Series-Parallel Hybrid SOT-MRAM Computing-in-Memory Macro with Multi-Method Modulation for High Area and Energy Efficiency

Weiliang Huang[1], Jinyu Bai[1], Wang Kang[1], Zhaohao Wang[1], Kaihua Cao[1], Hongxi Liu[2], He Zhang[1,*], Weisheng Zhao[1]
[1] School of Integrated Circuit Science and Engineering, Beihang University, Beijing, China.
[2] Truth Memory Corporation, Beijing, China. *Corresponding author.
{ weiliang_huang, jinyu.bai, wang.kang, zhaohao.wang, kaihua.cao, zhanghe, weisheng.zhao }@buaa.edu.cn, hongxi_liu@tmc-bj.cn

ABSTRACT

Computing-in-memory (CIM) shows its superiority in lots of applications like neural network inference. Recently, there are lots of exploration of the application of Magnetic Random-Access Memory (MRAM) in CIM. This paper aims to investigate the potential of Spin-Orbit-Torque-MRAM (SOT-MRAM) in CIM and proposes a high area and energy efficiency SOT-MRAM CIM macro based on a 6T-4J weight group. The bit-cell array adopts series-parallel hybrid architecture, which combines both serial and parallel configurations of Magnetic Tunnel Junction (MTJ) to solve the problem of high energy cost and low flexibility caused by MRAM-series and MRAM-parallel architecture, respectively. Additionally, the proposed SOT-MRAM CIM macro incorporates a multi-method modulation scheme, ranging from input unit to array, which meanwhile allows for configurable input precision (2/4/6/8-bit). The SOT-MRAM CIM macro is designed and verified in both 180nm and 28nm nodes, based on the verified electrical performance of the SOT-MRAM array in a 200-nm wafer pre-fabricated. The simulation results in 28nm show that this macro can achieve energy efficiency of 23.7~29.6 Tops/W at 8-bit input and output precision.

KEYWORDS

SOT-MRAM, Computing-in-memory, Series-Parallel hybrid, Configurable precision

1 Introduction

In recent years, Deep Neural Networks (DNNs) have played a significant role in numerous edge applications [1-2]. CIM is regarded as a promising scheme for the acceleration of neural networks. However, there is no consensus on which storage media, including FLASH, SRAM, DRAM, MRAM, RRAM and PCRAM, has the most superior performance in CIM applications [3-8]. Notably, there is still a gap for the discussion of SOT-MRAM CIM. This paper aims to fill this void by investigating SOT-MRAM in CIM circuits design and simulation, offering a new perspective to advance CIM technology progress.

Fig. 1(a) shows the architecture of a conventional MRAM-based CIM architecture [9]. External inputs are converted into analog signals through a Digital-to-Analog Converter (DAC) and then input into the array. After the weight data stored in each unit

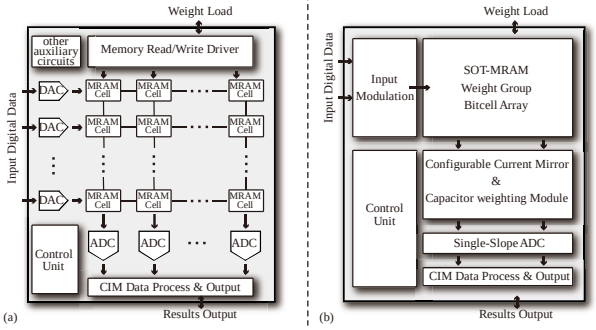


Fig. 1. (a) Typical structure of an MRAM CIM macro; (b) Structure of the proposed SOT-MRAM CIM macro.

perform computations with the inputs, the intermediate results of each column are accumulated on the Source Line (SL), and then read out through an Analog-to-Digital Converter (ADC). However, MRAM exhibits relatively low resistance in existing process technology, which leads to large operating currents resulting increased power consumption while employing the traditional crossbar array architecture. Samsung adopts the resistance summing strategy to represent the computation results, replacing the current summing strategy [10]. This scheme mitigates the impact of low readout resistance on power consumption. Nevertheless, it cannot effectively accommodate multi-bit input modulation and weight representation. Conversely, Interuniversity Microelectronics Centre (IMEC) uses the conductance difference between SOT devices to store and represent different weight levels [11]. This design allows for flexible adjustment of weight levels, catering to various computational requirements. However, this structure requires specialized bit-cell design and necessitates higher top electrode resistance and operating voltages. On the other hand, the pursuit of higher input precision introduces significant delays or peripheral circuit overhead, like Modulation Pulse Counting (MPC) [12] and proportional capacitor weighting [13].

In this paper, we illustrate a high energy-efficient and low area-cost SOT-MRAM CIM paradigm, as shown in fig. 1(b), based on the SOT-MRAM array pre-fabricated. We also put forward a series-parallel hybrid bit-cell array architecture to address the issues arising from a single SOT device-connectivity form, like energy consumption and flexibility concerns. A related hybrid-input-precision scheme is proposed to achieve a configurable-precision SOT-MRAM CIM macro, which enables 2/4/6/8-bit input precision and 5-level weight. Specifically, the multi-bit input

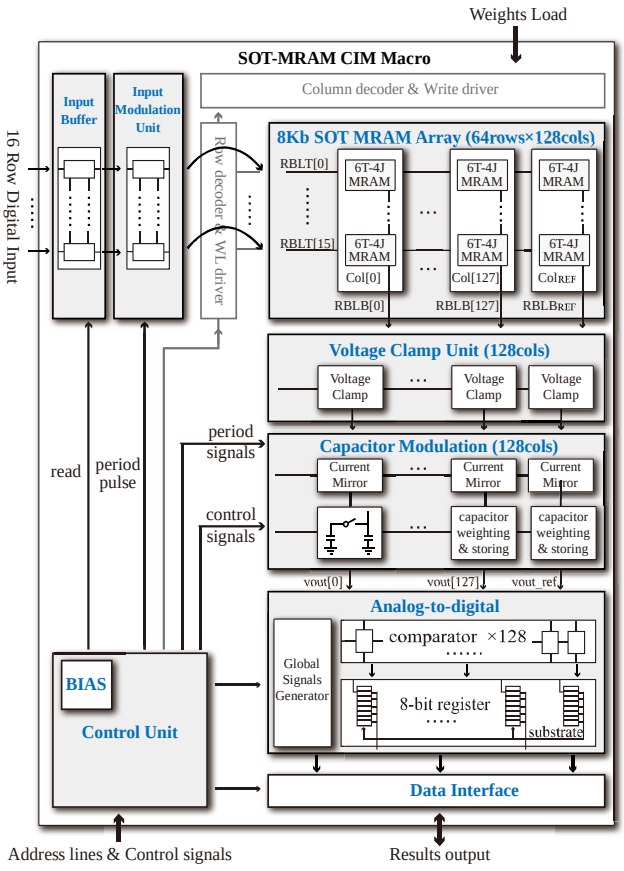


Fig. 2. Overall structure of the SOT-MRAM CIM macro, including a global control unit, an input buffer, an input modulation unit, an 8-Kb SOT-MRAM CIM array consisting of 128 columns, a voltage clamp unit, a storing and weighting capacitor unit and an analog-to-digital converter.

modulation is achieved through three components: a feedback modulation circuit, a configurable current mirror and capacitor split-cycle weighting module. At the output, the high-area-efficient data conversion is accomplished using a Single-Slope-ADC (SS-ADC). Finally, the proposed macro is designed and verified at both the 180nm and 28nm process nodes. In the 28nm process node, the energy efficiency reached 23.7-29.6 Tops/W in full-precision testing scenario.

2 Proposed SOT-MRAM CIM macro

2.1 Illustration of the fabricated SOT-MTJ array

We pre-fabricate a 1-Kb SOT-MRAM array based on "type-Y" MTJ [14]. Fig. 3(a) shows a magnified image of the array, captured by a scanning electron microscope (SEM). Fig. 3(b) presents a cross-section transmission electron microscope (TEM) image of a selected SOT-MTJ obtained from the fully integrated 200 mm CMOS wafer. Fig. 3(c) illustrates the mean resistance distribution of the device in two different magnetic torque parallelism states. With the increasing of the voltage applied at the both ends of the MTJ device, the MTJ resistance degenerated, resulting in the decline of the tunneling magnetoresistance (TMR) ratio, defined as $(\mathbf{R_{AP}} - \mathbf{R_P})/\mathbf{R_P}$, as shown in fig. 3(d). TMR ratio exceeds 70 at

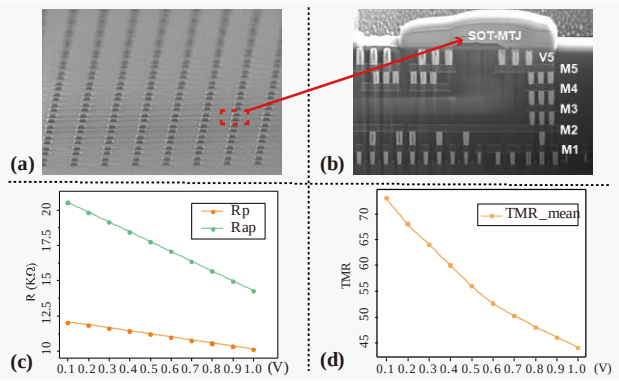
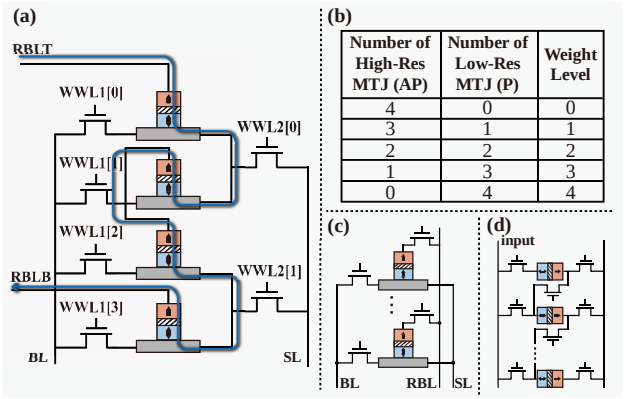


Fig. 3. (a) A zoom-in SEM image of the SOT-MRAM array; (b) The cross-section TEM of a selected MTJ in the array using common 0.18 μm CMOS process technology node [14]; (c) Resistance distribution with different voltages on MTJ; (d) MTJ distribution with different voltages on MTJ.



Number of High-Res MTJ (AP)	Number of Low-Res MTJ (P)	Weight Level
4	0	0
3	1	1
2	2	2
1	3	3
0	4	4

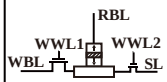
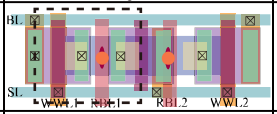
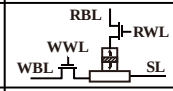
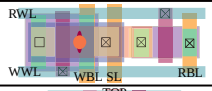
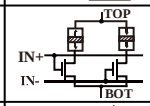
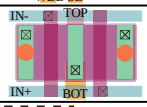
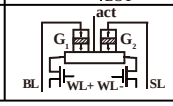
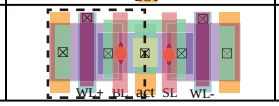
Fig. 4. (a) Structure of a 6T-4J weight group;(b) Parallel-bit-cell CIM structure; (c) Series-bit-cell CIM structure;(d) Mapping between 5-level weights and MTJ resistance states distribution.

low operating voltages, which is important for a robust calculating and readout performance. The design and simulation validations of this work are based on these verified experimental data.

2.2 Reconfigurable CIM macro architecture

Fig. 2 shows the overall architecture of the proposed SOT-MRAM CIM macro, mainly including an 8-Kb SOT-MRAM CIM bit-cell array, an input modulation module, a storing and weighting module, an output unit and a control unit. The macro can perform Multiply-Accumulate (MAC) operations with a mixed-precision input modulation scheme. The SOT-MRAM CIM bit-cell array consists of 64 rows and 128 columns, in addition to one column used as a reference column. Four rows within each column constitute a weight group and a total of 16 weight groups are accommodated per column. Each weight group integrates 4 SOT devices and 6 transistors, thereby enabling a 5-level weight characterization. Following single-bit normalization, each cell adopts a 1.5T-1J structure. In the computing mode, 16 sets of 8-bit external input data are fed into the array through a feedback modulation circuit. The arithmetic current generated in each column is modulated by configurable current-mirror selectivity and split-period capacitance weighting. This modulation process can be tailored to meet diverse input accuracy requirements. The analog MAC resulting voltage

Table 1. Several bit-cells' schematic and layout design

Structure	Schematic	Layout	area
This Work 1.5T-1J			1×
Traditional 2T-1J			1.68×
Samsung 2T-2J			1.21×
IMEC 1T-1J			0.89×

are accumulated on the storing capacitors in each row and are split-period weighted. Finally, multiple 8-bit precision column-wise SS-ADC work and subtract the output from the reference column output to obtain the final digital output.

2.3 The series-parallel hybrid SOT-MRAM CIM bit-cell structure

The proposed SOT-MRAM CIM macro combines both serial and parallel configurations of MTJ to reduce area overhead. As shown in fig. 4(a), each bit-cell consists of six transistors and four SOT devices, and it can represent 5-level weight values, as indicated in fig. 4(b). On average, each MTJ corresponds to 1.5 transistors, meaning that one weight-level needs a 1.5T-1J bit-cell overhead. In the operational mode, 16 bit-cell in one column share a common RBLB which is clamped to a fixed voltage level V_{cm} through peripheral feedback circuit. Sixteen distinct input modulation voltages are applied to RBLT, leading to different current values based on the storage weight level and input voltage. This current value is aggregated on RBLB and received by the external feedback clamping circuit, generating output voltages of varying amplitudes. Consequently, the cumulative results of 16 sets of 8-bit inputs with 5-level weight values can be represented based on the voltage levels.

Fig. 4(c) shows a typical parallel-CIM architecture. The lower top electrode resistance of the SOT device leads to large cumulative current on the SL, which results in higher power consumption. Although this can be avoided by introducing a high resistance MTJ, the high process difficulty cannot be avoided. Fig. 4(d) illustrates a series-CIM architecture. A fully MTJ-series architecture can only perform one multiplication operation in a cycle and cannot perform accumulating operation in one column. Moreover, the series connection of a large number of devices leads to a large set-up time, which results in lower efficiency. Therefore, this work proposes the series-parallel hybrid CIM array architecture. At the bit-cell level, this architecture solves the problem of power consumption due to lower MTJ resistance values in parallel structures by connecting SOT devices in series. At the array level, the accumulation of multiplication results within a single bit-cell is achieved by connecting 6T-4J weight groups in parallel.

Table 1 shows several representative CIM bit-cell circuit and layout designs, including the traditional 2T-1J structure, Samsung's proposed 2T-2J series resistance summing structure, and IMEC's proposed SOT differential structure [10-11]. Each of these schemes is designed based on either parallel or serial bit-cell configurations, exhibiting distinct characteristics in terms of computational speed, energy efficiency and reliability. The area efficiency of the computing cell can be compared by the bit-cell layout area. Specifically, two layout area is normalized by one MTJ. The 1.5T-1J bit-cell proposed in this work combines the advantages of resistance parallel and serial structures, which not only ensures low operating current but also leads in terms of area efficiency.

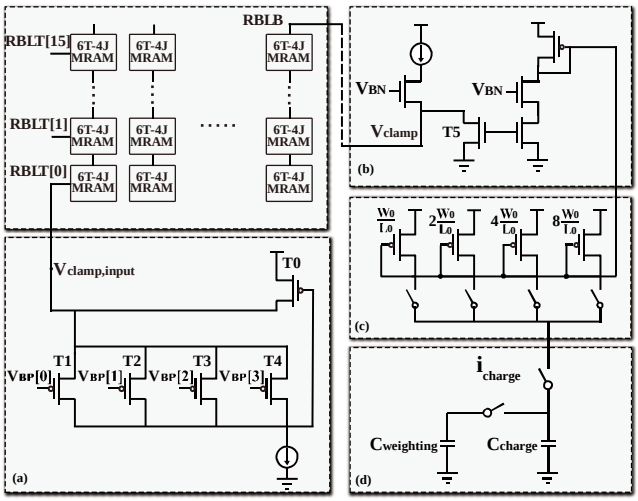


Fig. 5. Structure of input modulation scheme: (a) Input modulation unit; (b) Voltage clamp module; (c) Configurable current mirror; (d) Storing and weighting capacitors.

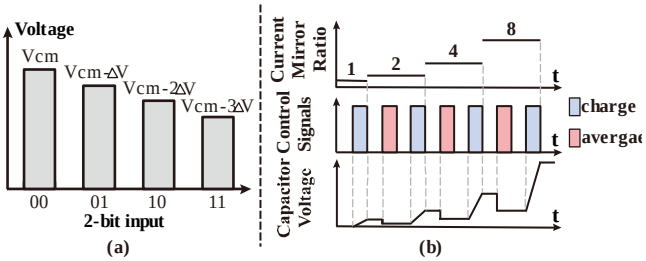


Fig. 6. (a) Amplitude modulation of per 2-bit input data; (b) The amplifying ratio of the current mirror, key control signals and capacitor voltage.

2.4 Multi-method input modulation scheme

Fig. 5 presents the schematic diagram of the input modulation scheme, which is implemented through the integration of a feedback modulation circuit, a configurable current mirror and a voltage weighting module. The fundamental principle involves the segmentation of an 8-bit input data into four 2-bit segments for modulation. Each 2-bit input segment is modulated to generate four distinct amplitude voltages, resulting in the formation of four 4-level pulse-modulated voltages. During each pulse cycle, different proportions of current mirrors are activated, and capacitive weighting is performed between successive pulse cycles. Ultimately, this process achieves the characterization of input weights in a split-cycle manner.

Fig. 5(a) represents the feedback modulation circuit. This circuit

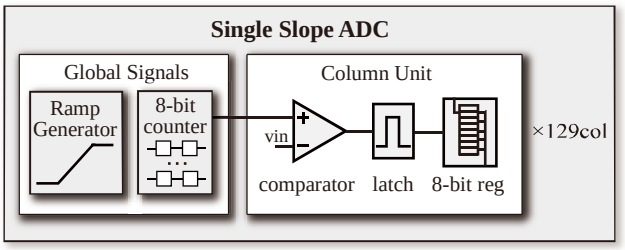


Fig. 8. Overall structure of Single-Slope ADC.

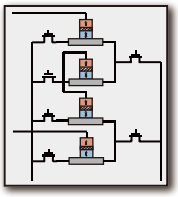
allows for the selective activation of bias transistors T1 to T4 based on the 2-bit input data. The different aspect ratios of these four bias transistors enable the clamping of the source voltage at various levels. The voltage at the $V_{clamp,input}$ point is modulated to yield four distinct amplitude values and input to the RBLT. Variations in voltage at the ends of the memory computing unit result in changes in the current flowing out of the RBLB. Transistor T0 is designed to receive and accommodate these current variations, thereby maintaining a consistent current through the bias transistors, ensuring the stability of input voltage modulation. For the RBLB side, it is clamped to a reference voltage value V_{cm}, as depicted in fig. 5(b), through a symmetric feedback modulation circuit (where N-type MOS and P-type MOS transistors alternate). Fig. 6(a) illustrates the amplitudes of the four input voltages and the reference voltage V_{cm}. Since RBLB is shared among 16 bit-cell within a column, the current exiting from RBLB represents the sum of 16 MAC outcomes. Variations in this current are subsequently mirrored in the drain-source current of feedback transistor T5, which is the result of a single pulse-cycle computation.

The drain-source current of the feedback tube is copied and input to the configurable transistor. For bits 0-1/2-3/4-5/6-7 of the 8-bit input data, different current mirrors are selected to realize the scaling of the current values, as shown in fig. 5(c) The weighted current values are charged to the storage capacitors in each column in cycles, as shown in fig. 5(d). The scaling ratio of the current for different pulse cycles and the relationship between the change of the capacitor voltage and the control signal are shown in fig. 6(b). When the charging pulse is high, the current mirror is selected to charge the capacitor according to the operation cycle. When the weighting signal is high, the storage capacitor and the equivalent capacitor is shorted to realize the halving and weighting of the storage voltage. In other words, the combination of configurable current mirrors and capacitor averaging enables 2-bit weighting between two pulse cycles. The scheme achieves high energy and area efficiency under 8-bit precision input modulation by combining the three-part modulation circuit.

2.5 Low overhead output unit

The area and power overhead caused by ADC in output unit remains a challenging issue in analog CIM. Some approaches have attempted to reduce the overhead of the output unit by reusing readout circuitry [15]. However, such methods compromise circuit parallelism, resulting in decreased operational efficiency. To minimize the overhead introduced by the ADCs, this work employs SS-ADCs with smaller area overhead, as illustrated in fig. 8. Apart from the global ramp generator, each column's output unit incurs only a comparator, an RS-latch and an 8-bit register. This approach ensures maximum operational efficiency while minimizing area and energy consumption. Furthermore, a subtraction module is introduced at the output level to facilitate the computation of the difference between column data and reference column data, thus mitigating the impact of the low TMR ratio of SOT-MRAM.

Table 2. 8-bit MAC between an 8-bit input and a 5-level weight.

8-bit input IN[0:7] (Pulse 1 Pulse 2 Pulse 3 Pulse 4)		**5-level Weight: W** 	
Input Data	Current Mirror Magnification		
IN[0:1]	1	I_{charge}	$\alpha \times IN[0:1] \times W$
		$V_{capacitor}$	$\alpha\beta \times \frac{1}{2} IN[0:1] \times W$
IN[2:3]	2	I_{charge}	$2\alpha \times IN[2:3] \times W$
		$V_{capacitor}$	$\alpha\beta \times (IN[2:3] + \frac{1}{4} \times IN[0:1]) \times W$
IN[4:5]	4	I_{charge}	$4\alpha \times IN[4:5] \times W$
		$V_{capacitor}$	$\alpha\beta \times (2 \times IN[4:5] + \frac{1}{2} \times IN[2:3] + \frac{1}{8} \times IN[0:1]) \times W$
IN[6:7]	8	I_{charge}	$8\alpha \times IN[6:7] \times W$
		$V_{capacitor}$	$\alpha\beta \times 4 \times IN[6:7] + IN[4:5] + \frac{1}{4} \times IN[2:3] + \frac{1}{16} \times IN[0:1]) \times W$

3 Hybrid-precision input modulation scheme

To support MAC operations with multi-bit input precision, we can implement a configurable 2/4/6/8 input precision reconfigurable scheme by modifying the circuit control signals.

For a 2-bit input precision, the input unit modulates the input signal into a four-amplitude pulse within a single period. To uphold maximum resolution and precision, the storage capacitor is charged with an $8\times$ current mirror amplification, while the capacitor weighting module remains inactive. After completing a cycle of charging, the voltage value stored on the capacitor is the result of the MAC operation of 2-bit input and 5-level weight.

To achieve a 4-bit/6-bit input precision, the input unit modulates the input into two or three 4-amplitude pulse signals. Between each charging pulse period, the weighting module is activated to halve the capacitor voltage. By adjusting the current mirror ratio and controlling signals, the circuit can calculate 4-bit or 6-bit input precision with 5-level weight within 2 or 3 pulse cycles.

For an 8-bit input precision modulation, Table 2 presents an overview of the workflow for full 8-bit input precision. An 8-bit input IN[7:0] is divided into four 2-bit portions, namely IN[7:6], IN[5:4], IN[3:2] and IN[1:0]. Subsequently, the input modulation unit modulates each 2-bit data into four pulse signals. A 5-level weight is stored within the 6T-4J bit-cell. The MAC operation between 8-bit inputs and 5-level weights can be divided into the following four steps:

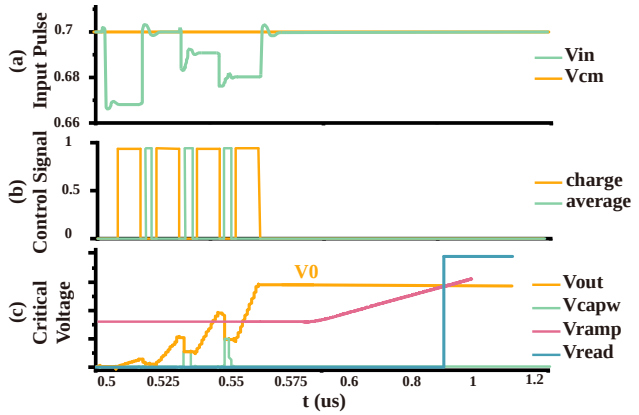


Fig. 9. Transient simulation of the proposed SOT-MRAM CIM macro: (a) Modulated input pulse; (b) Control signals; (c) Voltage of critical nodes, including capacitor voltage, global ramp voltage and readout signal.

(a) Pulse period 1: The multiplication of IN[1:0] with W and weighting. After modulating IN[1:0] into an analog input voltage, the input voltage and clamping voltage act simultaneously on the RBL at both ends of the bit-cell to obtain the charging current of the capacitor. Meanwhile, the charging enable signal governs the current to charge the storage capacitor. The charging current and the capacitor voltage are denoted as follows:

$$i_{charge,1} = \alpha \times IN[0:1] \times W \quad (1)$$

$$v_{cap,charge} = \alpha\beta \times IN[0:1] \times W \quad (2)$$

α and β are conversion coefficients from sampling gate voltage to charging current and from charging current to capacitor voltage, respectively. After charging is completed, the weighting enable signal is raised. At this point, the storage capacitor and the equivalent capacitor are short-circuited, causing the capacitor voltage to halve:

$$v_{cap,1} = \alpha\beta \times \frac{1}{2} \times IN[0:1] \times W \quad (3)$$

(b) Pulse period 2: The multiplication of IN[3:2] with W, application, and weighting. IN[3:2] are modulated into four-amplitude pulse inputs. In the first pulse period, the current mirror's input magnification factor is set to 1. In the second pulse period, the current mirror magnification factor needs to be adjusted to 2. At this point, the charging current is given by:

$$i_{charge,2} = 2\alpha \times IN[2:3] \times W \quad (4)$$

Next, the capacitor is charged and the storing voltage is weighted, resulting in a capacitor voltage of:

$$v_{charge,2} = \alpha\beta \times \left(IN[2:3] + \frac{1}{4} \times IN[0:1]\right) \times W \quad (5)$$

(c) pulse period 3: The multiplication of IN[5:4] with W, application and weighting. Adjust the current mirror amplification ratio to 4, and the charging current and capacitor voltage are:

$$i_{charge,3} = 4\alpha \times IN[4:5] \times W \quad (6)$$

$$v_{charge,3} = \alpha\beta \times \left(2 \times IN[4:5] + \frac{1}{2} \times IN[2:3] + \frac{1}{8} \times IN[0:1]\right) \times W \quad (7)$$

Table 3. Chip specifications

Technology	28 nm
Supply Voltage	1 V
Array Size	8-Kb
Bit-Norm Cell Structure	1.5T-1J
Input Precision	8-bit
Weight Precision	5-level
Output Precision	8-bit
Energy Efficiency	23.7-29.6 Tops/W

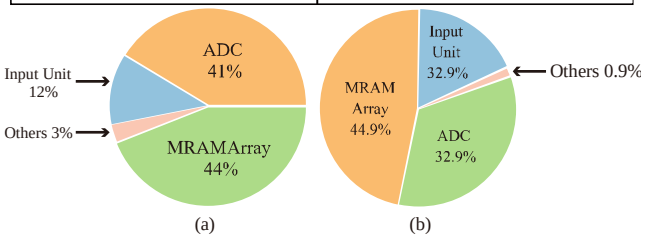


Fig. 10. (a) Energy breakdown; (b) Area breakdown.

(d) pulse period 4: The multiplication of IN[7:6] with W and application. In the charging of the highest 2 bits, the current mirror amplification ratio needs to be adjusted to 8, and the enable control of capacitor voltage weighting needs to be cancelled. The charging current and final capacitor voltage are:

$$i_{charge,4} = 8\alpha \times IN[6:7] \times W \quad (8)$$

$$v_{charge,4} = \alpha\beta \times \left(4 \times IN[6:7] + IN[4:5] + \frac{1}{4} \times IN[2:3] + \frac{1}{16} \times IN[0:1]\right) \times W \quad (9)$$

Overall, the traditional MPC modulation scheme requires using 64 input pulses to process an 8-bit input data of all ones. However, in the design of this scheme, by combining the three parts of modulation means, only 4 pulse periods are needed to achieve the same modulation. Compared with MPC, the input modulation scheme of this work can reduce the delay by about 70%.

4 Experimental Results

4.1 Function validation

To evaluate the performance, we design an 8-Kb SOT-MRAM CIM macros at the 180nm and 28nm process nodes respectively and conducted post-simulation verification. The results presented in this section primarily stem from simulations carried out at the 28nm process node. Take an MAC operation with a full-precision 8-bit input and 5-level weight as an example, fig. 9 displays the transient waveforms of the proposed SOT-MRAM CIM macro and depicts voltage variations at critical nodes within the circuit. Fig. 9(a) and 9(b) show the pulse modulation voltage for the four 2-bit inputs and key control signal waveforms, respectively. V_{cm} denotes the clamping voltage applied to RBLB. In fig. 9(c), V_{out} and $V_{cap,w}$ denote the output voltage and the average weighting capacitor voltage, respectively. V_{ramp} signifies the global ramp voltage generated by the SS-ADC for comparison with the capacitor voltage. During the charging and weighting phases of the capacitor,

Table 4. Performance comparison

	ISSCC'22 [7]	NAT. '21 [10]	VLSI '22 [16]	ISSCC '23 [17]	This Work
Memory Type	ReRAM	STT MRAM	STT MRAM	STT MRAM	SOT MRAM
Process Node	22nm	28nm	22nm	28nm	28nm
Cell Structure	1T-1R	2T-2J	1T-1J	2T-2J	1.5T-1J
Row Parallelism	1024	64	256	2-Mb	64
Col Parallelism	256	64	512		128
Output Precision	19-bit	4-bit	6-bit	1-bit / 5-bit	8-bit
Frequency (MHz)	69.4*2	11.1	\	20	18.2
Energy Efficiency (1b-Tops/W)	21.6	262-405	19.5-41.6	22.4-35.2	23.7-29.4
FOM*1	1382.4	262-405	78-166.8	1814-2851	440.2-546.1

*1: FOM= Input Precision × Weight Precision × Energy Efficiency;*2: Performance @VDD=0.8V

V_{out} and $V_{cap,w}$ are initially reset. Subsequently, the input pulse voltage for the first preceding 2 bits is raised. Once the voltage in the feedback modulation circuit is stable, the charging enable signal is raised, and the configurable current mirror is activated based on the computation period to charge the capacitor. Following the initial charging, the capacitor weighting signal is raised. Another equivalent capacitor, which has been reset, is shorted to the storage capacitor, halving the voltage across the storage capacitor. Then the capacitor weighting signal is lowered, and $V_{cap,w}$ is reset in preparation for the next weighting operation. Over the subsequent three computation pulse periods, the circuit performs similar charging and weighting operations on the storage capacitor based on the voltage levels of different input pulse signals (no weighting operation is required in the final pulse period). At the end of the operation cycle, the voltage value V_0 characterizing the result of the operation is obtained and the conversion from analog voltage to digital output is performed. As shown in fig. 9(c), V_{read} is raised when the ramp voltage surpasses the value of V_0. At this point, the column's register reads the counter's value, completing a full set of data computation and conversion. The chip specifications for the SOT-MRAM CIM macro are presented in the table 3. For 8-bit full-precision computation, the peak energy efficiency reaches 29.6 Tops/W, with a worst-case efficiency of 23.7 Tops/W.

4.2 Energy and area breakdown

Fig. 10(a) shows the energy breakdown of each part of the macro. It can be seen that the input unit accounts for about 12% of the total energy, which includes the energy consumption of the input modulation of each part. The energy cost of the SOT-MRAM CIM array occupies the highest proportion, more than 41%. The use of SS-ADC with low energy and area overhead solves the problem of large output unit power consumption to some extent. Compared with more than 50% of the energy cost of the output unit, this scheme has greatly cut down the energy cost while performing 8-bit quantization. Fig. 10(b) shows the area overhead of each part of the macro. The area overhead of the SS-ADC is extremely low, occupying only 32.9% of the total area of the circuit. At the same time, the input modulation efficiency of input modulation and voltage clamping using feedback structure is much higher than that using DAC in the area of input unit. Table 4 compares this work with advanced CIM-related work based on MRAM and other non-volatile memory in the past two years. It can be seen that there is little discussion of the feasibility of SOT-MRAM in the previous work. This work has reached high efficiency comparable to other work, while exploring the promise of SOT-MRAM.

5 Conclusion

In this paper, we propose a high energy and area efficiency SOT-MRAM CIM structure based on the series-parallel hybrid bit-cell structure and the electrical performance of the pre-fabricated SOT-MTJ array. We also illustrate a multi-method modulation scheme, realizing hybrid-input precision scheme to achieve configurable precision. The verification based on the 180nm node and 28nm node validate the functionality of this work and shows the prospects of SOT-MRAM. Our results in 28nm node can achieve energy efficiency of 23.7~29.6Tops/W at 8-bit precision. Under the same technology node and precision, this work shows the promising future of SOT-MRAM in CIM architecture design.

REFERENCES

[1] G. Yoo et al., "Implementing Practical DNN-Based Object Detection Offloading Decision for Maximizing Detection Performance of Mobile Edge Devices," *IEEE Access*. Vol. 9, pp. 140199-140211, Aug. 2021.

[2] J. Li et al., "Throughput Maximization of Delay-Aware DNN Inference in Edge Computing by Exploring DNN Model Partitioning and Inference Parallelism." *IEEE Trans. on Mobile Comput.*, vol. 22, no. 5, pp. 3017–3030, May. 2023.

[3] H. -W. Hu. et al., "A 512Gb In-Memory-Computing 3D-NAND Flash Supporting Similar-Vector-Matching Operations on Edge-AI Devices," in *IEEE ISSCC*, pp. 138–140, Feb. 2022.

[4] H. Wang. et al., "A Charge Domain SRAM Compute-in-Memory Macro With C-2C Ladder-Based 8-Bit MAC Unit in 22-nm FinFET Process for Edge Inference," in *IEEE J. Solid-State Circuits*, vol. 58, no. 4, pp. 1037–1050, Apr. 2023.

[5] Y. -C. Kwon. et al., "25.4 A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications," in *IEEE ISSCC*, pp. 350–352, Feb. 2021.

[6] H. Cai. et al., "33.4 A 28nm 2Mb STT-MRAM Computing-in-Memory Macro with a Refined Bit-Cell and 22.4 - 41.5TOPS/W for AI Inference," in *IEEE ISSCC*, pp. 500–502, Feb. 2023.

[7] J.-M. Hung. et al., "An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4-21.6TOPS/W for Edge-AI Devices," in *IEEE ISSCC*, pp. 1–3, Feb. 2022.

[8] W.-S. Khwa. et al., "A 40-nm, 2M-Cell, 8b-Precision, Hybrid SLC-MLC PCM Computing-in-Memory Macro with 20.5 - 65.0TOPS/W for Tiny-Al Edge Devices," in *IEEE ISSCC*, pp. 1–3, Feb. 2022.

[9] Y. Li. et al., "A Survey of MRAM-Centric Computing: From Near Memory to In Memory" *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 1–12, Oct. 2022.

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

[11] J. Doevenspeck. et al., "SOT-MRAM Based Analog in-Memory Computing for DNN Inference," *IEEE Symp. VLSI Technol.*, pp. 1-2, Jun. 2020.

[12] Q. Dong et al., "15.3 A 351TOPS/W and 372.4GOPS compute-in memory SRAM macro in 7 nm FinFET CMOS for machine-learning applications," in *IEEE ISSCC*, pp. 242–244, Feb. 2020.

[13] X. Si et al., "A Twin-8T SRAM Computation-in-Memory Unit-Macro for Multibit CNN-Based AI Edge Processors," in *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 189-202, 2020.

[14] Jiang et al., "Demonstration of a manufacturable SOT-MRAM multiplexer array towards industrial applications," in J. Semicond., vol.44, no.2, Dec. 2023.

[15] J.-H. Yoon. et al., "29.1 A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute-in-Memory/Digital RRAM Macro with Active-Feedback-Based Read and In-Situ Write Verification," in *IEEE ISSCC*, pp. 404–406, Feb. 2021.

[16] P. Deaville et al., "A 22nm 128-kb MRAM Row/Column-Parallel In-Memory Computing Macro with Memory-Resistance Boosting and Multi-Column ADC Readout," in *IEEE VLSI Technology and Circuits*, pp. 268–269, Jun. 2022.

[17] H. Cai. et al., "33.4 A 28nm 2Mb STT-MRAM Computing-in-Memory Macro with a Refined Bit-Cell and 22.4 - 41.5TOPS/W for AI Inference," in IEEE ISSCC, pp. 500–502, Feb. 2023.