku.edu.cn, caiyimao@pku.edu.cn)
† Ling Y T and Wang Z W have the same contribution to this work.

Ohm's and Kirchhoff's laws [14], demonstrating great potential to accelerate AI computation. As a result, RRAM-based IMC chips have been extensively researched in recent years [15, 16].

2 Array architectures and weight mapping scheme

One of the key enablers of implementing RRAM to execute MAC is the weight mapping scheme, which transfers the algorithm parameters (e.g., the weights in a neural network) into the RRAM array. In a typical feedforward neural network, the strength between each neuron is determined by the signed weights [2]. For the mapping scheme of signed weight, two RRAM cells (W^+ and W^-) are usually implemented to represent the signed weights, and the output is obtained by subtracting the current of the two RRAMs [17]. This mapping scheme can be expressed by (1), where the former and latter parts of the weight are mapped into positive and negative parts of the IS-2T2R structure, respectively. According to the placement of W^+ and W^-, there are two typical formations of the RRAM array (Figure 1). The operating method varies with the array formation.

$$W = \begin{cases} 0 - |W|, & W < 0, \\ |W| - 0, & W \geqslant 0. \end{cases} \tag{1}$$

One typical array architecture stores W^+ and W^- at different source lines (SL) (column subtraction 2T2R, CS-2T2R) [18], as shown in Figure 1(a). The vector ("0" or "1") is input through word lines (WL) while constant read voltages are applied to bit lines (BL) [19]. After completing MAC operation in both SLs, peripheral subtraction circuits are needed to obtain the final result. The multi-bit input is enabled by time-division multiplexing (TDM) [13].

Another array architecture stores W^+ and W^- at different BLs (row subtraction 2T2R, RS-2T2R), as shown in Figure 1(b). The vector is input at BLs where either binary read voltage or analog read voltage can be applied [20]. The signed weights (binary or multi-bit) are mapped into pairs of differential BLs. For instance, two memory cells are placed between two adjacent BLs on the same SL, enabling in situ subtraction in the SL. Compared with the CS-2T2R, the RS-2T2R with the differentially paired BLs has several advantages. First, RS-2T2R does not require peripheral subtractors, reducing the hardware overhead of the peripheral circuit. Second, the output current (I_{out}) of RS-2T2R is much smaller than that of CS-2T2R, which also contributes to the reduction of power consumption as well as heat generation of the memory array [20]. Third, the SL dimension in the array layout can be further slimed, contributing to the density enhancement of the array.

3 Challenges of asymmetrical weight sensing

As mentioned above, the row-wise mapping scheme in RS-2T2R has several advantages on circuit overhead and power consumption over CS-2T2R. However, RRAM cells in RS-2T2R also suffer from challenges due to the connection topology between the transistor and memory cell, which may have been overlooked by the previous research. In the subsequent contents, the RRAM-based IMC array is exemplified to demonstrate the issues in RS architecture.

Conventional 2T2R structure faces the problem of non-zero source bias caused by the asymmetry of the 2T2R bias scheme. As shown in Figure 1(b), since the SLs are always grounded, the actual sources of the transistors in the W^+ and W^- cells are connected to the RRAM and SL, respectively. As a result, the V_{gs} of the W^+ cell is fixed since the read voltage (V_{read}) and global reference voltage (V_{ref}) are fixed, while the V_{gs} of W^- cell varies with the conductance of the RRAM. This phenomenon will cause the asymmetrical sensing of W^+ and W^- cells, that is the positive and negative weights with the same absolute value exhibit different $|I_{\text{out}}|$ due to the deviation of V_{gs}, particularly when using transistors with small gate width/length (W/L) ratio in the high-density IMC array. Figure 1(c) shows the I_{out} and its difference between W^+ and W^- cells with different W/L transistors and different weight-conductance mapping values. The symbols are the measured results from the 2T2R cell using the RS architecture. The solid lines are the simulation results, which accord well with the experimental results. The non-linearity in the left part of Figure 1(c) occurs with analog input when the inference voltage enters the transistor's saturation region. This phenomenon aggravates when aggressively scaling the transistor size. Ideally, the I_{out} of W^+ and W^- in two rows with the same W/L should be identical because they represent the same

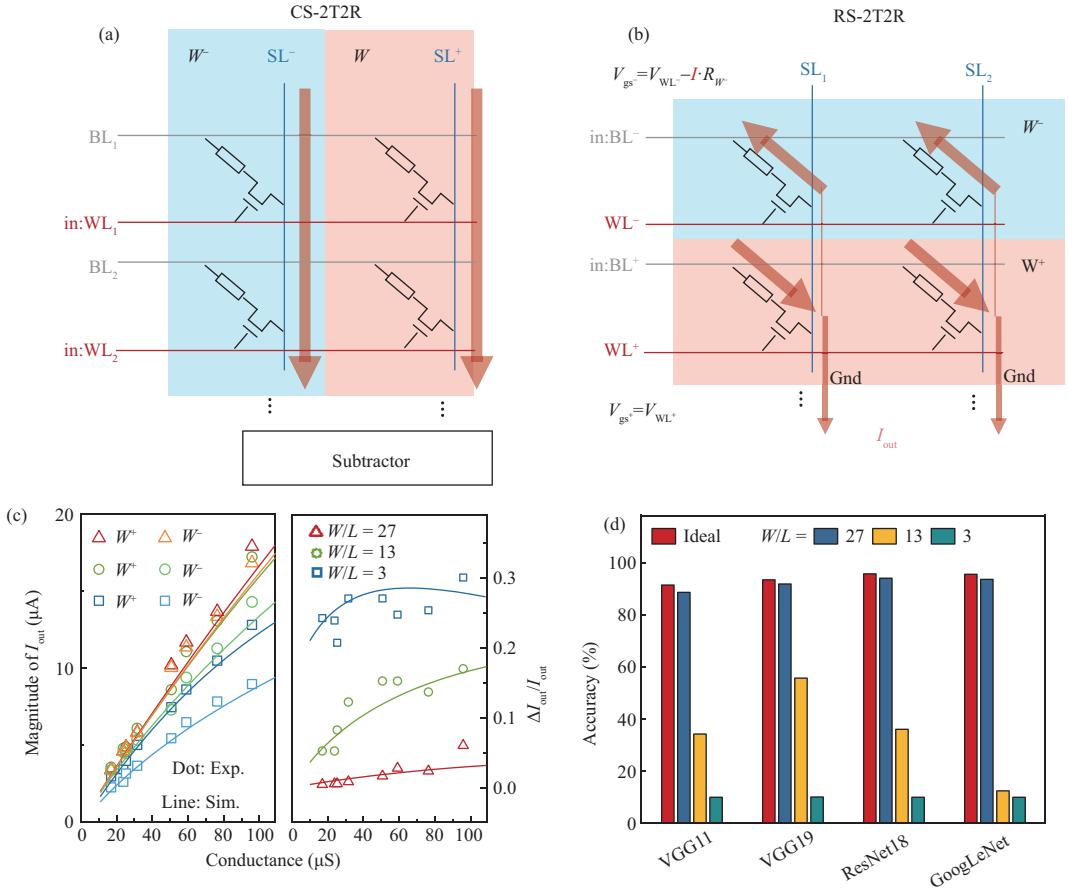

Figure 1 (Color online) (a) Column subtraction architecture and (b) row subtraction architecture for storing the positive part and negative part of the weight. (c) Output current for different conductance states (left) in RS 2T2R and the difference between W^+ and W^- (right) when W/L = 27, 13, and 3. The corresponding V_{read} is 0.2 V. Solid lines represent simulated data and symbols represent experimental data. (d) Cifar-10 recognition accuracy for different networks with various W/L in conventional 2T2R structure.

absolute weight. Yet it can be seen that when the W/L becomes larger ($W/L \geqslant 13$ in this case), there are significant deviations of I_{out}. The deviation can be explained by (2) and (3).

$$I_{\text{out}} = \frac{1}{2}\mu C_{\text{ox}}\frac{W}{L}\left((V_{\text{gs}} - V_{\text{th}})V_{\text{ds}} - \frac{1}{2}V_{\text{ds}}^2\right), \tag{2}$$

$$V_{\text{gs}^-} = V_{\text{WL}^-} - I_{\text{out}} \times R_{W^-}, \tag{3}$$

where I_{out} is the output current of the 1T1R structure. μ is the carrier mobility. C_{ox} is the gate oxide capacitance per area. W/L is the width-to-length ratio. V_{gs} is the voltage across the gate and the source. V_{th} is the threshold voltage. V_{ds} is the voltage between the drain and the source. V_{gs^-} and V_{WL^-} are the V_{gs} and WL voltage in W^- cell. R_{W^-} is the resistance of W^- cell. Eq. (1) describes transistor output current in linear region [21]. For a given RRAM conductance, reducing the W/L results in a smaller output current (I_{out}) and higher transistor resistance. Eq. (2) shows that the V_{gs} in W^- cells is influenced by I_{out}. Thus, the deviated V_{gs} in scaled W^- cells can lead to a severe mismatch between W^+ and W^- cells. Although simply increasing W/L can mitigate the asymmetrical sensing effect, the area cost would be too high for high-density integration, and may eventually counteract the area efficiency of IMC.

Figure 1(d) demonstrates the impact of AWS on the accuracy of typical ANN algorithms. The recognition result of the cifar-10 dataset is examined with four network structures, VGG11, VGG19, ResNet18, and GoogLeNet [22–25] considering the deviations in Figure 1(c). The trend of $\Delta I_{\text{out}}/I_{\text{out}}$ in the right part of Figure 1(c) has also been taken into account in this simulation. The impact of weight distortion from the asymmetrical sensing of W^+ and W^- is introduced under different W/L. All networks demonstrate catastrophic degradation of accuracy at $W/L = 3$ while there is still severe performance

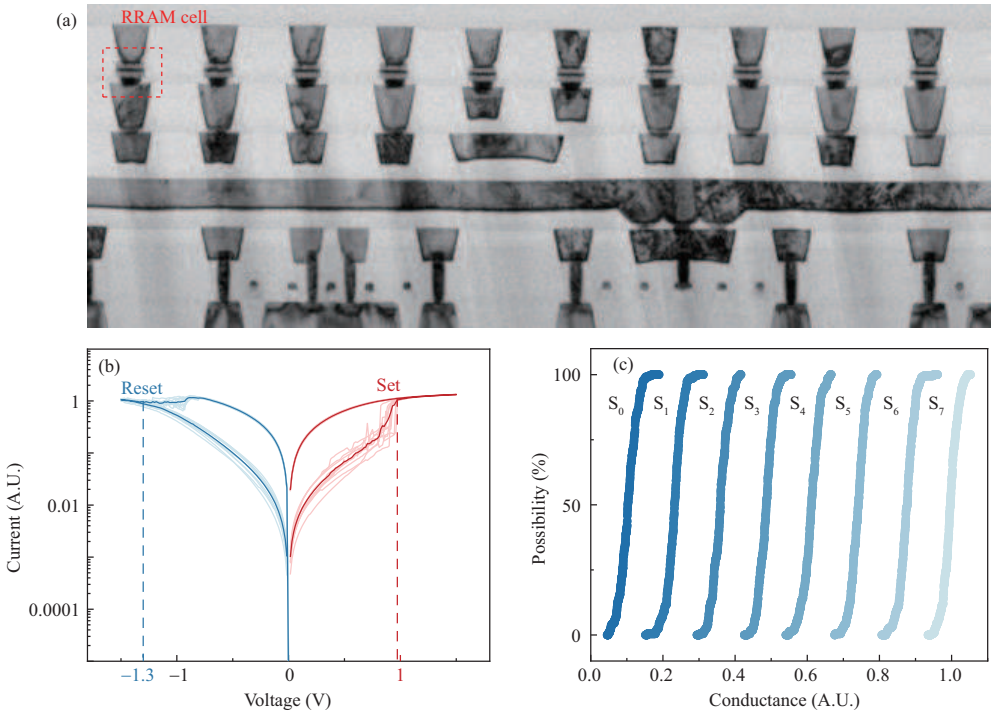

Figure 2 (Color online) (a) TEM image of the fabricated IS-2T2R using 40 nm technology node; (b) typical switching characteristics of the RRAM device; (c) cumulative distribution of 8 conductance levels.

degradation at $W/L = 13$. The functionalities of networks recover when $W/L = 27$, however, a 1%–2% accuracy drop still exists compared to the ideal case. It is important to note that although the simulation is based on the 8-level RRAM, the AWS also exists in IMC networks with binary RRAMs. This is because the values of I_{out} and R_{W-} in (2) are determined by the input feature map and the weight, which also vary in IMC networks with binary RRAMs. All IMC systems based on emerging memories are facing AWS, as long as a signed weight scheme is utilized and the basic memory cell is composed of serially integrated transistor and resistive devices, such as RRAM, PCM, MRAM, and ferroelectric memory. Simply enlarging the transistor size or overdriving the transistor cannot fully rescue the loss of accuracy and may cause other problems such as higher power consumption and reliability issues.

4 Isolated symmetrical 2T2R structure

In this paper, we propose and fabricate an isolated symmetrical (IS) 2T2R cell and array by introducing deep N-well isolation on a standard commercial CMOS 40 nm manufacturing platform. When mapping and inferring weights in the proposed 2T2R cell, the AWS in RS-2T2R can be eliminated in a very compact RRAM-based weight cell. The network accuracy can be improved to ∼90% in the proposed IS-2T2R while maintaining a very small W/L. When the same recognition accuracy is achieved, the proposed cell accomplishes a 42.2% enhancement of the array integration density compared with traditional 2T2R cells for IMC [19, 20].

Figure 2(a) shows the cross-sectional transmission electron microscopy (TEM) image of the fabricated IS-2T2R cell using a commercial 40 nm CMOS process. The electrical characteristics of the basic IS-2T2R cell are shown in Figures 2(b) and (c). Figure 2(b) shows the forming curve and 100 DC switching cycles. The RRAM cells show low operation voltage, which is compatible with the core transistor voltage (V_{dd} = 1.1 V) in the 40 nm technology node. To realize multi-level storage, devices are programmed with the write-and-verify method. Figure 2(c) shows the cumulative probability distribution of 8-level conductance ($S_0 - S_7$), demonstrating a 3-bit storage ability. Combining a 3-bit unsigned value (8-levels RRAM) and a 1-bit sign represented by WL input, 4-bit weight precision can be achieved, which is favorable for the IMC system.

The schematic of the IS-2T2R cell is shown in Figure 3. Figures 3(a) and (d) represent two different operating schemes of the proposed structure. The IS-2T2R cell is composed of two complementary 1T1R

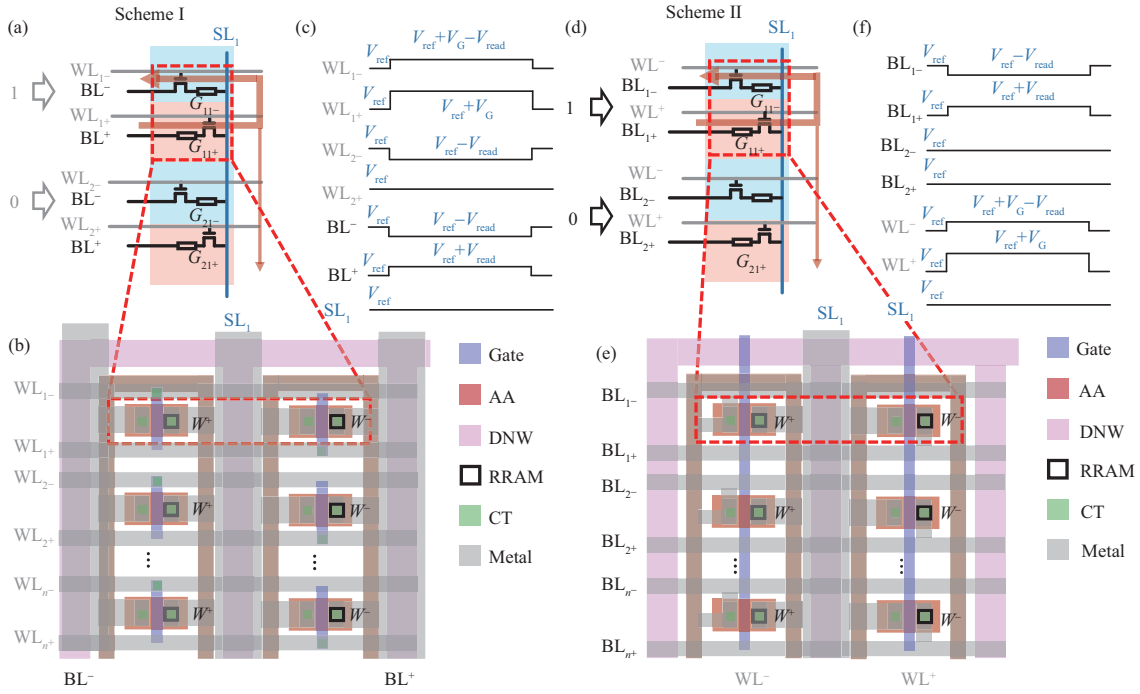

Figure 3 (Color online) Schematic of IS-2T2R in Scheme I (a) and Scheme II (d). Input signals are applied on WLs. Exemplary input of "1" and "0" is applied on top and bottom cells. The layout of IS-2T2R for Scheme I (b) and Scheme II (e). Schematic operation voltage for Scheme I (c) and Scheme II (f).

structures where the RRAM cells are integrated into the opposite sides of the transistor, ensuring that the sources of the IS-2T2R are always connected to a fixed voltage potential while the drains are connected to the input.

Figure 3(a) illustrates Scheme I: the binary input signals are applied to WLs. When sensing the weight, the read voltages are applied to the IS-2T2R structure storing W^+ and W^-. The electrical characteristics of the IS-2T2R are symmetric since $V_{\rm gs} = V_G + V_{\rm ref}$, where W^+ and W^- represent the applied positive and negative input to transistor gates, respectively, and $V_{\rm read}$ is the read voltage, $V_{\rm ref}$ is the global bias, V_G is the gate voltage. The absolute value of $V_{\rm read}$ is constant. Figure 3(c) shows the schematic of the weight-sensing process in Scheme I. When the input is "1", $V_{\rm ref} + V_G$ is applied on W^+, and $V_{\rm ref} + V_G - V_{\rm read}$ is applied on W^-. When the input is "0", $V_{\rm ref}$ is applied to W^+, and $V_{\rm ref} - V_{\rm read}$ is applied to W^-. The fixed read voltage is applied to BLs: $V_{\rm BL+} = V_{\rm ref} + V_{\rm read}$, $V_{\rm BL-} = V_{\rm ref} - V_{\rm read}$. Thus, AWS can be eliminated with the proposed cell structure and bias scheme.

The layout of Scheme I is shown in Figure 3(b). To achieve complete symmetry, the transistors in the IS-2T2R need to be isolated with additional n-well (NW) and deep n-well (DNW), which inevitably causes area overhead compared with the conventional 2T2R structures when using transistors of the same W/L. Also, the W^+ and W^- counterparts are applied with different substrate voltages ($V_{\rm sub}$). However, as shown in Figures 1(c) and (d), the conventional 2T2R structure will have to use transistors with large W/L to avoid AWS. In the proposed structure, the transistor can be scaled to the minimum W/L allowed by the technology node, thus strongly compensating the area overhead.

Scheme II is similar to Scheme I in topology, as shown in Figure 3(d). However, the input signals are applied to BLs. Fixed WL voltage is applied to W^+ and W^- respectively. Different from Scheme I, scheme II supports both binary and analog input from the BLs. However, the analog mapping of the conductance range concerning the weight should be carefully designed to match the transistor's linear region so that the input voltage and output current can roughly follow Ohm's law. The layout of Scheme II is shown in Figure 3(e), the metal wires and transistors are rearranged to fit the signal input type. The schematic of the weight-sensing process of Scheme II is shown in Figure 3(f). Regardless of the input on BLs (binary or analog), $V_{\rm ref} + V_G$ is applied on $\rm WL^+$, and $V_{\rm ref} + V_G - V_{\rm read}$ is applied on $\rm WL^-$. When the input is "0", $V_{\rm BL+} = V_{\rm BL-} = V_{\rm ref}$. When the input is x, $V_{\rm BL+} = V_{\rm ref} + V_{\rm in}$, $V_{\rm BL-} = V_{\rm ref} - V_{\rm in}$. The value of $V_{\rm in}$ represents input x: $V_{\rm in} = x \cdot V_{\rm read}$.

Figure 4(a) shows the performance enhancement after adopting the IS-2T2R cell. In Figure 4(a), both

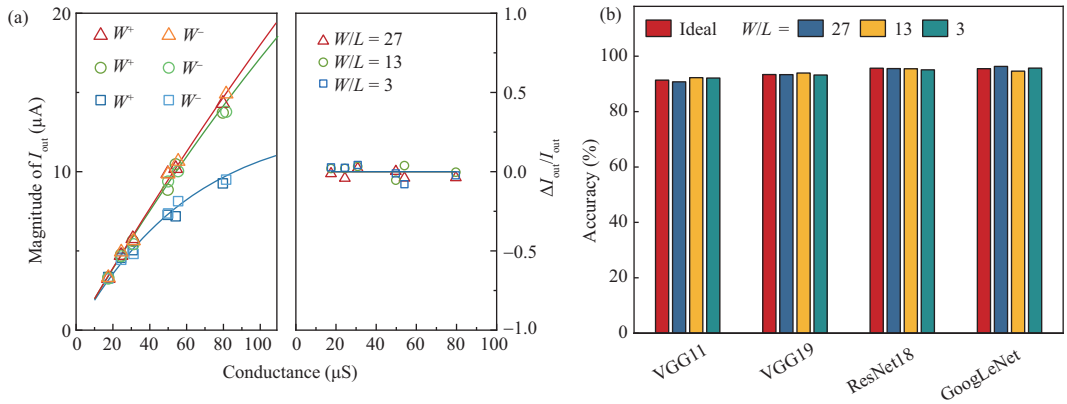

Figure 4 (Color online) (a) Our 2T2R structure output current for different stored conductance states. Corresponding V_{read} = 0.2 V. Lines represent simulated data and symbols represent experimental data. (b) Cifar-10 recognition accuracy for different networks with various gate widths W in our 2T2R structure.

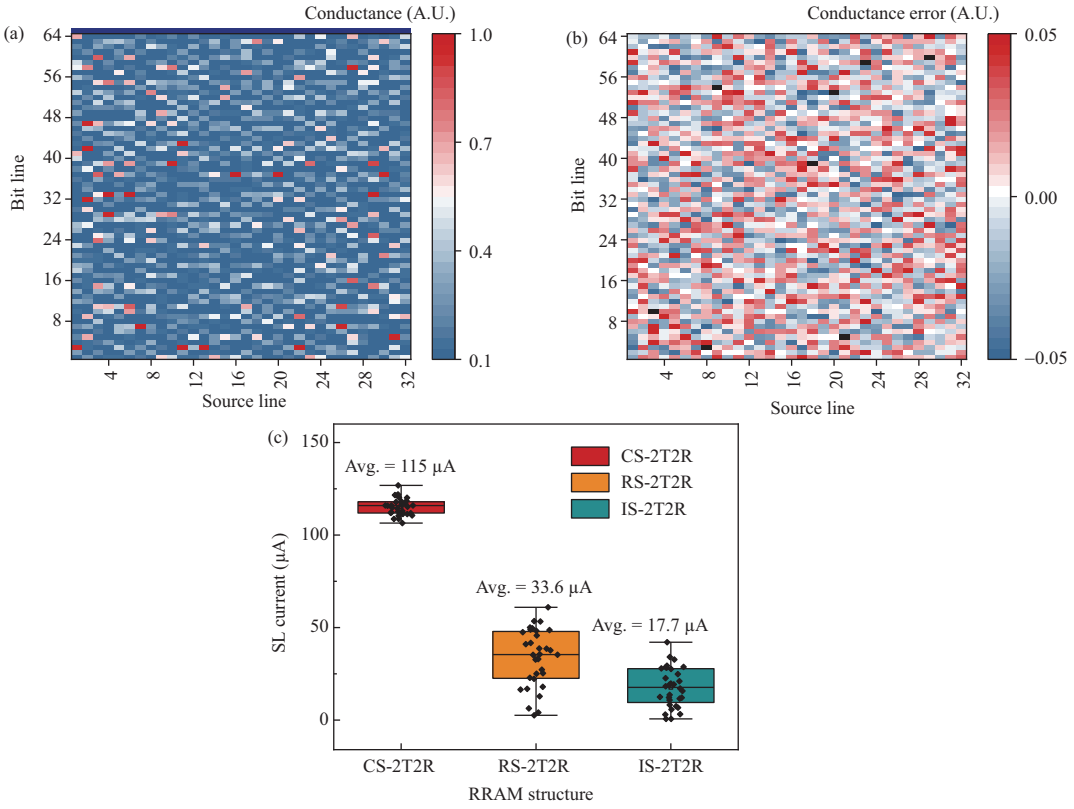

Figure 5 (Color online) (a) Conductance map of a 64×32 2T2R array written with a 32×32 weight matrix; (b) error map of the 2T2R array; (c) SL currents for different RRAM structures.

simulated results and experimental data demonstrate that the deviation of I_{out} is eliminated, indicating the successful removal of the AWS. The network simulation results are shown in Figure 4(b). The results indicate that all networks have high recognition accuracy after implementing the IS-2T2R cell. Even at extremely small W/L, the networks can achieve high recognition accuracy, which allows dense integration of the IMC array by shrinking W/L. An improvement of 42.2% in integration density can be achieved in both Scheme I and Scheme II.

To further demonstrate the advancements and validity of this work. A weight matrix with the dimension of 32 × 32 is written into a 64 × 32 IS-2T2R array. The resistance map of this array is measured and plotted in Figure 5. It is evident that the majority of RRAMs are in a low conductance state, and one of the two adjacent RRAMs at the same SL must be in the high resistance state (HRS), which is consistent with the mapping scheme in (1). The error map is also provided to show the mapping accuracy. The error

Table 1 Comparison with existing studies

	CS-2T2R [19]	RS-2T2R [20]	IS-2T2R (this work)
AWS ratio ($\Delta I_{\text{out}}/I_{\text{out}}$)	∼0	0.019 ($W/L = 27$) 0.115 ($W/L = 13$) 0.253 ($W/L = 3$)	∼0
Accuracy (%, @Vgg11)	91.38	88.56 ($W/L = 27$) 34.24 ($W/L = 13$) 10.12 ($W/L = 3$)	91.38
Area (μm^2)	0.265	0.265	0.184
Operation energy (fJ/op)	57.7	16.8	8.8

is defined as error $= G_m - G_i$, where G_m stands for the measured conductance and G_i stands for the ideal conductance calculated from the weight map. Black lattice means that the |error| < 0.05. Figure 5(b) shows a mapping accuracy that the error of most cells is within $(-0.05, 0.05)$ which is enough for level distinguishing. Only very few cells (<10) in the array exhibit |error| larger than 0.05.

Based on this array, SL currents are measured and compared with other RRAM structures, and the result is shown in Figure 5(c). It could be seen that the SL current of RS-2T2R and IS-2T2R is significantly lower than CS-2T2R because in these structures the current subtraction takes place within each 2T2R cell. Furthermore, the SL current of IS-2T2R also has a noteworthy improvement compared to RS-2T2R because there is AWS in RS-2T2R that brings asymmetry to the current of positive and negative parts, while in IS-2T2R the AWS is eliminated.

The performance of the proposed IS-2T2R structure is summarized and compared with existing studies in Table 1. The cell size is calculated based on the 40-nm node, and the operation energy is estimated using a reference voltage of 0.05 V and a frequency of 100 MHz. It can be seen that the IS-2T2R structure shows advantages in integration density and energy efficiency for IMC applications.

5 Conclusion

In this study, the impact of AWS on conventional RS architecture is investigated. For the first time, an IS-2T2R cell is proposed to improve weight-sensing precision and array integration density in IMC systems. The IS-2T2R array is fabricated using the commercial 40 nm CMOS process. Compared with conventional RS-2T2R cells, the IS-2T2R cell achieves a 42.2% improvement in integration density while maintaining the same recognition accuracy.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2019YFB2205401), National Natural Science Foundation of China (Grant Nos. 61834001, 62025401, 61927901), Beijing Natural Science Foundation (Grant No. L223004), Beijing Nova Program (Grant No. 20220484113), and "111" Project (Grant No. B18001).

References

1 Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 3730–3738
2 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
3 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
4 Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents. 2022. ArXiv:220406125
5 Sebastian A, Le Gallo M, Khaddam-Aljameh R, et al. Memory devices and applications for in-memory computing. Nat Nanotechnol, 2020, 15: 529–544
6 Zou X Q, Xu S, Chen X M, et al. Breaking the von Neumann bottleneck: architecture-level processing-in-memory technology. Sci China Inf Sci, 2021, 64: 160404
7 Wan T Q, Ma S J, Liao F Y, et al. Neuromorphic sensory computing. Sci China Inf Sci, 2022, 65: 141401
8 Chen Q, Wang Z, Lin M, et al. Homogeneous 3D vertical integration of parylene-C based organic flexible resistive memory on standard CMOS platform. Adv Elect Mater, 2021, 7: 2000864
9 Yu Z, Wang Z, Bao S, et al. A new insight and modeling of pulse-to-pulse variability in analog resistive memory for on-chip training. IEEE Trans Electron Dev, 2022, 69: 3100–3104
10 Zheng Q, Wang Z, Gong N, et al. Artificial neural network based on doped HfO_2 ferroelectric capacitors with multilevel characteristics. IEEE Electron Dev Lett, 2019, 40: 1309–1312
11 Zhao Y L, Yang J L, Li B, et al. NAND-SPIN-based processing-in-MRAM architecture for convolutional neural network acceleration. Sci China Inf Sci, 2023, 66: 142401
12 Arnaud F, Ferreira P, Piazza F, et al. High density embedded PCM cell in 28nm FDSOI technology for automotive microcontroller applications. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2020
13 Ielmini D, Wong H S P. In-memory computing with resistive switching devices. Nat Electron, 2018, 1: 333–343

14 Wang Z, Zheng Q, Kang J, et al. Self-activation neural network based on self-selective memory device with rectified multilevel states. IEEE Trans Electron Dev, 2020, 67: 4166–4171
15 Zheng Q, Li X, Wang Z, et al. Mobilatice: a depth-wise DCNN accelerator with hybrid digital/analog nonvolatile processing-in-memory block. In: Proceedings of the 39th International Conference on Computer-Aided Design, 2020
16 Zheng Q, Li X, Guan Y, et al. PIMulator-NN: an event-driven, cross-level simulation framework for processing-in-memory-based neural network accelerators. IEEE Trans Comput-Aided Des Integr Circ Syst, 2022, 41: 5464–5475
17 Yu Z, Wang Z, Kang J, et al. Early-stage fluctuation in low-power analog resistive memory: impacts on neural network and mitigation approach. IEEE Electron Dev Lett, 2020, 41: 940–943
18 Chen W H, Li K X, Lin W Y, et al. A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2018. 494–496
19 Zhou Z, Huang P, Xiang Y, et al. A new hardware implementation approach of BNNs based on nonlinear 2T2R synaptic cell. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018
20 Wan W, Kubendran R, Gao B, et al. A voltage-mode sensing scheme with differential-row weight mapping for energy-efficient RRAM-based in-memory computing. In: Proceedings of IEEE Symposium on VLSI Technology, 2020
21 Razavi B. Design of Analog CMOS Integrated Circuits. New York: McGraw-Hill Education, 2002
22 Krizhevsky A. Learning multiple layers of features from tiny images. 2009. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf
23 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
24 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
25 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1–9