TAM: A Computing in Memory based on Tandem Array within STT-MRAM for Energy-Efficient Analog MAC Operation

Jinkai Wang[1,2], Zhengkun Gu[1], Hongyu Wang[1], Zuolei Hao[1], Bojun Zhang[1], Weisheng Zhao[1,3], Yue Zhang[1,3*]

[1]Fert Beijing Institute, MIIT Key Laboratory of Spintronics, School of Integrated Circuit Science and Engineering, Beihang University, Beijing 100191, China
[2]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China.
[3]Nanoelectronics Science and Technology Center, Hefei Innovation Research Institute, Beihang University, Hefei 230012, China
*Email: yz@buaa.edu.cn

Abstract—Computing in memory (CIM) has been demonstrated promising for energy efficient computing. However, the dramatic growth of the data scale in neural network processors has aroused a demand for CIM architecture of higher bit density, for which the spin transfer torque magnetic RAM (STT-MRAM) with high bit density and performance arises as an up-and-coming candidate solution. In this work, we propose an analog CIM scheme based on tandem array within STT-MRAM (TAM) to further improve energy efficiency while achieving high bit density. First, the resistance summation based analog MAC operation minimizes the effect of low tunnel magnetoresistance (TMR) by the serial magnetic tunnel junctions (MTJs) structure in the proposed tandem array with smaller area overhead. Moreover, a read scheme of resistive-to-binary is designed to achieve the MAC results accurately and reliably. Besides, the data-dependent error caused by MTJs in series has been eliminated with a proposed dynamic selection circuit. Simulation results of a 2Kb TAM architecture show 113.2 TOPS/W and 63.7 TOPS/W for 4-bit and 8-bit input/weight precision, respectively, and reduction by 39.3% for bit-cell area compared with existing array of MTJs in series.

Keywords—Computing in memory, tandem array, analog MAC, resistance summation, STT-MRAM

I. INTRODUCTION

The rise of data-intensive applications, such as machine learning, artificial intelligence and Internet of Things, has resulted in processors requiring more frequent data transfers and access to memory. Von Neuman architecture is difficult to balance low latency and energy efficiency because it requires data transfer between separate memories and processors through a limited-bandwidth, which leads to the "memory wall" [1], as shown in Fig.1 (a).

Computing in memory (CIM) is a promising solution to overcome the memory wall issue by embedding computing capabilities in memory, as shown in Fig.1 (b) [2]-[6]. At present, various CIM architectures have employed mature memory technology, such as SRAM and DRAM. SRAM based CIM architecture have successfully demonstrated the high efficiency of performing multiply-and-accumulate (MAC) operation that dominate the computation of neural networks [7]-[11]. However, the capacity of SRAM has been restricted due to the low density of the standard 6-transistor bit-cell, resulting in limited application of SRAM based CIM in large-scale data operations. Benefitting from a larger bit density, DRAM-based CIM architecture may avoid the above disadvantages but at the cost of increased energy consumption induced by its constant and repeated refresh mechanism.

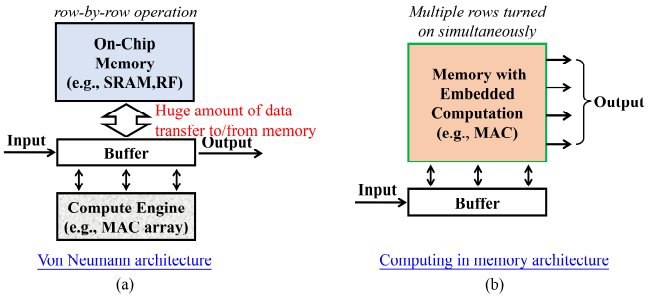

Fig. 1. (a) Von Neumann architecture. (b) Computing in memory architecture.

Emerging non-volatility memory (NVM) utilizes the intrinsic physical properties of the device to store data, which allows its bit-cell to consist of one transistor and one device, i.e., high bit density similar to DRAM. Moreover, the performance of CIM architecture can be greatly improved due to non-volatility with the realization of near-zero leakage and static power consumption. Among different NVMs, spintronic memories offer higher speed of read and write operations and have been commercialized in areas such as automotive electronics and wearable devices. Therefore, it is considered to be one of the most promising next-generation memories. Recently, a variety of CIM architectures based on magnetic random access memory (MRAM) have been proposed [12]-[16]. [12] presented spin transfer torque CIM (STT-CIM) architecture based on STT-MRAM to perform Boolean logic, arithmetic and complex vector operations. [14] presented a time-domain CIM (TD-CIM) scheme using spin-orbit torque MRAM (SOT-MRAM), which can be applied to construct the energy-efficient convolution neural network (CNN).

The current CIM architecture based on MRAM are almost all digital CIM schemes, where the arithmetic operations are executed by scheduling Boolean logic [16]. Although the digital CIM schemes achieve highly parallel and reliable computation, the energy efficiency of them are low compared with analog CIM schemes (i.e., highly parallel analog MAC operations based on Ohm's law in crossbar array) due to more complex timing and control. The main reason that MRAM cannot apply analog CIM scheme is low tunnel magnetoresistance ratio (TMR) of magnetic tunnel junction (MTJ) device, which makes it difficult to generate computational signal margins that meet the demand of analog computations in conventional crossbar array. In order to

978-3-9819263-7-8/DATE23/© 2023 EDAA

perform analog CIM in MRAM to achieve higher performance of computing, [17] reported a new MRAM crossbar array to implement analog MAC operation by using resistance summation. However, the existing problems of this crossbar array limit the further improvement and wider application: (1) The bit-cell structure of 3T2MTJ greatly reduces the bit density; (2) Bit-cell structure only supports the weights that are binarized to −1 or +1, i.e., binary neural networks, which greatly limits its application field; (3) Different resistance distributions of MTJs in series lead to different biases for the equivalent resistances, i.e., data-dependent error, which affects the accuracy of computing.

To solve above problems, we propose a CIM scheme using tandem array within STT-MRAM (TAM) to perform multi-bit analog MAC operations, which improves the area and energy efficiency. First, in the proposed tandem array, the resistance summation is achieved by only adding one transmission gate to the standard cell of STT-MRAM (1T1MTJ), which effectively reduces the area consumption. Second, a read scheme of resistance represented MAC value is proposed by designing count and shift circuit for time-domain converter (TDC), which can efficiently complete the conversion of resistance to binary number and realize the shift operation according to the weight quickly and easily. Third, a dynamic selection mechanism of the counting signal has been designed to completely eliminate the data-dependent error. Finally, we built 2Kb TAM architecture and evaluate its performance by performing MAC operations with 4-bit and 8-bit inputs and weights, achieving energy efficiency of 113.2 TOPS/W and 63.7 TOPS/W. Meanwhile, the area of bit-cell in tandem array is decreased by 39.3%.

The rest of this article is organized as follows. Section II introduces the structure of the proposed tandem array and the principle of performing MAC operation. Section III describes the proposed dynamic selection mechanism of the counting signal and the structure of column partition in the TAM architecture. Section IV presents the performance of TAM architecture. Conclusions are presented in Section V.

II. Tandem Array for MAC Operation

A. 3T1MTJ Bit-cell

In CIM scheme, the bit-cell needs to provide paths for both carrying out the write/read operations and logical computations. The read path of standard bit-cell is adopted to the computing path in conventional crossbar array, which avoids adding the transistors for rebuilding the computing path. However, to avoid the effect of low TMR, it is necessary to build a new computing path to perform concatenation of MTJs in array. As shown in Fig.2 (a), [17] presents the structure of 3T2MTJ bit-cell, where the data stored by the two MTJs are complementary, e.g., if one of the MTJs stores '1', the other must be '0'. When performing computing operation, the transistors connected to the word lines (WLs) in the operated bit-cells are turned off. Meanwhile, one of the transistors connected directly to the two MTJs is turned on according to the input of WLL, which forms a series path with the MTJs in the adjacent bit-cells to perform computing. However, this structure increases the difficulty of write operations. First, the current that performs the write operation (I_{write}) for MTJ requires flow through three transistors. To complete the write operation in the corresponding time (within 10 ns), the size of these transistors must be large enough to ensure that the current through the MTJ is sufficient (~80 μA), which will undoubtedly increase the area overhead. Second, the energy of write operation in the 3T2MTJ bit-cell is much greater than 1T1MTJ bit-cell and the write operation is also more complex.

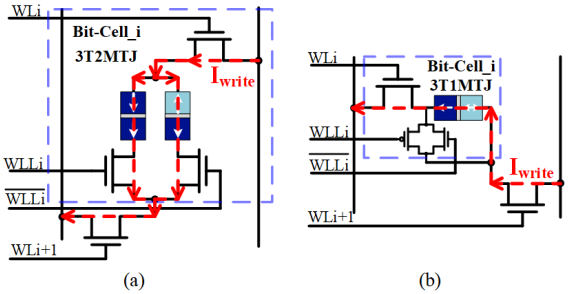

Fig. 2. (a) Structure of 3T2MTJ bit-cell in [17]. (b) Proposed structure of 3T1MTJ bit-cell.

We proposed a new 3T1MTJ bit-cell structure, which can perform the tandem computing of MTJ in array while maintaining high bit density and write energy similar to the 1T1MTJ bit-cell, as shown in Fig.2 (b). The proposed 3T1MTJ bit-cell consists of one transferring transistor, one transmission gate and one MTJ, where the transferring transistor is large in size to meet the requirement of transferring large current in write operations. However, the size of the two transistors that make up the transmission gate can be small as they function to short the MTJ. When performing write operation, the write current only needs to flow through one MTJ and two transferring transistors, almost similar to the standard 1T1MTJ bit-cell. When performing computing, if the input is '1', the transmission gate is turned off, i.e., the computing current will flow through the MTJ, which produces corresponding result of dot product according to the datum stored in MTJ. If the input is '0', the transmission gate is turned on, shorting the MTJ. At this point, regardless of the datum in the MTJ, the result of dot product is '0'. Thereby, the realization of the dot product operation in the proposed 3T1MTJ bit-cell is achieved.

B. Structure of Tandem Array and Principle of performing MAC Operation

Based on the proposed 3T1MTJ bit-cell and the operation of dot product, we construct the tandem array to achieve MAC

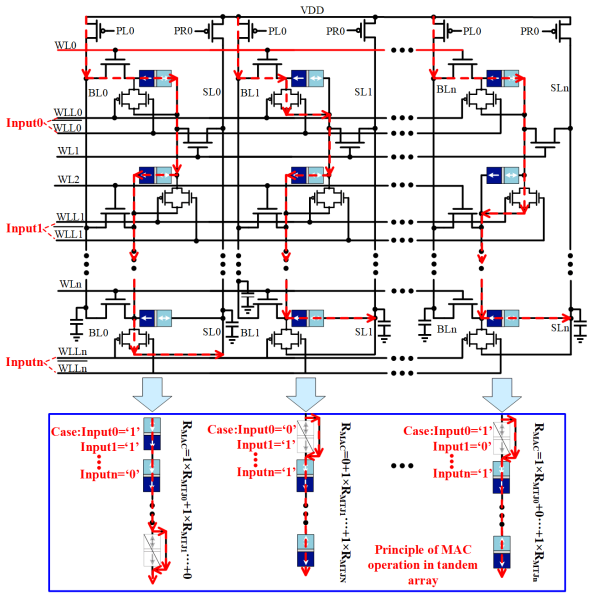

Fig. 3. Structure of tandem array and the principle of performing MAC operation.

operation by utilizing the resistance summation, as shown in Fig.3. When performing MAC operation in tandem array, the transistor of the first bit-cell on a column is turned on by activating WL0 signal. Meanwhile, the inputs on the WLLs turn on or off the transmission gates in each bit-cell on a column. Then, the transistor connecting the supply voltage and the bit-line (BL) is activated to generate the computing current, which will flow through the MTJs and transmission gates in series. Note that the initial voltages of BL and source line (SL) are both low voltage and the voltage of SL will gradually increase to a stable voltage due to the existence of parasitic capacitance of BL and SL. Besides, according to Ohm's Law, the equivalent resistance is equal to the sum of the resistances of all MTJs in series, as shown in Fig.3.

With this set-up, the number of MTJs in series on a column can ideally be unlimited. However, as the number of MTJs in series increases, the delay generated by the MAC operation also increases. The main reason is that the big resistance will increase the time for SL to reach the stable voltage. Besides, the resistance ratio of a single MTJ will decrease as the number of MTJs in series increases on a column, which reduces the signal margin between adjacent MAC values and thus lessens the accuracy of computing. Therefore, considering the above factors, the bit-cells on a column are divided into multiple groups in this work, where each group contains 9 bit-cells connected in series.

III. TAM Architecture Based on Tandem Array

To apply the tandem array to neural network computing, we design a TAM architecture, which consists of voltage-to-time circuit, count pulse dynamic selection circuit, counting and shifting circuit and full adder tree, as shown in Fig.4. In integrated circuit, the equivalent resistance represented MAC value cannot be detected directly and needs to be read indirectly through the current or voltage. Based on the above principle of tandem array, it is known that the uptrend of SL is related to the equivalent resistance. Therefore, the MAC value can be read by detecting the voltage trend of the SL.

We adopt the voltage-to-time circuit (i.e., TDC) to measure the uptrend of the SL voltage from zero to a reference voltage (i.e., the charging of the parasitic capacitors). Note that the TDC proposed in [14] is adopted in this work, where the reference voltage is set to the threshold voltage of NMOS

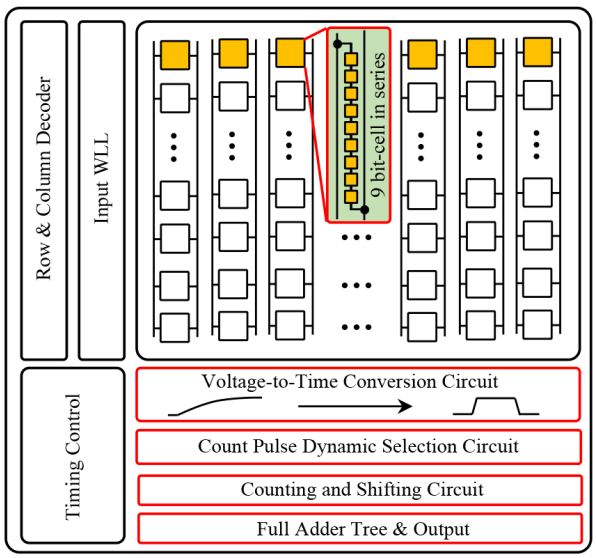

Fig. 4. Architecture of TAM scheme.

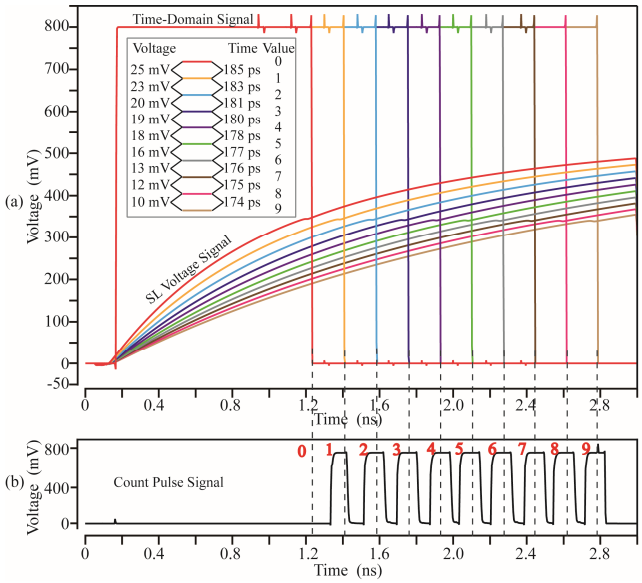

Fig. 5. (a) Transient simulation results of voltage-to-time circuit. (b) Count pulse signal of counting circuit.

transistor. Compared with a definite reference voltage, this set-up can reduce the voltage fluctuations caused by PVT variation, which improves the accuracy of computing. Fig.5 (a) shows the transient simulation results of TDC in the case that 9 inputs are all '1'. There are 10 cases for the uptrend of SL voltage according to the data in the 9 MTJs in series. If the data stored in the 9 MTJs are all '0', the uptrend of SL voltage is the fastest and the pulse width after it is converted to the time domain is the shortest, representing '0'. Meanwhile, the uptrend of SL voltage will be slower as the number of MTJs storing '1' increases. Besides, it is obvious that the TDC improve the signal margin between adjacent MAC values. In voltage domain, the signal margin between adjacent MAC values is only 10-25 mV, which is difficult to distinguish by using analogue-to-digital converter. After converting to the TD signal, this signal margin is improved to 174-185 ps, which is more than enough for the counting and shifting circuit composed of D flip-flops (DFFs).

Counting and shifting circuit consists of edge-triggered DFFs connected in series, i.e., DFFs chain, as shown in Fig.6 (a). When implementing counting operation, the count pulse (CP) signal is transmitted to the clk terminal of the first DFF in the DFFs chain through the transmission gate controlled by the TD signal. Therefore, the pulse width of TD signal determines the number of rising edges transmitted to the clk

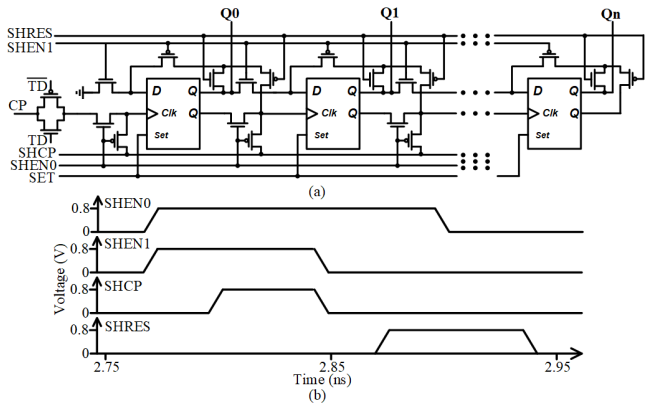

Fig. 6. (a) Counting and shifting circuit. (b) Timing sequences during shifting operation

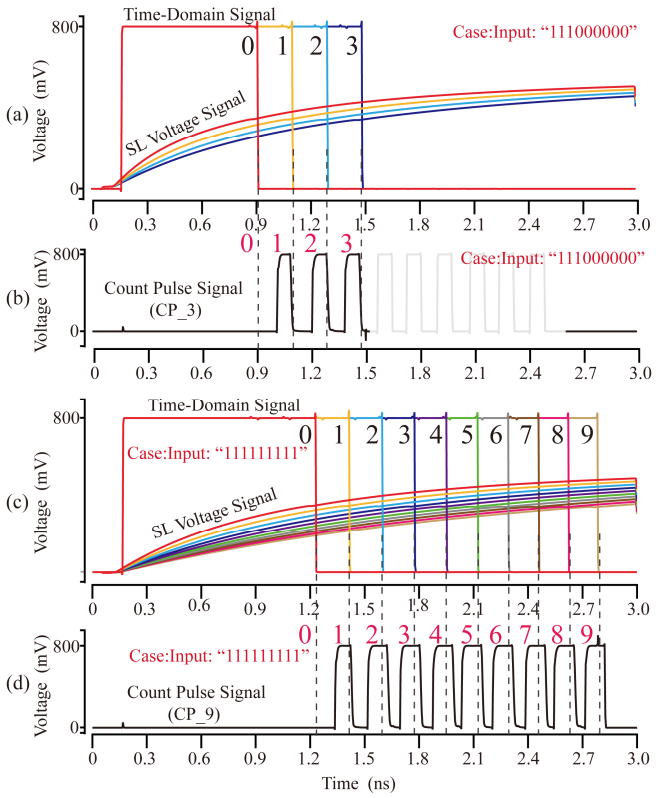

Fig. 7. Transient simulation results of TAM architecture. (a) TD and SL voltage signals in the cases "111000000". (b) CP_3 signal in the cases "111000000". (c) TD and SL voltage signals in the cases "111111111". (d) CP_9 signal in the cases "111111111".

terminal, as shown in Fig. 5(b). For example, 5 rising edges are sent to the clk terminal of the first DFF in the DFFs chain when the MAC value represented by the pulse width of TD signal is 5. This realizes the conversion of TD signals to digital domain, i.e., binary numbers. Note that the signals of SHCP, SHEN0, SHEN1 and SHRES are all zero voltages during counting operation.

For the matrix multiplication, each column in array represents a different weight. Therefore, the results of MAC operations on a column first require to be shifted according to the weight, and then the results of each column are added by utilizing full adder tree to obtain the final result. In existing CIM scheme, the shifting operation is implemented by full adder, which increases the area overhead and cannot be reconfigured (i.e., the number of bits shifted by the same circuit is fixed). In the proposed counting circuit, the shifting operation is simpler. By activating the signals of SHCP, SHEN0, SHEN1 and SHRES, the working mode of DFFs is switched from "counting" to "shifting" via conversion of connections, as shown in Fig.6 (b). Note that the number of the SHCP signal pulses determines the number of bits to shift right, which realizes the reconfigurability. Therefore, the proposed shifting scheme has higher efficiency and lower area overhead.

In the MRAM crossbar array proposed by [17], there is a problem that the different distributions of MTJ resistances used to produce the same expected equivalent resistance on a column lead to different biases for the uptrend of SL voltage, which causes extreme fluctuation in the TD signal, and affects the accuracy of computing. This is considered as a purely data (input)-dependent error. After a detailed circuit analysis, we found that the main reason of this problem was the series connection of CMOS. Due to the threshold loss of CMOS, different positions of CMOS in the series MTJs circuit will cause different voltage drops to leading to the biases for the uptrend of SL voltage. Although the proposed tandem array does not add CMOS to the series MTJs, the TD signals representing the same MAC value do not coincide for different input data, as shown in Fig.7. Obviously, as shown in Fig.7 (a) and (c), the TD signals representing the MAC values of '0', '1', '2' and '3' in the case "111000000" are different from the TD signals representing the MAC values of '0', '1', '2' and '3' in the case "11111111". Note that "111000000" means that there are three '1' in the 9 inputs. Therefore, it is difficult for the same counting pulse signal to accurately convert the TD signal into a binary number in the proposed counting and shifting circuit.

To solve this problem, we propose the dynamic selection circuit of CP signal, as shown in Fig.8, which consists of input dynamic converter, self-generated of count pulses circuit and count pulses selection circuit. Input dynamic converter will translate the 9 inputs into 10 outputs (D0~D9) based on the number of '0' in it. For example, if there are 6 inputs of '1' (e.g., "111111000" or "101101101"), D6 outputs "1" while other bits output '0' regardless of the input permutation. At this time, in count pulses selection circuit, the transistor connected to D6 output is turned on, which connects to the CP_3 and CP, i.e., the output of the dynamic selection circuit is CP_3, as shown in Fig.7 (b). Fig.7 (d) shows the CP_9 that is selected in the case "111111111". Self-generated of count pulses circuit generates 10 CP signals to apply in 10 cases (i.e., 9 inputs define 10 cases according to the number of '0' in it) by delaying the CP_1 which is set. Note that there is a linear relationship between the TD signals in all cases. Therefore, CP_1 is set to a periodic pulse. Meanwhile, its period and pulse width are related to the time difference between adjacent TD signals. Hence, the dynamic selection circuit eliminates the huge biases of the TD signal due to different inputs, i.e., data-dependent error.

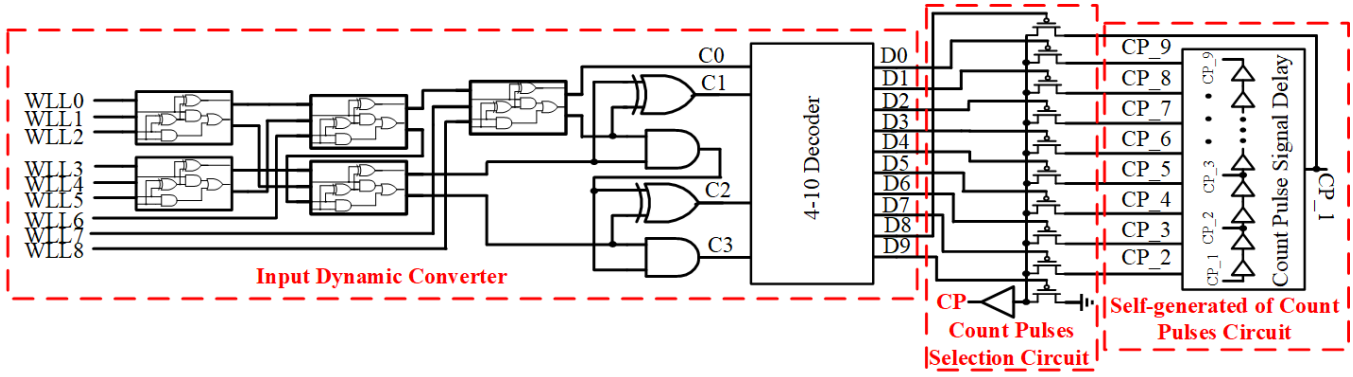

Fig. 8. Count Pulse Dynamic selection circuit.

TABLE I
KEY PARAMETERS OF THE PERPENDICULAR-MAGNETIC-ANISOTROPY MTJ

Parameter	Value
MTJ area	40 nm x 40 nm x π/4
Oxide barrier height of MTJ	0.85 nm
free layer height of MTJ	1.3 nm
Temperature	300 K
Nominal R_{MTJ} at R_L (R_H) of MTJ	3.98 KΩ (8.75 KΩ)
Critical switching current R_H - R_L of MTJ	28 μA
Critical switching current R_L - R_H of MTJ	72 μA
TMR	200%

Based on the voltage-to-time circuit, the proposed counting and shifting circuit and dynamic selection circuit, the MAC operations are efficiently and reliably realized in the proposed tandem array.

IV. PERFORMANCE EVALUATION AND ANALYSIS

To demonstrate the performance advantage of the proposed TAM scheme, we construct a 2Kb TAM architecture by applying 14 nm CMOS process technology and perpendicular-magnetic-anisotropy MTJ compact model [18]-[19]. Table I summarizes the key parameters of the MTJ, which are dependent on physical models and experimental measurements [20]-[21].

As reliability is crucial for implementing computing operations, we first analyze the reliability of the TAM circuit. In addition to the voltage biases due to the input described above, the resistance of MTJ differs from bit-cell to bit-cell due to process variations, because every MTJ is slightly different in its physical properties due to imperfections in the fabrication processes, as is every CMOS. The process variation factors can be simulated by global Monte Carlo simulation which can span over different process corners [22]. Therefore, we carry out the Monte Carlo simulations of 10^4 samples for the proposed TAM scheme to evaluate its computation accuracy at 1V, 25℃, as shown in Fig. 9. Here, the process deviation of the MTJ resistance follow a Gaussian distribution with 5% variability [23]. Simulation results show that the lowest computation accuracy appears in the case of '9' (~92.92%). This is because the large equivalent resistance in the series circuit makes the uptrend of SL voltage more gradual, resulting in higher sensitivity to the process variation. However, the accuracy of 92.92% is normally sufficient for the NN algorithm.

Fig.10 shows the area overhead of 6T SRAM, 3T2MTJ and 3T1MTJ. It is obvious that the bit-cell area of the proposed 3T1MTJ is the smallest. In 3T2MTJ bit-cell, three transferring transistors are used to transmit the write current of MTJ. However, in MRAM, the write current required by

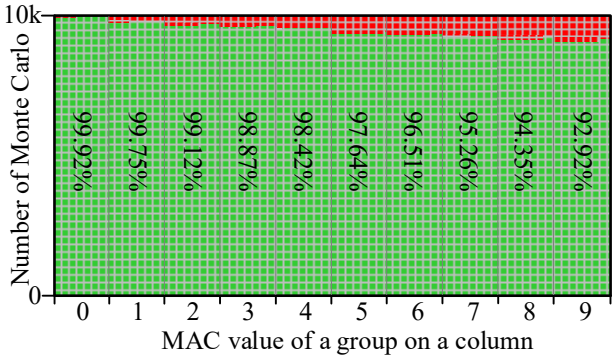

Fig. 9. Accuracy of computing in TAM architecture.

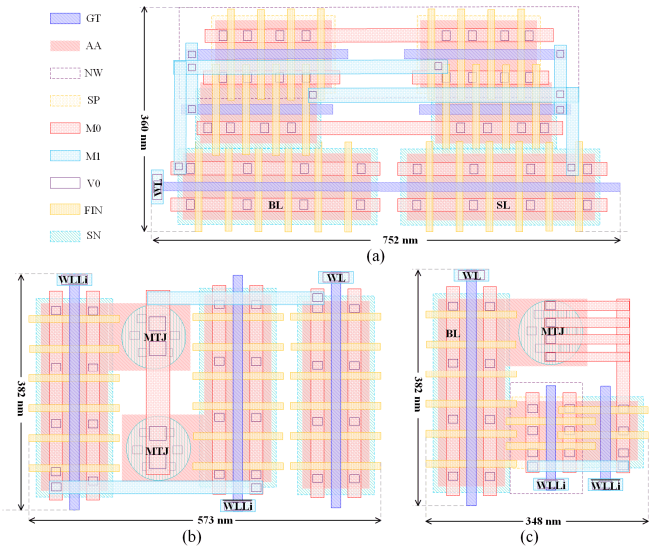

Fig. 10. (a) Layout of 6T bit-cell in SRAM. (b) Layout of bit-cell in [17]. (c) Layout of bit-cell in TAM architecture.

MTJs is relatively large due to its physical properties. The three transistors have to be larger in size to carry the large currents required for write operations. While in the proposed 3T1MTJ bit-cell, only one transistor is used when performing write operation. Therefore, the area of 3T1MTJ bit-cell is reduced by 39.3% and 51 % compared with the 3T2MTJ bit- and the standard 6T bit-cell in SRAM, respectively, which demonstrates that tandem array achieves an efficient MTJ in series by consuming minimal area. It can be concluded that the TAM scheme has the advantage of achieving higher area energy efficiency compared with SRAM based CIM.

Table II compares the proposed TAM scheme with state-of-the-art CIM architectures published in the recent years. Compared with reference [16], the TAM exhibits the excellent advantages of performance. First, in reference [16], the MTJs on a column are connected in parallel, which limits the number accumulated in one MAC operation. Although the authors designed the amplification scheme of TMR to activate more bit-cell, the failure rate of computing is high. Besides, it adopts ADC circuit to realize the conversion of voltage to binary number, which increases the energy compared with TDC. For reference [24], it adopts the charge-recycling voltage-type small-offset sense amplifier (CR-VSA). Although CR-VSA can reduce read energy consumption in MAC operation, its MAC computing delay of 22 ns restricts its performance. Besides, compared with CIM based on SRAM and RRAM with higher switching ratio, TAM scheme realizes the analog computing similar to them through the MTJs in series. Meanwhile, we design the count pulse dynamic selection circuit and counting and shifting circuit to further improve the computing efficiency.

V. CONCLUSION

This article proposes a TAM scheme to achieve energy efficiency and high bit density at the same time. The proposed tandem array based on 3T1MTJ bit-cell realizes the MTJ in series, which implements the analog MAC operation by using resistance summation to carry out the energy efficiency similar to SRAM or RRAM based CIM. To further improve the energy efficiency, a fast and convenient read scheme for MAC result from tandem array is designed using the natural counting and shifting functions of flip-flops. Moreover, we

TABLE II COMPARISON WITH PREVIOUS WORKS

		This work		TCAS-I 2022 [16]	ISSCC 2022 [24]		ISSCC 2020 [25]		ISSCC 2022 [11]	
CMOS Technology		**14 nm**		28 nm	22 nm		22 nm		28 nm	
Memory Type		**STT-MRAM**		STT-MRAM	STT-MRAM		RRAM		SRAM	
Cell Type		**3T1MTJ**		1T1MTJ	1T1MTJ		1T1R		6T	
Array Size		**2Kb**			4Mb		2Mb		1Mb	
Supply Voltage (V)		**0.8**		0.9	0.9		0.7 ~ 0.9		0.65-0.9	
Bit Precision (bit)	Input	**4**	**8**	2	8	8	2	4	4	8
	Weight	**4**	**8**	1	4	8	4	4	4	8
MAC computing delay (ns)		**11.2**	**22.4**	4	22.1	22.1	13.1	18.3	6.5	6.6
Energy Efficiency (TOPS/W)		**113.2**	**63.7**	9.47	127	55.1	45.52	28.93	148.1	37.75

propose a dynamic selection mechanism for count pulse applied to eliminate the voltage biases due to the input and the different distributions of MTJ resistance, which improves the accuracy of computing. Finally, we achieve 113.2TOPS/W and 63.7 TOPS/W in the 2Kb TAM architecture for 4-bit and 8-bit precision, respectively. Meanwhile, the bit-cell area of tandem array is also saved by 39.3%. Therefore, this work exhibits significance for further research on MRAM to realize high-performance and high-density computing systems.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61971024, 62122008, and 51901008, and in part by the International Mobility Project under Grant B16001.

REFERENCES

[1] M. M. Waldrop, "The Chips Are Down for Moore's Law," Nature, vol. 530, pp. 144-147, Feb. 2016.

[2] M. Ali, A. Jaiswal, S. Kodge, A. Agrawal, I. Chakraborty and K. Roy, "IMAC: In-Memory Multi-Bit Multiplication and ACcumulation in 6T SRAM Array," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 67, no. 8, pp. 2521-2531, Aug. 2020.

[3] J. Wang *et al*., "A 28-nm Compute SRAM With Bit-Serial Logic/Arithmetic Operations for Programmable In-Memory Vector Computing," IEEE Journal of Solid-State Circuits, vol. 55, no. 1, pp.76-86, Aug. 2019.

[4] A. Sayal, S. S. T. Nibhanupudi, S. Fathima and J. P. Kulkarni, "A 12.08-TOPS/W All-Digital Time-Domain CNN Engine Using Bi-Directional Memory Delay Lines for Energy Efficient Edge Computing," IEEE Journal of Solid-State Circuits, vol. 55, no. 1, pp. 60-75, Jan. 2020.

[5] W. Wan et al., "A compute-in-memory chip based on resistive random-access memory," Nature, vol. 608, pp. 504–512, Aug. 2022.

[6] J.M. Hung et al., "A four-megabit compute-in-memory macro with eight-bit precision based on CMOS and resistive random-access memory for AI edge devices," Nat. Electron, vol. 4, pp. 921–930, Dec. 2021.

[7] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," IEEE Journal of Solid-State Circuits, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[8] M. Kang, S. K. Gonugondla, A. Patil and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6T SRAM array," IEEE Journal of Solid-State Circuits, vol. 53, no. 2, pp. 642–655, Feb. 2018.

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

[10] 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 Journal of Solid-State Circuits, vol. 53, no. 1, pp. 188-198, Jan. 2021.

[11] P. C. Wu et al., "A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices," IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, Feb. 2022, pp. 190–191.

[12] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in memory with spin-transfer torque magnetic RAM," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 26, no. 3, pp. 470–483, Mar. 2018.

[13] M. Zabihi, Z. I. Chowdhury, Z. Zhao, U. R. Karpuzcu, J.-P. Wang, and S. S. Sapatnekar, "In-memory processing on the spintronic CRAM: From hardware design to application mapping," IEEE Transactions on computer, vol. 68, no. 8, pp. 1159–1173, Aug. 2019.

[14] Y. Zhang et al., "Time-Domain Computing in Memory Using Spintronics for Energy-Efficient Convolutional Neural Network," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 3, pp. 1193-1205, March 2021.

[15] J. K Wang et al., "Reconfigurable Bit-Serial Operation Using Toggle SOT-MRAM for High-Performance Computing in Memory Architecture," IEEE Transactions on Circuits and Systems I: Regular Papers, in press.

[16] H. Cai et al., "Proposal of Analog In-Memory Computing With Magnified Tunnel Magnetoresistance Ratio and Universal STT-MRAM Cell," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 4, pp. 1519-1531, Apr. 2022.

[17] Jung, S., et al., "A crossbar array of magnetoresistive memory devices for in-memory computing," Nature vol. 601, pp. 211–216, Jan. 2022.

[18] Y. Zhang et al., "Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions," IEEE Transactions on Electron Devices, vol. 59, no. 3, pp. 819–826, Mar. 2012.

[19] Y. Wang, et al, "Compact model of dielectric breakdown in spin transfer torque magnetic tunnel junction," IEEE Transactions on Electron Devices, vol. 63, no. 4, pp. 1762-1767, Apr. 2016.

[20] Z. Wang, W. Zhao, E. Deng, J.-O. Klein, and C. Chappert, "Perpendicular-anisotropy magnetic tunnel junction switched by spin-Hall-assisted spin-transfer torque," Journal of Physics D: Applied Physics, vol. 48, no. 6, Jan. 2015, Art. no. 065001.

[21] M. Wang et al., "Field-free switching of a perpendicular magnetic tunnel junction through the interplay of spin–orbit and spin-transfer torques," Nature Electron., vol. 1, pp. 582–588, Nov. 2018.

[22] T. H. Choi, H. Jeong, Y. Yang, J. Park, and S.-O. Jung, "SRAM operational mismatch corner model for efficient circuit design and yield analysis," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 64, no. 8, pp. 2063–2072, Aug. 2017.

[23] J. K. Wang et al., "A self-matching complementary-reference sensing scheme for high-speed and reliable toggle spin torque MRAM," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 67, no. 12, pp. 4247–4258, Dec. 2020.

[24] Y. C. Chiu et al., "A 22nm 4Mb STT-MRAM Data-Encrypted Near-Memory Computation Macro with a 192GB/s Read-and-Decryption Bandwidth and 25.1-55.1TOPS/W 8b MAC for AI Operations," IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, Feb. 2022, pp. 178–179.

[25] C. X. Xue et al., "A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices," IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, Feb. 2020, pp. 244–249.