A 28 nm 81 Kb 59–95.3 TOPS/W 4T2R ReRAM Computing-in-Memory Accelerator With Voltage-to-Time-to-Digital Based Output

Keji Zhou, Xinru Jia, *Graduate Student Member, IEEE*, Chenyang Zhao, Xumeng Zhang, *Member, IEEE*, Guangjian Wu, Chen Mu, Haozhe Zhu, *Member, IEEE*, Yanting Ding, *Graduate Student Member, IEEE*, Chixiao Chen, *Member, IEEE*, Xiaoyong Xue, *Member, IEEE*, Xiaoyang Zeng, *Member, IEEE*, and Qi Liu

***Abstract*— Computing-in-memory (CIM) based on Resistive RAM (ReRAM) can effectively improve the energy efficiency and throughput of artificial intelligence (AI) edge devices. However, due to the complex hardware structure and the non-ideal factors of the circuit, improving the processing precision will sharply reduce the energy efficiency and reliability of AI computing. In this work, a high-performance ReRAM-based CIM accelerator is presented to solve the above problems using: 1) a 4T2R cell to replace the traditional 2T2R cell for weight storage with higher on/off ratio and smaller computing current; 2) a pulse width modulation converter to realize linear input with lower performance cost; 3) a voltage-to-time-to-digital based converter to reduce the power consumption and area of the output circuit. An 81Kb ReRAM based accelerator was designed using 28nm process with 1-4b input/weight/output. To verify the reliability of the accelerator, non-ideal factors are added in training and testing. For evaluation, a network is built for CIFAR-10 based on the proposed accelerator. The proposed accelerator achieves a high processing frequency of 167-500 MHz and an energy efficiency of 95.3-59 TOPS/W for 1-4b precision operation with an FoM (input-precision × weight-precision × energy efficiency) 3.6× higher than prior work.**

***Index Terms*— ReRAM, 4T2R, neural network (NN), computing-in-memory (CIM), artificial intelligence (AI).**

Manuscript received 25 April 2022; revised 7 July 2022; accepted 2 August 2022. Date of publication 5 August 2022; date of current version 19 December 2022. This work was supported in part by the National Key Research and Development Program under Grant 2018YFA0701500; in part by the National Natural Science Foundation of China under Grant 62104040, Grant 61732020, Grant 61825404, Grant 61804167, and Grant 62104044; in part by the Major Science and Technology Special Project of China under Grant 2017ZX02301007-001; in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB44000000; in part by the China Postdoctoral Science Foundation under Grant 2022M710723; in part by Biren Technology; and in part by the Ministy of Education Innovation Platform. This article was recommended by Guest Editor S. Yu. *(Corresponding author: Qi Liu.)*

Keji Zhou, Xumeng Zhang, Guangjian Wu, Chixiao Chen, and Qi Liu are with the Frontier Institute of Chip and System and the State Key Laboratory of ASIC and System, Fudan University, Shanghai 201203, China, and also with the Shanghai Qi Zhi Institute, Shanghai, Xuhui 200232, China (e-mail: qi_liu@fudan.edu.cn).

Xinru Jia, Chen Mu, and Yanting Ding are with the Frontier Institute of Chip and System and the State Key Laboratory of ASIC and System, Fudan University, Shanghai 201203, China.

Chenyang Zhao, Haozhe Zhu, Xiaoyong Xue, and Xiaoyang Zeng are with the State Key Laboratory of ASIC and System, Fudan University, Shanghai 201203, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JETCAS.2022.3196678.

Digital Object Identifier 10.1109/JETCAS.2022.3196678

I. Introduction

DEEP neural network (DNN) in artificial intelligence (AI) devices has developed rapidly in recent years. With the massive parameters and computations, DNN obtains much higher accuracy and robustness than traditional machine learning algorithms [1]–[3]. However, the energy efficiency and throughput of the Von Neumann architecture are not enough to meet the development needs of neural networks, especially for the edge device with limited energy budget and rapid response requirement [4], [5]. The bottleneck between the Von Neumann architecture and neural networks is caused by the massive movement of parameters and intermediate data, which is usually referred to as memory wall.

To break the memory wall, computing in memory (CIM) architecture is a promising solution. CIM can combine data access and computing operations by embedding computing functions into the memory, which can reduce the overhead of data movement, and realize large-scale parallel operation [6],[7]. In the past, the CIM design in DNN was mainly based on SRAM or emerging non-volatile memory (NVM) devices [8]–[14]. However, the SRAM cell structure for storing weights is usually composed of 6T-10T, causing a large area overhead. In addition, owing to the volatile characteristics, SRAM-based DNNs need a constant power supply to store the weight data, which will consume a lot of standby power.

Storing data with the NVM device can effectively reduce the area of weight array under the same technology node, and the nonvolatile characteristic of NVM device is also helpful to save data after power off. Due to the high speed, small area, and process compatibility of ReRAM (Resistive RAM), a lot of work on ReRAM-based CIM has been put forward for edge devices. As shown in Fig. 1, the primary components of a ReRAM-based CIM macro are weight array, Digital-to-Analog Converter (DAC), and Analog-to-Digital Converter (ADC). Among these, the weight cell usually adopts 2T2R or crossbar structure, utilizing different resistance states to represent weight [24], [25]. When multiple rows of cells are activated at the same time, the sum of the current or voltage variation on the bit line can characterize the multiply-and-accumulate (MAC) operation result. As for the input circuit, the selected word line needs to be charged and discharged for

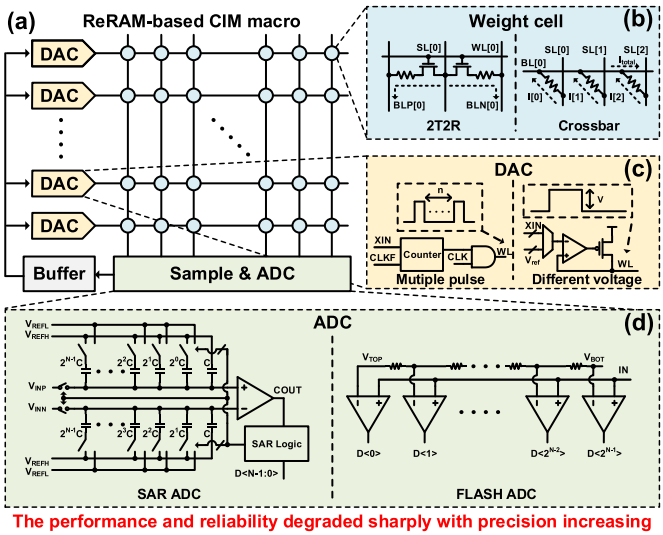


Fig. 1. Concept of ReRAM-based CIM macro: (a) macro, (b) weight cell, (c) DAC, (d) ADC.

many times, or each word line needs to be designed with a high gain clamping circuit, which can transfer multi-bit input data into multiple pulses or different voltages [14], [15]. The output circuit needs to sample the calculation result from bit lines first, and then use SAR ADC or FLASH ADC to convert the calculation result into digital output [15], [25]. With the combination of NVM and CIM, the performance of neural network processing has been optimized greatly.

With the success of AI, the application scenarios continue to expand, and the processing targets become more complex. To deal with the complex targets, CIM needs high computing precision to ensure sufficient accuracy. However, existing ReRAM-based CIM aiming at optimizing ultra-low precision computing, which has two challenges on higher precision computing for edge devices: (1) The energy and area efficiency degraded sharply with increased precision. The interface circuits of existing research schemes are mainly realized by high overhead mixed-signal circuits, such as DAC/ADC, which leads to a sharp increase in the overhead of area, power consumption and delay when improving the computing precision; (2) The existing ReRAM-based CIM circuit has poor reliability in the higher precision analog calculation. The calculation results of ReRAM-based CIM are affected by non-ideal factors, such as nonlinearity and process variation, and the calculation error will be further increased after improving the precision, resulting in a sharp decline in the calculation accuracy of the neural network.

In this work, several techniques are proposed to deal with the above-mentioned performance and reliability issues, which are as shown below.

1) A 4T2R Weight Cell to Enhance the Sense Margin and Reduce the Converging Current

2) A Pulse Width Modulation Converter to Achieve Linearity Input Without Multiple Charging and Discharging Operations

3) A Voltage-to-Time-to-Digital Based Converter to Sample Higher Precision Results With Lower Energy and Area Overhead

The rest of this paper is organized as follows. Section II presents the structure of high-performance ReRAM-based CIM accelerator. Section III shows the simulation results of the proposed accelerator. Conclusions are drawn in Section IV.

II. Structure of Proposed CIM Accelerator

The architectural block diagram of the proposed ReRAM-based CIM accelerator is shown in Fig. 2. The accelerator is composed of computing in memory macro (CIMM), test circuit, dataflow bus, and global control circuit. Due to the influence of IR drop and leakage current on long interconnect lines, the size of the array is limited, so the accelerator is divided into multiple macros. Therefore, the calculation of the same layer network can be distributed to different macros, and the results are transmitted and summarized through the dataflow bus. As for CIMM, each macro is composed of 4T2R computing-in-memory array (CIMA), pulse width modulation input (PWM), voltage-to-time-to-digital converter based output (VTDC), row decoder, column decoder, read circuit, write circuit, and sub-control circuit.

As mentioned above, this accelerator supports higher precision processing with lower performance overhead. The input data are transferred into different width pulses by shared PWM input in the global control circuit, the corresponding MUX array is turned on by the decoder and the pulses are transmitted to the specified word lines. Multi-bit weights are realized by sampling the charge variation on multiple BLs at different wire lengths after charge sharing. Lastly, the calculation results are converted into different width pulses by VTDC, and then converted into digital output. In the following, the designed details of these critical components are introduced.

A. High Sense Margin 4T2R Array for Weight Storage

To enhance the sense margin and reduce the converging current of the weight array, we utilize the 4T2R cell to replace the traditional 2T2R cell for weight storage. Two similar 4T2R structures were used for search operation and in-memory dot operation in previous works [17], [18]. However, they apply voltage from the direction of ReRAM to form the partial voltage to obtain the large on/off ratio, and this will cause two problems. The first is that applying a high voltage on the ReRAM is easy to cause pseudo-reset. The second is that applying the high voltage on the ReRAM will activate a whole column of cells, which is easy to form excessive current and voltage drop on the interconnection line. Therefore, we apply voltage on the access transistor in our design for inference operation, trading off part of the on/off ratio for higher reliability. In addition, we have designed a weight extension scheme based on the small converging current of 4T2R, which uses existing line capacitance instead of large additional independent capacitance.

Fig. 3 shows the proposed high sense margin 4T2R cell. As shown in Fig. 3(a), each 4T2R cell consists of four NMOS transistors and two ReRAMs. Two access transistors (N3 and N4) and two ReRAMs (R1 and R2) form a 2T2R cell for normal access operation, two additional computing transistors (N1 and N2) are added for an inference operation. Fig. 3(b) is

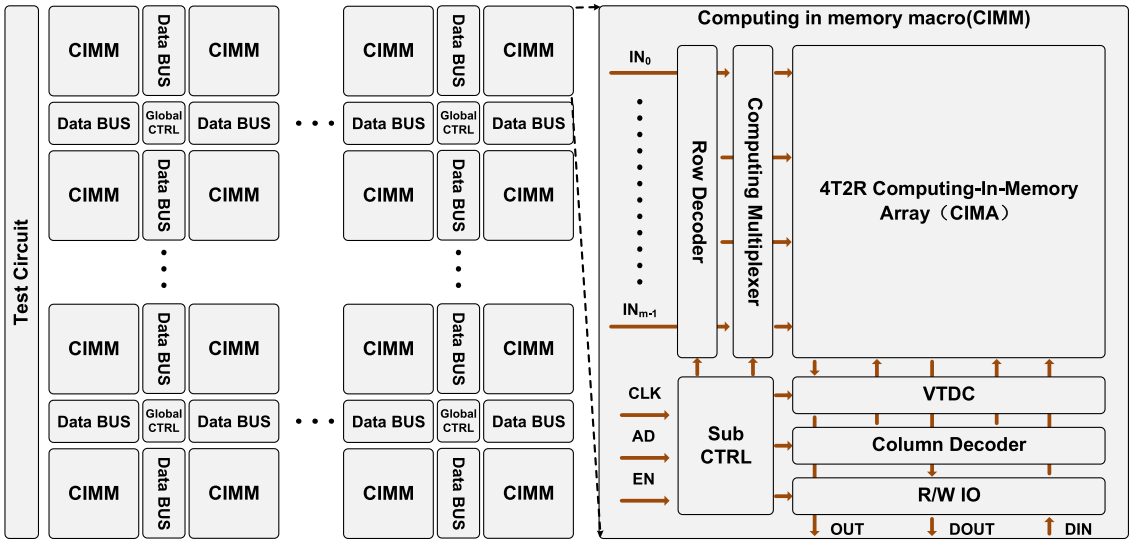
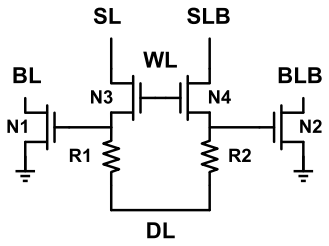


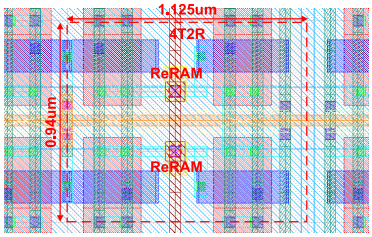
Fig. 2. The architectural block diagram of the proposed ReRAM-based CIM accelerator, comprising computing in memory macro (CIMM), test circuit, dataflow bus, and global control circuit.



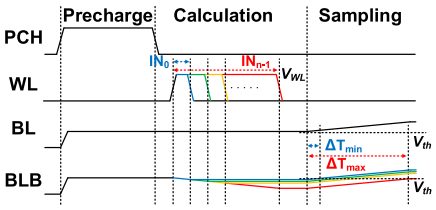
(a)

Ternary Weight	R1	R2
-1	HRS	LRS
0	HRS	HRS
1	LRS	HRS

(b)



(c)



(d)

Write data=1	R1	R2	WL	SL	SLB	DL
1st Step	X=>LRS	X	VDD	Vset	VSS	VSS
2nd Step	LRS	X=>HRS	VDD	Vp	VSS	Vreset
Write data=-1						
1st Step	X=>HRS	X	VDD	VSS	Vp	Vreset
2nd Step	HRS	X=>LRS	VDD	VSS	Vset	VSS
Write data=0						
1st Step	X=>HRS	X	VDD	VSS	Vp	Vreset
2nd Step	HRS	X=>HRS	VDD	Vp	VSS	Vreset
Read						
	X	X	VDD	VSS(Iread1)	VSS(Iread-1)	Vread

(e)

Fig. 3. (a) The structure of 4T2R cell, (b) the truth table of 4T2R cell, (c) the layout of 4T2R cell, (d) the waveform for the inference operation of activating one cell with weight value "1," (e) the access operation condition of 4T2R cell.

the truth table of the 4T2R cell. LRS and HRS represent low resistance state and high resistance state, respectively. When the weight value is 1(-1), the R1 and R2 are LRS(HRS) and HRS(LRS), respectively. When both ReRAMs are HRS, the weight is 0. Fig. 3(c) shows the layout of the 4T2R cell. The area of 4T2R is about 1.06um^2, which is about 2.7 times that of the traditional 2T2R cell. However, owing to the small convergence current, the weight array composed of 4T2R can directly use line capacitance for MAC operation instead of large independent capacitance, and the area cost will be further degraded with weight precision increased.

The inference operation is realized by activating multiple rows of cells through the word lines (WL) for MAC calculations. Take the multiplication operation of activating just one cell that has the weight value of "1" as an example, and the waveform is shown in Fig. 3(d). The inference operation of the array can be split into three phases, which are precharge, calculation, and sampling. Before activating the selected row for calculation, the selected bit lines are precharged first, then the complementary search lines SL and SLB of selected columns are charged to VSL, the WL and data line (DL) are discharged to the VSS. During the calculation phase, the input data are converted into pulses to charge the selected WLs to VWL with different pulse widths, where the IN_0 and IN_{n-1} represent the minimum and maximum input data. By appropriately adjusting the amplitude of VSL and VWL, the partial voltage between the access NMOS transistor and ReRAM can open or close the computing transistors N1 and N2. When the cell stores weight "1", N2 is turned on to discharge BLB as the word line is activated. Therefore, utilizing the input pulses with different widths, the multiplication operation result is transmitted to the voltage difference between the pair of bit lines. Then, by charging BL and BLB through the sampling circuit, the voltage difference can be converted into the time difference when BL and BLB exceed the threshold voltage V_{th}.

In particular, compared with the 2T2R cell, the computing transistors of the 4T2R cell can isolate the bit line voltage and ReRAM device. Owing to the I-V nonlinearity of ReRAM, the resistance of ReRAM is different under different read voltage. During the calculation phase, as for the 2T2R cell, the conductance of ReRAM will be changed with the discharging of bit lines. Therefore, the calculation result will deviate from the ideal value, which will affect the reliability of the network. On the contrary, as for the 4T2R cell, because the partial voltage between the access transistor and the ReRAM is stable, the calculation results on the bit lines will not be directly affected by the I-V nonlinearity of the ReRAM, which is helpful to improve the reliability. Similar to the multiplication operation, the MAC operation is achieved by activating multiple rows at the same time, and the calculation result is corresponding to the sum of the voltage difference between the pair of bit lines. After the calculation phase, the charge sharing operation is applied on the multiple bit lines corresponding to different weights, and then the voltage difference between the pair of bit lines can be sampled and converted into digital output with the proposed VTDC.

The conditions of write and read operation are shown in Fig. 3(e). The write operation is divided into two steps. Taking the writing of "−1" as an example, the R1 is reset to HRS in the first step, and then the R2 is set to LRS in the second step. To set R1, the access transistor N3 is activated by WL, Vset is applied on SL while the DL and SLB are grounded. To reset R1, the access transistor N3 is also activated by WL, Vreset is applied on DL while SL is grounded. To restrain write disturbance, a protection voltage Vp is applied on SLB, which can make the voltage difference between the two ends of R2 much less than Vreset. The set/reset operation of R2 is similar to R1. To ensure the reliable access and computing operation of the proposed accelerator, programming ReRAM to the corresponding resistance state accurately is very important. The accurate HRS/LRS transition of ReRAM can be achieved by adjusting set/reset conditions such as compliance current, programming pulse, and employing the interactive write-verify method [19], [20]. For read operation, the access transistor N3 and N4 are activated by WL, Vread is applied on DL while SL and SLB are grounded, and the differential read current Iread1 and Iread-1 will be derived from SL and SLB.

The sense margin of the ReRAM-based weight cell mainly depends on the current on/off ratio $I_{on/off}$. As for the 2T2R cell, the MAC operation is mainly realized by applying a constant voltage to ReRAM with different resistance states, and then comparing the current flowing through ReRAM. The $I_{on/off}$ of the 2T2R cell directly depends on the ratio of high and low resistance states of ReRAM. Due to the process of the ReRAM, the sense margin of 2T2R cell is probably not enough for higher precision in-memory-computing. As for the proposed 4T2R cell, the gate voltage of computing transistors N1, N2 is the partial voltage between the access transistor and the ReRAM. By adjusting the VWL and VSL, the computing transistors can be used as an amplifier to amplify the difference of the calculated current through different resistance ReRAM. Therefore, the $I_{on/off}$ of the weight cell can be effectively improved, which can increase the reliability of higher precision

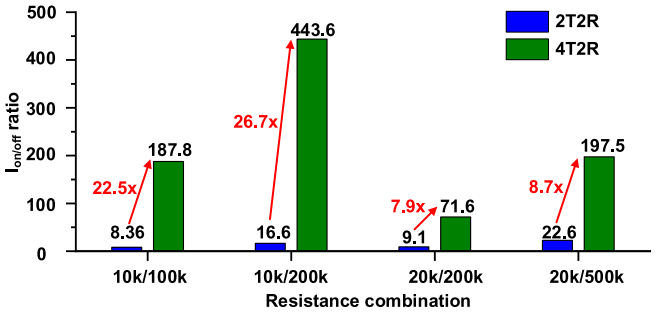


Fig. 4. The comparison of sense margin between 4T2R-based and 2T2R-based voltage sense methods under different resistance combination.

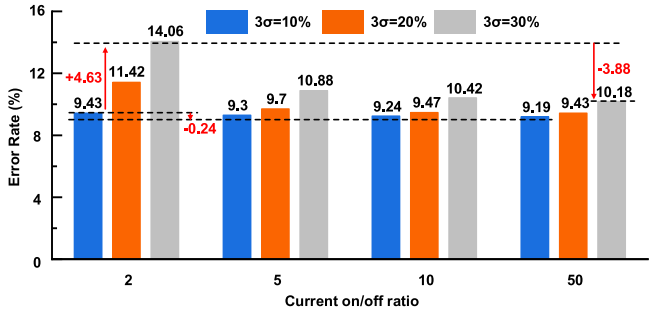


Fig. 5. The MC simulation results of error rate with different current on/off ratio and ReRAM variation.

computing. Fig. 4 gives the sense margin comparison between 4T2R-based and 2T2R-based voltage sensing methods under different resistance combinations. The simulation results show that, by using the proposed 4T2R cell to replace the 2T2R one, when the typical value of ReRAM high and low resistance state is 200K/20K, the current on/off ratio $I_{on/off}$ can be increased from 9.1 to about 71.6, and the $I_{on/off}$ can be further raised with resistance combination changed.

To evaluate the reduction of error rate before and after applying the 4T2R structure, the Monte Carlo (MC) simulations of error rate are carried out under different current on/off ratios and ReRAM variation, the mean error rate is shown in Fig. 5. This evaluation is based on a neural network with 8 layers and 2bit processing precision for CIFAR 10, and the specific structure of the neural network is shown in Table I. The simulation results show that when the local variation is 30%(3σ), the error rate can be reduced by about 3.88% by increasing the current on/off ratio from 2 to 50. However, when the local variation is 10%(3σ), the error rate can be reduced by about 0.24% by increasing the current on/off ratio from 2 to 50. The reduction of error rate is mainly due to the fact that increasing the on/off ratio can reduce the influence of variation on high resistance paths. Although 2T2R and 4T2R adopt differential outputs, the influence of current in the high resistance path cannot be completely offset under the influence of local variation, especially when the on/off ratio is relatively low. Therefore, the 4T2R structure is suitable for improving the accuracy of the network when the device has a low on/off ratio and large variation.

As mentioned above, apart from the higher on/off ratio, another advantage of the proposed 4T2R cell is the small convergence current. With the small convergence current, the weight extension operation can utilize the line capacitance instead of large independent capacitors, which can make up

TABLE I
PARAMETERS OF NETWORK

Layer	Configuration	In-Channels	Output-Channels
Conv-1	Conv 3×3	3	128
Conv-2	Conv 3×3	128	128
Conv-3	Conv 3×3	128	256
Conv-4	Conv 3×3	256	256
Conv-5	Conv 3×3	256	512
Conv-6	Conv 3×3	512	512
Conv-7	Conv 3×3	512	1024
FC-1	FC	1024	10

the area overhead caused by the large cell. Fig. 6 shows an example to illustrate the weight extension based on the 4T2R weight cell, the weight precision is set to 4bit for convenient illustration, in which BLP0 and BLN0 correspond to the bit lines of the least significant bit, BLP3 and BLN3 correspond to the bit lines of the most significant bit, the global bit line GBLP and GBLN correspond to the inputs of VTDC. At the beginning of the inference operation, all selected BLs and BLBs are precharged firstly. Then, during the calculation phase, multiple rows of cells are activated with WL pulse of different widths, the BLP and BLN are discharged, and the discharge difference on the pair of bit lines corresponds to the MAC operation result. After the calculation phase, the WL is discharged to shut down the cells, and the voltage difference between the multiple pair of bit lines needs to be sampled and weighted summed. Therefore, the charge share switches CS0-CS3 are closed, and the weight extension switches S2-S4 are opened. Take BLP for example, 1/8 Q_{BLP0}, 1/4 Q_{BLP1}, 1/2 Q_{BLP2}, and Q_{BLP3} are summarized and shared across the selected bit lines with different lengths, which is equivalent to multiplying the corresponding input by 1/8 w_{i0}, 1/4 w_{i1}, 1/2 w_{i2}, and w_{i3}, where the w_{ij} represents the weight in the i-th row and j-th column. Thus, the weight extension is achieved by utilizing the line capacitance. When the weight precision changes, such as changing from 4bit to lower precision, the other columns can be multiplexed by controlling switches CS0-CS3 and S2-S4 to avoid idling the weight cell.

Owing to the big convergence current of 2T2R, the voltage sensing method based on 2T2R needs the binary-weighted capacitor to hold and sample the MAC result. However, the cost of the capacitor is increased exponentially with the increase of precision, which is not suitable for higher precision calculation. By contrast, utilizing the 4T2R to extend weight only need extra switch transistors, which is much smaller than the large independent capacitors. In order to evaluate the compensation effect on area overhead by utilizing line capacitance, we compare the area overhead of 4T2R and 2T2R with different weight extension and sampling schemes. The sampling scheme of 4T2R is based on line capacitance, and the sampling scheme of 2T2R uses SAR ADC with independent capacitances, the structure of SAR ADC refers to [21]. In the simulation, the smallest sampling reference capacitance of SAR ADC is set to 1fF, and the number of storage cells on the single bit line is set to 144. The average area cost of each cell includes the area of cells, independent capacitances, and switch transistors. The comparison result under different

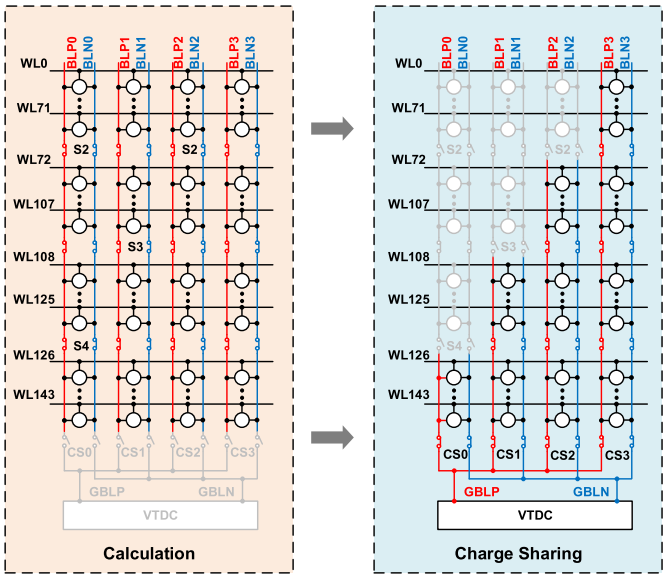


Fig. 6. An example to illustrate weight extension with 4bit weight and 144 rows of cells, in which charge sharing is implemented based on line and device capacitance instead of large independent capacitance.

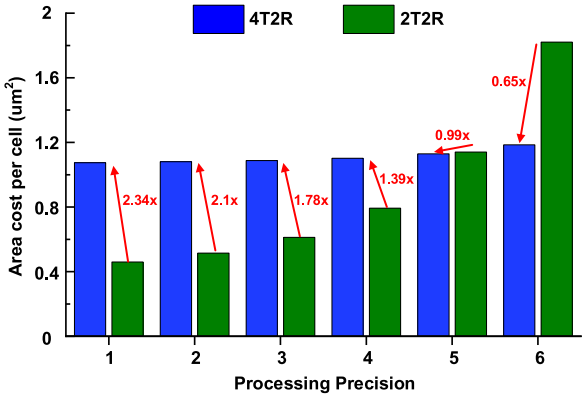


Fig. 7. The area cost per cell comparison of 4T2R and 2T2R under different processing precision, and the area evaluation includes the area of cells, independent capacitances, and switch transistors.

processing precision is shown in Fig. 7. We can observe that with the increase of precision, the area overhead per 2T2R cell increases sharply, exceeding that of 4T2R at 5-bit precision. This result proves using existing line capacitance instead of large additional independent capacitance is a valuable method for higher precision calculation.

B. Linear Pulse Width Modulation Input

The fundamental of multi-bit input is to transform the digital input into linear analog signals, such as time domain and voltage domain. As for the voltage domain, the common way is to use different WL voltage levels for representing multi-bit input [14]. This method can represent multi-bit input in a fixed cycle, and only needs to switch the WL just once. However, this method needs a DAC for each row, resulting in a large area and power consumption overhead. In addition, this method also suffers from linearity issues in the different load currents and the presence of process, voltage, and temperature (PVT) variations. As for the time domain, there are two classical methods. The first method is to use multiple WL pulses to represent multi-bit input. As mentioned in [15], the multiple

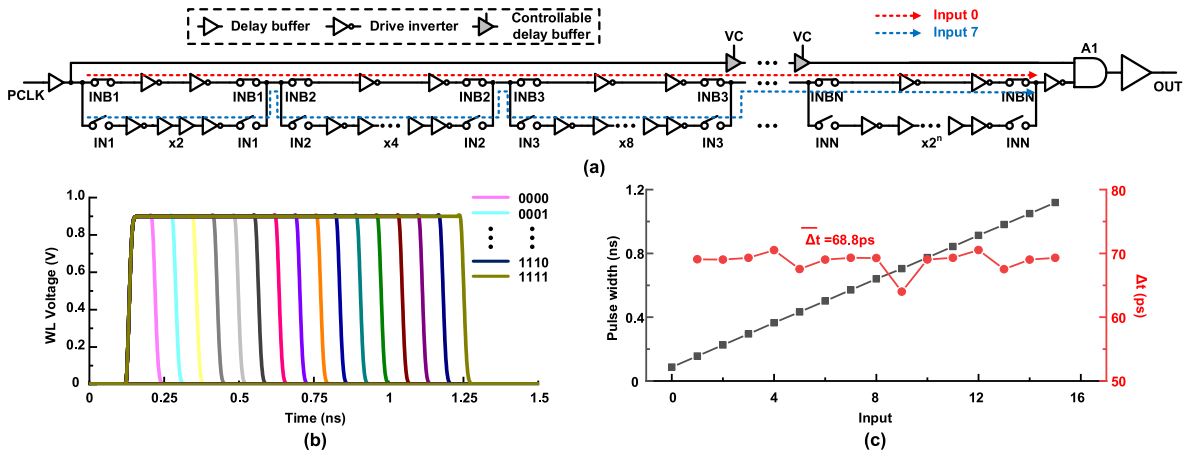


Fig. 8. (a) The structure of proposed linear pulse width modulation input circuit, (b) the waveform of word line pulse corresponding to 4bit different input data, (c) the statistical diagram of pulse width and the increase of width under different input data.

input pulses can be implemented by fully digital components, which shows great linearity. However, this approach needs a counter for each row, which caused hardware overhead. In addition, owing to the large line capacitance, the frequent switch of WL will sharply increase the power consumption and latency of the input circuit. The second method is to realize the pulse width modulation through the delay chain [16]. Compared with representing multi-bit input with multiple pulses, this method only needs to charge and discharge the WL once, which has lower power consumption and faster throughput. However, the linearity of this method is worse than the above method with multiple pulses.

In order to represent time-domain multi-bit input with lower cost and high linearity, the optimized pulse width modulation input circuit is utilized to in our design, the structure of the proposed input circuit is shown in Fig. 8(a). By adopting the design of the programmable delay chain, the length of the delay chain can be changed by controlling different switches signals IN1-INN and their complementary signals INB1-INBN. Therefore, the output pulses are realized with different widths corresponding to n bit input, which reduces the performance overhead caused by multiple charging and discharging of word lines. Owing to the different rising and falling actions of the output signal under zero and other inputs, the traditional pulse width modulation circuit will degrade the linearity of the output signal. In order to improve the linearity, the proposed input circuit utilizes the timing competition to produce different width pulses. When PCLK is switched to high, the switch signal will be transmitted to the inputs of AND gate A1 through two paths with different delays. For smaller delay, the switch signal will arrive at A1 first through the upper path, and the output of A1 will be turned to 1. When the switch signal is transmitted to the lower input of A1 through another path, A1 is closed and output is turned to 0. Since the delay of the upper path is less than that of the lower path, even if the input is 0, a small pulse will be output, which can ensure that all outputs have the same rise and fall action. However, since the 0 input pulse will lead to unnecessary charge and discharge on the array, the controllable delay buffer is added to adjust the delay on the upper path to reduce the unnecessary power consumption. In addition, the delay buffer is used to control the difference of different pulse widths, and the drive buffer has a stronger driving ability, which is used to reduce the influence of parasitic capacitance change on linearity.

To understand how the delay chain works, let us consider two input values of 0 and 7. For input value of 0, IN1-INN are all low level, and other INB1-INBN signals are high level, which means the PCLK will reach the lower input of A1 through the shortest delay path. The transmission path is shown by the red dotted line, and the output signal is a narrow pulse. For input value of 7, IN1-IN3 are changed to the high level, and INB1-INBN signals are changed to the low level, remaining signals keep unchanged. The 14 delay buffers will be added to the long transmission path of PCLK. The transmission path is shown by the blue dotted line, and the width of the output signal will be increased accordingly.

Instead of adding a clamp or a counter on each word line, the chip only needs to connect a small number of proposed input circuits to the activated rows through the multiplexer and bus design. Therefore, as long as the number of proposed input circuits is more than the maximum number of activated rows, the calculation needs of the whole macro can be satisfied, which can reduce the hardware overhead. Fig. 8(b) is the waveform of word line pulses corresponding to different 4bit inputs, and Fig. 8(c) is the statistical diagram of pulse width and the increase of width Δt under different input data. The simulation results show that the mean of Δt is about 68.8ps, and the maximum deviation of Δt is 4.8ps. Therefore, the input circuit has good linearity under different input data, which can effectively represent high precision input data, and realize the CIM input circuit with low overhead.

C. High Performance Voltage-to-Time-to-Digital Based Converter

The output of the CIM accelerator usually adopts complex analog to digital converter to convert the calculation result into digital output, such as SAR ADC and FLASH ADC. However, for SAR ADC, it needs a large amount of sampling capacitors, and multiple comparison operations according to processing precision [25]. On contrary, FLASH ADC can complete the multi-bit comparison operations in just one cycle,

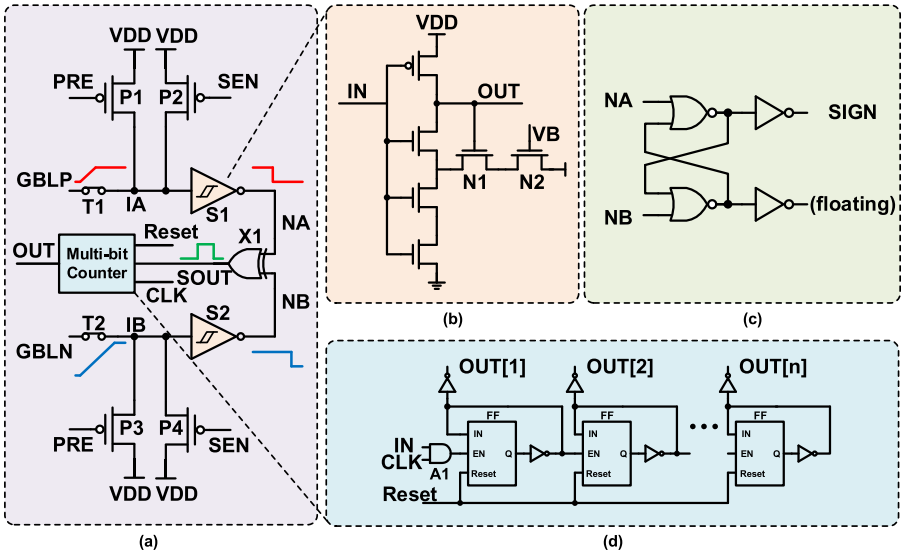


Fig. 9. (a) The schematic of the proposed voltage-to-time-to-digital based converter, (b) the semi Schmitt trigger with controllable hysteresis, (c) the schematic of NOR-based RS-latch., (d) the schematic of the multi-bit counter.

but at the cost of using multiple comparators [15]. Because the calculation output is an analog signal, each column needs an ADC to sample and convert the calculation results accurately. Therefore, in neural network processing, ADC is the main source of power consumption and area overhead of the chip, and the overhead will increase sharply for higher precision. In order to solve the above problems, we propose the voltage-to-time-to-digital based converter (VTDC) to realize the output circuit. The proposed output circuit utilizes the Schmitt trigger and XOR gate to convert the voltage into pulses with different widths, and then converts the pulse into a digital output signal through a multi-bit counter. Compared with sense amplifier and operational amplifier, Schmitt trigger can continuously compare input signals without constant bias current, and has a simpler structure with less energy consumption. In addition, utilizing the XOR gate can easily deal with the subtraction of positive and negative results. Therefore, our proposed output circuit does not need multi-cycle comparison or using a large number of comparators to sample signals, which reduces the energy consumption and area overhead caused by higher precision processing of neural network.

The schematic of the proposed output circuit is shown in Fig. 9(a). The VTDC includes voltage to time converter (VTC) and time to digital converter (TDC) two parts. The main components of VTC are sampling circuits S1, S2, and an XOR gate X1, and TDC is realized by a multi-bit counter.

In the VTC, we use a Schmitt trigger to sample the voltage on the bit line instead of the traditional operational amplifier (OPA) or sense amplifier (SA). The structure of the adopted semi-Schmitt trigger is shown in Fig. 9(b). By adding transistor N1 to the pull-down path, the semi Schmitt trigger structure is formed. Compared with the traditional Schmitt trigger, it can complete a one-way sampling operation with less overhead. In order to flexibly adjust the threshold of the sampling circuit to satisfy different processing precision, we add the control transistor N2. By controlling the gate voltage VB of N2, we can flexibly adjust the threshold voltage of the trigger, which can flexibly adapt to the needs of different processing precision. Based on this sampling circuit, the voltage on the bit lines can be converted into pulses with different falling edges and transmitted to the input of X1, which are NA and NB. Due to the different voltages on the global bit line GBLP and GBLN, NA and NB will discharge to the VSS at different times. When one of the nodes is discharged to the VSS first, the output of X1 will switch to VDD. When both nodes are discharged to the VSS, the output of X1 will switch to VSS. Based on the above design, the VTC can transform the voltage difference on the pair of bit lines into a pulse with different widths, providing linear voltage to time conversion with high noise immunity and less performance overhead. In addition, an RS latch is added for the output sign signal, which is shown in Fig. 9(c). Its input terminals are connected with NA and NB, and the sign of output is determined by which node switches first.

To transform the pulses with different widths into digital output, we have designed a multi-bit counter as TDC, as shown in Fig. 9(d). It is mainly composed of an N-bit flip-flop (FF) and an AND gate A1. The output of VTC and clock signal CLK are connected to the inputs of gate A1. The output of A1 is connected to the enabling end of the first FF, and the enabling end of the other FF is connected with the inverse signal of the output of the previous FF. The output of each FF is the output of the corresponding n-bit accumulator. The reset terminals of all flip flops are connected to the same reset signal Reset. Before sampling, the Reset signal is discharged to VSS to reset all FFs to 0. During the sampling phase, the reset signal is charged to VDD, and the CLK signal inputs a high-frequency clock signal. When the terminal IN of TDC is 0, the output of A1 is 0. When terminal IN switches to 1, the output of A1 will be switched followed by the CLK signal. By adjusting the speed of CLK, the pulse width can be linearly converted into the enable counts of the converter. Therefore, the output pulses of VTC with different widths are converted into digital output signals.

To better illustrate the operation of the proposed VTDC, we have shown the waveform in Fig. 10. The whole operation

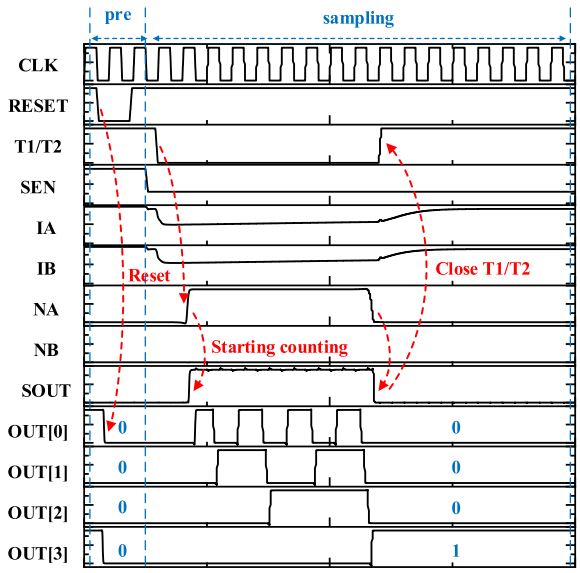


Fig. 10. The waveform of proposed VTDC.

can be divided into two phases, which are the preparation phase, and sampling phase. During the preparation phase, the transmission transistors T1 and T2 remain closed, the inputs IA and IB of S1 and S2 are precharged to VDD through the PRE signal, the multi-bit counter will also be reset to zero through the RESET signal. During the sampling phase, the calculation results on GBLP and GBLN will be transmitted to the IA and IB through T1 and T2. The sense enable signal SEN is discharged to VSS, and the charging transistors P2 and P4 are enabled to slowly charge the IA and IB. Because the initial potentials on GBLP and GBLN are different, the outputs of S1 and S2 will be discharged to the VSS at different times. Therefore, the voltage difference between the pair of bit lines can be converted into pulses with different switch times, and then input to the XOR gate X1. The reset signal is invalid, and a high-frequency signal CLK is an input to the clock terminal of TDC. When the input of TDC IN is at a high level, a high-frequency signal will be generated through the A1 applied on the enable terminal of the first stage of the counter. This counter will keep the accumulation operation until the output of X1 switches to zero. When the output of X1 becomes 0, T1 / T2 is turned off, IA and IB are quickly charged to VDD, which can cut off the competitive current in S1 and S2 to reduce power consumption overhead. Therefore, by utilizing the proposed VTDC, the proposed output can continuously sample high-precision signals without multiple comparison operations or using multiple comparators.

III. Simulation

To demonstrate the advantages of our design, an 81kb accelerator with 1-4 bit weight and interface precision is designed and simulated in the 28nm process. The accelerator contains 4 macros, the size of each macro is 144×144 bit, the size of the smallest process element is 9×8 bit, which means every four columns cells shared one VTDC. To reduce the influence of IR drop and avoid interconnection line fusing, at most 9 rows of cells can be activated in a macro for calculation in each cycle. In order to evaluate the accuracy and energy

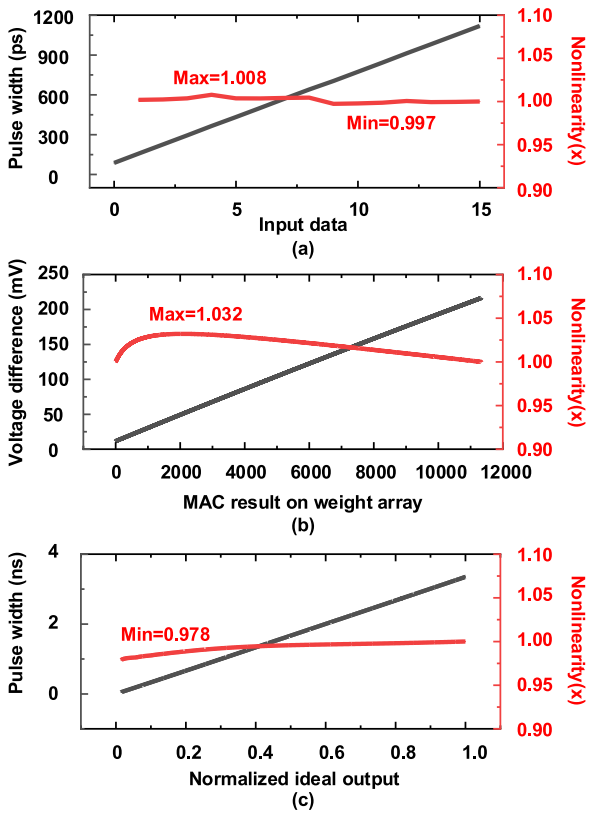


Fig. 11. The effect of the nonlinearity: (a) the nonlinearity of input, (b) the nonlinearity of weight, (c) the nonlinearity of output.

consumption, a neural network with 8 layers is implemented to carry out the training and test of the CIFAR 10, the structure is shown in Table I. During calculation operation, the activation voltage of WL and SL is 1.15V and 0.6V. The HRS and LRS resistance of ReRAM are about 124kΩ and 15kΩ, and the corresponding $I_{on/off}$ of 4T2R is about 78x, which is 10 times compared with the 2T2R structure.

The non-idealities associated with the circuits have a serious effect on inference accuracy. We have evaluated the effect of the nonlinearity on 4-bit processing, which is shown in Fig. 11. The nonlinearity acting on the components can be mainly characterized by the deviations of ideal input pulse width, weight array output voltage difference, and output pulse width. The maximum nonlinear drift of input, weight array, and output is about 0.8%, 3.2%, and 2.2%, respectively. Therefore, the small nonlinear drift of the circuit proves that the proposed circuit has good linearity.

To eliminate the influence of nonlinearity on inference accuracy, we bring the nonlinearity of each component into the network training and test, taking the fully connection layer as an example to illustrate, as shown in Fig. 12. The network diagram consists of the input layer, hidden layers and, output layer. There are q neurons in the input layer, t neurons in the output layer, r and s neurons in the hidden layer l and $l+1$ layer, respectively. The w_{ji} and b_{ji} are represented the weights and bias between the i-th neuron in the l-th layer and the j-th neuron in the $(l+1)$-th layer. The z and a are the linear MAC result and activation result, and the corresponding activation function is denoted as δ. To reflect the nonlinearity acting on each component, the nonlinear functions f(x), g(x), and h(x)

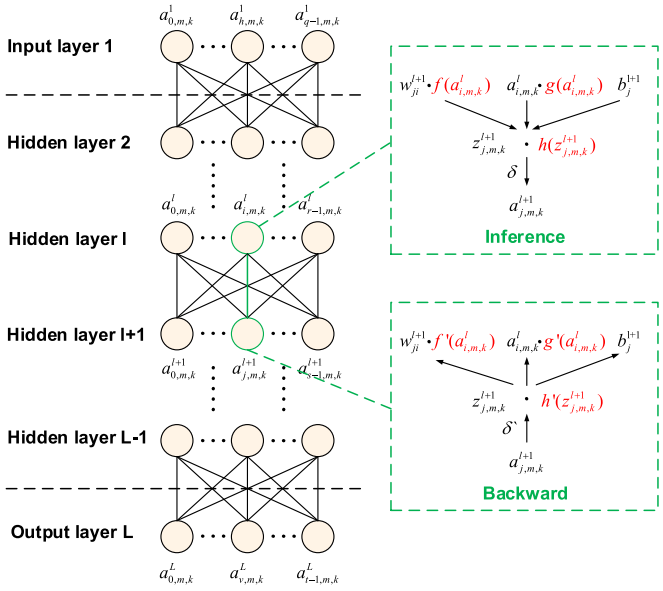


Fig. 12. Network derivation diagram against nonlinearity.

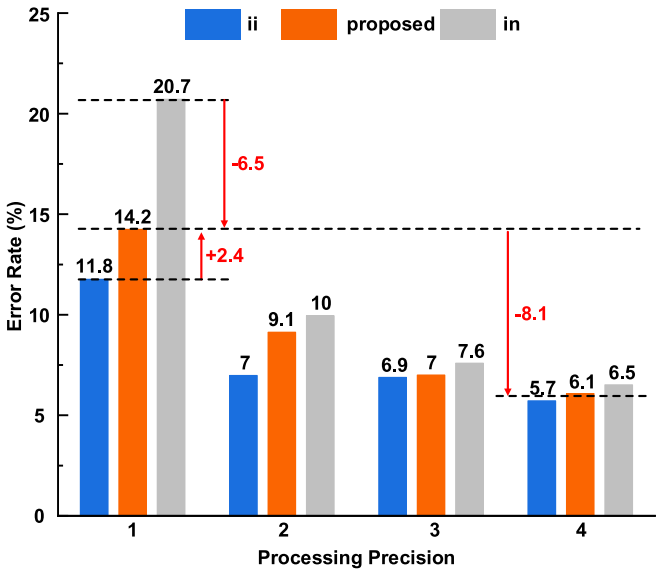


Fig. 13. The accuracy of the proposed accelerator under 1–4bit processing precision for CIFAR 10.

are fitted as an analytical equation according to the measured nonlinear drift of weight, input, and output. Here, the nonlinear function can be viewed as an attenuation coefficient on the output of each component, which can be acted as an extra custom activation function layer in training and test.

To evaluate the accuracy of the design accelerator and the effect of nonlinearity, the accuracy was simulated under 1-4 bit processing precision for CIFAR 10, which is shown in Fig. 13. The blue bar is the error rate under state ii, which represents ideal training and ideal test. Under state ii, the network is trained and tested based on the ideal circuit without considering the nonlinearity mentioned above. The grey bar is the error rate under the state in, which represents ideal training and test with the real nonlinearity of the circuit. Under state in, the network is trained based on the circuit without considering the nonlinearity, but tested based on the circuit with the nonlinearity. The orange bar is the error rate under

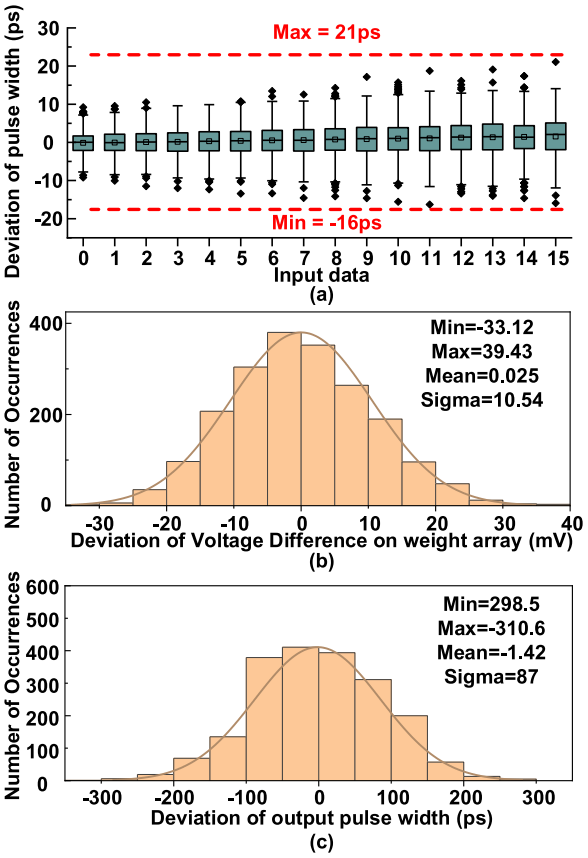


Fig. 14. The effect of the nonlinearity: (a) the nonlinearity of input, (b) the nonlinearity of weight, (c) the nonlinearity of output.

our proposed training and test method. Under this state, the network is trained and tested based on the circuit with the nonlinearity. For 1-bit processing precision, the error rate of ii is 11.8%, which will be raised to 20.7% after considering the nonlinearity. However, based on our proposed method, the error rate can be reduced to 14.2%, which is 6.5% less than the state in and 3% more than the state ii. In addition, the error rate based on our proposed method can be further reduced to 6.1% from 14.2% when using 4-bit precision, realizing 57% optimization compared with the 1-bit precision. These results show that the error rate based on our method can be reduced by about 32% at most compared with ideal training without considering the nonlinearity (state in). Therefore, our proposed training and test method can effectively reduce the influence of nonlinearity. Moreover, these results also show that improving the processing precision is a direct way to improve the accuracy of neural networks. Therefore, in order to better deal with complex targets in edge devices, it is necessary to reduce the power consumption and area overhead caused by higher precision.

Besides the nonlinearity, the process variation with the analog circuits also affects the reliability of the accelerator. To evaluate the influence of the process variation on each component, the MC simulations are carried out under 4-bit processing precision, which is shown in Fig. 14. The influence of process variation on input is the deviation of input pulse width, and the maximum deviation is about 21ps. As for the voltage difference on the weight array, the local variation of

TABLE II
COMPARISON WITH STATE-OF-THE-ART DESIGNS

Metric	ESSCIRC'21 [22]	TCASI'20 [23]	ISSCC'19 [24]	JSSC'22 [26]	This Work
Technology	40nm	45nm	55nm	40nm	**28nm**
Weight type	1T1R	1T1R	1T1R	1T1R	**4T2R**
Capacity	16kb	64Kb	1Mb	64kb	**81Kb**
Clock frequency	100MHz	25MHz	85.1MHz(1b input) 68.5MHz(2b input)	20-100MHz	**500MHz(1b)** **167MHz(4b)**
Precision (input, weight, output)	(N/A,1/2/4/8,3b)	(2b,3b,2b) (8b,8b,8b)	(1/2b,3b,3b)	(1-8b,1-8b,1-20b)	**(1-4b,1-4b,1-4b)**
Power (mW)	N/A	26.8/199.7	N/A	N/A	**0.92 (4b)**
Core Area (mm^2)	0.25	N/A	7.5	N/A	**0.093**
In/out type	Flash ADC(out)	Binary input(in) SAR ADC(out)	2b serial input(in) DSWCT+3b CSA(out)	Pulse width input (in) Flash ADC (out)	**PWM(in)** **VTDC(out)**
weight Size(um^2)	N/A	N/A	0.2025	N/A	**1.06**
Energy efficiency (TOPS/W)[1]	36.4(1b)	38.8(2b,3b,2b) 0.61(8b)	53.17(1b input) 21.9(2b input)	4.15-56.67(1b)	**95.3(1b)** **59(4b)**
FoM[2]	36.4(1b)	39.04	197.1	265.6	**944**

1) 1 MAC = 1multiplication + 1 addition = 2Ops
2) FoM = input-precision × weight-precision × energy efficiency

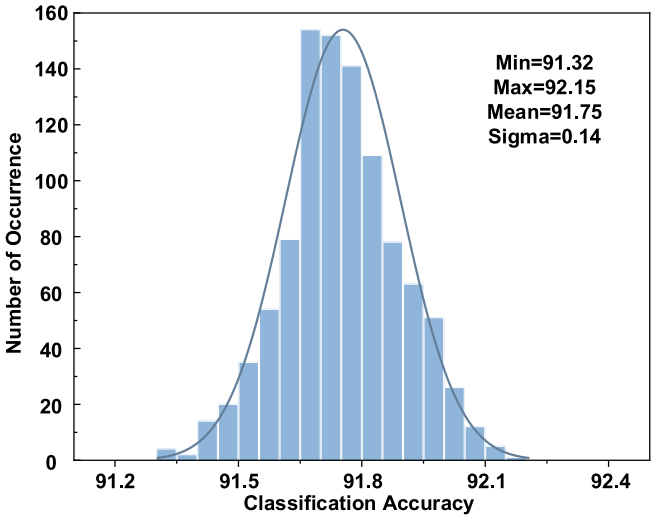


Fig. 15. The MC simulation results of accuracy with ReRAM variation under 4-bit processing precision: 20% (3σ) local variation.

ReRAM is also considered for both HRS and LRS, which is 20%(3σ). The maximum deviation voltage difference on the weight array is about 39.43mV, which is about 18% of the ideal maximum voltage difference. For the output pulse width, the maximum deviation is 310.6ps, which is about 9.3% of the ideal maximum output pulse width.

Fig. 15 shows the MC simulation results of accuracy with the above non-idealities under 4-bit processing precision. It can be observed that the average error rate increases by 2.15% compared with the ideal case, and the error rate in the worst case will increase by 2.58%. This result shows that the non-idealities have an impact on the accuracy of neural network calculation, but the network still maintains high performance in the worst case with our proposed accelerator and training method.

Table II presents the comparison with the state-of-the-art CIM accelerators based on ReRAM. By using the 4T2R cell and corresponding weight extension scheme, the processing precision can be greatly improved. With the proposed PWM input and VTDC output, the time-consuming conversions of each layer in the inference can be greatly reduced, helping to achieve a high processing speed of about 167-500MHz and high throughput of 216-648 GOPS, at least 2X improvement

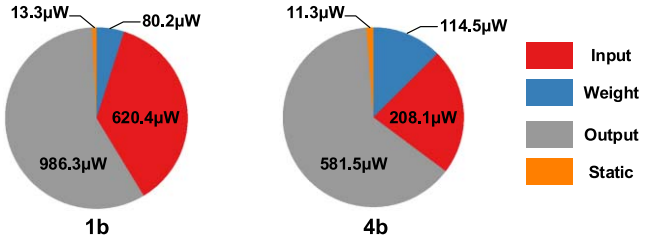


Fig. 16. The pie-chart of power consumption of the proposed accelerator for 1-bit and 4-bit precision.

over the prior CIM-based accelerators [24]. Because the number of inputs is reduced by sharing, and the output circuit does not need large reference capacitance, the core area of the designed accelerator is only 0.093 mm^2. In addition, the low-cost peripheral circuit design also brings about 0.92mW power consumption under high processing speed. The simulated energy efficiency is 95.3 - 59 TOPS/W under 1-4 bit processing precision, achieving at least 54% improvement compared to [23] with higher precision. To better understand the power consumption of each component, the pie-charts of power consumption of the proposed accelerator for 1-bit and 4-bit precision are shown in Fig. 16. Here, the power consumption includes the input, output, weight, and static power consumption of the network. More than half of the power consumption comes from the output circuit, which proves the importance of designing high energy efficiency output circuit. In order to compare the energy efficiency of accelerators with different precision more accurately, refer to the comparison method of [27], we have computed the figure of merit (FoM) as the product of input precision, weight precision, and energy efficiency. The comparison result shows that the FoM of our work is about 3.6× higher than prior work.

IV. CONCLUSION

In this work, a high-performance ReRAM-based accelerator for higher precision computing is proposed. The 4T2R cell can offer a higher on/off ratio and smaller computing current, and achieve weight extension without the usage of large independent capacitors. The PWM input can switch the WL just once with good linearity, reducing the power consumption overhead from higher precision processing. The proposed VTDC output

utilizes the semi Schmitt trigger and XOR gate to converter the calculation results with lower energy consumption and area overhead. In addition, the nonlinearity and process variation of the circuit and ReRAM devices are tested, and the results show that the network on the accelerator has good accuracy. The results indicate that the proposed high-performance ReRAM-based accelerator is suitable for edge AI applications.

REFERENCES

[1] S. Kang, G. Park, S. Kim, S. Kim, D. Han, and H.-J. Yoo, "An overview of sparsity exploitation in CNNs for on-device intelligence with software-hardware cross-layer optimizations," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, pp. 634–648, Dec. 2021.

[2] C.-X. Xue *et al.*, "Embedded 1-Mb ReRAM-based computing-in-memory macro with multibit input and weight for CNN-based AI edge processors," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 203–215, Jan. 2020.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017.

[4] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[5] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[6] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.

[7] M. F. Ali, A. Jaiswal, and K. Roy, "In-memory low-cost bit-serial addition using commodity DRAM technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 1, pp. 155–165, Jan. 2020.

[8] 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.

[9] X. Si *et al.*, "A dual-split 6T SRAM-based computing-in-memory unit-macro with fully parallel product-sum operation for binarized DNN edge processors," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 11, pp. 4172–4185, Nov. 2019.

[10] W. Haensch, T. Gokmen, and R. Puri, "The next generation of deep learning hardware: Analog computing," *Proc. IEEE*, vol. 107, no. 1, pp. 108–122, Jan. 2019.

[11] Y. Xi *et al.*, "In-memory learning with analog resistive switching memory: A review and perspective," *Proc. IEEE*, vol. 109, no. 1, pp. 14–42, Jan. 2021.

[12] Q. Zheng *et al.*, "Lattice: An ADC/DAC-less ReRAM-based processing-in-memory architecture for accelerating deep convolution neural networks," in *Proc. 57th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2020, pp. 1–6.

[13] Y. Zha, E. Nowak, and J. Li, "Liquid silicon: A nonvolatile fully programmable processing-in-memory processor with monolithically integrated ReRAM," *IEEE J. Solid-State Circuits*, vol. 55, no. 4, pp. 908–919, Apr. 2020.

[14] P. Chi *et al.*, "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 27–39.

[15] M. E. Sinangil *et al.*, "A 7-nm compute-in-memory SRAM macro supporting multi-bit input, weight and output and achieving 351 TOPS/W and 372.4 GOPS," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 188–198, Jan. 2021.

[16] S. K. Gonugondla, M. Kang, and N. Shanbhag, "A 42 pJ/decision 3.12 TOPS/W robust in-memory machine learning classifier with on-chip training," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 490–492.

[17] X. Wang *et al.*, "A 4T2R RRAM bit cell for highly parallel ternary content addressable memory," *IEEE Trans. Electron Devices*, vol. 68, no. 10, pp. 4933–4937, Oct. 2021.

[18] Y. Chen, L. Lu, B. Kim, and T. T.-H. Kim, "A reconfigurable 4T2R ReRAM computing in-memory macro for efficient edge applications," *IEEE Open J. Circuits Syst.*, vol. 2, pp. 210–222, 2021.

[19] S. R. Lee *et al.*, "Multi-level switching of triple-layered TaO_x RRAM with excellent reliability for storage class memory," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. 71–72.

[20] X. Xu *et al.*, "Degradation of gate voltage controlled multilevel storage in one transistor one resistor electrochemical metallization cell," *IEEE Electron Device Lett.*, vol. 36, no. 6, pp. 555–557, Jun. 2015.

[21] Y. Zhang *et al.*, "A 40 nm 1 Mb 35.6 TOPS/W MLC NOR-flash based computation-in-memory structure for machine learning," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.

[22] W. Li, X. Sun, H. Jiang, S. Huang, and S. Yu, "A 40 nm RRAM compute-in-memory macro featuring on-chip write-verify and offset-cancelling ADC references," in *Proc. IEEE 47th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2021, pp. 79–82.

[23] S. Zhang, K. Huang, and H. Shen, "A robust 8-bit non-volatile computing-in-memory core for low-power parallel MAC operations," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 6, pp. 1867–1880, Jun. 2020.

[24] C.-X. Xue *et al.*, "A 1 Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge processors," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 388–389.

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

[26] 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.

[27] S. Xie, C. Ni, A. Sayal, P. Jain, F. Hamzaoglu, and J. P. Kulkarni, "eDRAM-CIM: Compute-in-memory design with reconfigurable embedded-dynamic-memory array realizing adaptive data converters and charge-domain computing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2021, pp. 248–250.

Keji Zhou received the B.S. degree from the South China University of Technology, Guangzhou, China, in 2013, the M.E. degree from Ningbo University, Ningbo, China, in 2016, and the Ph.D. degree from Fudan University, Shanghai, China, in 2021.

He is currently a Post-Doctoral Researcher with Fudan University. His current research interests include computing-in-memory, neuromorphic computing, and nonvolatile memory for neural networks.

Xinru Jia (Graduate Student Member, IEEE) received the B.S. degree in microelectronics from the Dalian University of Technology, Dalian, China, in 2020. She is currently pursuing the M.S. degree with the Frontier Institute of Chip and System, Fudan University.

Her research interests include keyword spotting, computing-in-memory, and low-power neural-networks specific accelerator design.

Chenyang Zhao received the B.S. degree from the North University of China, Taiyuan, China, in 2019. She is currently a bachelor-straight-to-doctorate student with the State Key Laboratory of ASIC and Systems, Fudan University, Shanghai, China.

Her current research interest is in chip design of neural networks accelerator based on computing-in-memory architecture.

Xumeng Zhang (Member, IEEE) received the Ph.D. degree from the Institute of Microelectronics, Chinese Academy of Sciences, in 2020.

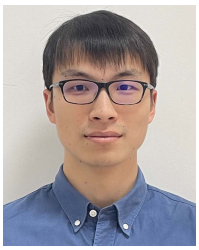
From 2018 to 2019, he was a visiting Ph.D. student with the University of Massachusetts Amherst. He is currently a Post-Doctoral Researcher with Fudan University. His research interests include memristive devices and their applications on spiking neuron circuits and neuromorphic computing.

Guangjian Wu received the B.S., M.E., and Ph.D. degrees from Nanjing University, China, in 2014, 2017, and 2020, respectively.

He is currently a Post-Doctoral Researcher with Fudan University. His current research interests include low-dimensional devices and high-sensitivity infrared photodetector.

Chen Mu received the B.S. degree from Southeast University, Nanjing, China, in 2020. He is currently pursuing Ph.D. degree with the Frontier Institute of Chip and System, Fudan University, Shanghai, China.

His current research interests include computing-in-memory, high-efficient AI accelerator, and chiplets.



Haozhe Zhu (Member, IEEE) received the B.E. degree in microelectronics from Fudan University, Shanghai, China, in 2017, where he is currently pursuing the Ph.D. degree.

His research interest includes VLSI systems design for intelligent applications.

Yanting Ding (Graduate Student Member, IEEE) received the B.S. degree from Nanchang University in 2020. She is currently pursuing the Ph.D. degree in microelectronics and solid-state electronics with Fudan University, under the supervision of Ming Liu and Qi Liu.

Her current research interests include memristive devices, process optimization, and their applications on SNN as well as neuromorphic computing.

Chixiao Chen (Member, IEEE) received the B.S. and Ph.D. degrees in microelectronics from Fudan University, Shanghai, China, in 2010 and 2015, respectively.

He was an exchange student with the University of California at Davis, Davis, during 2008 to 2009. In 2015, he worked at Calterah Inc. as an Analog/Mixed Signal Circuit Design Engineer. From 2016 to 2018, he was a Post-Doctoral Research Associate with the University of Washington, Seattle. Since 2019, he has been an Assistant Professor with Fudan University. He is also an Adjunct Research Associate with the State Key Laboratory of ASIC and Systems, Fudan University. His research interests include mixed-signal integrated circuit design and custom intelligent software-hardware co-designs. He received the Outstanding Ph.D. Research Support Award from Fudan University and the IEEE Solid-State STGA Award from ISSCC 2014.

Xiaoyong Xue (Member, IEEE) received the Ph.D. degree in microelectronics from Fudan University, Shanghai, China, in 2011.

In 2011, he joined the Department of Microelectronics, Fudan University, as a Post-Doctoral Research Fellow, where he is currently an Associate Professor. His research interests include memory circuit design, in-memory, and neuromorphic computing algorithm and circuit for machine learning acceleration.

Xiaoyang Zeng (Member, IEEE) received the B.Sc. degree from Xiangtan University, Xiangtan, China, in 1996, and the Ph.D. degree (Hons.) from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2001.

He was a Post-Doctoral Researcher with Fudan University, Shanghai, China, from March 2001 to February 2003, where he has been a Faculty Member since 2003 and the Chair Professor and the Executing Director of the State-Key Laboratory of ASIC and System. He has published over 200 articles in international journals and conferences, and applied for over 120 patents. His research fields include information security chips, baseband processing technologies for wireless communication, and mixed-signal IC designs and ultra-low power IC methodology. He serves as a TPC Member for the ISSCC, a member for the Steering Committee of ASPDAC and the A-SSCC Technical Committee, the Chair for the Shanghai Chapter of the IEEE SSCS, the Co-Chair for the Circuit and System Division of the Chinese Institute of Electronics, and the TPC Chair for ASICON 2009 and 2013.

Qi Liu received the Ph.D. degree from Anhui University in 2010.

He then joined the IMECAS and became a Professor in 2016. He is currently a Professor with the Frontier Institute of Chip and System, Fudan University; and a Guest Professor with the IMECAS. His research interests focus on the fabrication, characterization, and mechanism of the emerging memristor devices for nonvolatile memory; logic circuit; and neuromorphic computing applications.