Input Sparsity Aware In-Memory Computing Macro Based on SOT-MRAM Multi-Level Cell for Efficient Deep Neural Network Acceleration

Chao Wang[*], Qihang Gao[†‡§], Xianzeng Guo[*†‡§], Zhongzhen Tong[*†‡§], Zhaohao Wang[*†‡§]*, and Weisheng Zhao[*†‡§]
[*]State Key Laboratory of Spintronics, Hangzhou International Innovation Institute, Beihang University
[†]Fert Beijing Research Institute, Beihang University
[‡]MIIT Key Laboratory of Spintronics, Beihang University
[§]School of Integrated Circuit Science and Engineering, Beihang University
Hangzhou 311115, China. *Corresponding Author Email: zhaohao.wang@buaa.edu.cn

Abstract—**Deep neural network (DNN) technology has gained widespread applications, but its high energy demands continue to drive the advancement of low-power computing architectures, particularly in in-memory computing (IMC) architectures based on non-volatile memory. Among these, spin-transfer torque magnetic random-access memory (STT-MRAM)-based IMC architectures have achieved some progress, but their performance remains constrained by limited resistance and binary characteristics. By contrast, the next-generation spin-orbit torque MRAM (SOT-MRAM) offers superior magnetic tunnel junction (MTJ) resistance and more flexible cell structures, presenting significant potential for energy-efficient IMC implementation. In this work, leveraging the ultra-high MTJ resistance and the separation of read/write paths in SOT-MRAM, we propose a multi-level cell (MLC) structure-based high energy-efficiency IMC architecture (MLC-SOT-IMC), which performs standard multiplication operations by optimizing the conductance mapping paradigm. The proposed architecture not only maintains high inference accuracy but also significantly enhances integration density and reduces the overhead per bit. Additionally, a self-terminating time-to-digital converter (TDC) readout circuit, which is dependent on input sparsity, is introduced to eliminate the excess power consumption associated with ineffective pulses after readout completion. Ultimately, the proposed MLC-SOT-IMC architecture achieves an inference energy efficiency of 6388.98 1-bit TOPS/W under an input sparsity of 50%, with the peak energy efficiency reaching 8426.19 1-bit TOPS/W at an input sparsity of 90%.**

Keywords—**In-memory computing, Spin-orbit torque magnetic random access memory, Multi-level cell, Time-to-digital converter, High energy-efficiency**

I. Introduction

With the rapid development of deep neural networks (DNN), there has been an explosive growth in data volume, which has led to processors frequently accessing memory and transferring large amounts of data, resulting in significant increases in power consumption. In-memory computing (IMC) technology effectively addresses the issue of frequent data movement in traditional computing architectures and breaks through the performance bottleneck known as the "memory wall" [1], [2]. Among various emerging memory technologies, magnetic random access memory (MRAM) stands out for its non-volatile, high endurance, high-speed writing and low power consumption, and IMC based on MRAM has garnered widespread attention in recent years [3].

Among the diverse MRAM technologies, spin-transfer torque MRAM (STT-MRAM) has become one of the most promising memory technologies due to its high technology maturity, CMOS compatibility, and integration density [2], [4]–[7]. However, the magnetic tunnel junction (MTJ) resistance (R_{MTJ}) in STT-MRAM is typically limited to below 10 kΩ due to the current-induced writing mechanism. Lower R_{MTJ} results in higher operating currents and increased power consumption, thereby hindering the further advancement of STT-MRAM in the domain of high energy-efficient analog IMC [2], [8], [9].

In contrast, benefiting from the separation of its read and write paths [10]–[12], the R_{MTJ} of spin-orbit torque MRAM (SOT-MRAM) can theoretically be freely adjusted within the kΩ-MΩ range [10], effectively suppressing inference currents and energy consumption while achieving sub-nanosecond write speeds and requiring low write voltages. Meanwhile, the SOT-MRAM exhibits greater potential for achieving higher on/off ratio, mitigating the low inference accuracy caused by the limited resistance contrast in STT-MRAM [13], [14].

Leveraging the ultra-high R_{MTJ} and the separation of read/write paths in SOT-MRAM, this paper proposes a high energy-efficiency IMC architecture design based on the multi-level cell (MLC) structure of in-plane SOT-MRAM (MLC-SOT-IMC) which performs standard multiplication operations with a self-terminating readout circuit based on time-to-digital converter (TDC). The main contributions are as follows:

1. A novel IMC bit-cell employing the MLC structure based on in-plane SOT-MRAM is proposed. By appropriately setting the cross-sectional areas of the MTJs and the SOT layer, the two MTJs achieve different R_{MTJ}s and obtain distinct write threshold currents, which enables the storage of multi-bit data and stepwise writing, realizing high density and efficiency IMC.

2. A conductance mapping paradigm and crossbar array design based on the MLC structure is proposed to address the non-correspondence between conductance outcomes and inference results during standard multiplication operations in traditional

978-3-9826741-1-7/DATE26/© 2026 EDAA

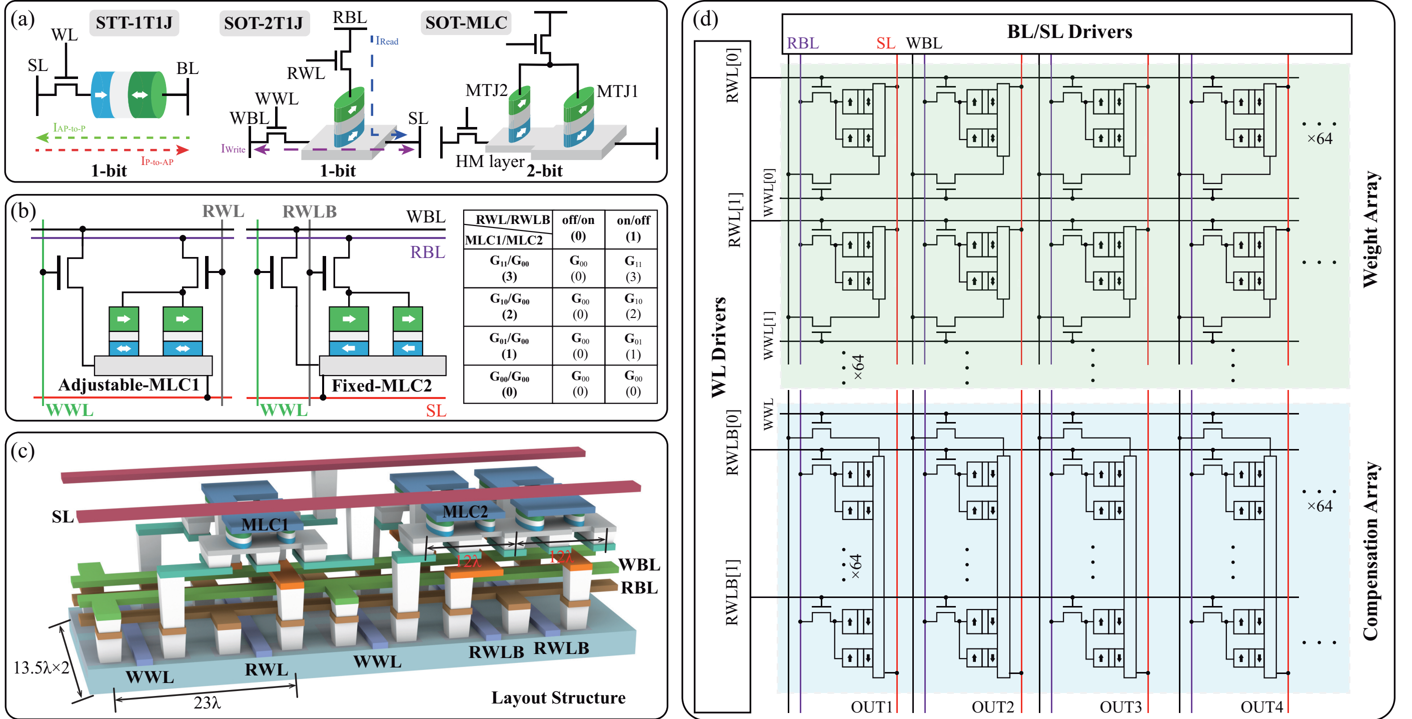

Fig. 1. (a) Schematics of the conventional STT-MRAM, SOT-MRAM and SOT-MLC bit-cells. (b) The proposed MLC-SOT-IMC bit-cell structure and the Truth Table for standard multiplication operations. (c) The proposed layout structure of the MLC-SOT-IMC bit-cell. The widths 12 λ and 23 λ correspond to the internal bit-cells of the array. The illustration depicts the edge of the array which is slightly larger than the internal bit-cells. (d) The proposed IMC crossbar array structure.

MRAM. Each bit-cell is composed of a weight MLC whose conductance state can be freely adjusted, and a compensation MLC that is fixed in a low-conductance (G_L) state. The crossbar array consists of a weight array, formed by weight MLCs, and a compensation array, formed by compensation MLCs. The compensation array integrates all compensation bit-cells on the same SOT layer and is collectively written by a single write transistor, which not only reduces the write power consumption but also saves area overhead.

3. A readout circuit based on TDC equipped with input sparsity awareness and self-terminating functionality is proposed. Once the array computation results are read out, the self-terminating module deactivates the readout circuit to prevent power consumption associated with redundant pulse signals. Based on 28 nm CMOS simulation, when the input sparsity is 50%, the inference energy efficiency of the proposed IMC architecture employing the self-terminating TDC circuit reaches 6388.98 1-bit TOPS/W, with the peak energy efficiency reaching 8426.19 1-bit TOPS/W at an input sparsity of 90%.

II. The Proposed MLC-SOT-IMC Architecture with Standard Multiplication Paradigm

A. Multi-level cell structure of SOT-MRAM

Leveraging its distinct advantages such as the separation of read/write paths and ultrafast read/write speeds, SOT-MRAM has garnered extensive attention in the realm of IMC [12], [15]. Nevertheless, compared with resistive random access memory (RRAM) and phase change memory (PCM) that can perform multi-bit operations in single bit-cell [16], [17], traditional SOT-MRAM has certain disadvantages in terms of data integration density. Additionally, when compared with the conventional 1-transistor-1-MTJ (1T1J) structure of STT-MRAM, each fundamental SOT-MRAM bit-cell necessitates two transistors to separately govern the read and write paths, as shown in Fig. 1(a)&(b). This requirement incurs substantial area overhead in the application of high-parallelism analog IMC arrays, thereby significantly impeding the further advancement of SOT-MRAM in the field of IMC.

To address this issue, the MLC structure of SOT-MRAM is adopted. Each MLC cell can store 2-bit data and still requires only two transistors to control read and write operations, as shown in Fig. 1(a), thereby markedly enhancing data storage density compared with traditional structures [18]–[21]. The in-plane MTJ structure employed consists of an MTJ layer and a heavy metal (HM) layer, which offers advantages such as a low critical switching current, mature fabrication processes, high tunneling magnetoresistance (TMR) and field-free switching capability [13], [14], [22]. The cross-sectional areas of the HM layer and the MTJs can be independently adjusted to achieve optimal write reliability and weight mapping.

B. Standard multiplication paradigm

In efficient quantized DNN algorithms, input vectors IN and weight vectors W are typically represented as integer (INT) type. In IMC applications, integer multiplication is commonly decomposed into binary standard multiplications [23]. Specifically, for high-parallelism binary multiplication vectors, if the IN vector contains N "1"s, the column accumulation process following the standard multiplication by the bit-cells involves at

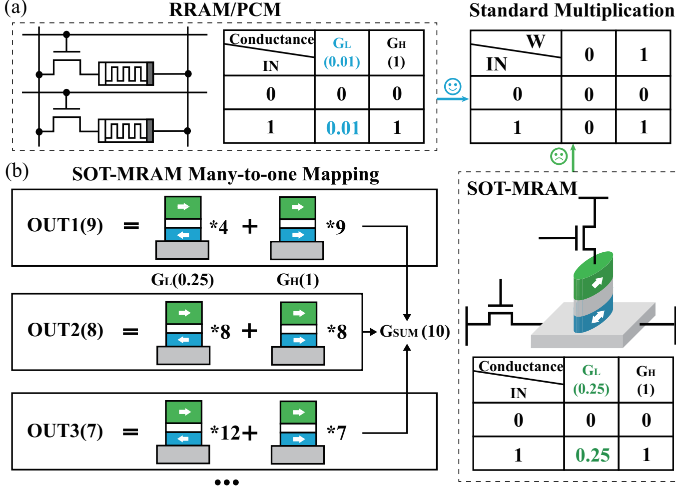

Fig. 2. (a) Schematic of conductance computation and Truth Table mapping for traditional RRAM/PCM and SOT-MRAM in standard multiplication paradigm. (b) The many-to-one mapping between G_{SUM} and output result OUT in MRAM based standard multiplication implementation.

most N "1"s being added together, with the inference result not exceeding "N". When passing the inference result to the readout circuit, the readout circuits can be selectively activated based on the number of "1"s in the IN vector, which is related to input sparsity, thereby conserving unnecessary power consumption.

Fig. 2(a) illustrates the implementation of mainstream standard multiplication in IMC [23], while RRAM and PCM are usually adopted due to their high on/off ratios which can achieve a $G_{\mathrm{L}} \leq$ "0.01" when high-conductance (G_{H}) is quantized to "1", close to the off-state with zero conductance. In contrast, MRAM has a lower on/off ratio, while the conductance values of G_{L} and off-state, both representing "0", show significant differences, making it impossible for the sum of the column conductances (G_{SUM}) to accurately map to the output results [24]. For instance, with TMR = 300%, the G_{H} is normalized to "1" while the G_{L} is "0.25". When performing standard multiplication, if the G_{SUM} is "10", the result may be "10"(10 G_{H}s), or "9"(9 G_{H}s and 4 G_{L}s), or "8"(8 G_{H}s and 8 G_{L}s), and could even be some other number as shown in Fig. 2(b). Therefore, MRAM cannot directly support standard multiplication without precise post-processing, which would significantly increase design complexity [24]. Consequently, mainstream MRAM-based IMC architectures still rely on XNOR logic, which exhibits significant application limitations and is independent of input sparsity [23]–[25].

C. Proposed MLC-SOT-IMC architecture

To realize standard multiplication operations and achieve high energy efficiency, this paper proposes an IMC architecture with array size of 64×64 based on the SOT-MLC. A 64×64 compensation array with all bit-cells being fixed at G_L states is introduced to implement standard multiplication operations, combined with a readout method related to input sparsity to achieve extremely high inference energy efficiency.

The bit-cell structure and operation Truth Table are shown in Fig. 1(b), with the layout structure being shown in Fig. 1(c).

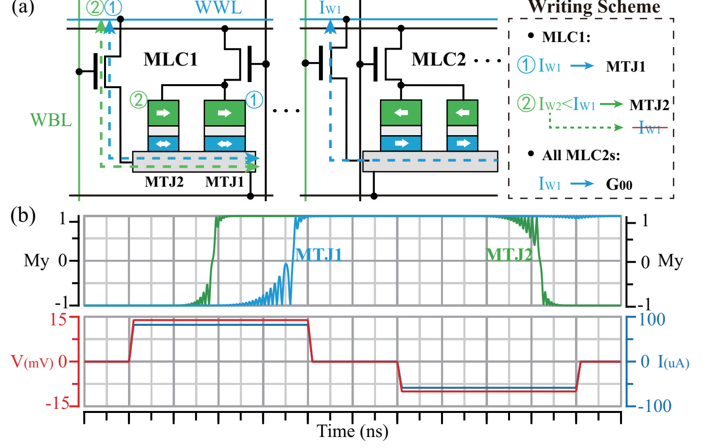

Fig. 3. (a) Schematic diagram of the write operation for the proposed MLC-SOT-IMC bit-cell. (b) The practical write waveforms involve the y-axis component of the magnetization (My) of MTJ1/MTJ2 in the MLC and the write voltage/current across the heavy metal layer.

The layout structure employs the mainstream SOT-MRAM design rules and utilizes a staggered layout optimization method to reduce the cell width [26]. Each bit-cell consists of a weight MLC1 and a compensation MLC2. In each MLC structure, the different sizes of the two MTJs result in different R_{MTJ}s. This allows the weight MLC structure stores four different conductance states G_{11}, G_{10}, G_{01}, and G_{00}, which are mapped to "3", "2", "1", and "0", respectively, in the standard multiplication logic, while the conductance state of the compensation MLC structure is fixed at G_{00}. This conductance allocation, which maps the G_{00} state solely to logic "0", resolves the issue of many-to-one correspondence between computation results and output results. When the resistance ratio of the two MTJs $R_{\mathrm{MTJ2}}/R_{\mathrm{MTJ1}}$ is set as 2, the four different conductance values can be linearly distributed. This linear distribution allows for uniform distribution of different column results in subsequent quantization, thereby achieving the highest possible readout accuracy. For example, when the TMR is 300%, the ratio of the four conductance states is calculated to be 4:3:2:1. The relationship between the bit-cell on/off ratio and TMR is as follows:

$$\text{On/off ratio} = \frac{G_{\mathrm{L}}}{G_{\mathrm{H}}} = \mathrm{TMR} + 1. \tag{1}$$

In each MLC cell, the different cross-sectional areas of the HM layers in the two MTJs lead to distinct write threshold currents, namely I_{W1} and I_{W2} ($< I_{\mathrm{W1}}$). The write operation for a weight MLC proceeds in two potential steps, as shown in Fig. 3(a):

1) **Writing to MTJ1:** Initially, I_{W1} is applied to write the desired state to MTJ1. During this first step, due to the coupled nature of the write operation, the magnetization direction of MTJ2 is also changed.
2) **Correcting MTJ2 (if needed):** Subsequently, I_{W2} is applied to overwrite any erroneous write that occurred in MTJ2 during the first step, thereby ensuring MTJ2 achieves its correct target state. Naturally, if the initial write operation to MTJ2 in the first step was already correct, the application of I_{W2} is omitted.

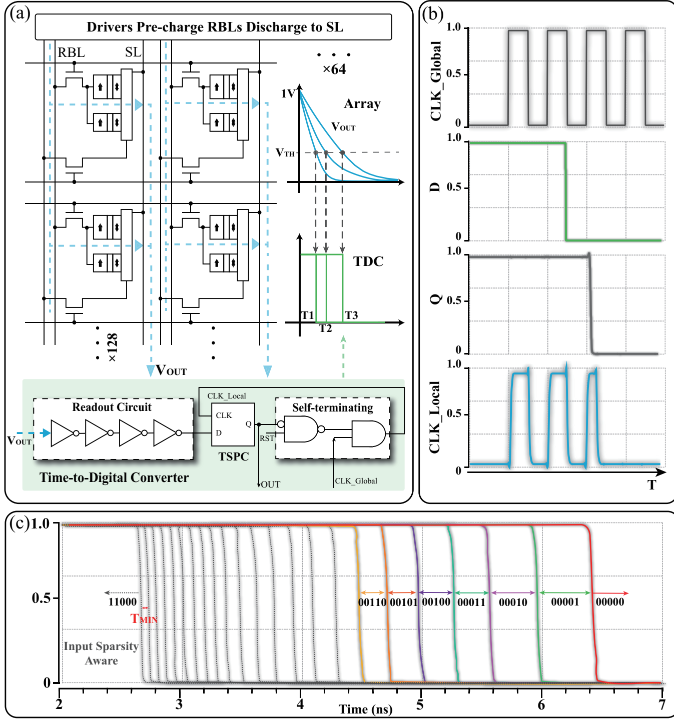

Fig. 4. (a) Schematics of the transformation of the inference result from the voltage-domain to the time-domain and the proposed self-terminating TDC readout circuit. (b) Practical pulse waveforms of the self-terminating circuit. (c) Waveforms of 23 different readout results corresponding to input sparsity.

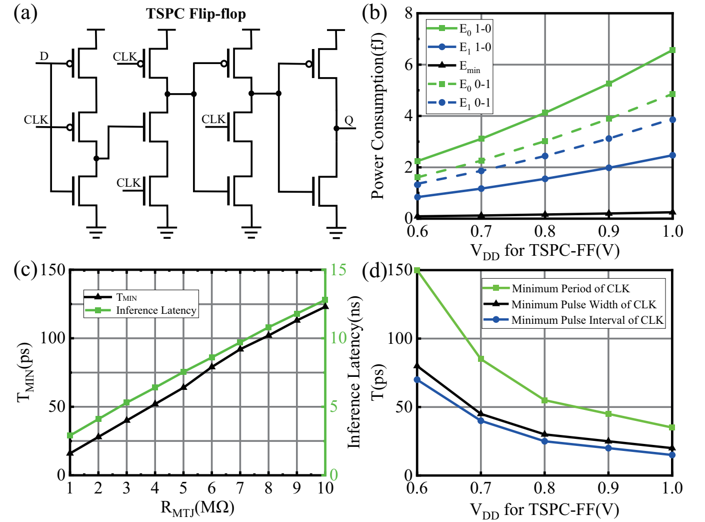

Fig. 5. (a) Schematic of the TSPC-FF circuit. (b) The energy consumption of TSPC-FF with different V_{DD}. E_0 quantifies power consumption during output-0 pulses, E_1 during output-1 pulses, and $E_{\min}$ represents the minimized power state after circuit termination. The notation "0-1" refers to the input voltage of the TSPC-FF transitioning from low to high, while "1-0" indicates a transition from high to low. (c) The minimum time interval and inference delay for different quantization results in TDC-based IMC among various R_{MTJ}. R_{MTJ} denotes the resistance of the MTJ in the MLC structure that represents the smaller resistance value R_{MTJ1}. (d) The shortest operating clock cycle with different V_{DD}.

The practical write implementation waveform is shown in Fig. 3(b). In the compensation array, only I_{W1} is required and the write current for MTJ is approximately 80 μA. Owing to the ultra-low resistivity of the doped-W material, the total HM resistance of the two SOT layers is about 170 Ω [22]. Therefore, a voltage of approximately 14 mV is required to drive a single MLC, and only 0.9 V is needed to drive the entire compensation array. This allows all compensation bit-cells to be integrated into the same HM layer and controlled by a single write transistor, which greatly saves architecture area overhead. Once written to the G_{00} state, the magnetization states of the compensation cells remain unchanged.

The proposed MLC-SOT-IMC array is shown in Fig. 1(d), which consists of a 64×64 weight array and a 64×64 compensation array. During the execution of standard multiplication operations, the conductance states stored in the MLC bit-cells are represented as the weight vector W. The input vector IN selects the corresponding MLCs by turning on the inference transistors through read word line (RWL/RWLB). The weights stored in the selected MLCs within each bit-cell are represented as the output vector OUT of the bit-cell. The G_{SUM} of 64 bit-cells in each column serves as the output result of the standard multiplication operation. During the inference operation, the read bit line (RBL) capacitor is precharged, and then the charge is released to the grounded source line (SL) through the parallel bit-cells. The discharge rate is positively correlated with G_{SUM}, thereby G_{SUM} is converted to the voltage domain V_{OUT}, which is input to the readout circuit for quantization.

III. TDC-Based Self-Terminating Readout

Fig. 4(a) illustrates the array discharge schematic and the proposed self-terminating TDC readout circuit. V_{OUT} of inference column discharges exponentially toward GND, with the time required to reach the inverter threshold voltage V_{TH} exhibiting linear dependence on G_{SUM}. For each inference column, when $V_{\mathrm{OUT}} < V_{\mathrm{TH}}$, the inverter output transitions to a low voltage state, and the precise flip time serves as the basis for quantizing different output voltages into discrete digital results.

Recognizing that continued time-domain readout after the inverter state transition is functionally redundant and incurs unnecessary power consumption, a self-terminating control scheme has been developed to automatically disable the readout process. The proposed self-terminating module incorporates a NAND gate for resetting: when terminal RST is "1", terminal Q gates the CLK signal through an AND logic operation; when RST is "0", the NAND output remains "1", ensuring the CLK passes through. Specifically, the true single phase clock flip-flop (TSPC-FF) continuously monitors the pulse sequence of the inverter output. When the inverter output transitions to a low voltage level, the output pulse of the TSPC-FF also transitions to a low voltage level and is transmitted to the self-terminating module. Upon receiving the inverter state transition signal, the self-terminating module controls the CLK pulse to deactivate the TSPC-FF, thereby completing the self-termination scheme to conserve inference energy. The practical pulse waveform of the self-terminating scheme is shown in Fig. 4(b).

Under the 64×64 MLC-SOT-IMC architecture, the standard Multiply-Accumulate (MAC) operation of bit-cells with 1-bit input and 2-bit weights can produce 193 possible output results ranging from "0" to "192". The readout process maps 193

TABLE I
DEVICE PARAMETERS USED IN OUR SIMULATIONS

Symbol	Parameter	Value	Unit
-	MTJ1 area	$105\times42\times\pi/4$	nm^2
-	MTJ2 area	$75\times30\times\pi/4$	nm^2
t_{HM}	HM layer thickness	4.5 [22]	nm
ρ_{HM}	HM layer resistivity	80 [22]	$\mu\Omega$·cm
θ_{SH}	Spin Hall angle	0.6 [22]	-
t_{FL}	Free layer thickness	1.5	nm
α	Damping constant	0.02	-
M_s	Saturation magnetization	1.2×10^6	A/m
T	Temperature	300	K
R_L	Low resistance of MTJ1	5	MΩ
TMR_0	TMR with 0 V bias	300% [13]	-
V_h	Voltage bias when $TMR=TMR_0/2$	0.8	V
-	Wiring Parasitic Capacitance	0.288 [27]	$fF/\mu m$
-	Wiring Parasitic Resistance	16.2 [27]	$\Omega/\mu m$

distinct outputs to 24 readout states ("00000"–"11000") using 23 precisely timed pulses, with every 8 consecutive distinct outputs being encoded into the same binary output. Specifically, the outputs "9", "10", "11", "12", "13", "14", "15", and "16" are mapped to "00001". The pulse timing in the readout circuit is adaptively adjusted based on input sparsity, that is, the number of "0"s in the input vector. For example, when a 64-bit binary input vector containing 18 "1"s is subjected to a MAC operation with a MLC array that stores 2-bit data per cell, the maximum possible outcome is "54", which is read out as "00110" in TDC circuit. The output flips at the 17th pulse, allowing the omission of the first 16 pulses to save energy, as shown in Fig. 4(c). For ideal weight distributions in DNN, the number of pulses can be further reduced by approximately 50%. Compared to conventional voltage-domain readout approaches using sense amplifier (SA) or analog-to-digital converters (ADCs), this TDC-based implementation features substantially reduced circuit complexity and power consumption. Due to the small area overhead of the TDC circuit, the architecture enables dedicated readout circuits for each column while maintaining high precision, eliminating WL power waste through parallel operation.

Fig. 5(a) illustrates the employed TSPC-FF structure in TDC circuit. When the input voltage to the TSPC-FF is at a high voltage level, the internal voltage state remains essentially unchanged, resulting in energy consumption E_1 that is lower than the energy consumption E_0 when the input voltage is at a low level. Therefore, a four-stage inverter structure is adopted in the readout circuit, which not only enhances the driving capability but also ensures that the D pulse remains at a high voltage level for most of the time under high input sparsity. Once the voltage signal of the TSPC-FF flips, the self-terminating module regulates its deactivation and only one E_0 is required, thereby reducing power consumption. The different power consumption of the TSPC-FF under different driving voltages is shown in Fig. 5(b).

Fig. 5(c) illustrates the minimum time interval between different output results and the inference latency of the proposed architecture under various R_{MTJ} and Fig. 5(d) shows the minimum clock period of the TSPC-FF under different driving voltages. Considering both the timing margin requirements for different inference results and the fundamental relationship between R_{MTJ} and the minimum inversion interval of TDC, the stable operation requires minimum inversion intervals exceeding 60 ps. Therefore, $R_{MTJ} = 5$ MΩ is selected as the optimal value to meet the delay requirements of the proposed TDC readout circuits while ensuring a certain level of architectural inference speed. To achieve a balance between energy efficiency and inference latency in the MLC architecture, we optimized the self-terminating circuit for 0.8 V operation with a demonstrated 55 ps minimum period to maintain optimal performance-energy balance.

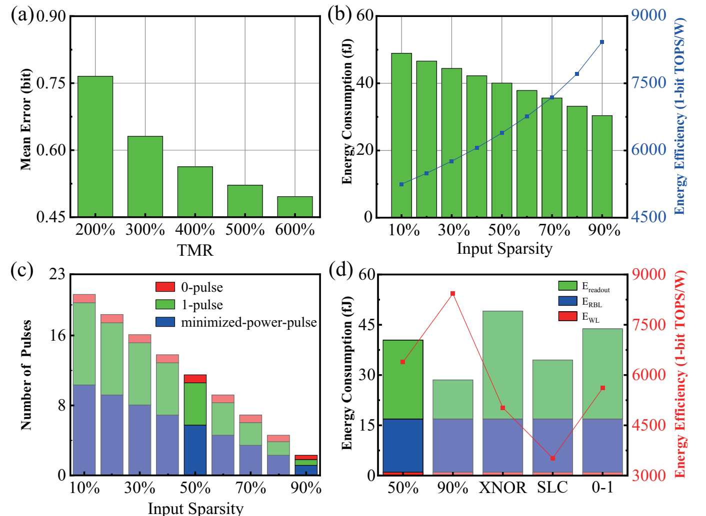

Fig. 6. (a) Inference mean error of the proposed architecture under different TMR. (b) Energy consumption and efficiency of the proposed architecture under different input sparsity. (c) Number of pulses required for each readout operation at different input sparsity, where the minimized-power-pulse refers to the minimum power consumption required after the circuit termination. (d) The energy consumption distribution and energy efficiency in the proposed architecture, where "0-1" refers to the case with an input sparsity of 50% and "SLC" refers to the IMC architecture based on SLC-SOT cells.

IV. SIMULATION AND ANALYSES

In this section, we evaluate the performance of MLC-SOT-IMC architecture via extensive Monte Carlo simulations, employing in-plane SOT-MRAM devices from a 28 nm CMOS node. Key parameters include a 5% MTJ variation (1σ), a read transistor width of 56 nm, and a wire parasitic capacitance of 0.288 fF/μm [27]. The RBL voltage is 1 V, and the control voltage of RWL is selected to be 0.6 V. Table I presents the complete parameter set for the proposed architecture.

The inference accuracy of the proposed IMC architecture is quantified by the mean error [2], defined as the average discrepancy between ideal and actual quantized readout results across all outputs. For example, an ideal "00110" for "54" versus an actual "00100" signifies an error of 2 least significant bits (LSBs). This mean error is calculated by weighting and averaging the probabilities of all error cases, then averaging over all output scenarios. In practical circuits, the effect of 20 ps clock jitter is considered to simulate real-world circuit conditions. Fig. 6(a) illustrates the inference errors of the proposed architecture for different TMR when $R_{MTJ1} = 5$ MΩ. The TMR of in-plane SOT-MRAM can currently reach 300%,

TABLE II
COMPARISON OF THE PROPOSED DESIGNS WITH THE PREVIOUS WORKS

	Nature22 [2]	VLSI23 [4]	ISSCC23 [5]	JSSC24 [6]	TCAS-I24 [28]	DAC24 [12]	This Work
Memory Type	STT-MRAM	STT-MRAM	STT-MRAM	STT-MRAM	STT-MRAM	SOT-MRAM	SOT-MRAM
Process Node	28 nm	28 nm	28 nm	22 nm	28 nm	28 nm	28nm
Cell Structure	3T-2J	2T-2J	4T-2J	2T-2J	2T-2J	1.5T-1J	3T-4J[c]
Logic Type	XNOR	XNOR	XNOR	XNOR	XNOR	AND[b]	AND[b]
Parallelism	64	32	9	512	512	512	128
Output Precision (bit)	4	3	1	6	22	8	4[d]
Mean error (bit)	0.47-0.83	1.665	–	–	–	–	0.63[d]
Readout Circuit	TDC	SAR-ADC	CR-C^2R	ADC	SAR-ADC	ADC	TDC
Frequency (MHz)	11.1	200	20	50	–	18.2	133
Energy Efficiency[a]	262-405	709.3	35.2	37.2-84.2	2329.2	23.7-29.4	6388.98[e] - 8426.19[f]

[a]1-bit TOPS/W. [b]Standard multiplication. [c]The write transistor in the compensation array can be ignored.
[d]For most significant bit (MSB). [e]Input sparsity of 50%. [f]Input sparsity of 90%.

with a maximum value approaching 600% [13], [14]. Under TMR = 300%, the inference error is 0.63-bit.

Inference energy consumption is a crucial metric for evaluating the performance of an IMC architecture. In the proposed MLC-SOT-IMC architecture, the vector product calculation involves a 64-bit input vector and a 128-bit weight vector. A single readout operation of the TDC can perform 128 bitwise multiplications and 128 accumulations, totaling 256 operations. Under TMR = 300%, the proposed MLC-SOT-IMC architecture achieves a power efficiency of 6388.98 1-bit TOPS/W with 50% input sparsity. Fig. 6(b) illustrates the impact of varying input sparsity levels on the inference energy efficiency of the architecture. Given the inherent nature of input sparsity in neural networks, it is inevitable that within high-volume, 64-bit binarized input vectors, some instances will have fewer than eight "1"s. In these specific cases, the input quantizes to "00000", and the IMC architecture circuit remains inactive, thus conserving energy. Fig. 6(c) illustrates the number of pulses required under different input sparsity. In practical DNN systems, input sparsity can reach higher levels, thereby further reducing inference energy consumption. At 90% input sparsity, the peak energy efficiency reaches 8426.19 1-bit TOPS/W.

The power consumption distribution and efficiency of the proposed IMC architecture is shown in Fig. 6(d). Compared to direct current method, the discharge-based readout method can significantly reduce the inference column power consumption. By dedicating a quantization circuit to each column, power waste at the WL is avoided, and the high R_{MTJ} allows for the use of lower WL voltages, thereby effectively saving energy while maintaining computational accuracy. Moreover, the MLC structure further boosts density and inference energy efficiency. While traditional single-level cell (SLC) requires 128 operations (64 multiplications + 64 accumulations) per readout, and their TDC circuits consume less power, the proposed MLC architecture effectively doubles operations per readout, leading to an 81.3% improvement in inference energy efficiency. Additionally, the compensation array sharing the same HM layer and controlled by a single write transistor further reduce the power consumption waste caused by transistor capacitance and metal line capacitance, resulting in a 7.6% improvement in efficiency.

The proposed input sparsity-related readout paradigm effectively enhances inference energy efficiency. When performing XNOR operations, the input sparsity-related paradigm can not be executed. Consequently, the TDC readout circuit will require higher power consumption. Compared to XNOR operations, the self-terminating TDC readout circuit achieves a power savings of 21.3% during standard multiplication operations, which results in an overall improvement of 27.1% in the inference energy efficiency of the proposed architecture. Moreover, to further enhance the energy efficiency, a four-stage inverter structure is employed, which allows the TSPC-DFF voltage to transition from high to low. Compared to the opposite voltage signal generated by a three-stage inverter, the four-stage inverter structure reduces the power consumption of the TDC readout circuit by 12.2%, as shown in Fig. 6(d).

Table II compares this work with advanced IMC implementations using STT-MRAM and SOT-MRAM from the past three years. Most of them are due to the low R_{MTJ} and low on/off ratio, which lead to excessive inference current and thereby reduce inference energy efficiency. The proposed IMC array, integrating high R_{MTJ}, multi-bit computation using MLC bit-cell, excellent array design, and high energy-efficient readout circuit design, achieves extremely high inference energy efficiency while maintaining high computational accuracy.

V. CONCLUSION

This work introduces a high energy efficiency IMC architecture leveraging in-plane SOT-MRAM MLCs. By exploiting the ultra-high resistance and separate read/write paths of SOT-MRAM, the architecture achieves high inference accuracy, increased integration density, and reduced per-bit overhead. Standard multiplication is realized via optimized bit-cell conductance mapping. A proposed self-terminating TDC readout circuit, sensitive to input sparsity, further enhances energy efficiency by eliminating power from ineffective post-quantization pulses. Based on 28 nm CMOS simulation, this MLC-SOT-IMC architecture achieves 6388.98 – 8426.19 1-bit TOPS/W inference energy efficiency under 50% – 90% input sparsity.

VI. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grants 62401031, National Program for Support of Top-notch Young Professionals, and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnology*, vol. 15, no. 7, pp. 529–544, Jul 2020.

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

[3] Z. Guo *et al.*, "Spintronics for energy- efficient computing: An overview and outlook," *Proceedings of the IEEE*, vol. 109, no. 8, pp. 1398–1417, 2021.

[4] W. Xie *et al.*, "A 709.3 TOPS/W event-driven smart vision SoC with high-linearity and reconfigurable MRAM PIM," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2023, pp. 1–2.

[5] 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 *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, 2023, pp. 500–502.

[6] P. Deaville, B. Zhang, and N. Verma, "A fully row/column-parallel in-memory computing macro in foundry MRAM with differential readout for noise rejection," *IEEE Journal of Solid-State Circuits*, vol. 59, no. 7, pp. 2070–2080, 2024.

[7] ——, "A fully row/column-parallel MRAM in-memory computing macro with memory-resistance boosting and weighted multi-column ADC readout," *IEEE Journal of Solid-State Circuits*, pp. 1–11, 2024.

[8] S. Cosemans *et al.*, "Towards 10000TOPS/W DNN inference with analog in-memory computing – a circuit blueprint, device options and requirements," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 22.2.1–22.2.4.

[9] Y. Luo *et al.*, "Performance benchmarking of spin-orbit torque magnetic RAM (SOT-MRAM) for deep neural network (DNN) accelerators," in *2022 IEEE International Memory Workshop (IMW)*, 2022, pp. 1–4.

[10] J. Doevenspeck *et al.*, "SOT-MRAM based analog in-memory computing for DNN inference," in *2020 IEEE Symposium on VLSI Technology*, 2020, pp. 1–2.

[11] C. Wang, Z. Wang, S. Li, Z. Zhang, and Y. Zhang, "Variation aware evaluation approach and design methodology for SOT-MRAM," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 71, no. 4, pp. 1651–1664, 2024.

[12] W. Huang *et al.*, "Series-parallel hybrid SOT-MRAM computing-in-memory macro with multi-method modulation for high area and energy efficiency," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024, pp. Article 193, 1–6.

[13] H. Yoda *et al.*, "Excellent temperature dependence of retention energy and large tunnel magnetoresistance of MTJs with strain-induced magnetic anisotropy for SOT-MRAMs with high write efficiency," *IEEE Transactions on Magnetics*, vol. 60, no. 9, pp. 1–5, 2024.

[14] S. Ikeda *et al.*, "Tunnel magnetoresistance of 604% at 300K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature," *Applied Physics Letters*, vol. 93, no. 8, p. 082508, 08 2008.

[15] Z. Tong *et al.*, "A high throughput in-MRAM-computing scheme using hybrid p-SOT-MTJ/GAA-CNTFET," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 71, no. 2, pp. 606–619, 2024.

[16] C. Tsai *et al.*, "A CMOS-compatible 12nm 8Mb MLC RRAM enabling producible 2-bit per cell for high energy efficiency Compute-In-Memory in edge AI applications," in *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2025, pp. 1–3.

[17] 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-AI edge devices," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3.

[18] Z. Wang *et al.*, "Demonstration of spin orbit torque multi-level cell with enhanced state distinction," *IEEE Electron Device Letters*, pp. 1–1, 2025.

[19] K.-W. Kwon and Y. Seo, "Hybrid multi-level cell spin-orbit torque memory for fast and robust memory operations," *IEEE Transactions on Nanotechnology*, 2025.

[20] S. Shreya and B. K. Kaushik, "Modeling of voltage-controlled spin–orbit torque MRAM for multilevel switching application," *IEEE Transactions on Electron Devices*, vol. 67, no. 1, pp. 90–98, 2019.

[21] Y. Kim, X. Fong, K.-W. Kwon, M.-C. Chen, and K. Roy, "Multilevel spin-orbit torque MRAMs," *IEEE Transactions on Electron Devices*, vol. 62, no. 2, pp. 561–568, 2015.

[22] C. Jiang *et al.*, "Demonstration of 128 kb SOT-MRAM chip with 5 ns write and 15 ns read speed, high endurance over 1010 and low ECC-on bit error rate," in *2024 IEEE International Electron Devices Meeting (IEDM)*, 2024, pp. 1–4.

[23] X. Peng, S. Huang, H. Jiang, A. Lu, and S. Yu, "DNN+NeuroSim V2.0: an end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 11, pp. 2306–2319, 2021.

[24] Y. Luo *et al.*, "A variation robust inference engine based on STT-MRAM with parallel read-out," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[25] J.-L. Cui, Y. Guo, J. Chen, B. Liu, and H. Cai, "Sparsity-oriented MRAM-Centric computing for efficient neural network inference," *IEEE Transactions on Emerging Topics in Computing*, vol. 12, no. 1, pp. 97–108, 2024.

[26] C. Wang *et al.*, "Layout aware optimization methodology for SOT-MRAM based on technically feasible top-pinned magnetic tunnel junction process," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 5, pp. 1463–1476, 2023.

[27] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012.

[28] Y. Zhou *et al.*, "A cfmb STT-MRAM-based computing-in-memory proposal with cascade computing unit for edge AI devices," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 71, no. 1, pp. 187–200, 2023.