FDAIMC: A Fully-Differential Analog In-Memory-Computing for MAC in MRAM with Accuracy Calibration under Process and Voltage Variation

Xiangyu Li[1], Weichong Chen[1], Ruida Hong[1], Jinghai Wang[1], Ningyuan Yin[1,2,*], Zhiyi Yu[1,2,*]
[1] School of Microelectronics Science and Technology, Sun Yat-sen University, China
[2] Guangdong Provincial Key Laboratory of Optoelectronic Information Processing Chips and Systems, China
E-mail: {lixy928, chenwch36, hongrd, wangjh335}@mail2.sysu.edu.cn
{yuzhiyi, yinny5}@mail.sysu.edu.cn

***Abstract*—Analog in-memory-computing (AIMC) is adopted extensively in non-volatile memory for multibit multiply-and-accumulate (MAC) operation. However, the low-on/off-ratio feature of magnetic tunnel junction (MTJ) impedes a high-performance AIMC macro based on spin transfer torque magnetic random access memory (STT-MRAM). Secondly, because of the uncertainty feature of a mixed-signal system under process and voltage variation, a calibration support is indispensable. Moreover, the incompatibility between a nonlinear analog signal and a linear digital signal hinders accurate computation and calibration support. To overcome these challenges, this work proposes a STT-MRAM-AIMC macro featuring: 1) a 2-level-differential cell array and a linear computing scheme with a calibration support in analog domain; 2) an analog-digital-conversion (ADC) system, including a slew-rate-independent voltage-to-time converter (SRIVTC) scheme and a self-triggered time-to-MAC value converter (STTMC) scheme; 3) a compact layout design for high area efficiency. Finally, an average accuracy of 95.44% is obtained under the TT&0.9V corner. By using the calibration strategy, the average accuracy of 97.8% and 88.6% are obtained under FF&0.945V and SS&0.855V separately, with over 30% enhancement. Furthermore, a 1.64~21.18 times area FoM than state of the art is obtained. An energy efficiency of 87.2~312.4 TOPS/W is obtained.**

Keywords*—*In-memory-computing (IMC), analog multiply-and-accumulate (MAC) operation, TDC, calibration, PVT, STT-MRAM

I. Introduction

Artificial intelligence (AI) edge devices intended for artificial neural networks (ANN) desire highly parallel MAC to achieve low inference latency and low energy consumption. In-memory-computing draws extensive interest [1]-[6] relying on its high performance versus conventional von Neumann-based computing circuitry. These computation strategies can be classified into two modes: all-digital and mixed-signal. The all-digital strategy is also referred to as near-memory-computing (NMC) and the mixed-signal strategy is referred to as analog in-memory-computing (AIMC). The no-accuracy loss and easy implementation advantages make NMC attractive. However, limited by the low parallelism of the readout circuit, the energy efficiency of NMC can hardly surpass that of AIMC in the same technology.

* Corresponding authors: Zhiyi Yu; Ningyuan Yin.
This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62334014; in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2023B0303030004.

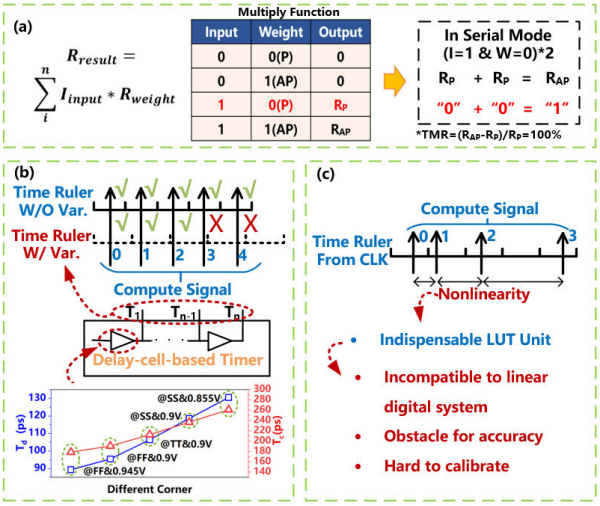


Fig. 1.Challenges in MRAM-based analog in-memory-computing(IMC) for MAC operation. (a) The "0 + 0 =1" error, because of low on/off-ratio obstacle of MTJ. (b) The time variation induced error under process and voltage variation. T_d is the time of inner clock cycle based on delay cells and T_c is the time of 1-LSB in ADC system. (c) The nonlinearity in analog domain is incompatible to linear digital system.

Extensive efforts had been dived into analog in-memory-computing macro based on RRAM [3][5]. However, MRAM featuring faster write/read speed and higher endurance lacks a high-energy-efficiency and high-throughput AIMC design. This is mainly because of two obstacles of the MTJ device. The first is low resistance, and the second is low on/off-ratio. The low resistance characteristic leads to high current in the current summation strategy that is used extensively in RRAM AIMC [7][8]. Thus, considerable energy consumption is inevitable. A resistance summation strategy is proposed in [9] and produces high energy efficiency, which provides a luminous idea to solve the low-resistance-low-energy-efficiency problem. But the low-on/off-ratio problem is rarely mentioned in the binary-neural-network (BNN) design adopted in [9]. The low-on/off-ratio feature brings a great challenge to realize multibit MAC operation in AIMC. This challenge can be called a "0 + 0 =1" paradox because of the ~100% tunnel magnetoresistance ratio (TMR) of MTJ, as shown in Fig. 1(a). (R_P represents the low resistance of MTJ stored a "0", R_{AP} represents the high resistance of MTJ stored a "1." TMR = (R_{AP} - R_P) / R_P.) Some efforts [10] [11] tried to solve the on/off-ratio problem by a SAR-ADC system but only obtained an unsatisfactory performance. RSACIM[12] tries to solve the two problems at once by a dynamic delay selection unit based on a serial delay cell and obtains a satisfactory performance. However, a calibration support is normally necessary in mixed-signal system because of the uncertainty feature under process-voltage-temperature (PVT)

978-3-9826741-0-0/DATE25/© 2025 EDAA

variation, which is lacked in RSACIM. The delay time of delay cells can be dramatically affected (> ±20%) by process and voltage variation, as shown in Fig. 1(b). Under fixed serial delay cells, the variation of the time system can be accumulated and hardly calibrated. Furthermore, the incompatibility between a linear digital system and a nonlinear analog signal brings a lot of inconvenience during analog-digital conversion and calibration, as shown in Fig. 1(c). A linear analog-to-digital system is more attractive for convenience in architecture design and robustness enhancement.

To overcome the problems mentioned above at once, a fully-differential AIMC (FDAIMC) macro for MRAM is proposed. The main contributions of this macro are as follows:

- We propose a 2-level-differential cell array structure. By using the differential concept in cell and between columns, the low-on/off ratio obstacle is solved, and an accurate linear relationship between analog signal and digital code is obtained. Benefited from the linear relationship, a calibration support is adopted in the analog domain without linearity deterioration.
- We propose a time-based ADC (TDC) system, including a slew-rate-independent voltage-to-time converter (SRIVTC) and a self-triggered time-to-MAC value converter (STTMC). The SRIVTC and a pseudo-differential structure between two SRIVTCs are delicately designed to maintain the linearity between analog signal and digital code and enhance the computing robustness under process and voltage variation. For high energy efficiency, the STTMC is composed of a self-triggered timer and a sharing decoder. Finally, an average accuracy of 95.44% is obtained under TT&0.9V. Furthermore, the average accuracy of 97.8% and 88.6% are obtained under FF&0.945V and SS&0.855V separately, with over 30% enhancement.
- The narrowness and compactness of the layout are considered throughout the design process. Thus, a 2:1 multiplexer can be applicable. Finally, a high throughput of 376 GOPS/bank and an energy efficiency of 312.4 TOPS/W are achieved in 1b-1b mode (87.2 TOPS/W @ 1b-8b). A 1.64 ~ 21.1 times area FoM than state of the art [9] [13] [14] is realized.

The rest of this article is organized as follows: Section II introduces the structure of the proposed 2-level-differential memory array and the principle of performing multibit MAC operation. Section III describes the workflow, circuit detail and layout design of the proposed FDAIMC macro. Section IV presents the performance of FDAIMC macro. Conclusions are presented in Section V.

II. 2-Level-Differential Memory Cell and Computing Principle

A. Background of resistance summation strategy

To overcome the low-resistance obstacle, a resistance summation strategy was adopted in [9] [10] [12] and obtained a satisfactory energy performance. The basic concept is to form a computing path with a series connection of MTJs. Then, the computation result is contained in the sum of resistance along the computing path. Thus, by using signal processing in analog and digital domains, the computing result can be

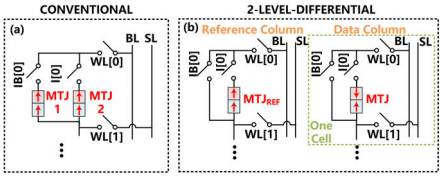


Fig. 2. Structure of Memory cell. (a) Conventional 3S2R cell in[9]. (b) Proposed 3S1R 2-level-differential cell. Only one reference column is needed in an array.

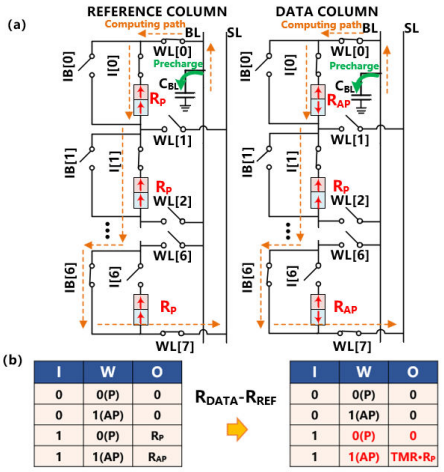


Fig. 3. The 2-level-differential computing principle for MAC operation. (a) Computing path in the reference column and the data column in a 7-row-activition compute. (b) Solution for low on/off-ratio obstacle by using resistance-differential computing in a row.

extracted and converted to a digital code. In this work, this resistance summation strategy is used as a basic computing principle, and a fully differential concept is merged.

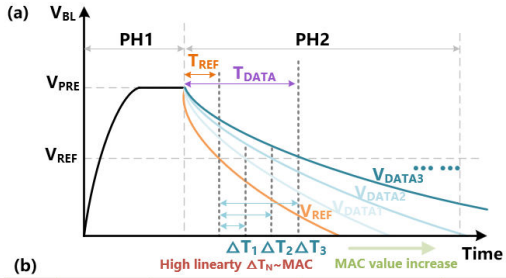
B. 2-level-differential structure of memory cell

In the cell level, as shown in Fig. 2(b), a 3-switch-1-resistive-memory-cell (3S1R) structure is proposed against the conventional 3S2R structure, which can effectively reduce the energy consumption and control logic during write operation. The WL switch between two adjacent cells in a same column is shared by them two. A pair of same-size NMOS is used as I/IB switch. The I[n] and IB[n] switches are complementary during computing event. Notice that the series of MOS and MTJ may cause insufficiency of write voltage, which is dependent on the MTJ device characteristic. Thus, for a conservative design, transmission gates are used as WL switches.

In the column level, a reference column is leveraged to overcome the low-on/off-ratio obstacle during realizing MAC operation. The MTJ cells in the reference column are fixed in R_p state and share the same WL[n] with the data column.

C. Computing Principle for MAC operation

A precharge-discharge workflow is used instead of conventional strategy in [9]. In brief, a computation capacitor (or the stary capacitance) on the bit-line (BL) is precharged to V_{PRE} and discharged through a series connection of MTJ (a computing path), producing an analog signal that contains the computation result. Then, the computation result is extracted in the analog domain and converted to a digital MAC value through an TDC system. Details of the computation principle are demonstrated below.



(b)

MAC value	Case	T_{REF}	T_{DATA}	$\Delta T = T_{DATA} - T_{REF}$
7	IN:1111111 W:1111111	$(7*R_L+2*R_W)*A$	$(7*R_H+2*R_W)*A$	$7*(R_{AP}-R_P)*A$
6	IN:0111111 W:1111111	$(R_{mos}+6*R_L+2*R_W)*A$	$(R_{mos}+6*R_H+2*R_W)*A$	$6*(R_{AP}-R_P)*A$
	IN:1111111 W:0111111	$(7*R_L+2*R_W)*A$	$(6*R_H+R_L+2*R_W)*A$	$6*(R_{AP}-R_P)*A$
	IN:0111111 W:0111111	$(R_{mos}+6*R_L+2*R_W)*A$	$(R_{mos}+6*R_H+2*R_W)*A$	$6*(R_{AP}-R_P)*A$
... ...				
0	IN:0000001 W:0000000	$(6*R_{mos}+R_L+2*R_W)*A$	$(6*R_{mos}+R_L+2*R_W)*A$	0
0	IN:0000000 W:0000000	$(7*R_{mos}+2*R_W)*A$	$(7*R_{mos}+2*R_W)*A$	0

*$R_L = R_{mos} + R_P$; $R_H = R_{mos} + R_{AP}$; R_W means the resistance of WL switch and BL switch

Fig. 4.MAC operation principle in time domain. (a)Timing waveform in analog domain. (b) The mapping relationship between MAC value and time signal. The case "1111111" means there are seven "1".

When the input signal I[n] is high (representing "1"), the I[n] switch is closed and the IB[n] switch is opened. Then, the summation resistance of R_{MTJ} and R_{MOS} is counted into the whole resistance along computing path. For inverse case, I[n] is "0" and IB[n] is "1", only R_{MOS} makes a contribution to the sum of resistance. Thus, R_{MTJ} is the difference between these two cases. This feature forms the first-level difference. When the input signal I[n] (IN) = "1" and activated weight (W) = "0" in a row, the contribution of resistance from this row in reference column and data column are the same (R_P). Thus, the resistance difference between the two columns is zero. When IN = "1" and W = "1," the resistance difference between the two columns is R_{AP} - R_P = R_P * TMR, as shown in Fig. 3(b). Thus, the second-level difference is formed. Furthermore, the position related nonideal effect observed in [9] can also be alleviated. Two factors play key roles in this nonideal phenomenon. The first is the R_{MOS} increasement induced by V_{GS} loss, and the second is the stray capacitance in middle nodes. Because of the less row activation in this design, the factor of effective V_{GS} loss of MOS is predominant. The position in the data column, which induces the effect, can be traced in the reference column because of the same input activation. Thereby, the difference of two columns can effectively alleviate this deterioration and thus enlarge the signal margin.

As mentioned above, the computing result is contained in the differential resistance between the reference column the and data column. The next step is to convert this resistance information into a time signal, realizing a sampling for TDC. As shown in Fig. 4(a), BL is precharged to V_{PRE} and then discharged through a computing path. In essence, this is a process of RC discharge. Thus, we have the relationship:

$$T = R_{sum} * C * \ln\left(\frac{V_{PRE}}{V_{REF}}\right) \quad (1)$$

R_{sum} denotes the total resistance along the discharge path, C is the computing capacitance on the BL, V_{PRE} denotes the precharge voltage on the BL, V_{REF} denotes the detect voltage of the voltage-to-time converter (VTC), and T denotes the time for which V_{BL} decreases from V_{PRE} to V_{REF}.

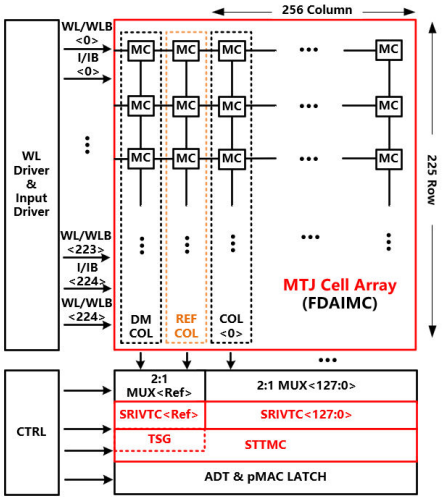


Fig. 5.The whole sheme of FDAIMC macro.

By using (1), we can get the differential time between the data column and reference column in formula (2).

$$\Delta T = T_{DATA} - T_{REF} = (R_{DATA} - R_{REF}) * C * \ln\left(\frac{V_{PRE}}{V_{REF}}\right) \quad (2)$$

$$C * \ln\left(\frac{V_{PRE}}{V_{REF}}\right) = A \quad (3)$$

R_{DATA} denotes the total computing resistance along the data column, R_{REF} denotes the total computing resistance along the reference column, and A is an algebraic sign. By using formulas (2) and (3), a clear map of ΔT vs. MAC value can be obtained, as shown in Fig. 4(b).

As mentioned above, PVT variation can lead to a deviation on delay time or computing resistance. A PVT tracking technology [15] may be the best choice but with very high difficulty in design and area/power cost. The calibration method in [14] can be an alternative option. Given the formula (2), V_{PRE} and V_{REF} provide the possibility to tune all the output channels at once in analog domain rather than a one-by-one mode in digital domain as used in [14]. Thus, we adopt a fixed V_{REF} and a float V_{PRE} to achieve the calibration simply.

III. FDAIMC Macro

A. FDAIMC Macro Scheme

To realize the computation principle in circuitry, the FDAIMC macro is proposed, as shown in Fig. 5. The FDAIMC macro consists of two core parts. The first part is the MTJ cell array, which offers the computing signal in the analog domain, as demonstrated in the previous section. The second part is the ADC system, in which a time-based ADC is leveraged to achieve high energy efficiency and area efficiency. The ADC system is composed of SRIVTC and STTMC. Other non-core parts include drivers, multiplexer (MUX), controller (CTRL), and adder tree (ADT). To reduce latency and enhance throughput, a fully asynchronous control logic based on delay elements is used. The timing sequence waveform is given in Fig. 6. The workflow of macro is described as below.

When external actuating signal EN arrives, an initialization phase is activated and INI is pulled high. The BL will be precharge to V_{PRE}, and the trigger circuit will be initialized. After initialization, CEN is pulled high, which

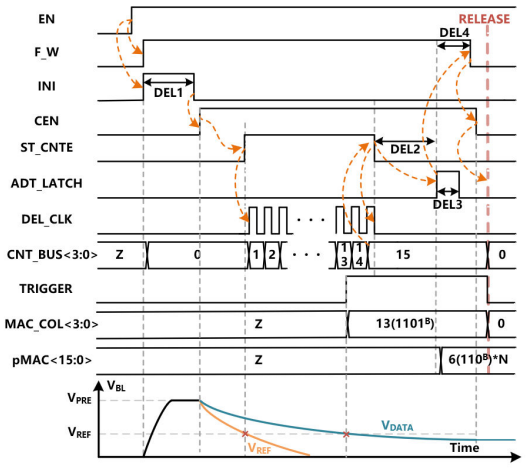


Fig. 6.Timing sequence waveform of FDAIMC macro.

means the computing process is enabled. As described in the previous section, a decreasing V_{BL} is offered by a cell column. A SRI-VTC is introduced to produce a flip signal when V_{BL}=V_{REF}. The first flip signal is always from the reference column, which will trigger ST_CNTE to high. And the timer is activated to work. During computing, inner clock DEL_CLK is generated, which will drive a counter to count. The output code of the counter is sent to a BUS<3:0> through drivers. Each data column has a SRI-VTC circuitry. When V_{BL} of a data column crosses V_{REF}, a flip signal, TRIGGER, will be activated. Through a synchronizer, which is designed to avoid error-code latch during bus code refreshing, TRIGGER will activate a LAS signal. Then, LAS latches the bus code in a dynamic-logic-based latch. This latched code MAC_COL<3:1> is the MAC value in a data column. When the counter is full, STP_CNTE is pulled down. At this moment, compute in all data columns is done. The MAC values are sent to ADT for summation during a delay time (DEL2). Finally, the output of ADT is latched in the output terminal. The whole macro will be released in a while after F_W is pulled down.

B. SRIVTC

The proposed SRIVTC is shown in Fig. 7(a). The core part of this module is a threshold-cross detector (TCD), shown in Fig. 7(b). A design of TCD with high speed, low power, low and stable latency, PVT immune ability, and low mismatch is very challenging. A conventional strategy [15] of differential amplifier is used as a reference. To achieve low power consumption, a sub-threshold technology is leveraged to input terminals. And M9~M14 is introduced to preset the correlated node to around VDD/2 with a negligible leakage increasement. A similar digital gain-tunning strategy in [16] is adopted. The M0/M1[1:0] means the effective width of M0/M1 can be adjusted by a 2-bit digital signal and a simple combinatorial logic.

During the idle phase (INI = CEN = “0”), M1 is off and only a tiny leakage current (~0.06μA) exists. In the initialization phase (INI = “1” & CEN = “0”), M0 and M1 are on. The amplifier begins to work and calibrates itself. In the computing phase (INI = “0” & CEN = “1”), M0 and M1 are on and V_{OUT} is pulled low at the beginning. The inverter is working in a waiting state until V_{OUT} goes high. When V_{BL}=V_{REF}, V_{OUT} is pulled high. Then, M7 is closed and VSPB is pulled low. Finally, VSP (equal to TRIGGER) is pulled high to open M0 and cutoff the bias current.

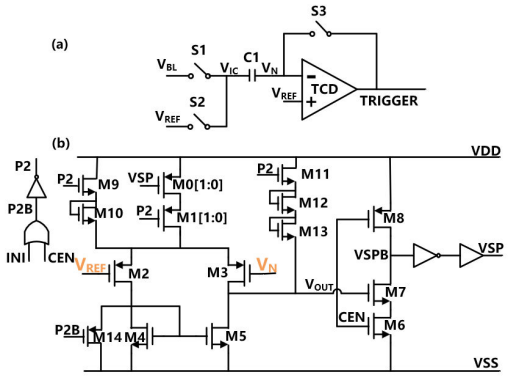


Fig. 7.The schematic of SRIVTC. (b) The schematic of threshold-cross detector (TCD) in VTC.

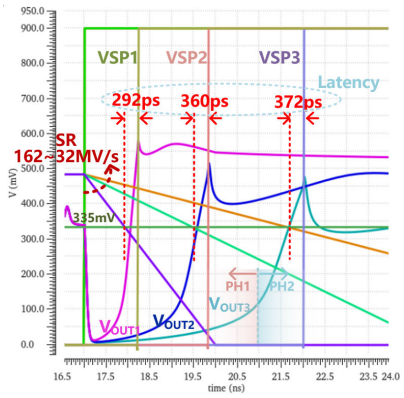


Fig. 8.The small latency varation of the proposed TCD under different SR.

A great challenge for accurate computing is the mismatch between different columns. For a SR of 50 MV/s, a 5 mV mismatch of the detecting point will lead to a 100 ps error. A self-calibrate strategy is employed to alleviate this problem, as shown in Fig. 7(a). During the calibration phase, S1 is opened and S2 and S3 are closed. Then the offset information will be stored in C1 in an analog format. During the computing phase, S1 is closed and S2 and S3 are opened. V_{BL} will be coupled with V_N by C1. By using this strategy, a 10 ~ 40 ps (under different SR) standard deviation of latency can be obtained in Monte Carlo simulation. For a simple control logic, S1 is CEN and S2/S3 is INI.

Because the linearity of ΔT vs. MAC in formula (2) is clear. Any latency produced by TCD will be an error in TDC system, which will be converted to an incorrect digital code. Thanks to the pseudo-differential structure formed by the reference column and data column, the net value of time is suppressed to a negligible level in the same condition.

Another error source is SR. Because the latency of the TCD is SR-related, the different SR of V_{BL} in reference and data columns will induce a latency difference, which will be a time error. The latency is produced in two phases. In the first phase, V_N is higher than V_{REF} and continues to decline. Before V_N = V_{REF}, V_{OUT} will have a small increasement, which will decrease the latency that V_{OUT} needed to turn on M7. A lower SR leads to a higher V_{OUT} increasement. Thus, in this phase, SR has a negative correlation with latency. In the second phase, when V_N = V_{REF}, V_{OUT} will be pulled high quickly. A higher SR will induce a larger ΔV (= V_{REF} -V_N) in the same time interval. Thus, a higher SR induces a higher charging current, leading to a low latency. In this phase, SR has a

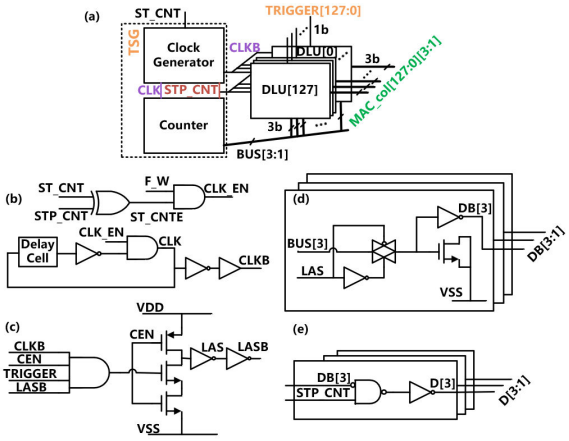


Fig. 9.The scheme of (a) STTMC, (b) clock generator, (c) synchronizer, (d) dynamic-logic-based latch, and (e) isolation unit.

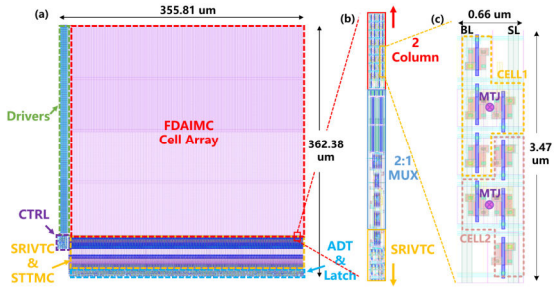


Fig. 10. (a) The layout of an FDAIMC macro. (b) The implementation of 2:1 tight layout. (c) The layout of two cells.

positive correlation with latency. As shown in Fig. 8, the summation latency of these two phases can exhibit an SR-independent characteristic in a range of SR by proper circuit tuning. Based on simulation, the availability of this feature is related to the bias current. To overcome the PVT influence, a digital calibration strategy is used to keep a suitable bias.

C. STTMC

The proposed STTMC is shown in fig. 9(a). It consists of a time scale-plate generator (TSG) and a set of data latch units (DLU). The inner clock generator is activated by ST_CNT the trigger signal from the reference column. The clock generator is based on delay cell and control logic, as shown in Fig. 9(b). A synchronizer, shown in Fig. 9(c), is introduced to coordinate with the code-sharing system. The moment of arrival of TRIGGER from a data column is uncertain, because of its analog feature. Thus, if the latch occurs at the moment at which BUS code is refreshing, an error code may be latched. This synchronizer is driven by DEL_CLKB, which is the complementary signal of DEL_CLK. In this strategy, the BUS code will have a half cycle time to refresh without error latch action. Furthermore, the BUS<0> is designed as a redundancy bit. To decrease the energy consumption and area consumption, a transmission-gate-based dynamic logic latch is used, as shown in Fig. 9(d). Furthermore, a set of logic gates is inserted between latches and ADT to avoid meaningless flips of ADT, shown in Fig. 9(e).

D. Layout design

A tight layout design can effectively enlarge the throughput by increasing output channels at the same array size. The narrowness of layout of SRIVTC and STTMC is concerned during device parameter selection, e.g., the length

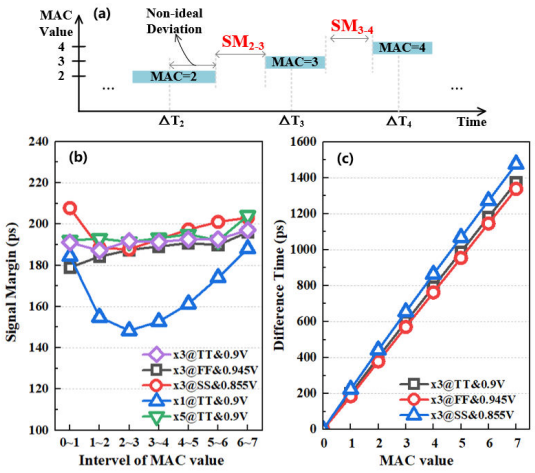


Fig. 11. (a) The definition of SM. (b) The SM analysis under different corner and MOS size. "xN" means the MOS in a cell has a "N" times width than the minimum size MOS. (c) The linear relationship between difference time and digital code under different corner.

TABLE I KEY PARAMETERS OF MTJ

Parameter	Value
MTJ area	40 nm x 40 nm x π/4
Oxide barrier height of MTJ	0.85 nm
Free layer height of MTJ	1.3 nm
Temperature	300K
Nominal R_{MTJ} at R_P(R_{AP}) of MTJ	4KΩ (9.3KΩ)
TMR	132%

of M2 and M3 in SRIVTC. Thus, a 2:1 multiplexer can be adoptable. To evaluate the area advantage of FDAIMC architecture, the layout of a macro is designed, as shown in Fig. 10(a). The tight layout implementation in one output channel is shown in Fig. 10(b). A standard MOS layout is used, as shown in Fig. 10(c).

IV. PERFORMENCE EVALUATION

A. Roubustness Analysis

To evaluate the performance of the proposed FDAIMC, a 28nm CMOS technology and a perpendicular-magnetic-anisotropy MTJ compact model [17] are used. The CMOS-MTJ hybrid simulation is demonstrated on the Cadence Virtuoso platform. The basic parameter of MTJ is shown in TABLE I. The corner featuring TT, 0.9V, and 27°C is used as the normal condition. A ±5% variation of power supply is adopted to evaluate the robustness under voltage variation. Thus, the fastest corner is FF&0.945V and the slowest corner is SS&0.855V.

Firstly, the signal margin (SM), which is seen as the performance of the cell array in the analog domain, is analyzed. The SM is the time interval between two MAC values considering the non-ideal deviation, as shown in Fig. 11(a). The V_{PRE} and V_{REF} are set to 484 mV and 335 mV, respectively. Given the position-dependent error (reported in [9] [12] and mainly contributed by V_{GS} loss of MOS in this design) will deteriorate the signal margin, MSB type and LSB type of weights distribution for each MAC value are considered to analyze the worst case of SM, e.g., "1100000" and "0000011." (Here, MSB means the nearest position to the head of the series structure, and LSB means the farthest position.) As shown in Fig. 11(b), the SM keeps a high

TABLE II COMPARATION WITH PREVIOUS WORKS

		NATURE 2022[9]	*VLSI 2023[13]*	*JSSC 2023[14]*	*TCASI 2024[12]*	**This work**
CMOS Technology		28nm	40nm	22nm	14nm	**28nm**
Memory Type		MRAM	RRAM	RRAM	MRAM	**MRAM**
Array Size(row x col.)		64 x 64	256 x 256	1024 x 256	2Kb(no detail)	**225 x 256**
Supply Voltage		0.8/1.0 V	0.9~1.1 V	0.8 V	0.8 V	**0.9 V**
Precision	Input;Weight	1; 1	1; 1	1~8; 1~8	1or4or8; 1or4or8	**1or 8; 1or 8**
Computing Mode		Approximate	Accurate	Accurate	Accurate	**Accurate**
Num. of Activated Rows		64	8	8	11	**7**
Calibration Support		√	√	√	X	**√**
Computing Latency@1b		90ns	11.19ns	3.2ns	4.5ns	**4.77ns**
Throughput @1b		90.9 GOPS/bank	22.88 GOPS/bank	160 GOPS/bank	NA	**376 GOPS/bank**
[a]Area FoM@1b		4.4(x1)	56.82(x12.9)	30.58(x6.95)	NA	**93.52(x21.1)**
Energy Efficiency@1b		405 TOPS/W	15.47 TOPS/W	416.5 TOPS/W	92.9 TOPS/W@8b	**312.4 TOPS/W**

[a]Area FoM =Throughput/Area*(Num. of rows in an array)/ (Num. of activated rows) @ scaling to 28 nm

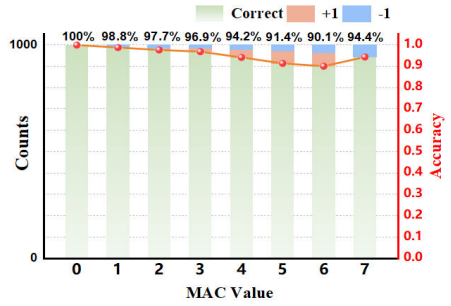


Fig. 12. The computing accuracy of FDAIMC macro under CMOS-MTJ hybrid Monte Carlo simulation at TT&0.9V. (1K points for each MAC value)

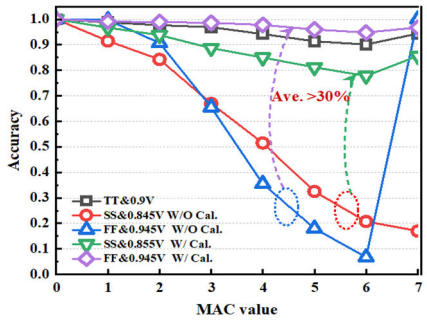


Fig. 13. The computing accuracy enhancement with calibration under process and voltage variation.

uniformity for every MAC value, which benefits from the 2-level-differential structure. Moreover, the variation of process and voltage can hardly affect this uniformity. It's worth noting that a smaller MOS can deteriorate the SM, which is consistent with our "V_{GS} loss with resistance increasement" theory. Figure 11(c) gives another important information that time interval only has a small change under process and voltage variation (vs. delay time of delay cell).

To verify the robustness of FDAIMC, a model of MTJ with 5% resistance variation [18]-[20] is used. A Monte Carlo simulation is employed to verify the computing accuracy. The mismatch of the MTJ and the CMOS circuit are active together in a local-mismatch mode. Each MAC case is divided into an MSB type and a LSB type. Each type has a 500-times simulation. As shown in Fig. 12, an over 90% accuracy under each MAC case is obtained. The average accuracy is 95.44%. Given the sparsity reported in [12], an over 94% accuracy for 0~4 MAC value is sufficient.

As described in Fig. 1(b) and Fig. 11(c), the variation of latency of the delay cell is more intensive than the variation of computing time under process and voltage variation. Thus, a dramatic accuracy decrease will occur, as shown in Fig. 13. With proper V_{PRE} trimming, the accuracy can be recovered effectively. The average accuracy is enhanced from 64.5% to 97.8% at FF&0.945V corner (with V_{PRE} = 462 mV)and from 58.1% to 88.6% at SS&0.855V corner (with V_{PRE} = 510 mV).

B. Area and Energy Efficiency

Thanks to the tight layout of the TDC system (2:1 MUX) and code sharing strategy, a high throughput of 376 GOPS can be obtained. Because the area efficiency is related to array size, parallelism degree of rows (whether accurate compute or not), and technology node. A FoM is used to evaluate the area efficiency with all these aspects in consideration. As shown in TABLE II, the proposed FDAIMC can offer a 1.64~21.18 times area FoM than [9][13][14]. Two modes of weight precision is supported, i.e., 1-bit and 8-bit. Multibit precision of input can be achieved by multicycle computing with a shift and accumulation circuit. The energy efficiency is 312.4 TOPs/W in 1b-1b-3b mode and 87.2 TOPs/W in 1b-8b-16b mode with about 85% sparsity.

V. CONCLUTION

In this article, an analog in-memory-computing macro FDAIMC that contains a calibration strategy for high-accuracy maintenance under process and power voltage variation is proposed to achieve an accurate multibit MAC operation with high performance. A fully differential concept is realized throughout the whole design in the cell array and ADC system. The 2-level-differential cell array solves the low-on/off-ratio obstacle during multibit MAC operation and enables a linear computing scheme with a calibration support in the analog domain. The ADC system, composed of SRIVTC and STTMC, is proposed to realize a high-speed and area/energy efficient analog-digital conversion while suppressing any time error induced by the converter. Finally, an average accuracy of 95.44% under TT&0.9V corner is obtained. By using the calibration strategy, the average accuracy of 97.8% and 88.6% are obtained under FF&0.945V and SS&0.855V separately, with over 30% enhancement. Furthermore, a 1.64~21.18 times area FoM is obtained, benefiting from the compact layout design. An energy efficiency of 312.4 TOPs/W is obtained in 1b-1b-3b mode and 87.2 TOPs/W in 1b-8b-16b mode.

REFERENCES

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

[2] Q. Dong, "A 351 TOPS/W and 372.4 GOPS compute-in-memory SRAM macro in 7 nm FinFET CMOS for machine-learning applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 242–243.

[3] J.-H. Yoon *et al.*, "A 40 nm 64 Kb 56.67 TOPS/W read-disturb-tolerant compute-in-memory/digital RRAM macro with active-feedback-based read and *in-situ* write verification," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2021, pp. 404–406.

[4] Y.-C. Chiu *et al.*, "A 22-nm 1-Mb 1024-b read data-protected STTMRAM macro with near-memory shift-and-rotate functionality and 42.6-GB/s read bandwidth for security-aware mobile device," *IEEE J. Solid-State Circuits*, vol. 57, no. 6, pp. 1936–1949, Jun. 2022.

[5] M. F. Chang, "Embedded 1 Mb ReRAM in 28 nm CMOS with 0.27V to 1V read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 332–334.

[6] R. Khaddam-Aljameh *et al.*, "HERMES core—A 14 nm CMOS and PCM-based in-memory compute core using an array of 300ps/LSB linearized CCO-based ADCs and local digital processing," in *Proc. Symp. VLSI Circuits*, Jun. 2021, pp. 1–2.

[7] C.-X. Xue *et al.*, "24.1 a 1Mb multibit ReRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN based AI edge processors," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 388–390.

[8] J.-M. Hung *et al.*, "An 8-Mb DC-current-free binary-to-8b precision ReRAM nonvolatile computing-in-memory macro using time-space readout with 1286.4–21.6 TOPS/W for edge-AI devices," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2022, pp. 1–3.

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

[10] Zizhao Ma, et al., "A 40nm 150 TOPS/W High Row-Parallel MRAM Compute-in-Memory Macro with Series 3T1MTJ Bitcell for MAC Operation," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May. 2023, pp. 1-5.

[11] Y. Luo, et al., "A Variation Robust Inference Engine Based on STT-MRAM with Parallel Read-Out," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May. 2019, pp. 1-5.

[12] J. Wang, et al., "RSACIM: Resistance Summation Analog Computing in Memory With AccuracyOptimization Scheme Based on MRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 71, no. 3, pp. 1014–1024, March 2024.

[13] S. D. Spetalnick, et al., "A 2.38 MCells/mm2 9.8 -350 TOPS/W RRAM Compute-in-Memory Macro in 40nm CMOS with Hybrid Offset/I_{OFF} Cancellation and $I_{CELL}R_{BLSL}$ Drop Mitigation," in *Proc. Symp. VLSI Circuits*, Jun. 2023, pp. 1–2.

[14] Je-Min Hung et al., "8-b Precision 8-Mb ReRAM Compute-in-Memory Macro Using Direct-Current-Free Time-Domain Readout Scheme for AI Edge Devices," *IEEE J. Solid-State Circuits*, vol. 58, no. 1, pp. 303–314, Jan. 2023.

[15] M. Zhang, et al., "A 0.6-V 13-bit 20-MS/s Two-Step TDC-Assisted SAR ADC With PVT Tracking and Speed-Enhanced Techniques," *IEEE J. Solid-State Circuits*, vol. 54, no. 12, pp. 3396–3409, Dec. 2019.

[16] L. Brooks, Hae-Seung Lee,"A 12b 50MS/s Fully Differential Zero-Crossing-Based ADC Without CMFB," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 166–167.

[17] Y. Zhang et al., "Compact Model of Subvolume MTJ and Its Design Application at Nanoscale Technology Nodes," *IEEE Transactions on Electron Device*, 2015, vol. 62, no. 6, pp. 2048-2055, June 2015.

[18] 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

[19] Q.-K. Trinh, et al., "Time-based sensing for reference-less and robust read in STT-MRAM memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 10, pp. 3338–3348, Oct. 2018.

[20] A. Yang, et al., "Double-ended superposition anti-noise resistance monitoring write termination scheme for reliable write operation in STT-MRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 3, pp. 1147–1160, Mar. 2023