A High Throughput In-MRAM-Computing Scheme Using Hybrid p-SOT-MTJ/GAA-CNTFET

Zhongzhen Tong, Yilin Xu, Yunlong Liu, Xinrui Duan, Hao Tang, Suteng Zhao, Chenghang Li, Zhiting Lin, *Senior Member, IEEE*, Xiulong Wu, *Senior Member, IEEE*, Zhaohao Wang, *Senior Member, IEEE*, and Xiaoyang Lin, *Senior Member, IEEE*

***Abstract*— Silicon-based semiconductor transistors are approaching their physical limits due to shrinking feature sizes. Simultaneously, traditional silicon-based von Neumann architectures exhibit significant latency and power consumption issues in data-centric applications, such as the Internet of Things and artificial intelligence. To tackle these challenges, this study introduces a novel approach: Magnetoresistance Random Access Memory (MRAM) computing in-memory (CIM) using gate-all-around carbon nanotube field-effect transistors (GAA-CNTFET). The proposed MRAM array comprised three transistors and one perpendicular magnetic anisotropy spin-orbit torque magnetic tunnel junction (p-SOT-MTJ) (3T1M) cell and achieves full-array Boolean logic operations and half/full-adder operations. The calculated results can be stored *in-situ* during the computing phase without requiring additional peripheral circuits. A 16 Kb MRAM was simulated in both GAA-CNTFET/p-SOT-MTJ and 14-nm FinFET/p-SOT-MTJ technologies to examine the effectiveness of the proposed design. Compared to its 14-nm FinFET/p-SOT-MTJ counterparts, the write and computing latencies of the GAA-CNTFET/p-SOT-MTJ CIM macro were reduced by approximately 21% and 20.6%, respectively, while the read and computing energy consumption by approximately 45.3% and 24.7%, respectively. Moreover, the proposed in-memory Boolean logic throughput was 8192 GOPS, which was approximately 160–250 times higher than that of existing CIM solutions, in which only two rows of word lines can be activated.**

***Index Terms*— Boolean logic operation, magnetoresistance random access memory, computing in-memory, in-situ storing, half/full adder, gate-all-around carbon nanotube field-effect-transistor, von Neumann bottleneck.**

Manuscript received 26 April 2023; revised 14 October 2023; accepted 6 December 2023. Date of publication 20 December 2023; date of current version 30 January 2024. This work was supported in part by the Beijing Natural Science Foundation under Grant 4232070, in part by the National Natural Science Foundation of China under Grant 51602013 and Grant 52261145694, in part by the Young Elite Scientists Sponsorship Program by the China Association for Science and Technology (CAST) under Grant 2018QNRC001, in part by the International Mobility Project under Grant B16001, and in part by the Fundamental Research Funds for Central Universities of China. This article was recommended by Associate Editor C. Liu. *(Corresponding author: Xiaoyang Lin.)*

Zhongzhen Tong and Xiaoyang Lin are with the National Key Laboratory of Spintronics, Institute of International Innovation, Beihang University, Yuhang, Hangzhou 311115, China, and also with the MIIT Key Laboratory of Spintronics, School of Integrated Circuit Science and Engineering, Fert Beijing Institute, Beihang University, Beijing 100191, China (e-mail: xylin@buaa.edu.cn).

Yilin Xu, Xinrui Duan, Hao Tang, Suteng Zhao, Chenghang Li, and Zhaohao Wang are with the School of Integrated Circuit Science and Engineering, Beihang University, Beijing 100191, China.

Yunlong Liu, Zhiting Lin, and Xiulong Wu are with the School of Integrated Circuits, Anhui Provincial High-Performance Integrated Circuit Engineering Research Center, Anhui University, Hefei 230601, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2023.3341608.

Digital Object Identifier 10.1109/TCSI.2023.3341608

I. Introduction

OVER the years, the advancement of complementary metal-oxide semiconductor (CMOS) technology for large-scale integrated circuit applications has experienced significant growth, largely attributable to Moore's Law [1], [2]. However, as silicon-based semiconductor transistors approach their physical limits, issues such as the short-channel effect intensify [3], [4]. This means that as transistors are scaled down in size, their performance improvements become increasingly limited, while their leakage current rises [5]. Consequently, Moore's Law is confronted with extraordinary challenges [6].

Additionally, owing to the development of artificial intelligence (AI) applications [7], the scale of algorithmic models is increasing, resulting in considerable computing power and storage requirements. A large volume of frequent memory access tasks limits the energy efficiency of AI chips because of the von Neumann bottleneck [8], [9], a problem that is prominent among resource-constrained devices such as smart watches, mobile phones, and wireless earphones, amongst others [10]. Therefore, scholars have focused on exploring devices with new physical characteristics [11], [12], [13], [14] and integrated storage and computing architectures. To improve energy efficiency and overcome existing limitations, the concept of computing-in-memory (CIM)—a technique that performs part of the calculations in memory [6], [15]—has been introduced. CIM techniques can effectively reduce the power consumption and latency associated with data migration. However, existing silicon-based CIM macros have reached their performance limit [10], [16], [17], [18].

Because of the lower leakage currents and higher I_{ON}/I_{OFF} of gate-all-around carbon nanotube field-effect transistors (GAA-CNTFETs) [19], [20], [21] compared to silicon-based transistors, GAA-CNTFETs are being used to design high-speed and low-power-consumption circuits. Furthermore, spintronic devices, such as perpendicular magnetic anisotropy spin-orbit torque magnetic tunnel junction (p-SOT-MTJ) devices, offer several advantages. First, a small spin-transfer torque (STT) current (I_{STT}) has the capability to break the symmetry of SOT, enabling field-free switching in p-SOT-MTJ devices. Second, due to the orthogonal orientation of SOT to

the magnetic moment of the free layer (FL) in equilibrium, applying the SOT current (I_{SOT}) immediately disrupts the stable magnetic state of the FL. This unique feature results in enhanced write endurance and reliability for magnetoresistance random access memory (MRAM), improved operational speed, and reduced power consumption [22]. Moreover, the back-end integration of the MTJ is compatible with the CNTFET process [3], [23] and does not require a large die area. Consequently, GAA-CNTFET and p-SOT-MTJ devices are considered promising candidates for building future energy-efficient CIM macros.

At present, research on hybrid CNTFET and MTJ is focused on logic operations [5], [24], [25], [26], [27], ternary circuits [3], [28], non-volatile flip-flops [29], [30], neural network accelerators [21], [31], [32] and true random number generators [33]. However, the above studies on hybrid CNTFET and MTJ have designed computing or storage units, including [5] combined CNTFETs and spin-transfer torque MTJ to build near-memory computing circuits, a nonvolatile universal ternary flip-flop was achieved using multi-threshold CNTFETs [30] and in [32], a fully nonvolatile spin-based synapse was obtained by utilizing CNTFET and MTJ. Although numerous studies have focused on the design of circuits using CNFETs/MTJs, most of them only implemented small-scale circuits, resulting in limited throughput and energy efficiency.

To overcome the aforementioned limitation, this study proposes a structure comprising three GAA-CNTFETs and one p-SOT-MTJ (3T1M) cell for full-array Boolean logic operations. The advantages of the proposed structure are as follows:

1. A 16 Kb MRAM can be achieved using p-SOT-MTJ/GAA-CNTFET and 14-nm FinFET/p-SOT-MTJ technology. Compared to its 14-nm FinFET/p-SOT-MTJ counterparts, the proposed CIM macro based on the p-SOT-MTJ/GAA-CNTFET exhibits better performance for read/write and computing operations.
2. The 3T1M cell realizes a full-array Boolean logic operation, and *in-situ* storage of the operation results in a computing cycle, thereby eliminating the need for additional storage space and peripheral circuits.
3. The proposed circuits can support full-array half-adder operations by changing the bit line (BL) encoding mode. Based on half-adder operations, the proposed structure uses fewer devices and cycles to implement full-adder operations.

The remainder of this paper is organized as follows: **Section II** introduces related studies on MRAM-based in-memory Boolean logic. **Section III** presents the structure and operating principles of the proposed 3T1M CIM macro. **Section IV** validates the performance of the structure, analyses the energy consumption and delay, and compares the results with those of other studies. Lastly, **Section V** presents the conclusions.

II. Related Studies

To solve the problems of energy efficiency and latency, CIM technology has received much attention from the scientific community. MTJ technology is widely used in CIM solutions owing to its nonvolatility and endurance in writing operations [34]. MRAM-based logic implementation schemes can be divided into read-based [35], [36], [37], [38], [39], [40] and write-based [9], [41], [42], [43], [44], [45] schemes, as shown in **Fig. 1**.

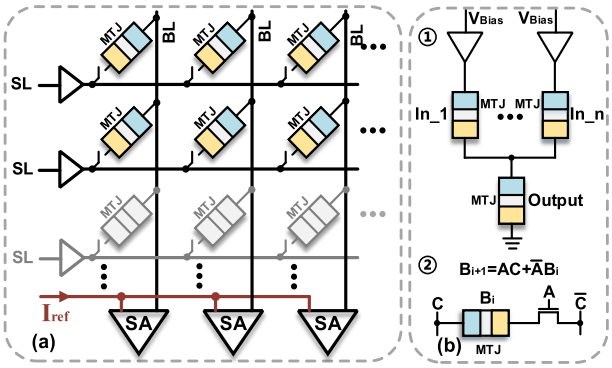


Fig. 1. MRAM-based logic implementation scheme: (a) read-based; (b) write-based.

A. Read-Based CIM Scheme

Pinatubo [46] proposed a method that could execute Boolean logic such as OR, AND, XOR, and INV functions in memory to realize batch-bit Boolean logic operations in a non-volatile storage architecture; multiple rows could be activated simultaneously, the sense amplifier (SA) outputting bitwise computing results by sensing the BL voltage or current. During the read operation, the SA was used to distinguish the BL voltage difference caused by the discharge of cells with high (antiparallel state, R_H) or low resistance (parallel state, R_L). However, in the CIM mode, as shown in **Fig. 1(a)**, Pinatubo configured multiple reference circuits into the SA; therefore, the following three states could be distinguished: $R_H//R_H$ (for logic "11"), $R_H//R_L$ (for logic "01" or "10"), and $R_L//R_L$ (for logic "00"), which could be used for bitwise AND/OR operations. However, if the XOR or XNOR logic was to be realized, additional NOR gates and SAs had to be added, suggesting that one column contained two SAs [36], [40]. Additionally, if the tunneling magnetoresistance ratio (TMR) of the MTJ was sufficiently high, it could support MAJORITY logic [36].

In [40], a 1T1M cell was used, wherein multiword line activation enabled the Boolean logic. The proposed CIM macro was integrated into a general-purpose computing system. Similarly, Monga et al. proposed a 2T+1MTJ memory cell-based MRAM array to enable logic operations such as NAND, NOR, and MAJORITY [36], with multireference voltages being required. In [37], a 4T1M cell was proposed to achieve 16 Boolean logic functions and arithmetic operations, a dynamic bit being stored using the gate capacitance of the transistor. These studies used SAs to complete the read-based CIM operations. However, sophisticated peripheral circuits—including reference circuits and SAs—reduced the energy efficiency and required additional area overhead. Moreover, a limited TMR could result in a smaller sensing margin and BL voltage swing, reducing the computing accuracy and limiting the operation throughput.

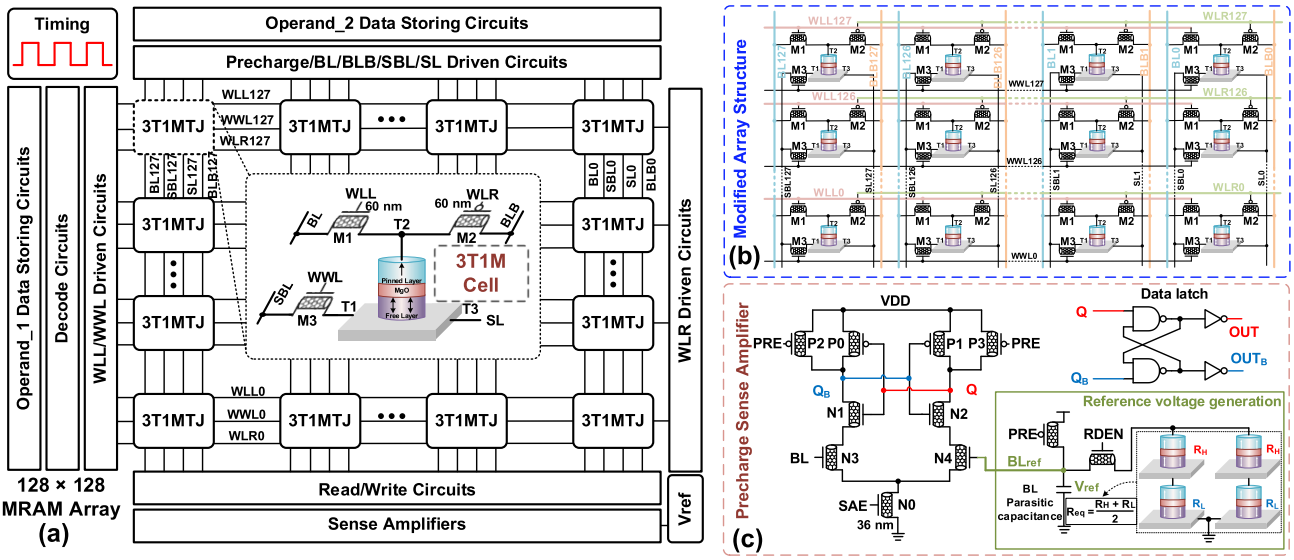


Fig. 2. (a) Overall MRAM-CIM architecture and schematic of the 3T1M cell. The schematic of (b) the modified array structure and (c) the precharge sense amplifier.

B. Write-Based CIM Scheme

Fig. 1(b) shows the two write-based CIM schemes. First, different write currents could be generated based on the different input MTJ resistance states. Therefore, the state of the input MTJ determined the magnitude of the calculated current. It then determined whether the state of the output MTJ had switched [9], [41], [44]. Furthermore, different operands could be mapped to the word line (WL), BL voltage, and initial state of the MTJ. The computational result was represented by the next state of the MTJ [42].

In [9], a reconfigurable bit-serial CIM operation (BSCIM) based on the toggle SOT-MRAM was proposed. The BSCIM scheme was designed to implement Boolean logic by leveraging the write mechanism of a TSOT device. However, to achieve different logic operations, an additional decision TSOT device was selected, with multicycle operations (the preset cycle) being required, increasing the area overhead and computing latency. Moreover, a limited TMR could influence the calculation accuracy. Reference [41] proposed a computational RAM (CRAM) architecture to perform Boolean logic operations in the array. However, the carry for the full adder had to be copied to participate in the next computation, a peripheral circuit being required. Moreover, the complex timing control signal increased the implementation complexity. Consequently, the authors in [42] proposed an STT-CIM scheme that performed logic and arithmetic by modifying the array structure of a STT magnetic RAM (STT-MRAM).

As shown in **Fig. 1(b)**, in the 1T1M cell, the logic of $B_{i+1} = AC + \bar{A}B_i$ can be realized, the calculated results being stored *in situ*. However, owing to the inability to flexibly perform XOR operations, five read cycles and one write cycle each are required to realize a 1-bit full adder. Moreover, these methods cannot activate a full array to participate computing.

Compared to previous studies, the proposed design used only 3T1M cells, full-array Boolean logic (AND/NAND, OR/NOR, IMP/NIMP, and XOR/XNOR) could be enabled, the operation results being stored automatically in a computing cycle. Half- and full-adder operations with fewer devices and cycles could also be realized in memory. Furthermore, it did not require additional SAs or complex peripheral circuits, thereby simplifying the circuit design.

III. Proposed MRAM Macro for Memory and CIM Modes

In this section, we provide an overview of the proposed hybrid p-SOT-MTJ/GAA-CNTFET MRAM-CIM macro and present the principles of the memory mode, full-array *in-situ* storing Boolean logic, and half/full operations.

A. Macro Structure and Overview

Fig. 2(a) shows the structure of the proposed hybrid p-SOT-MTJ/GAA-CNTFET MRAM-CIM macro, which comprises a 3T1M cell array, input data-storing circuits for Boolean operation, decode circuits, precharge circuits, WL- and BL-driven circuits, reference voltage generation circuits, SAs, and timing control circuit. The 3T1M MRAM-CIM macro can operate in memory and CIM modes. In memory mode, the read functions are performed through single-ended precharge SAs. In CIM mode, full-array *in-situ* restoration of Boolean logic and half/full-adder functions can be realized, where the full adder is implemented on the logic of the half adder.

Fig. 2(a) illustrates the 3T1M MRAM cell configuration. In contrast to the conventional 2T1M SOT-MRAM cell, the proposed 3T1M cell includes an additional transistor, enabling both memory and CIM modes. In this figure, M1–M3 represent the access transistors, and T1–T3 correspond to the three terminals for the p-SOT-MTJ device. The generation of I_{SOT} is achieved through the application of the spin-Hall effect (SHE) current between T1 and T3. Conventional I_{STT} can be applied between T2 and T3 (or T1). The word lines left (WLL) and right (WLR) control the access transistor gates (M1) and (M2), respectively, for the read/write and CIM operations. Moreover, the WLL and WLR are mapped as operand data to participate in the computation. Similarly, BL and BL bars (BLBs) can achieve read/write operations and can be mapped as operand

data. The write word line (WWL) controls the access transistor gate (M3) and the source line (SL) and source BL (SBL) are used to generate the I_{SOT} for the write and compute operations.

Fig. 2(b) shows a schematic of the modified array structure comprising 128 rows × 128 columns of 3T1M cells. Operand 1 (O_1) is mapped as the voltage of WLL/WLR connected horizontally. Operand 2 (O_2) is mapped as the voltage of BL/BLB connected vertically. The calculated results are stored *in situ* without requiring a preset cycle, thereby reducing the number of operation cycles for Boolean logic.

Fig. 2(c) shows a schematic of the sense amplifier based on the GAA-CNTFET, wherein P0/P1 and N1/N2 comprise cross-coupled inverters that provide positive feedback. P2 and P3 are used to precharge the Q and Q_B nodes using the PRE signals. N3 and N4 serve as the input transistors, and N0 is a pull-down transistor. Before amplification, Q and Q_B are precharged to a high voltage. The SA begins to operate when the enabled SAE signal is at a high voltage. The N3 and N4 gates connect to the BL and BL_{ref}, respectively. V_{ref} is obtained by a capacitor precharged to VDD discharging through a resistance of (R_H+ R_L)/2, where the capacitance is consistent with the parasitic capacitance of the BL, and the resistance is obtained by paralleling R_H (the MTJ is kept in an antiparallel state) and R_L (the MTJ is kept in a parallel state) in series to obtain an equivalent resistance of (R_H+ R_L)/2. When the SA starts operating, the voltage at Q and Q_B decreases. The discharge speeds differ considering the corresponding input voltages of the SA (BL and BL_{ref}) are different. If the voltage of the BL is higher than V_{ref}, the discharge speed of Q_B will be greater than that of Q. Consequently, the voltage at node Q_B decreases to $VDD - |V_{THp}|$ faster than the voltage of Q, where V_{THp} is the threshold voltage of the p-CNTFET. P1 is turned on to pull up node Q to a high voltage; N1 is turned on to pull-down node Q_B to the VSS. Finally, the data read from the Q and Q_B nodes are stored in a latch (represented by the node voltage of OUT and OUT_B) to achieve the reading operation.

B. Memory Mode

1) Writing Operation: **Fig. 3(a)** shows the writing data "0" operation. First, the WWL is activated, and the SL and SBL are set to a high voltage and VSS, respectively. At this point, the I_{SOT} from the SL to the heavy metal and SBL is generated, the MTJ forming an in-plane magnetization in the FL of the MTJ owing to the SHE. Then, the WLL and WLR are set to a high voltage and VSS, respectively. The BL and BLB are configured to be VSS and floating, respectively. The I_{STT} is generated from the SL to BL. Notably, the I_{SOT} and I_{STT} persist for a short period. Finally, M3 is turned off, and the I_{STT} from the SL to MTJ, M1 (n-CNTFET), and BL sets MTJ to R_L, which completes the write "0" operation.

Similarly, **Fig. 3(b)** shows the writing data "1" operation. First, the WWL is activated, and the SL and SBL are set to VSS and a high voltage, respectively. Note that under this condition, due to a threshold voltage loss when an n-CNTFET passes a high voltage, the voltage at T1 can only reach $VDD - V_{THn}$, where V_{THn} denotes the threshold voltage of the n-CNTFET. Consequently, an I_{SOT} is generated from the

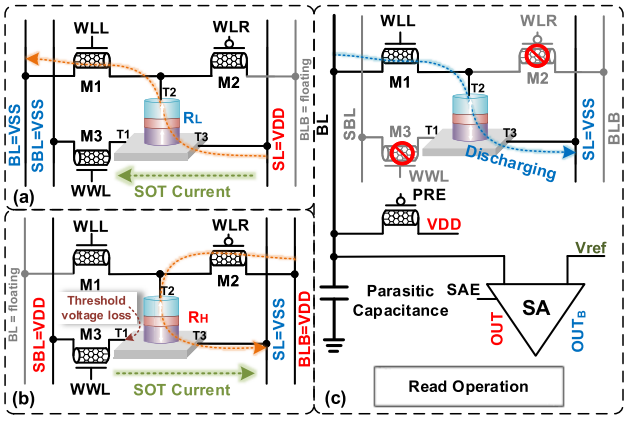


Fig. 3. Writing operation: (a) writing "0" and (b) writing "1." (c) Reading operation.

SBL to SL. Although the voltage difference between T1 and T3 is smaller than when writing data "0", the generated I_{SOT} is sufficient to initialize the p-SOT-MTJ to an intermediate state between R_H and R_L. Further details regarding this initialization are analyzed in **Section IV**. Then, the WLL and WLR are set to a high voltage and VSS, respectively. The BL and BLB are configured as floating and VDD, respectively. An I_{STT} from the BLB to SL is generated. Note that the I_{SOT} and I_{STT} exist simultaneously for some time. Finally, M3 is turned off, and the I_{STT} from the BLB to M2 (p-CNTFET), MTJ, and SL sets MTJ to R_H to complete the write "1" operation. The WLL and WLR are simultaneously enabled during write operations. Consequently, there is no threshold loss when writing "0" or "1". However, the traditional 2T1M cell—which uses NMOS to transmit a high voltage for write operations [47]—has a threshold loss. Therefore, the proposed 3T1M cell is faster in terms of the writing operations.

2) Reading Operation: In the read operation, the PRE signal is enabled to precharge the BL to a high voltage. Subsequently, the corresponding WLL is set to VDD, resulting in the BL discharge through the p-SOT-MTJ. Because the stored data has two states (R_H or R_L), there are two different voltage drop speeds for the BL. Additionally, the discharge speed of BL_{ref} ranges between "1" and "0". A voltage difference is formed between the BL and BL_{ref}. If the stored data is "1", the discharge speed of the BL will be slower than that of BL_{ref}, forming $\Delta V\_H = V_{BL} - V_{ref}$; if the stored data is "0", the discharge speed of the BL will be faster than that of BL_{ref}, forming $\Delta V\_L = V_{ref} - V_{BL}$. Finally, the SAE signal is enabled to complete the reading operation.

Fig. 4 shows the timing sequences of the proposed MRAM circuit for read/write operations. The first two cycles show the write and read operations of data bits "0". Similarly, the third and fourth cycles show the write and read operations of data bits "1".

Fig. 5 shows the transient simulation results of the entire write/read operation process for the proposed MRAM. By properly using I_{SOT} and I_{STT}, the state of the MTJ is switched from "1" to "0" and then back to "1." Here, the maximum I_{SOT} is 321 μA for writing "0" and 191 μA for writing "1" (represented by the T1 current). Similarly, the maximum I_{STT} is 92 μA for writing "0" and 70 μA

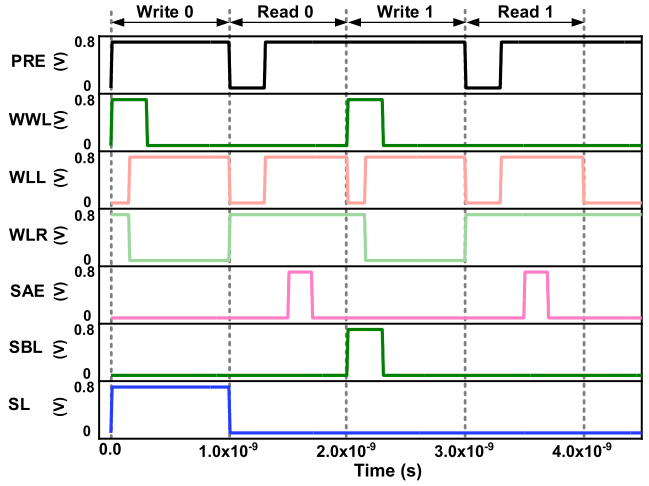


Fig. 4. Timing sequences of write and read operations for the proposed MRAM-CIM macro.

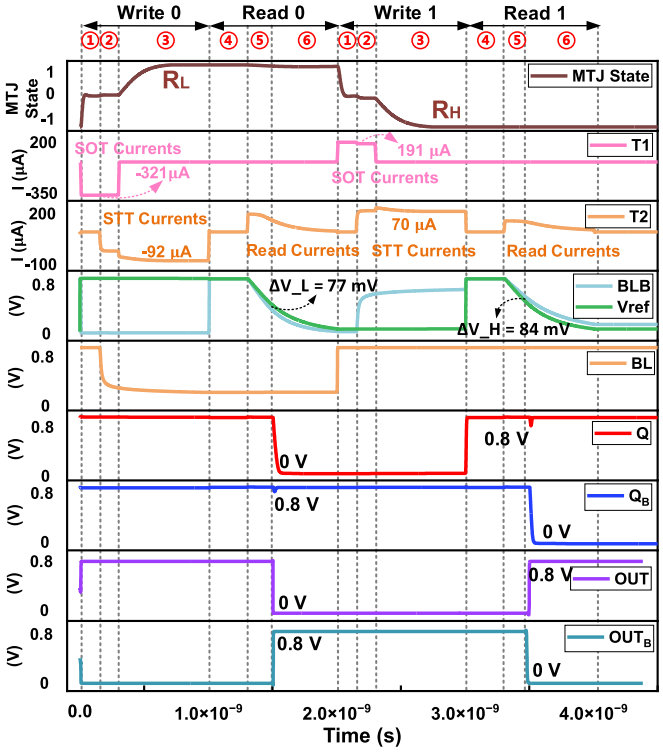


Fig. 5. Transient simulation results of the entire write and read process. ① SOT write phase. ② SOT and STT write phase. ③ STT write phase. ④ BL and SA pre-charge phase. ⑤ Voltage difference formation phase. ⑥ SA amplification phase.

for writing "1" (represented by the T2 current). Note that although the I_{SOT} flowing in the heavy metal is relatively large, the total energy consumption for the writing operation is effectively reduced owing to its short duration (300 ps). The large difference in the I_{SOT} between writing "0" and "1" is generated when the M3 passing a high voltage exhibits a threshold voltage loss. However, the I_{SOT} of 191 μA is sufficient to initialize the p-SOT-MTJ to an intermediate state. For reading "0" and "1," the ΔV_H and ΔV_L can reach 77 mV and 84 mV, respectively, having sufficient read margin

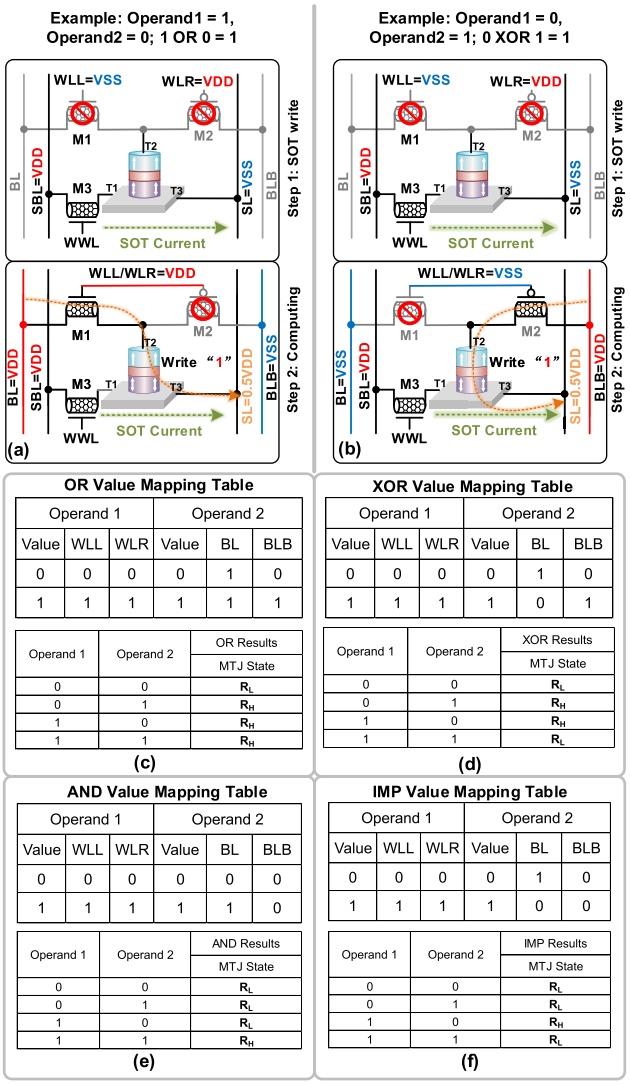


Fig. 6. Boolean operations and *in-situ* storing: (a) OR operations; (b) XOR operations; (c) OR value mapping table and calculation results; (d) XOR value mapping table and calculation results; (e) AND value mapping table and calculation results; (f) IMP value mapping table and calculation results.

for the SA (constrained mainly by the offset voltage discussed in **Section IV**).

C. CIM Mode

1) Boolean Operations and In-Situ Storing: The OR operations and *in-situ* storage of the calculation results can be divided into two steps, as shown in **Fig. 6(a)**.

(1) SOT write: The WWL is activated to turn on M3. The SBL and SL are configured as VDD and VSS, respectively. Therefore, the I_{SOT} flows from the SBL to the SL through M3 and the heavy-metal layer to initialize the MTJ to the intermediate state. At this point, the WLL and WLR are maintained at VSS and VDD, respectively. The BL and BLB remain floating.

(2) Computing: The WLL/WLR and BL/BLB of the 3T1M cell are changed based on the input data—for example, $O_1 = 1$ and $O_2 = 0$ (WLL/WLR = VDD; BL/BLB = VDD/VSS). Therefore, M2 is turned off. The voltage of the SL is set to 0.5 VDD. At this point, the I_{SOT} remains for a short time.

Subsequently, the I_{SOT} is removed. The voltage of the BL is higher than that of the SL, thereby realizing an I_{STT} flowing from the BL to the SL through M1, MTJ and a heavy-metal layer to set the MTJ to R_H. The MTJ state of R_H and R_L represent the computational results "1" and "0", respectively. Therefore, the computational result is stored in the MTJ, achieving OR logic and *in-situ* storage. Because the data read by the SA are complementary, NOR logic can be obtained from OUT_B.

Fig. 6(c) shows the OR value mapping table and calculation results, where WLL/WLR = 0 represents input O_1 0; WLL/WLR = 1 represents input O_1 1; BL/BLB = 1/0 represents O_2 0; and BL/BLB = 1/1 represents O_2 1. The OR calculation results are reflected in the MTJ state (R_H = 1 and R_L = 0). For example, if O_1 and O_2 are 0 (WLL = 0, WLR = 0, BL = 1, BLB = 0), the MTJ state becomes R_L and corresponds to the OR operation results. By contrast, if O_1 or O_2 is 1, the MTJ state becomes R_H.

The XOR operations are similar to the OR operations, including the two steps of SOT writing and computing, as shown in **Fig. 6(b).** Different logic operations can be realized by changing the code of the input operands. **Fig. 6(d)**, **6(e)**, and **6(f)** shows the different input codes for XOR, AND, and IMP computing. Similarly, the XNOR, NAND, and NIMP logic can be obtained from the OUT_B node in the SAs.

Given the consistency of the AND/NAND, OR/NOR, IMP/NIMP, and XOR/XNOR operations, only the transient simulation results of the XOR operation are shown in **Fig. 7**. The MTJ state is switched based on the voltages of WWL/WLR and BL/BLB, representing O_1 and O_2, respectively. For example, if O_1 = 1 and O_2 = 0, the WLL and WLR are configured as VDD, and the BL and BLB are configured as VDD and VSS, respectively. Here, the maximum current of I_{SOT} is 184 μA for initializing the MTJ to the intermediate state, and the minimum I_{STT} is 10 μA. In this case, the I_{STT} flows into the MTJ through M1, whereas the n-CNTFET exhibits a threshold loss when passing a high voltage. Additionally, the MTJ has a high resistance based on the calculation results, thereby increasing the resistance between the BL and SL. Consequently, I_{STT} is smallest when "1" XORing "0" is calculated as the worst case. Even if I_{SOT} and I_{STT} are relatively low when performing logic operations (compared to the write operations), I_{STT} can easily break the symmetry of the SOT, the calculation being completed within 2 ns (discussed in **Section IV**).

The WLs and BLs of the full array can be activated simultaneously during computation, which indicates that each 3T1M cell can perform logical operations. The calculation results are stored *in situ* without additional space or peripheral circuits. Moreover, by changing the input coding of the input operands, different Boolean logics can be realized in the full array, simultaneously. For instance, the XOR, OR, and IMP operations can be implemented in the first, second, and third columns, respectively, where different Boolean logic can be used to perform the function of an arithmetic logic unit (ALU), including a half adder and full adder. Consequently, the proposed hybrid p-SOT-MTJ/GAA-CNTFET MRAM-CIM

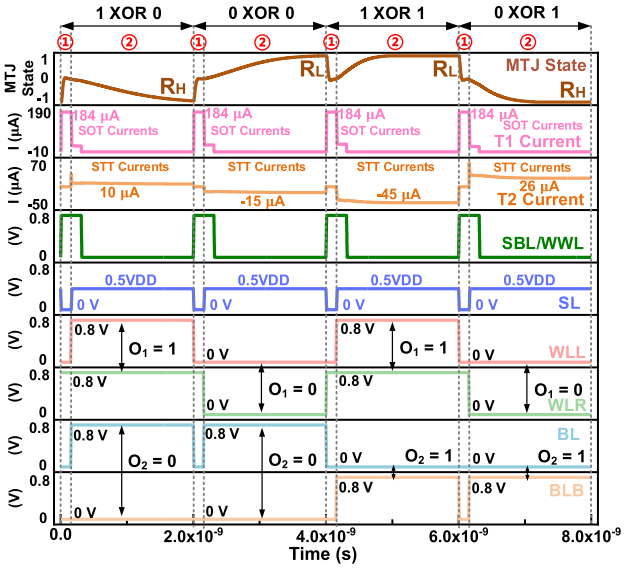


Fig. 7. Transient simulation results of XOR computing and *in-situ* storing process. ① SOT write phase. ② Computing phase.

macro solves the throughput bottleneck resulting from the conventional logic operations implemented by activating two WLs in the array.

2) Half-Adder Operations: **Fig. 8(a)** shows the principle of a half adder, which is an ALU that can add two one-bit binary numbers. It has two input terminals (addend A and addend B), output summation (S), and carry (C), the logical expressions of which are as follows:

$$S = A \oplus B$$
$$C = AB \tag{1}$$

where XORing and ANDing denote the primary logic operations.

The proposed MRAM-CIM macro can simultaneously realize different Boolean logics in memory. Thus, half-adder operations can be implemented easily using the proposed 3T1M array, as shown in **Fig. 8(b).** Addends A_n–A_0 are mapped as voltages of the WLLs and WLRs, and addends B_n–B_0 are mapped as voltages of the BLs and BLBs. The two 3T1M cells can be combined to form a half-adder. The XORing and ANDing calculations are implemented in the left and right cells, respectively. For example, B_n is mapped to BL_n/BLB_n and BL_{n-1}/BLB_{n-1}; A_n is mapped to WLL_n/WLR_n. The left cell can realize $A_n \oplus B_n$ for the output sum. Similarly, the right cell can realize A_nB_n for the output carry. Consequently, a full array can be activated to achieve a 1-bit half-adder. For the n × n MRAM array, $n^2/2$ half-adder operations can be implemented.

3) Full-Adder Operations: Unlike the half adder, the 1-bit full adder can handle low-order carry and output standard addition carry. By combining two half adders and adding OR logic, a 1-bit full-adder operation can be realized, as shown in **Fig. 9 (a)**. The full adder includes addend (A), addend (B), and carry (C). Because there are two half adders, two carriers are generated. Regardless of which half adder generates the carry, the full adder requires the output carry. Therefore, OR logic is required to combine the two carriers of the half adders. The

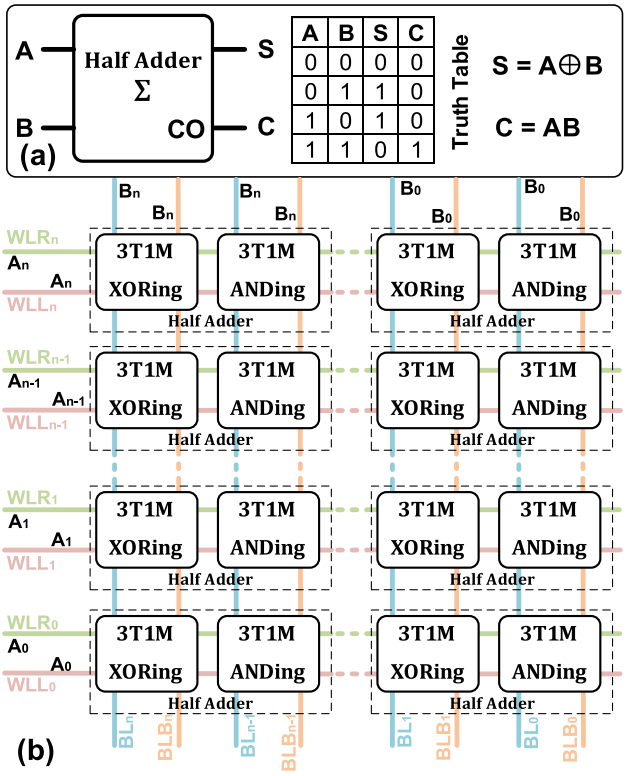


Fig. 8. (a) Principle of the half adder. (b) Principle of the full-array half adder implemented using the proposed MRAM-CIM macro.

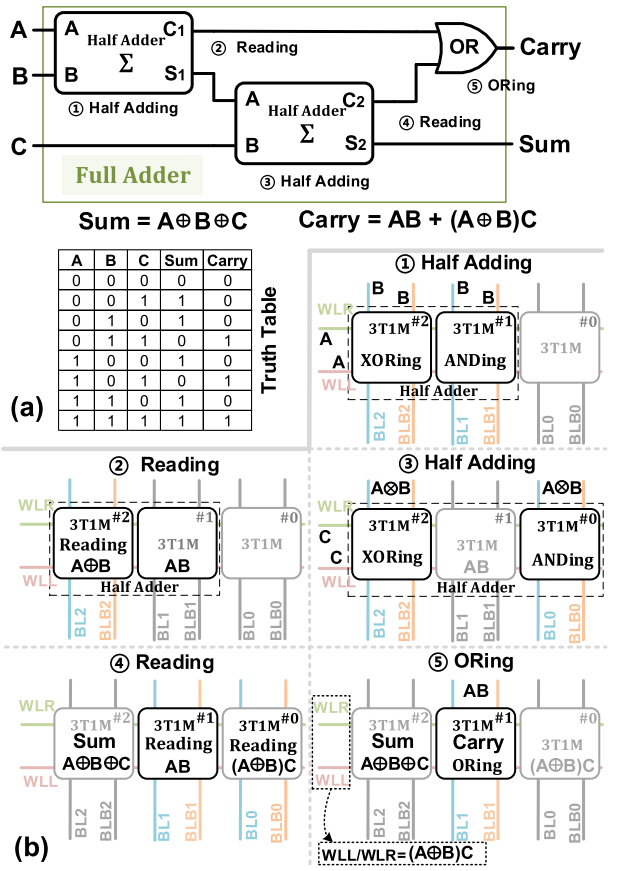


Fig. 9. (a) Principle of the full adder. (b) Basic operation of the 1-bit full adder implemented using the proposed half-adder operations.

sum of the full adder is generated by S_1 half adding input C, where S_1 is the sum of input A half adding input B. The logical expressions are as follows:

$$\begin{aligned} \text{Sum} &= A \oplus B \oplus C \\ \text{Carry} &= AB + (A \oplus B)\, C \end{aligned} \tag{2}$$

Fig. 9(b) illustrates the basic operation of a 1-bit full adder implemented using the proposed half-adder operations. Considering the three 3T1M cells as examples, the detailed steps are as follows:

Step 1:*A half adding B.* Input A is loaded into WLL/WLR, and input B is loaded into BL_2/BLB_2 and BL_1/BLB_1. The $3T1M^{\#2}$ and $3T1M^{\#1}$ cells are combined as a half adder. As a result, A XORing B and A ANDing B are achieved in the $3T1M^{\#2}$ and $3T1M^{\#1}$ cells, which represent S_1 and C_1, respectively. At this point, the $3T1M^{\#0}$ cell remains idle.

Step 2: *Reading operation.* The half-addition result of S_1 stored in the $3T1M^{\#2}$ cell is read. At this point, the $3T1M^{\#1}$ and $3T1M^{\#0}$ cells remain idle.

Step 3: *C half adding A⊕B.* The input C is loaded into WLL/WLR; The A ⊕B read from Step 2 is loaded into BL_2/BLB_2 and BL_0/BLB_0. $3T1M^{\#2}$ and $3T1M^{\#0}$ are combined as a half adder. Therefore, in the $3T1M^{\#2}$ and $3T1M^{\#0}$ cells, C XORing A⊕B and C ANDing A⊕B are achieved, which represent S_2 and C_2, respectively. S_2 is the sum of the full adders stored in the $3T1M^{\#2}$ cell. At this point, the $3T1M^{\#1}$ cell remains idle.

Step 4: *Reading operation.* The half adders of C_1 stored in the $3T1M^{\#1}$ cell and C_2 stored in the $3T1M^{\#0}$ cell are read out. At this point, the $3T1M^{\#2}$ cell remains idle.

Step 5: *(A⊕B)C ORing AB.* The C_2 ((A⊕B)C) read from the $3T1M^{\#0}$ cell is loaded into WLL/WLR; the C_1 (AB) read from the $3T1M^{\#1}$ cell is loaded into BL_1/BLB_1. The ORing operation is then implemented in the $3T1M^{\#1}$ cell to obtain the carry of the full adder. At this point, the $3T1M^{\#2}$ and $3T1M^{\#0}$ cells remain idle. Finally, the sum and carry of the full adder are stored in the $3T1M^{\#2}$ and $3T1M^{\#1}$ cells, respectively.

The transient simulation results of the full-adder operations are shown in **Fig. 10**, where the inputs of A, B, and C are 1, 1, and 1, respectively. In Step 1, the $3T1M^{\#2}$ and $3T1M^{\#1}$ cells are converted to R_L and R_H, respectively. BL_0 and BLB_0 are maintained in the VSS to maintain the state of the $3T1M^{\#0}$ cell. In Step 2, A⊕B is obtained for OUT2. In Step 3, the $3T1M^{\#2}$ and $3T1M^{\#0}$ cells are converted to R_H and R_L, respectively. BL_1 and BLB_1 are maintained in the VSS to maintain the state of the $3T1M^{\#1}$ cell, In Step 4, the results for AB and (A⊕B)C are obtained from OUT1 and OUT0, respectively. In Step 5, the $3T1M^{\#1}$ cell is exposed to R_H. Similarly, BL_2/BLB_2 and BL_0/BLB_0 are maintained at the VSS to maintain the states of the $3T1M^{\#2}$ and $3T1M^{\#0}$ cells. The entire full-adder operation is completed within 8 ns.

IV. Evaluation and Comparison Results

The proposed MRAM memory was simulated based on SPICE using a hybrid p-SOT-MTJ [22]/GAA-CNTFET [19], [20], and a commercia 14-nm FinFET/p-SOT-MTJ. The important parameters of the GAA-CNTFET and p-SOT-MTJ models are listed in **Table I**. Note that the minimum gate width of

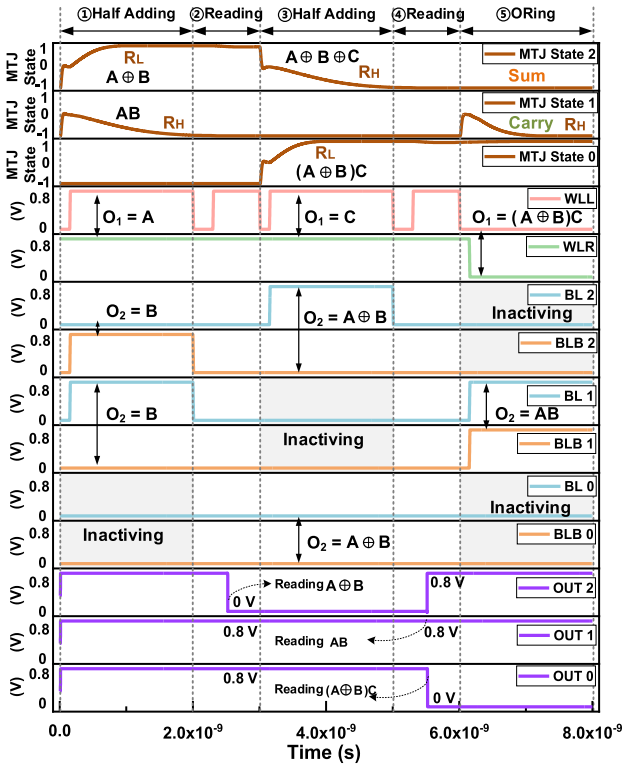


Fig. 10. Transient simulation results of the full-adder operations.

TABLE I

IMPORTANT PARAMETERS OF THE GAA-CNTFET AND MTJ MODELS

	Parameters	Value
GAA-CNTFET	Transistor width (W)	18nm
	Physical gate length (Lg)	12nm
	Contact length (Lc)	12nm
	Source/drain extension length (Lext)	3nm
	CNT diameter (d)	1.2nm
	Gate oxide thickness (tox)	3nm
	Gate height (Hg)	20nm
	Flat band voltage (Vfb)	±0.015V
	Spacing between the CNTs (s)	3nm
MTJ	Free layer thickness (tsl)	0.7nm
	MgO barrier thickness (tox)	0.85nm
	TMR ratio under zero bias voltage (TMR_0)	150%
	MTJ surface length (a)	40nm
	MTJ surface width (b)	40nm
	MTJ surface radius (r)	20nm
	Heavy-metal weidth (W)	40nm
	Heavy-metalthickness (d)	3nm
	Heavy-metal length (l)	60nm

the CNTFETs is 18 nm in the proposed MARM-based CIM macro. If a different CNTFET gate width is used, the gate width would be as shown in the schematic. The memory size is 128 × 128 to obtain a 16 Kb array.

A. Performance Analysis

To analyze the stability of read operations, we evaluated the read margin (ΔV_H and ΔV_L) under different TMRs, the evaluation results of which are shown in **Fig. 11**; the larger the TMR of the MTJ, the larger the read margin. When the TMR is 100%, 77 mV and 46 mV of ΔV_H and ΔV_L can

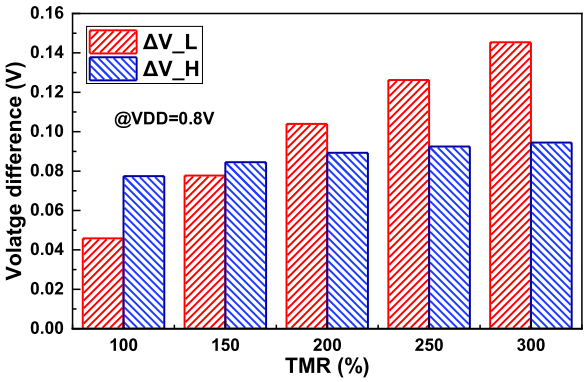


Fig. 11. Simulation results of the read margin with various TMRs.

TABLE II

MONTE CARLO PARAMETERS OF THE GAA-CNTFET AND MTJ MODELS

	Monte Carlo Parameters	±3δ Variation
GAA-CNTFET	Physical gate length (Lg)	10%
	Contact length (Lc)	10%
	Gate height (Hg)	10%
	Gate oxide thickness (tox)	10%
	Flat band voltage (Vfb)	10%
	The number of CNTFET (Ncnt)	10%
MTJ	MgO barrier thickness (tox)	3%
	Free layer thickness (tsl)	3%
	TMR ratio under zero bias voltage (TMR_0)	3%

be obtained, respectively; when TMR is up to 150%, 84 mV and 77 mV of ΔV_H and ΔV_L can be obtained, respectively. However, for the p-SOT-MTJ, numerous experimental studies have shown that the TMR at room temperature is approximately 180% to 250% [48]. In this study, the 150% TMR was selected to achieve the proposed circuit design.

The SA is critical in read operations. Consequently, we analyzed the influence of different parameters in the GAA-CNTFET on the SA performance, including the read delay (T_{SA}) which contains the precharge time, sensing time, and offset voltage. The 3δ variations for the physical gate length (L_g), contact length (L_c), gate height (H_g), gate oxide thickness (t_{ox}), flat band voltage (V_{fb}), and the number of CNTs [49] (N_{cnt}), as the most critical process parameters of GAA-CNTFETs are considered to be 10%, where N_{cnt} is calculated as the transistor width by the spacing (s) between the CNTs, as shown in **Table II**.

Fig. 12(a) and **12(b)** show the T_{SA} and offset voltage results of the 2k trial Monte Carlo simulations using different GAA-CNTFET parameters. The N_{cnt} greatly impacts the SA performance, providing the longest mean T_{SA} of 304.96 ps with a standard deviation of 140.28 fs, and a mean mismatch of 1.3 mV with a standard deviation of 18.11 mV. By contrast, V_{fb} has the least impact on the SA performance, providing the smallest mean T_{SA} of 304.92 ps with a standard deviation of 10.09 fs, and a mean mismatch of 0.19 mV with a standard deviation of 3.22 mV. It is noteworthy that the largest offset voltage is 18.11 mV, which is less than ΔV_H or ΔV_L. Consequently, there is sufficient read margin to guarantee the robustness of reading operations.

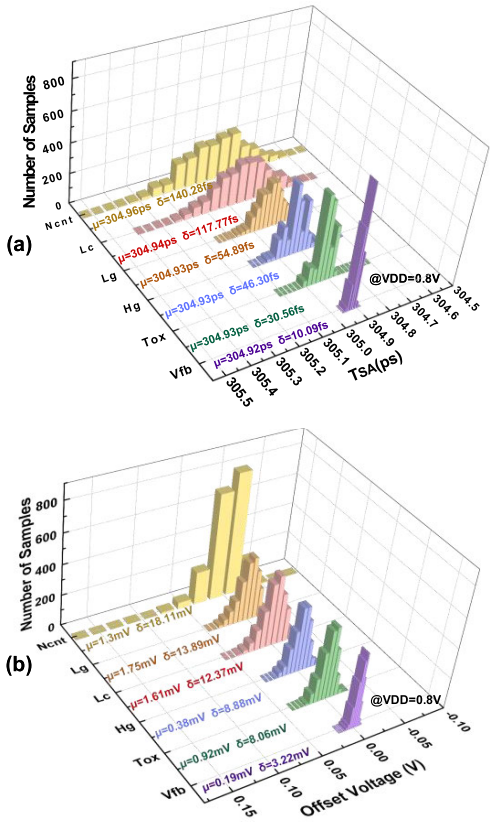
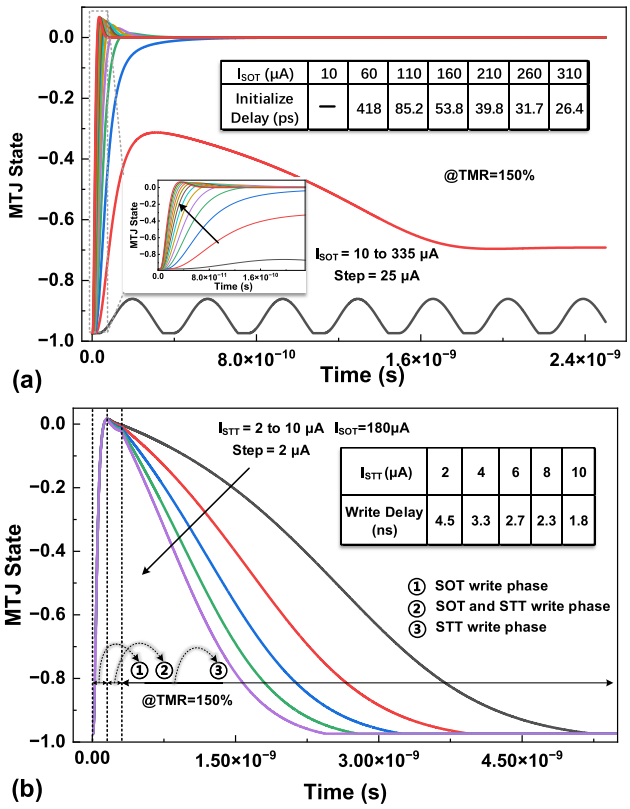


Fig. 12. Monte Carlo simulations of (a) the delay and (b) offset voltage for the precharge-SA at 0.8 V.



I_{SOT} (μA)	10	60	110	160	210	260	310
Initialize Delay (ps)	—	418	85.2	53.8	39.8	31.7	26.4

I_{STT} (μA)	2	4	6	8	10
Write Delay (ns)	4.5	3.3	2.7	2.3	1.8

Fig. 13. (a) I_{SOT} applied to T1 or T3 to initialize the p-SOT-MTJ to an intermediate state ranges from 10 to 310 μA in 25 μA. (b) I_{STT} applied to T2 to switch the p- SOT-MTJ to a R_H state ranges from 2 to 10 μA in 2 μA.

In the proposed CIM macro, during the writing "0", writing "1" and CIM operations, different I_{SOT} and I_{STT} are generated owing to different voltage differentials between T1 and T3, or T2 and T1 (T3). To understand the impact of these I_{SOT} and I_{STT} variations on the p-SOT-MTJ, we conducted simulation experiments, the results of which are depicted in **Fig. 13.** **Fig. 13(a)** illustrates the effects of different I_{SOT} values fed to the p-SOT-MTJ to initialize it from the R_H state to the intermediate state. The initialization operation cannot be completed when the I_{SOT} is below 60 μA. When the I_{SOT} is 60 μA, the initialization delay is 418 ps; when the I_{SOT} increases to 110 μA, the initialization delay shortens to 85.2 ps. The minimum I_{SOT} in the proposed circuit is 182 μA, ensuring that the initialization operation can be accomplished within 50 ps. Nevertheless, in this design, the 300 ps allocated in the SOT write phase is adequate to reliably achieve the initialization operation.

Fig. 13(b) demonstrates the impact of different I_{STT} values on the p-SOT-MTJ during the "writing 1" operation, with I_{SOT} fixed at 180 μA and a 300 ps delay. Owing to the assistance of I_{SOT}, a smaller I_{STT} can complete the write operation. However, the smaller the I_{STT}, the greater the write delay. When I_{STT} is 2 μA, the write delay is 4.5 ns; when it is 10 μA, the write delay is 1.8 ns. Consequently, even in the worst-case scenario (where 10 μA I_{STT} is generated), the proposed design can maintain a high computing speed.

To verify the reliability of Boolean logic operations, we conducted Monte Carlo simulations of them. In addition to the GAA-CNTFET process variations, a 3% variation was considered for the important process parameters of the MTJ, including the MgO barrier thickness (t_{ox}), FL thickness (t_{sl}), and TMR ratio under zero bias voltage [22], as shown in **Table II**. Boolean logic operations with 0 XORing 0 and 1 XORing 0 are shown in **Fig. 14**. All MTJ states were correctly reversed, implying that all the data were correctly calculated. **Fig. 14(a)** shows that when 0 XORing 0 is conducted, the state of the MTJ is rapidly reversed to R_L, indicating that the calculated result is zero. Owing to the relatively large I_{STT}, 0 XORing 0 is achieved with a mean delay of 0.7 ns, the corresponding standard deviation being 28 ps. Similarly, **Fig. 14(b)** shows that the state of the MTJ reverses to R_H when 1 XORing 0 is performed, indicating that the calculated result is 1, the I_{STT} flowing to MTJ from M1 to SL through the heavy-metal layer. Owing to the weak ability of the n-CNTFET to pass a high voltage, the I_{STT} is relatively low. Nonetheless, the results show that 1 XORing 0 is completed with a mean delay of 1.8 ns, the corresponding standard deviation being 92 ps.

B. Comparison of Energy Consumption and Delay With 14 nm FinFET/p-SOT-MTJ

To accurately demonstrate the advantages of the GAA-CNTFET, we used the commercial 14-nm FinFET process and the p-SOT-MTJ [22] modeled using Verilog-A to realize a 16-Kb MRAM-based CIM macro for comparison.

Fig. 15(a) shows the energy consumption and delay of the proposed MRAM in memory mode. Different delays and energy consumption are generated when writing different

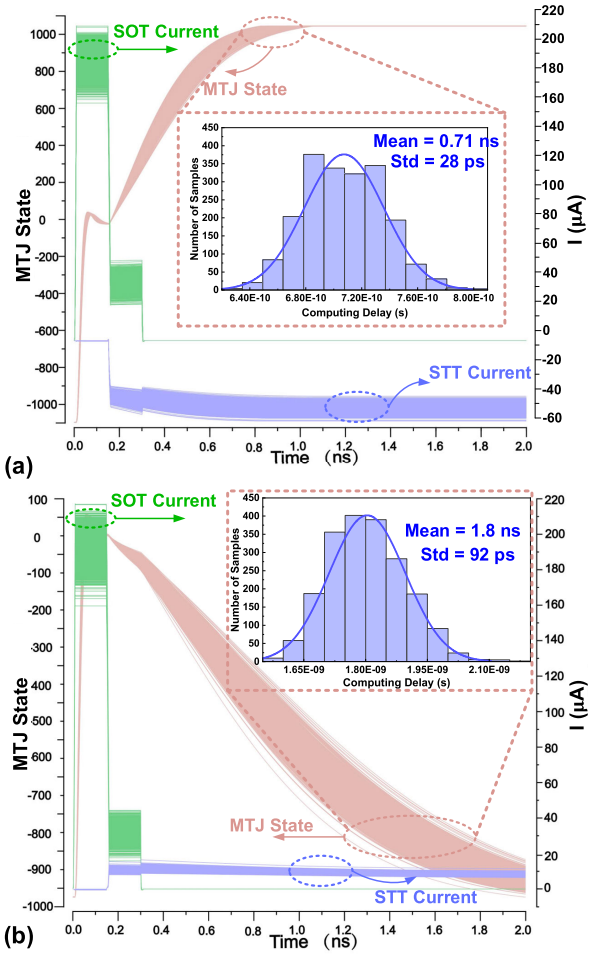


Fig. 14. Monte Carlo simulations of (a) 0 XORing 0 and (b) 1 XORing 0 for CIM operations at 0.8 V.

data (0 or 1), considering the I_{STT} differs when different resistance states of the MTJ are written. When the R_L is written, the I_{STT} is large, resulting in a lower delay and higher energy consumption. Conversely, when R_H is expressed, the I_{STT} is small, resulting in a higher delay and lower energy consumption. The write energy consumption is determined primarily by the intrinsic properties of the MTJ. Consequently, the improvement based on the GAA-CNTFET/p-SOT-MTJ design, with an average increase of 6.5%, is not significant. However, the write speed is determined primarily by the current generated by the transistor. Compared to the FinFET, the GAA-CNTFET has a larger on-state current [50], resulting in a 21% increase in the write speed.

For read operations, the read delay includes the precharge (T_{Pre}), discharge (T_{Dis}), and sensing time. The T_{Pre} and T_{Dis} are fixed; thus, the improvement (4.4%) in read delay brought by the GAA-CNTFET design comes primarily from the response speed of the SA. The read energy consumption comprises the discharge and SA. Compared to the FinFET/p-SOT-MTJ design, the SA based on the GAA-CNTFET has a higher response speed and lower energy consumption, decreasing the reading energy consumption by 45.3%.

Because Boolean logic operations are consistent, only the simulation results of the XOR operation are presented in **Fig. 15(b).** In the worst case (1 XORing 0), the MTJ is set

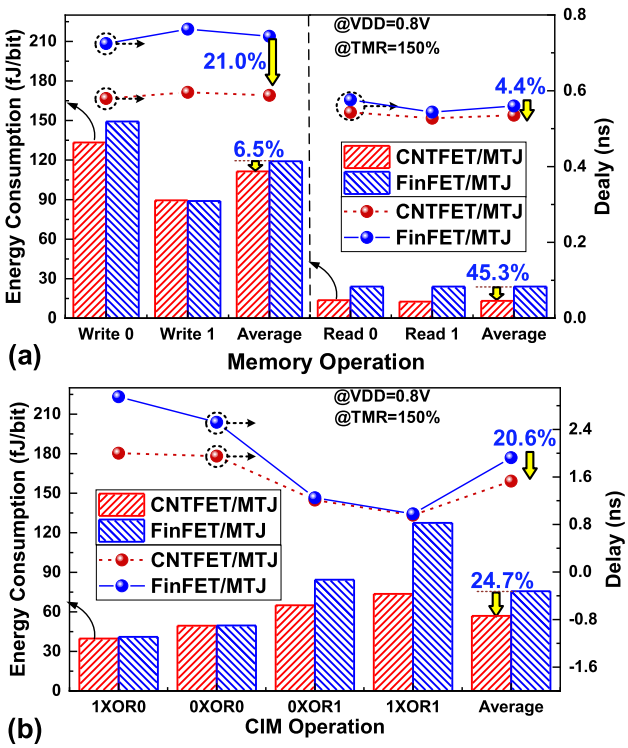


Fig. 15. Simulation results of the energy consumption and delay: (a) Memory operation. (b) CIM operation.

to a high-resistance state (R_H) through an n-type transistor. The GAA-CNTFET-based design can complete the calculation in 1.8 ns, whereas the FinFET-based design needs 2.95 ns to complete it. Consequently, the proposed MRAM-based CIM macro based on FinFETs requires a longer operational period. Although its on-state current is smaller than that of the GAA-CNTFET, the additional computing delay increases the energy overhead. For the CIM operation, the average energy consumption of the GAA-CNTFET/p-SOT-MTJ design and FinFET/p-SOT-MTJ design is 56.95 and 75.64 fJ/bit, respectively. The average computing delay of the GAA-CNTFET/p-SOT-MTJ design and FinFET/p-SOT-MTJ design is 1.53 and 1.93 ns, respectively. Therefore, compared with the FinFET/p-SOT-MTJ design, the proposed MRAM-based CIM macro based on the GAA-CNTFET/p-SOT-MTJ exhibits 20.6% and 24.7% reductions in computing delay and energy consumption, respectively.

To further demonstrate the performance of the proposed MRAM-based CIM macro, the simulation results for the 1-bit half-adder and 1-bit full-adder operations are shown in **Fig. 16**. Compared to the FinFET/p-SOT-MTJ design, the energy consumption of the 1-bit half-adder and 1-bit full-adder operations decrease by 34.6% and 30.5%, respectively, and the delay of the 1-bit half-adder and 1-bit full-adder operations decrease by 33.3% and 27.3%, respectively.

C. Comparison With Other Works

Recent literature detailing the implementation of Boolean logic and full-adder operations provides a basis for comparing devices, cycles, and delay results, as listed in **Table III**. Compared to state-of-the-art work [9], [42], [44], the proposed

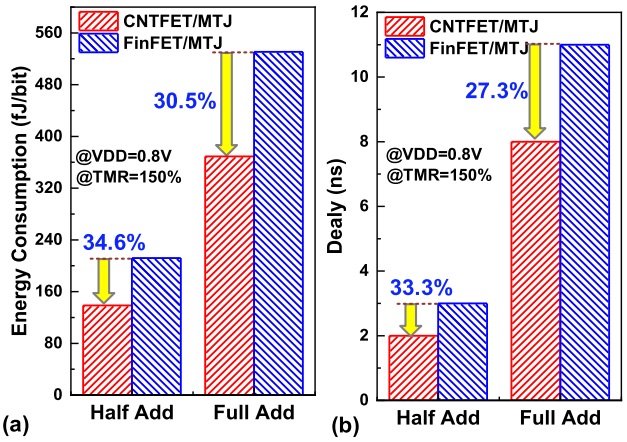


Fig. 16. Simulation results of 1-bit half-adder and 1-bit full-adder operations: (a) energy consumption. (b) delay.

architecture only requires one cycle (including the ① SOT write phase, ② SOT and STT write phase, and ③ STT write phase) to perform logic operations. Notably, compound logic operations such as XOR and XNOR can also be executed within one cycle, without the need for preset cycles. In contrast, previous works, including [9], [42], [44], necessitate preset cycles before logic operations, involving more intricate timing control and resulting in delays when implementing full-adder operations. Furthermore, the proposed design offers the flexibility to achieve various logic operations by merely altering the coding of the BLs. This approach reduces the overhead compared to other works, such as in [36] (4T+2M+SA) [17] (20T+SA), and [9] (15T+5M), where the overhead for logic operations is comparatively higher (3T+1M). Moreover, when comprehensively evaluating the overall computing operations' delay, it is essential to consider the writeback delay.

In the proposed CIM scheme, the computed result is directly stored in the 3T1M cell, effectively eliminating the need for a writeback operation. This results in a reduction in delay when compared to previous works referenced in [9], [17], [18], [42], [51], [52], and [53]. In a related study [26], it demonstrated faster calculation speeds during logic operations. However, this approach had a substantial circuit overhead (117T+6M) and included an SA structure in each computing unit, which reduced calculation parallelism. Additionally, it required additional writeback cycles to store the computed results. In summary, the proposed CIM macro offers advantages in terms of device overhead, operation cycles, and computing delay.

Table IV compares the proposed MRAM-based CIM scheme with state-of-the-art CIM schemes—that is, the BSCIM [9], CRAM [41], and hybrid compute static random access memory (HCSRAM) [16] schemes. In [9] and [41] a write-based CIM architecture was adopted, in which the calculated Boolean logic was stored in the target cell. However, the BSCIM scheme performed Boolean logic by increasing the number of additional decision cells, the preset cycle also requiring increased area overhead and computing delay. The throughput and energy efficiency of the BSCIM architecture were 51.2 GOPS and 9.9 TOPS/W respectively. Compared to the CRAM scheme, the XOR and XNOR in the proposed

TABLE III
COMPARISON OF THE LOGIC SCHEMES

Operation	Design	CIM Operation			Writeback	Total Delay (ns)
	References	Devices	Cycles	Delay (ns)	Delay (ns)	
AND	19'TED [42]	1T+1M	1	6	0	6
	22'TNANO [36]	4T+2M+SA	1	0.68	5.39	6.07
	21'JSSC [17]	20T+SA	1	3.3 [1]	2.7 [1]	6
	22'TCAS-I [9]	15T+5M	2 [3]	4 [3]	0	4
	Our	**3T+1M**	**1**	**1.8**	**0**	**1.8**
NAND	22'TNANO [36]	4T+2M+SA	1	0.68	5.39	6.07
	22'TCAS-I [9]	15T+5M	2	4	0	4
	Our	**3T+1M**	**1**	**1.8**	**0**	**1.8**
OR	19'TED [42]	1T+1M	1	6	0	6
	22'TNANO [36]	4T+2M+SA	1	0.68	5.39	6.07
	21'JSSC [17]	20T+SA	1	3.3 [1]	2.7 [1]	6
	22'TCAS-I [9]	15T+5M	2 [3]	4 [3]	0	4
	Our	**3T+1M**	**1**	**1.8**	**0**	**1.8**
NOR	22'TNANO [36]	4T+2M+SA	1	0.68	5.39	6.07
	22'TCAS-I [9]	15T+5M	2 [3]	4 [3]	0	4
	Our	**3T+1M**	**1**	**1.8**	**0**	**1.8**
XOR	23'TETC [44]	4T+2M	2	2.01	0	2.01
	17'HPCA [18]	12T+SA +Logic	1	2.65	2.42	5.07
	18'TVLSI [40]	2T+2M+SA +Logic	1	6.31	9.34	15.65
	19'TED [42]	1T+1M	2	10	0	10
	22'TCAS-I [9]	15T+5M	3 [3]	6 [3]	0	6
	Our	**3T+1M**	**1**	**1.8**	**0**	**1.8**
XNOR	23'TETC [44]	4T+2M	1	1.71	0	1.71
	17'HPCA [18]	12T+SA +Logic	1	2.65	2.42	5.07
	18'TVLSI [40]	2T+2M+SA +Logic	1	6.31	9.34	15.65
	22'TCAS-I [9]	15T+5M	3 [3]	6 [3]	0	6
	Our	**3T+1M**	**1**	**1.8**	**0**	**1.8**
1-bit FA	22'TMAG [26]	117T+6M	1	0.134	0.357	0.498
	17'TCAD [27]	23T+3M	1	7 [4]	0	7
	23'TETC [44]	12T+6M+SA+ 2 Delay Circuits	6	3.19	0	3.19
	17'HPCA [18]	18T+SA +Logic	1	2.65	2.42	5.07
	18'TVLSI [40]	3T+3M+SA +Logic	1	6.51	9.34	15.85
	19'TED [42]	6T+6M+SA +3 Registers	10	36 [2]	0	36 [2]
	22'TCAS-I [9]	30T+10M	4 [3]	8 [3]	0	8
	22'TNANO [36]	6T+3M+SA +Logic	1	0.7	5.39	6.09
	19'ISCAS [51]	14T+10R [5]	10	100 [2]	0	100
	17'J.Phys.D. [52]	8T+8R [5]	27	2000	0	2000
	18'ESSDERC [53]	9R [5]	43	345	0	345
	Our	**9T+3M+SA**	**5**	**8**	**0**	**8**

[1]Calculated form frequency of read and logic operations; [2]Calculated form transient simulation;
[3]Where contained a preset cycle; [4]Where the delay included write and read operations.
[5]Resistive random access memory.

MRAM-based CIM scheme can be implemented in one cycle, greatly reducing the number of operations required. In the read-based CIM architecture [16], [36], SAs were used to distinguish the BL voltage; when multiple rows were activated, the sensing margin decreased considerably. Moreover, the reference voltage-generating circuits also increased the system complexity. In the HCSRAM architecture, two rows were activated to achieve logic operations, where achieving XOR and XNOR still required the peripheral circuit to realize 32.7 GOPS of throughput and 5.27 TOPS/W of energy efficiency.

The proposed CIM architecture, based on MRAM, achieved an impressive maximum throughput of 8192 GOPS when considering only the computing operations, as illustrated in **Fig. 17(a)**. This represents a substantial improvement, being

TABLE IV
COMPARISON OF THE PROPOSED STUDY WITH PREVIOUS IN MEMORY BOOLEAN LOGIC STUDIES

	This work		BSCIM [9]	CRAM [41]	HCSRAM [16]	TNANO [36]			JSSC [17]
Technology	CNTFET/MTJ	14 nm FinFET/MTJ	28nm CMOS/MTJ	45nm CMOS/MTJ	28nm CMOS	45nm CMOS/MTJ	90nm CMOS/MTJ	180nm CMOS/MTJ	28nm CMOS
Cell type	3T1M		3T1M	2T1M	8T	2T1M			10T
Array size	16Kb		32KB	1Mb	16KB	1Kb			4Kb
Supply voltage (V)	0.8		1.1	—	0.6~1.1	—			0.7~0.9
Full-array Activated	Yes		No	No	No	No			No
Logic Energy (fJ/bit)	56.9	75.6	175~385	—	—	230	570	1230	15 (0.9V) 10.5 (0.7V)
Energy Efficency (TOPS/W)	17.6	13.2	9.9	3.58	5.27	4.3	1.8	0.8	66.67(0.9V) 95.2(0.7V)
Throughput (GOPS)	8192 [2]	5461 [2]	51.2	45.6	32.7	141 [1]	89 [1]	80 [1]	38.4
CIM Type	Write-based		Write-based	Write-based	Read-based	Read-based			Read-based
Type of supported functions	In-situ/Logic(AND/NAND/OR/NOR/IMP/XOR/XNOR)/Half Add/Full Add		In-situ/Logic/ Full Add/ Multiplication	In-situ/Logic/ Full Add/ Multiplication	Logic/Add/FP/ Sub/Mult/Div	Logic/Full Add			SRAM/CAM /Logic/

[1]Calculated as 32×3 Logic/cycle × 1 Ops/Logic × $1/T_{latency}$ cycles. Where $T_{latency}$ is computing latency;CAM, content-addressable memory.
[2]Only computing operation is considered.

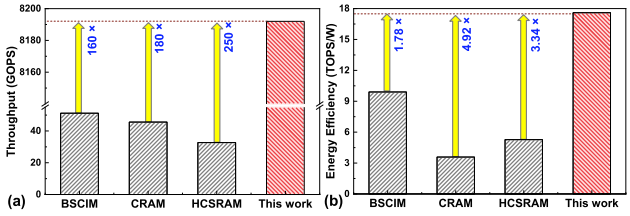


Fig. 17. Comparison of (a) throughput, (b) energy efficiency.

approximately 160–250 times higher than existing in-memory Boolean logic methods, taking into account the activation of the full array for calculations. In terms of the energy efficiency, **Fig. 17(b)** shows an improvement in the proposed MRAM-based CIM architecture. Compared to the BSCIM and CRAM schemes, its energy efficiency improves by 1.78–4.92 times owing to its flexible Boolean logic configuration.

V. CONCLUSION

In the post-Moore's Law era, it is important to explore devices with new physical characteristics and novel computing architectures. This study proposed an MRAM-based CIM macro using a hybrid p-SOT-MTJ/GAA-CNTFET method. The proposed method exhibits a flexible configuration coding mode based on a 3T1M cell to achieve various full-array Boolean logic without additional peripheral circuits. Additionally, the proposed structure uses fewer MTJs and cycles to realize full-adder operations based on the proposed half-adder operation. Hybrid simulation results demonstrated that compared to the 14-nm FinFET/p-SOT-MTJ design, the write and computing latencies of the GAA-CNTFET/p-SOT-MTJ CIM macro were reduced by approximately 21% and 20.6%, respectively, and the read and computing energy consumption were reduced by 45.3% and 24.7%, respectively. The throughput of the proposed MRAM-based CIM macro increased by 160–250 times to 8192 GOPS. The energy efficiency was 17.6 TOPS/W under a 0.8V supply voltage.

REFERENCES

[1] A. B. Kelleher, "Celebrating 75 years of the transistor a look at the evolution of Moore's law innovation," in *IEDM Tech. Dig.*, Dec. 2022, pp. 1.1.1–1.1.5, doi: 10.1109/IEDM45625.2022.10019538.
[2] F. Schwierz and J. J. Liou, "Status and future prospects of CMOS scaling and Moore's law—A personal perspective," in *Proc. IEEE Latin Amer. Electron Devices Conf. (LAEDC)*, Feb. 2020, pp. 1–4, doi: 10.1109/LAEDC49063.2020.9073539.
[3] F. Razi, M. H. Moaiyeri, R. Rajaei, and S. Mohammadi, "A variation-aware ternary spin-hall assisted STT-RAM based on hybrid MTJ/GAA-CNTFET logic," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 598–605, 2019.
[4] C.-X. Xue et al., "A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices," *Nature Electron.*, vol. 4, no. 1, pp. 81–90, Dec. 2020.
[5] N. Yang, X. Wang, X. Lin, and W. Zhao, "Exploiting carbon nanotube FET and magnetic tunneling junction for near-memory-computing paradigm," *IEEE Trans. Electron Devices*, vol. 68, no. 4, pp. 1975–1979, Apr. 2021.
[6] W. Zhang et al., "Neuro-inspired computing chips," *Nature Electron.*, vol. 3, no. 7, pp. 371–382, Jul. 2020.
[7] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. S36–S40, Apr. 2017.
[8] Y. Zhang et al., "Time-domain computing in memory using spintronics for energy-efficient convolutional neural network," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 3, pp. 1193–1205, Mar. 2021.
[9] J. Wang et al., "Reconfigurable bit-serial operation using toggle SOT-MRAM for high-performance computing in memory architecture," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 11, pp. 4535–4545, Nov. 2022.
[10] Z. Lin et al., "In situ storing 8T SRAM-CIM macro for full-array Boolean logic and copy operations," *IEEE J. Solid-State Circuits*, vol. 58, no. 5, pp. 1472–1486, May 2023.
[11] C. Lv et al., "Phase-change controlled magnetic tunnel junction for multifunctional in-sensor computing," *IEEE Electron Device Lett.*, vol. 43, no. 3, pp. 482–485, Mar. 2022.
[12] S. Li et al., "Phase-change-assisted spin-transfer torque switching in perpendicular magnetic tunnel junctions," *Appl. Phys. Lett.*, vol. 119, no. 12, Sep. 2021, Art. no. 122401.
[13] L. Wang et al., "Picosecond optospintronic tunnel junctions," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 24, Jun. 2022, Art. no. e2204732119.
[14] W. Yang, J.-C. Han, Y. Cao, X.-Y. Lin, and W.-S. Zhao, "Efficient spin injection in Fe_3GeTe_2/h-BN/graphene heterostructure," *Acta Phys. Sin.*, vol. 70, no. 12, Jun. 2021, Art. no. 129101
[15] 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.

[16] J. Wang et al., "A 28-nm compute SRAM with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 76–86, Jan. 2020.

[17] Z. Lin et al., "Two-direction in-memory computing based on 10T SRAM with horizontal and vertical decoupled read ports," *IEEE J. Solid-State Circuits*, vol. 56, no. 9, pp. 2832–2844, Sep. 2021.

[18] S. Aga, S. Jeloka, A. Subramaniyan, S. Narayanasamy, D. Blaauw, and R. Das, "Compute caches," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 481–492, doi: 10.1109/HPCA.2017.21.

[19] C.-S. Lee, E. Pop, A. D. Franklin, W. Haensch, and H.-S. P. Wong, "A compact virtual-source model for carbon nanotube FETs in the sub-10-nm regime—Part I: Intrinsic elements," *IEEE Trans. Electron Devices*, vol. 62, no. 9, pp. 3061–3069, Sep. 2015.

[20] C.-S. Lee, E. Pop, A. D. Franklin, W. Haensch, and H. P. Wong, "A compact virtual-source model for carbon nanotube FETs in the sub-10-nm regime—Part II: Extrinsic elements, performance assessment, and design optimization," *IEEE Trans. Electron Devices*, vol. 62, no. 9, pp. 3070–3078, Sep. 2015.

[21] M. T. Nasab, A. Amirany, M. H. Moaiyeri, and K. Jafari, "High-performance and robust spintronic/CNTFET-based binarized neural network hardware accelerator," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 527–533, Apr. 2023.

[22] 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," *J. Phys. D, Appl. Phys.*, vol. 48, no. 6, Feb. 2015, Art. no. 065001.

[23] M. M. Shulaker et al., "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature*, vol. 547, no. 7661, pp. 74–78, Jul. 2017.

[24] M. Krishna Gopi Krishna, A. Roohi, R. Zand, and R. F. DeMara, "Heterogeneous energy-sparing reconfigurable logic: Spin-based storage and CNFET-based multiplexing," *IET Circuits, Devices Syst.*, vol. 11, no. 3, pp. 274–279, May 2017.

[25] H. Thapliyal, F. Sharifi, and S. D. Kumar, "Energy-efficient design of hybrid MTJ/CMOS and MTJ/nanoelectronics circuits," *IEEE Trans. Magn.*, vol. 54, no. 7, pp. 1–8, Jul. 2018.

[26] P. Barla, V. K. Joshi, and S. Bhat, "Fully nonvolatile hybrid full adder based on SHE+STT-MTJ/CMOS LIM architecture," *IEEE Trans. Magn.*, vol. 58, no. 9, pp. 1–11, Sep. 2022.

[27] A. Roohi, R. Zand, D. Fan, and R. F. DeMara, "Voltage-based concatenatable full adder using spin Hall effect switching," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 12, pp. 2134–2138, Dec. 2017.

[28] M. BahmanAbadi, A. Amirany, K. Jafari, and M. H. Moaiyeri, "Efficient and highly reliable spintronic non-volatile quaternary memory based on carbon nanotube FETs and multi-TMR MTJs," *ECS J. Solid State Sci. Technol.*, vol. 11, no. 6, Jun. 2022, Art. no. 061007.

[29] A. A. Javadi, M. Morsali, and M. H. Moaiyeri, "Magnetic nonvolatile flip-flops with spin-Hall assistance for power gating in ternary systems," *J. Comput. Electron.*, vol. 19, no. 3, pp. 1175–1186, Sep. 2020.

[30] A. Amirany, K. Jafari, and M. H. Moaiyeri, "High-performance spintronic nonvolatile ternary flip-flop and universal shift register," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 5, pp. 916–924, May 2021.

[31] A. Amirany, M. H. Moaiyeri, and K. Jafari, "Nonvolatile associative memory design based on spintronic synapses and CNTFET neurons," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 1, pp. 428–437, Jan. 2022.

[32] A. Amirany, M. H. Moaiyeri, and K. Jafari, "Process-in-memory using a magnetic-tunnel-junction synapse and a neuron based on a carbon nanotube field-effect transistor," *IEEE Magn. Lett.*, vol. 10, pp. 1–5, 2019.

[33] A. Amirany, K. Jafari, and M. H. Moaiyeri, "True random number generator for reliable hardware security modules based on a neuromorphic variation-tolerant spintronic structure," *IEEE Trans. Nanotechnol.*, vol. 19, pp. 784–791, 2020.

[34] 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. 318–330, Apr./Jun. 2023, doi: 10.1109/TETC.2022.3214833.

[35] A. Musello, E. Garzón, M. Lanuzza, L. M. Prócel, and R. Taco, "XNOR-bitcount operation exploiting computing-in-memory with STT-MRAMs," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 3, pp. 1259–1263, Mar. 2023.

[36] K. Monga, N. Chaturvedi, and S. Gurunarayanan, "Design of a STT-MTJ based random-access memory with in-situ processing for data-intensive applications," *IEEE Trans. Nanotechnol.*, vol. 21, pp. 455–465, 2022.

[37] Z. Yang et al., "A novel computing-in-memory platform based on hybrid spintronic/CMOS memory," *IEEE Trans. Electron Devices*, vol. 69, no. 4, pp. 1698–1705, Apr. 2022.

[38] X. Wang et al., "Triangle counting accelerations: From algorithm to in-memory computing architecture," *IEEE Trans. Comput.*, vol. 71, no. 10, pp. 2462–2472, Oct. 2022.

[39] A. Jaiswal, R. Andrawis, A. Agrawal, and K. Roy, "Functional read enabling in-memory computations in 1transistor—1resistor memory arrays," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 12, pp. 3347–3351, Dec. 2020.

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

[41] 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 Trans. Comput.*, vol. 68, no. 8, pp. 1159–1173, Aug. 2019.

[42] H. Zhang, W. Kang, K. Cao, B. Wu, Y. Zhang, and W. Zhao, "Spintronic processing unit in spin transfer torque magnetic random access memory," *IEEE Trans. Electron Devices*, vol. 66, no. 4, pp. 2017–2022,3 Apr. 2019.

[43] L. Chang, X. Ma, Z. Wang, Y. Zhang, Y. Xie, and W. Zhao, "PXNOR-BNN: In/with spin-orbit torque MRAM preset-XNOR operation-based binary neural networks," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 11, pp. 2668–2679, Nov. 2019.

[44] B. Wu et al., "An energy-efficient computing-in-memory (CiM) scheme using field-free spin-orbit torque (SOT) magnetic RAMs," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 331–342, Apr./Jun. 2023, doi: 10.1109/TETC.2023.3237541.

[45] J. Feng, B. Wang, Z. Hou, C. Wang, Z. Wang, and W. Zhao, "In-memory computing paradigms by exploiting the intrinsic interactions between STT and SOT," *IEEE Trans. Electron Devices*, vol. 69, no. 12, pp. 6769–6775, Dec. 2022.

[46] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, and Y. Xie, "PinaTubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories," in *Proc. 53rd ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2016, pp. 1–6.

[47] J. Wang et al., "A self-matching complementary-reference sensing scheme for high-speed and reliable toggle spin torque MRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 12, pp. 4247–4258, Dec. 2020.

[48] A. Duluard et al., "Enhanced magnetoresistance by monoatomic roughness in epitaxial Fe/MgO/Fe tunnel junctions," *Phys. Rev. B, Condens. Matter*, vol. 91, no. 17, May 2015, Art. no. 174403.

[49] M. H. Moaiyeri and F. Razi, "Performance analysis and enhancement of 10-nm GAA CNTFET-based circuits in the presence of CNT-metal contact resistance," *J. Comput. Electron.*, vol. 16, no. 2, pp. 240–252, Jun. 2017.

[50] M. K. Q. Jooq, A. Bozorgmehr, and S. Mirzakuchaki, "A low power and energy efficient 4:2 precise compressor based on novel 14T hybrid full adders in 10 nm wrap gate CNTFET technology," *Microelectron. J.*, vol. 104, Oct. 2020, Art. no. 104892.

[51] Z. Yang and L. Wei, "Logic circuit and memory design for in-memory computing applications using bipolar RRAMs," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5, doi: 10.1109/ISCAS.2019.8702555.

[52] L. Cheng et al., "Reprogrammable logic in memristive crossbar for in-memory computing," *J. Phys. D, Appl. Phys.*, vol. 50, no. 50, Dec. 2017, Art. no. 505102.

[53] F. M. Puglisi, L. Pacchioni, N. Zagni, and P. Pavan, "Energy-efficient logic-in-memory I-bit full adder enabled by a physics-based RRAM compact model," in *Proc. 48th Eur. Solid-State Device Res. Conf.*, Sep. 2018, pp. 50–53, doi: 10.1109/ESSDERC.2018.8486886.

Zhongzhen Tong received the B.S. degree in electronic science and technology from the Tianjin University of Technology and Education, Tianjin, in 2019, and the M.Sc. degree in circuits and systems from Anhui University, Hefei, in 2022. He is currently pursuing the Ph.D. degree with the School of Integrated Circuit Science and Engineering, Beihang University, Beijing.

His research interests include hybrid carbon nanotube field effect transistor (CNTFET)/spintronic integrated circuit design and SRAM-based in-memory computing.

Yilin Xu is currently pursuing the B.S. degree in microelectronics science and engineering with Beihang University, Beijing, China. He majors in analog circuit design. His current research interests include the biomedical-oriented front-end design of analog circuits and hybrid CNTFET/spintronic integrated circuit design.

Yunlong Liu received the B.S. degree in electronic information engineering from the Qilu Institute of Technology, Jinan, China, in 2020. He is currently pursuing the M.S. degree in microelectronics and solid-state electronics with Anhui University. His research interests include in-memory computing circuits based on static random-access memory.

Xinrui Duan is currently pursuing the B.S. degree in microelectronics science and engineering with Beihang University, Beijing, China. She majors in analog circuit design. Her research interests include hybrid CNTFET/spintronic integrated circuit design.

Hao Tang is currently pursuing the B.S. degree in microelectronics science and engineering with Beihang University, Beijing, China. He majors in analog circuit design. His current research interests include emerging non-volatile memory technologies and hybrid low-power circuit designs.

Suteng Zhao received the B.S. degree in microelectronics science and engineering from Beihang University, Beijing, China, where he is currently pursuing the M.Sc. degree in integrated circuit science and engineering. He majors in the simulation and modeling of spin devices. His current research interests include magnetic recording devices and emerging non-volatile memory technologies.

Chenghang Li received the B.S. degree in material shaping and control engineering from the University of Science and Technology Beijing, Beijing, in 2023. He is currently pursuing the M.S. degree with the School of Integrated Circuit Science and Engineering, Beihang University, Beijing.

His research interests include hybrid carbon nanotube field effect transistor (CNTFET)/spintronic integrated circuits designs.

Zhiting Lin (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electronics and information engineering from the University of Science and Technology of China (USTC), Hefei, Anhui, China, in 2004 and 2009, respectively. From 2015 to 2016, he was a Visiting Scholar with the Department of Engineering and Computer Science, Baylor University, Waco, TX, USA. In 2011, he joined the School of Integrated Circuits, Anhui University, Hefei, where he is currently a Professor. He has published approximately 50 articles and holds more than 20 Chinese patents. His research interests include pipeline analog-to-digital converters and high-performance static random-access memory.

Xiulong Wu (Senior Member, IEEE) received the B.S. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2001, and the M.Sc. and Ph.D. degrees in electronic engineering from Anhui University, Hefei, in 2005 and 2008, respectively. From 2013 to 2014, he was a Visiting Scholar with the Department of Engineering, The University of Texas at Dallas, Richardson, TX, USA. He is currently a Professor with Anhui University. He has published approximately 60 articles and holds more than ten Chinese patents. His research interests include high-performance static random-access memory and mixed-signal ICs.

Zhaohao Wang (Senior Member, IEEE) received the B.S. degree in microelectronics from Tianjin University in 2009, the M.S. degree from Beihang University, China, in 2012, and the Ph.D. degree in physics from the University of Paris-Saclay, France, in 2015. His current research interests include the modeling of non-volatile nano-devices and the design of new non-volatile memories and logic circuits.

Xiaoyang Lin (Senior Member, IEEE) received the Ph.D. degree in physics from Tsinghua University, China, in 2015. He has been working on emerging spintronics materials and devices. He is currently an Associate Professor with Beihang University. He is the Chair of the Department of IC Materials and Devices, School of Integrated Circuit Science and Engineering, Beihang University. He has published more than 50 articles in SCI journals, such as *Nature Electronics*, *PNAS*, *Applied Physics Reviews*, and IEEE ELECTRON DEVICE LETTERS. He is an Editorial Board Member of *Scientific Reports*. He is an Associate Editor of *Advanced Devices and Instrumentation*.