CCF Transactions on High Performance Computing (2020) 2:282–296
https://doi.org/10.1007/s42514-020-00045-6

REGULAR PAPER

Interplay Bitwise Operation in Emerging MRAM for Efficient In-memory Computing

Hao Cai[1] · Honglan Jiang[2] · Yongliang Zhou[1] · Menglin Han[1] · Bo Liu[1]

Received: 17 February 2020 / Accepted: 22 July 2020 / Published online: 11 August 2020

Abstract

In order to realize high efficient magnetization switching in magnetic tunnel junction (MTJ), several potential mechanisms have been realized as the interplay effect to MTJ device, such as the interaction between spin orbit torque-spin transfer torque (STT) and voltage-controlled magnetic anisotropy (VCMA)-STT. The interplay mechanisms have been experimentally explored with improved switching energy efficiency comparing with traditional STT method. Considering the requirement of mixed-precision memory, we propose a novel write-only in-memory computing paradigm based on interplay bitwise operation in two terminal or three terminal MRAM bit-cell, which aims to reduce the layout overhead of peripheral computing circuits, as well as to eliminate read decision failure in the procedure of in-memory computing. Specifically, the proposed write-only bitwise in-memory computing is demonstrated with OR, AND, XOR, full adder operations. Four nonvolatile approximate full adders (AxFAs) are proposed and implemented in different MRAM bit-cells. The AxFAs can be easily reconfigured into memory units with simple connections. Image processing applications are used to demonstrate the in-memory computing, include FA, XOR operation. Comparing with traditional sensing based approach, more than 80% energy reduction is obtained using the proposed interplay writing-only in memory computing with approximation setup. A 61.4% energy reduction is achieved using VCMA mechanism interaction based XOR functions.

Keywords MTJ interplay writing · Mixed-precision memory · In-memory computing · Image processing

1 Introduction

As a promising candidate to replace traditional memories, the wide range application of spin transfer torque magnetic random access memory (STT-MRAM) is delayed due to its intrinsic limitations (Kanai et al. 2012; Wang et al. 2012; Maruyama et al. 2009; Wang et al. 2018a). Alternatively, magnetic tunnel junction (MTJ) with voltage-controlled magnetic anisotropy (VCMA) effect provides flipping of the magnetization upon a voltage pulse, irrespective of the initial state. Thus, this magnetoelectric random access memory achieves less energy consumption and higher density, as well as the improved switching latency thanks to the very little charge flow required to operate (Wang et al. 2012; Maruyama et al. 2009). Another switching method referred to as the spin orbit torque (SOT) has been well-developed recently with fast magnetization switching. The emerging SOT-STT interplay operations not only result in high density and switching efficiency, also overcomes the asymmetric switching and source degeneration (Wang et al. 2018a, b, 2019).

The mixed precision in-memory computing concept was firstly implemented based on the phase charge memory (Le Gallo et al. 2018), which reveals the significance of the nonvolatile approximate in-memory computing procedure. The data-dependent accuracy and approximate computing scheme is friendly to constraint-based collaborative design. Nonvolatile memory based process-in-memory (PIM) or computing-in-memory (CIM) schemes have been proposed to enhance bandwidth, massive parallelism and energy efficiency (Baek et al. 2018; Natsui et al. 2015; Li et al. 2016;

This work is funded with National Key R&D Program of China under Grant 2018YFB2202800 and National Natural Science Foundation of China under Grant 61904028.

✉ Hao Cai
hao.cai@seu.edu.cn

[1] National ASIC System Engineering Center, School of Electronic Science and Engineering, Southeast University, Nanjing 210096, China

[2] Institute of Microelectronics, Tsinghua University, Beijing 100084, China

Jain et al. 2018). Unfortunately, these techniques require the design of additional peripheral circuits, which are sensitive to variability issues from both nonvolatile devices and CMOS transistors.

Approximate approaches have been investigated at either data computing or storage phases (Mittal 2016; Ranjan et al. 2015; Sampson et al. 2013). The energy-accuracy tradeoff such as dynamic precision scaling (Yesil et al. 2018), error-energy adjustment (Monazzah et al. 2017; Frustaci et al. 2016; Ranjan et al. 2017; Zeinali et al. 2018) have been proposed for error-tolerant circuits and systems. Differ from previous the sensing and reference generation circuitry, such as Pinatubo (Li et al. 2016) and STT-CiM (Jain et al. 2018), this work is to use write operation as in-memory computing, so that data in MTJ is changed according to the computing data in the bitline.

It has been reported that MRAM writing efficiency can be significantly improved using VCMA and STT interaction in two-terminal MTJ devices (Kanai et al. 2014), as well as SOT and STT interaction, which have been experimentally demonstrated as the field-free switching in three-terminal perpendicular MTJs (Wang et al. 2018a, b, 2019). Considering the requirement of mixed-precision memory, we propose a novel write-only in-memory computing paradigm based on interplay bitwise operation in two terminal or three terminal MRAM bit-cell, which aims to reduce the layout overhead of peripheral computing circuits, as well as to eliminate read decision failure in the procedure of in-memory computing. Specifically, the proposed write-only bitwise in-memory computing is designed for OR, AND, XOR, full adder operations, and demonstrated with simulation results. Unlike traditional PIM methods, the proposed four NV-FAs utilize only the MTJ writing operation to complete the process-in-memory. In summary, we make the following key contributions:

- Novel writing-only in-MRAM bitwise processing schemes are proposed based on different switching mechanisms of MTJ device.
- In-memory bitwise logic operations (AND, OR, XOR, full adder) are realized using several disruptive implementations, including peripheral circuit design, interplay switching of MTJ and approximate computing.
- ***P***rocess-in-memory with ***J***oint magnetization switching for ***A***ppro***x***imate computing in MTJ, named $Pj-$ **AxMTJ** scheme is proposed for image sharpening application. Precessional VCMA-MTJ switching for XOR operation is used for image similarity examination operator.

The rest of this article is organized as follows. Section 2 discusses the mixed-precision and approximate memory. Section 3 proposes the interplay MTJ switching scheme for in-MRAM computing. Section 4 specifically evaluates the performance of in-MRAM approximate FA (AxFAs). Section 5 demonstrates the simulation results for the circuit metrics and applied the interplay bitwise operations to image processing applications. We provide our conclusion in Sect. 6.

2 Approximation in emerging memories

2.1 Mixed-precision memory

The spintronic devices offer memory elements as the capability of efficient nondestructive writing and sensing, logic compatible operation voltage, scaling to nanometer dimensions, as well as high density and endurance. However, due to the imperfection of MRAM manufacturing process, the global variations, local mismatch and reliability concerns seriously impact its performance (Ranjan et al. 2015; Wang et al. 2016). High reliability MRAM units rely on additional circuits, such as error-coding correction, redundancy circuits and writing booster to overcome above imperfections. Thus, the MRAM design-for-approximation has been studied in different memory hierarchies (Ranjan et al. 2017, 2015; Zhao et al. 2017; Zeinali et al. 2018).

Mixed-precision memory has been proposed in certain error tolerant applications to achieve overall high layout area and energy efficiency (Le Gallo et al. 2018). Figure 1 illustrates the block diagram of a MRAM based hybrid precision memory system. Accurate logic computing is performed with the classic Von-Neumann architecture, whereas low-precision computational memory can be configured with emerging spintronic devices based in-memory computing array. In this work, SOT, VCMA and STT switching mechanisms are jointly applied to the low-precision memory unit. Typical interplay mechanisms are SOT-STT, and VCMA-STT interactions.

2.2 Energy efficient in-MRAM/near-MRAM computing

In applications such as static/dynamic image compression, detection, neural networks and energy efficient computing, approximate memories enable lower-precision data storage for design trade-off of performance parameters (Sampson et al. 2013; Yamaga et al. 2018; Frustaci et al. 2016; Zeinali et al. 2018; Frustaci et al. 2016). Approximation MRAM techniques have been hierarchically proposed at different abstract levels for typical nonvolatile memories, through read disturbs, incomplete writes (over-scaled timing/voltage conditions) and re-configurable logic complexity (functionally approximate methods and circuits) (Teimoori et al. 2018; Ranjan et al. 2015; Cai et al. 2017). The writing energy efficiency can be largely improved at the cost of small

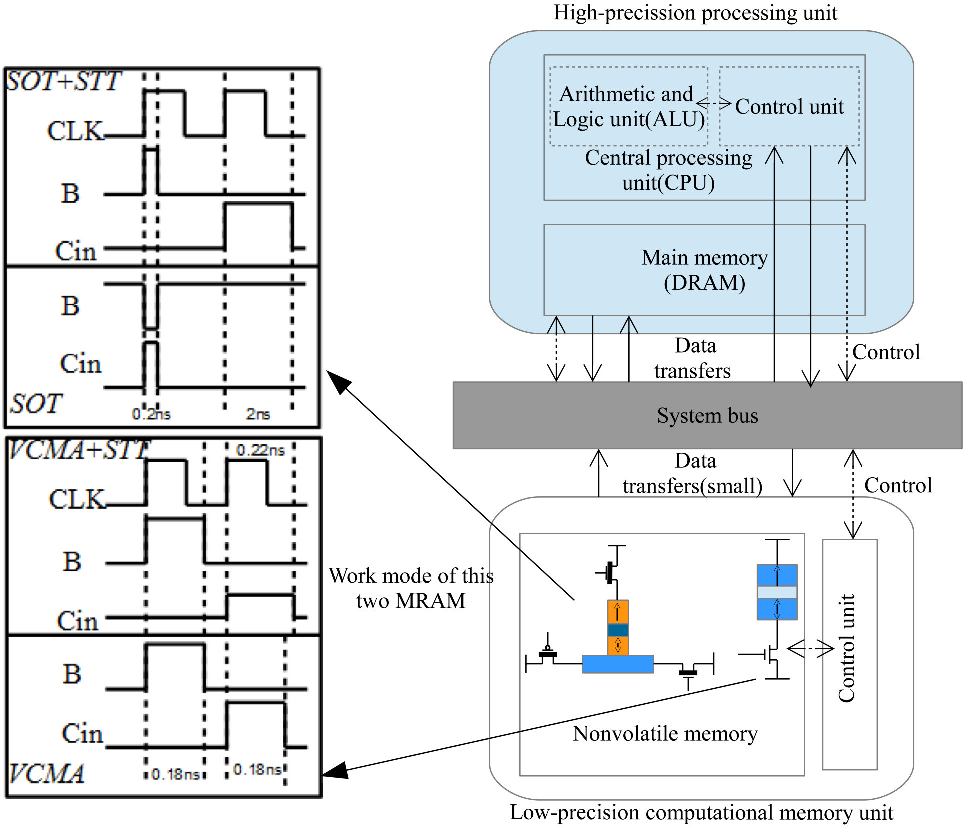


Fig. 1 Hybrid precision in memory processing and computing. Approximate computing can be configured for low-precision computational memory to achieve power–performance–accuracy trade-off

probabilities of sensing or writing errors. As MRAM suffers the PVT variations from both MTJ and CMOS transistors, the approximate memory design scheme can alleviate MRAM design constraints from three aspects:

- The TMR of MTJ is highly sensitive to temperature and process, the variation of TMR could be larger than 80%.
- The minimum writing voltage can be less than nominal supply voltage.
- Traditional error-coding correction and redundancy circuits could be the optional blocks since approximate MRAM is designed with error tolerance.

The approximate in-MRAM or near-MRAM computing schemes shows enhanced energy efficiency in fault tolerance applications (Cai et al. 2017; Oboril et al. 2016; Locatelli et al. 2018; Li et al. 2016; Jain et al. 2018). However, major problems for the existing designs are large layout area and lack of flexibility and variability.

MRAM writing efficiency can be significantly improved using VCMA and STT interaction in two-terminal MTJ devices (Kanai et al. 2014), as well as SOT and STT interaction, which have been experimentally demonstrated as the field-free switching in three-terminal perpendicular MTJs(Wang et al. 2018a, b, 2019). Therefore, simple 1T-1M and 3T-1M MRAM bit-cells are utilized to design approximate in-MRAM computing with small layout areas. As the schemes are based on the MTJ writing operation, which is implemented by one memory bit-cell without peripheral circuits. Thus, the in-MRAM operation can be reconfigured into a memory unit by simple wire connections.

3 MTJ interplay switching for in-memory computing

To assess the performance of Pj-AxMTJ as a new PIM platform, a comprehensive device-to-architecture evaluation framework along with two in-house simulators are developed. First, at the device level, we jointly use the University Paris Sud with spin Hall effect equations to model SOT-MRAM bit-cell (Spinmodel Library 2015). For the circuit level simulation, a VerilogA model of 3T1R SOT-MRAM device is developed to co-simulate with the interface CMOS circuits in Cadence Spectre and SPICE. TSMC 28 nm Product Development Kit (PDK) library is used in SPICE to verify the proposed design and acquire the performance. Second, an architectural-level simulator is built based on NVSim (Dong et al. 2014). Based on the device/circuit level results, our simulator can alter the configuration files corresponding to different array organization and report performance metrics for PIM operations. The controllers and add-on circuits are synthesized by Design Compiler ICC2 with an industry library. Third, a behavioral-level simulator is developed in Matlab calculating the latency and energy that GraphS spends on different graph processing tasks.

3.1 Behavioral of MTJ interplay switching

The proposed MTJ interplay switching based in-memory logic operations are based on the two typical joint switching structures shown in Fig. 2. The precessional VCMA and VCMA-STT interaction (Fig. 2a) are implemented with a 1T-1M bit-cell structure, which is similar with STT-MRAM. The SOT-STT interaction is realized with the three terminal device, which is consisted of a heavy metal strip with an MTJ located upon it. Table 1 lists the key metrics of SOT-MRAM and VCMA-MRAM used in MTJ behavioral modeling and bit-cell level simulations for in memory computing.

The working principle of the SOT erasing plus STT mechanism is that SOT writes ‘1’ (erase) and STT writes ‘0’ (program). Based on the setup of virtual node ‘*State*’ (for monitoring the state of MTJ), the output must be one of the two discrete voltage-levels: level ‘1’ (logic ‘0’) indicates the parallel state; level ‘-1’ (logic ‘1’) indicates the anti-parallel state. The state changes of SOT-STT-MTJ (using in Ax1) are shown in Fig. 3a, and the state other than ‘1’ and ‘-1’ is unstable.

Figure 3b shows the switching behavior of the joint MTJ driven by VCMA and STT mechanisms (using in Ax3). As can been seen, the 1st and the 2nd switches are affected by VCMA and STT, respectively. STT-assisted precessional VCMA-MTJ has two states. On the control of 1.2 V, the device is dominated by VCMA mechanism. When V_{MTJ} is setup to 0.6 V, the device works with the STT assisted mode. The output of the virtual node ‘*State*’ must be one of the two discrete voltage-levels: level ‘0’ indicates the parallel state; level ‘1’ indicates the anti-parallel state. Figure 3b shows the state changes for the STT-assisted precessional VCMA-MTJ at all eight different input combinations. Note that the state other than ‘0’ and ‘1’ is unstable.

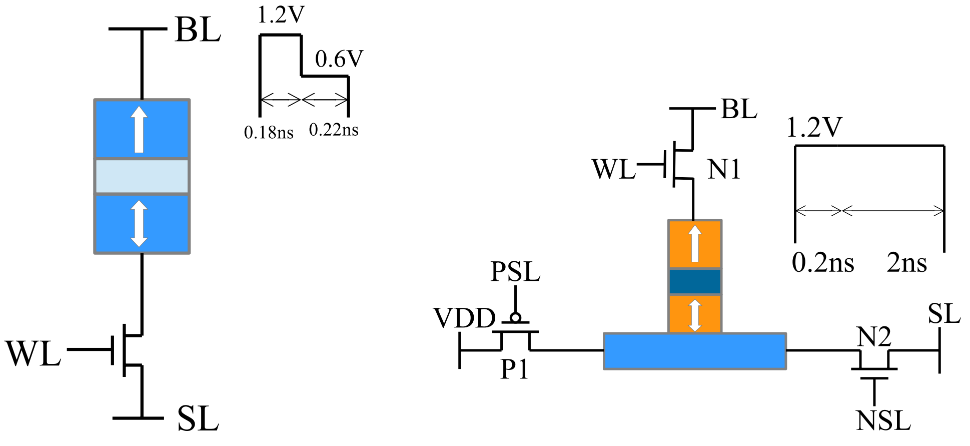
Table 1 Important parameters of SOT-MTJ and VC-MTJ model

	SOT-MTJ	VC-MTJ
MTJ Diameter	50 nm	50 nm
Free layer thickness	0.7 nm	1.1 nm
Oxide layer thickness	1.2 nm	1.4 nm
TMR ratio	120%	100%
Heavy metal dimension	50 * 60 * 3 nm^3	N/A
Gilbert Damping Factor	0.007	0.005
Spin hall angle	0.3	N/A
Structure	2T1M	1T1M
Access transistor	200 nm/30 nm	80 nm/30 nm
Supply voltage	1.2 V	1.2 V

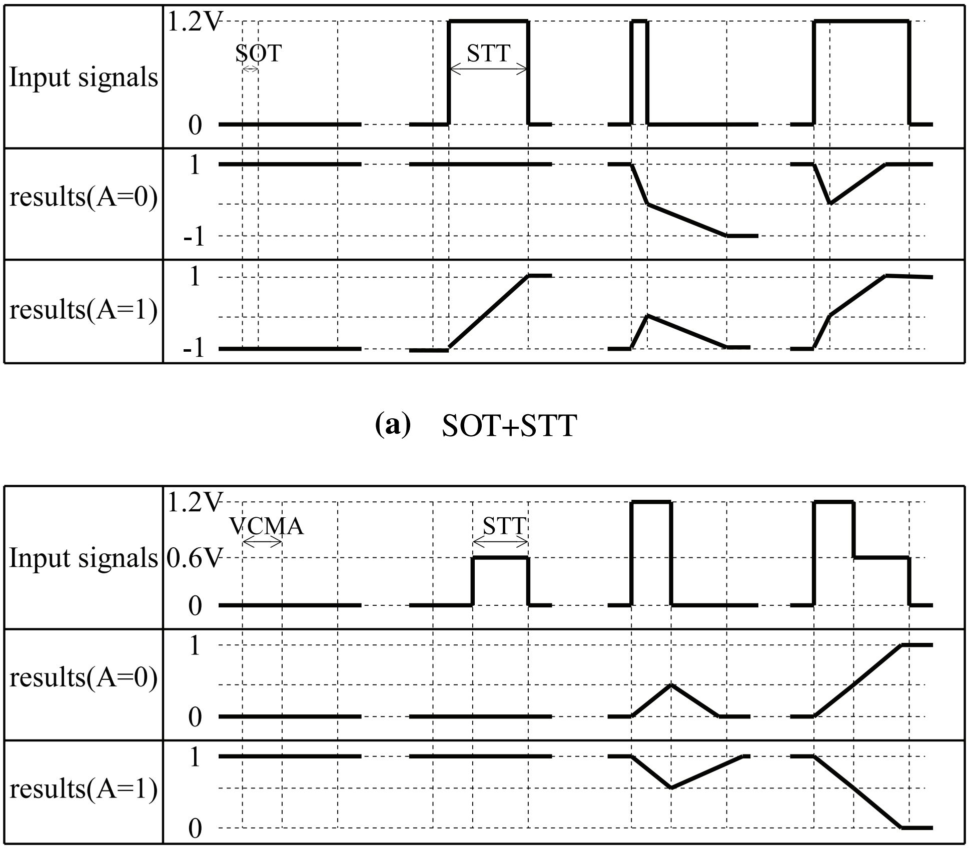
3.2 In-memory OR/AND operations

VCMA-MTJ for processing-in-memory was experimentally demonstrated in Wang et al. (2018), with three different Boolean logic operation: OR, AND, XNOR. Zhang et al. (2017) implemented stateful reconfigurable booeaen logic via a single VCMA-spin hall effect on three-terminal MTJ. In this work, the interplay of SOT and STT switching with two different voltages can achieve AND and OR Boolean logic function. Under normal voltage, the original data in MTJ is ‘0’. When input signals A and B are applied to the N1 NMOS and the N2 NMOS respectively, the data inMTJ becomes the result of the AND operation between A and B. The OR operation between A and B can be achieved when the supply voltage is up to a high level. Under this voltage, SOT and STT can accomplish write ‘1’ operation separately.

Fig. 2 Two typical interplay switching schemes for MTJ as writing/data storage process: **a** STT-assisted precessional VCMA, **b** SOT-STT interactions



(a) VCMA+STT **(b)** SOT+STT



(a) SOT+STT

(b) VCMA+STT

Fig. 3 The behavioral of MTJ interplay switching mechanisms, described with input signal in each step. **a** SOT–STT, **b** STT-assisted precessional VCMA

The energy-delay performance of interplay MTJ switching will be analyzed based on full adder operations.

3.3 In-memory approximate full adders (AxFAs)

3.3.1 Approximate adder and its design metrics

Conventional approximate adders based on CMOS technology are designed with reduced critical path or logic complexity. Approximate adders based on nonvolatile memory have been occurred by manipulating the data processing or storage phases. Several metrics can characterize the performance of an approximate adder (Liang et al. 2013). An important metric denoted as the error distance (ED) has been introduced to evaluate the accuracy of an approximate arithmetic design besides the power and latency performance. In an n-bit approximate adder, the inexact output S' and accurate output S is compared arithmetically, i.e., $ED(S', S) = |S' - S|$. The normalized mean error distance (NMED) that is the normalization of mean error distance (MED) by the maximum output of the accurate adder, is another error metric to evaluate the accuracy of an approximate adder. Also, the error rate (ER) is commonly used to evaluate the probability that an error occurs in an approximate design. The maximum error distance (ED_{max}) shows the maximal error magnitude that is possibly produced by an approximate arithmetic circuit. In this paper, ED_{max} is defined as the maximal ED normalized by the maximum output of the accurate design.

3.4 Proposed interplay switching for in-memory AxFAs

Controlled by the voltage pulse, an VC-MTJ can be at either parallel (logic ‘0’) or anti-parallel (logic ‘1’) states. Its writing operation is completed by switching the state of an MTJ. For the structure in Fig. 2a, the MTJ is switched by turning on the NMOS transistor (setting WL = ‘1’), and adding a specific voltage pulse between BL and SL. For example, a 1.2 voltage pulse with a duration of 0.18 ns followed by a 0.6 V voltage pulse with a duration of 0.22 ns results in an MTJ switching; this is referred to as the STT-assisted precessional VCMA. On the other hand, a 0.55 ns voltage pulse of 1.1 V can also switch the MTJ, which is denoted as the precessional VCMA (P-VCMA). In Fig. 2b, an erasing operation is performed before each writing by turning on P1 and N2 and turning off N1, and the MTJ is at anti-parallel state. The MTJ switches to parallel state by a positive voltage pulse between VDD and BL to complete the writing of '0', when P1 and N1 are on, N2 is off. Otherwise, the MTJ stays at anti-parallel state.

Unlike traditional In-MRAM computing methods: first sensing then processing or pass transistor logic based logic-in-memory, the proposed design relies on MTJ writing to complete the FA operations. Specifically, the input A of a FA is initially stored in an MTJ. The input B and C_{in} are consecutively fed into WL of a 1T-1M bit-cell shown in Fig. 2a, whereas B is connected to NSL (or PSL) and C_{in} is connected to WL (or NSL) when using the structure in Fig. 2b. By controlling the voltage on the MTJ, the 1-bit addition is approximately implemented with state switches of the MTJ. Finally, four AxFAs are obtained by applying different switching mechanisms, which are referred to as Ax1 (SOT+STT), Ax2 (SOT), Ax3 (VCMA+STT) and Ax4 (P-VCMA). The truth table of the proposed AxFAs is listed as Table 2. Ax1 and Ax2 are implemented by the structure of Fig. 2b, and Ax3 and Ax4 are implemented by the bit-cell in Fig. 2a.

- **Ax1** In this design, the inputs B and C_{in} are connected to NSL and WL (Fig. 2b), respectively. In step one, P1 is on and N1 is off, and B controls SOT current, i.e., the data in MTJ is '1' when $B = \}1'$, otherwise it is A. In step two, P1 is on and N2 is off, and C_{in} controls STT current flow, i.e., the data stored in MTJ is ultimately '0' if C_{in} is '1'.
- **Ax2** Unlike Ax1, the inputs B and C_{in} in Ax2 together control the SOT current and hence, only one step is required to obtain the sum result. For Ax2, B is connected to PSL, and C_{in} is fed into NSL. To use SOT mechanism, WL is always off. The data in MTJ is erased to '1' when B='0' and C_{in}='1', otherwise it stays as A.
- **Ax3** In Ax3, the inputs B and C_{in} are consecutively input to WL (Fig. 2a). In this case, B and C_{in} control VCMA and STT effects, respectively. Thus, each high input signal generates a voltage pulse. The data stored in MTJ will be flipped due to the VCMA-STT switching when B='1' and C_{in}='1', otherwise it does not change.
- **Ax4** As discussed in Sect. 2, P-VCMA without the assistance of STT also results in a state switching of MTJ driven by a different voltage pulse. Thus, Ax4 uses two P-VCMA effects; the inputs B and C_{in} are consecutively fed into WL. Ax4 achieves the same function as Ax3.

As shown in Table 2, the signal stored in the MTJ after the required operations for each design is output as the *Sum*. The carry-out of the four designs C_o is taken from the input signal B to avoid the carry propagation delay when using the AxFAs in a multi-bit adder. The "X" following each digit indicates an incorrect result. E_{Axi} shows the errors for Axi considering both Sum_{Axi} and C_o. Table 2 shows that the probability of generating an error C_o is pretty low (25%). Also, the C_o results in a very low average error because both positive and negative errors can be produced.

4 Simulation results

In this paper, the analysis is executed in Cadence Virtuoso with a VCMA-MTJ and a SOT-MTJ compact model, as well as a 28 nm CMOS technology (Amiri et al. 2015; Wang et al. 2018b). As discussed in Sect. 2, the MTJ depends on the voltage/bias condition of the different bit-cell structures for MRAM. Table 3 lists the setup of the proposed four AxFAs.

4.1 AxFAs energy-delay performance

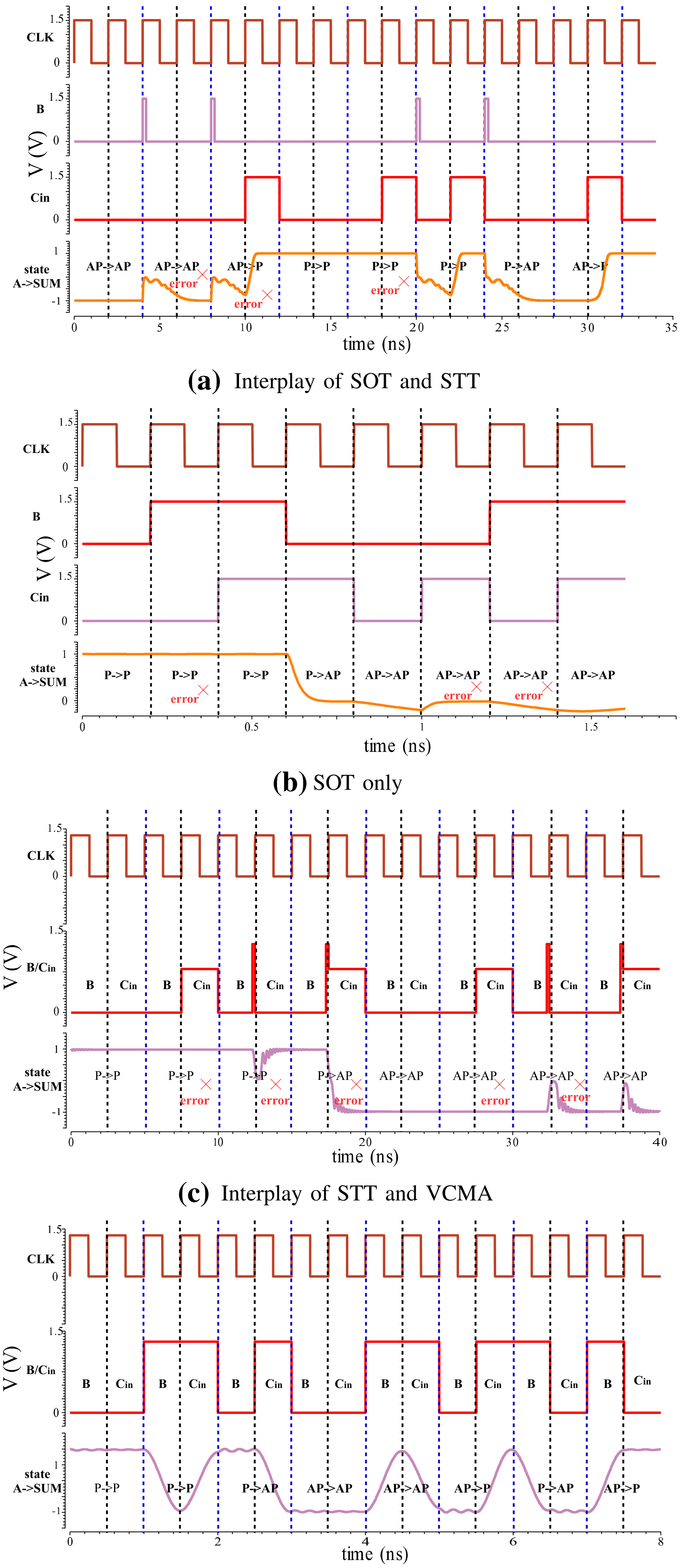
Figure 4 shows the simulation waveform of full adder Ax1 to Ax4. The error is highlighted according to the truth table. Table 4 shows the power consumption and latency of Ax1 and Ax2. Although Ax2 consumes a larger power than Ax1 when the state of MTJ switches, Ax2 has more input cases that no power is consumed. The average power of Ax1 is 186.7 µW, while it is 391.8 µW for Ax2. The maximum delay for both Ax1 and Ax2 is 2.56 ns.

Table 2 The truth table of the proposed joint switching approximate FA

FA inputs			SOT+STT		SOT		VCMA+STT		P-VCMA		C_o
A(InMTJ)	B	C_{in}	Sum_{Ax1}	E_{Ax1}	Sum_{Ax2}	E_{Ax2}	Sum_{Ax3}	E_{Ax3}	Sum_{Ax4}	E_{Ax4}	
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0 (X)	−1	1	0	0 (X)	−1	1	0	0
0	1	0	1	+2	0 (X)	+1	0 (X)	+1	1	+2	1 (X)
0	1	1	0	0	0	0	1 (X)	−1	0	0	1
1	0	0	1	0	1	0	1	0	1	0	0
1	0	1	0	−2	1 (X)	−1	1 (X)	−1	0	−2	0 (X)
1	1	0	1 (X)	+1	1 (X)	+1	1 (X)	−1	0	0	1
1	1	1	0 (X)	−1	1	0	1	0	1	0	1

The "X" indicates an incorrect output

Fig. 4 Simulation waveform of different proposed approximate full adder, the SUM operations **a** interplay of SOT and STT, **b** SOT, **c** interplay of STT and VCMA, **d** precessional VCMA



(a) Interplay of SOT and STT

(b) SOT only

(c) Interplay of STT and VCMA

(d) Precessional VCMA

Table 3 Switching setup of the proposed four approximate FAs

AxFA	1st step			2nd step		
	Mechanism	Pulse duration (ns)	V_{dd} (V)	Mechanism	Pulse duration (ns)	V_{dd} (V)
Ax1	SOT	0.2	1.2	STT	2	1.2
Ax2	SOT	0.2	1.2	No operation		
Ax3	VCMA	0.18	1.2	STT	0.22	0.6
Ax4	VCMA	0.55	1.1	VCMA	0.55	1.1

Table 4 Power-delay performance of SOT-STT and SOT based AxFAs

Full adder Input			SOT+STT (Ax1)		SOT (Ax2)	
A	B	C_{in}	Power (μW)	Delay (ns)	Power (μW)	Delay (ns)
0	0	0	0	2.2	0	0.2
0	0	1	238	2.2	1567	2.56
0	1	0	143.5	2.56	0	0.2
0	1	1	379.6	2.2	0	0.2
1	0	0	0	2.2	0	0.2
1	0	1	209.3	2.44	1567	2.56
1	1	0	143.5	2.56	0	0.2
1	1	1	379.6	2.2	0	0.2

Table 5 Power-delay performance of STT-Assisted VCMA and P-VCMA based AxFAs

Full adder input			VCMA+STT (Ax3)		P-VCMA (Ax4)	
A	B	C_{in}	Power (μW)	Delay (ns)	Power (μW)	Delay (ns)
0	0	0	0	1.1	0	0.5
0	0	1	5.63	1.15	0.086	0.5
0	1	0	5.575	1.1	4.931	1.15
0	1	1	10.98	1.4	6.719	1.19
1	0	0	0	1.1	0	0.5
1	0	1	5.458	1.15	0.087	0.5
1	1	0	5.525	1.1	4.378	1.18
1	1	1	11.21	1.4	6.094	1.09

The power and delay results for the eight different inputs of Ax3 and Ax4 are shown in Table 5. Compared with Ax3, Ax4 is more power-efficient with a shorter maximum latency. The average power of Ax3 (5.547 μW) is roughly the twice of that of Ax4. The maximum delay for Ax3 and Ax4 are 1.40 ns and 1.19 ns, respectively. Tables 4 and 5 illustrate that the VCMA-based AxFAs have smaller power dissipation and maximum delay than the SOT-based designs.

Comparing with traditional sensing based approach (sensing the data from MTJ, then calculate with 28-transistor CMOS-FA), more than 80% energy reduction is obtained using the proposed writing-only in memory computing (see Table 6).

Table 6 1-Bit energy comparison of proposed writing based and sensing based schemes

Energy consumption	Ax1	Ax2	Ax3	Ax4
Sensing based (fJ)	900.28	900.28	93.68 fJ	94.61
Writing-only (fJ)	742.89	210.41	16.25 fJ	21.27

4.2 Accuracy tradeoff

Assuming the inputs are equally likely to be ‘0’ and ‘1’, the ER, NMED, MRED and ED_{max} are calculated as shown in Table 8. Among the approximate designs for 1-bit addition,

Table 7 The MED and NMED metrics of the AxFAs

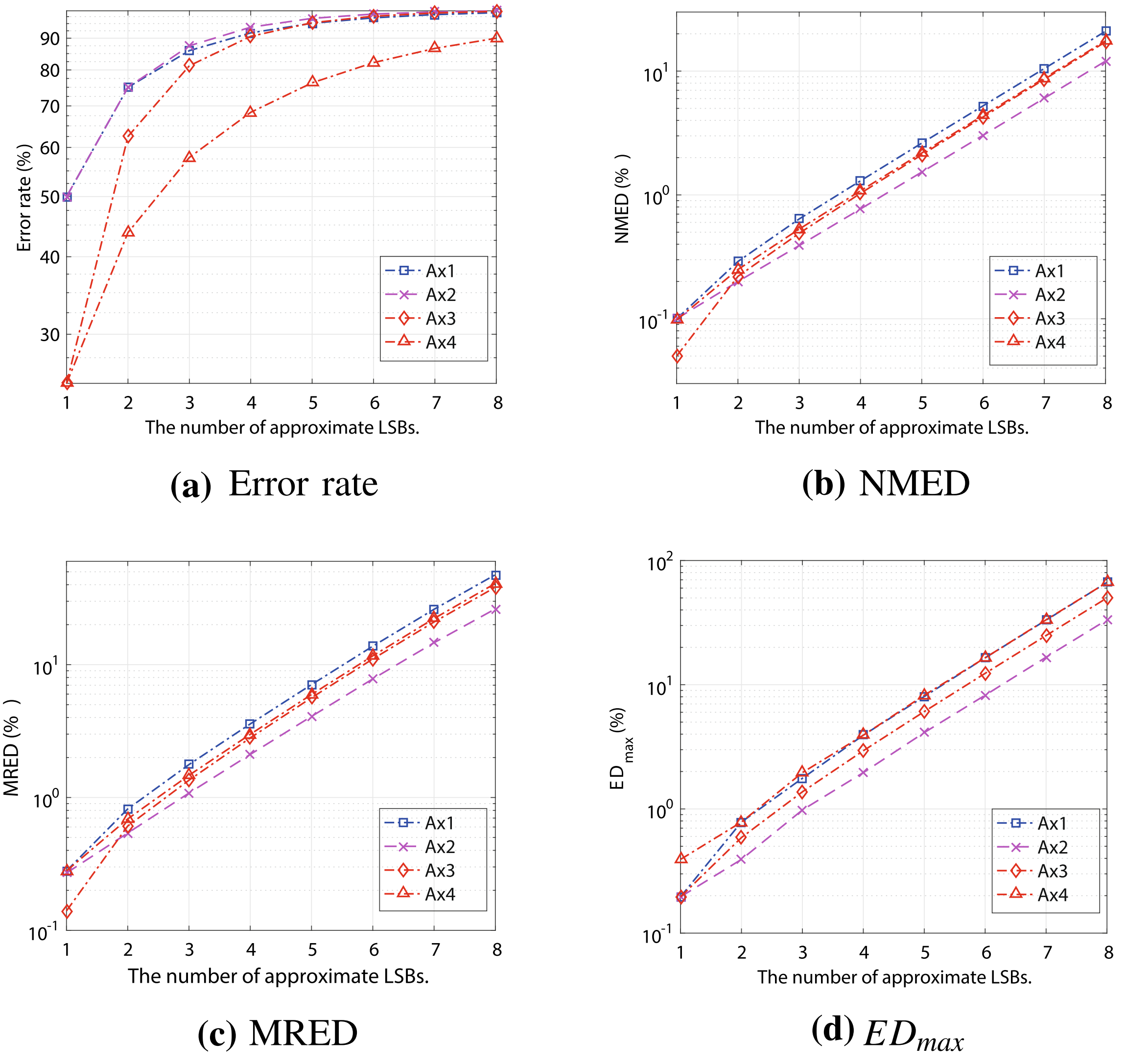
Error metric	Ax1	Ax2	Ax3	Ax4
MED	0.875	0.375	0.625	0.500
NMED	0.292	0.125	0.208	0.167

Ax2 is the most accurate in terms of NMED and MRED, whereas Ax1 has the poorest accuracy with the largest NMED and MRED. Ax4 has the lowest ERs but the largest maximum error distances (Table 7).

The proposed approximate PIM scheme can be further applied to multi-bit adders. To maintain a high-accuracy, k LSBs of an n-bit adder are usually approximated ($k < n$), while the more significant bits are accurately computed. The k LSBs can be processed by cascading k AxFAs. To further assess the accuracy of the designs, the functions of 8-bit and 16-bit approximate adders consisting of AxFAs are implemented in MATLAB. An exhaustive unsigned input combinations are used as the inputs for 8-bit adders. Monte-Carlo simulations are performed for 16-bit adders, where the inputs are 10 million input combinations in the range of [0, 65535] from an uniform distribution.

Figures 5 and 6 show the error characteristics of the 8-bit and 16-bit adders (using different approximate designs), respectively. These figures demonstrate that the comparison results with respect to ER, NMED, MRED and ED_{max} for the 8-bit and 16-bit approximate adders are the same. Figures 5a and 6a indicate that the adders based on Ax2 have the largest ERs Ax1 and Ax2 have the largest ERs, while the ones implemented by Ax4 shows the smallest ERs.

In terms of NMED and MRED, the adder using Ax2 results in the smallest values; the one based on Ax2 have similar small results when the number of approximate



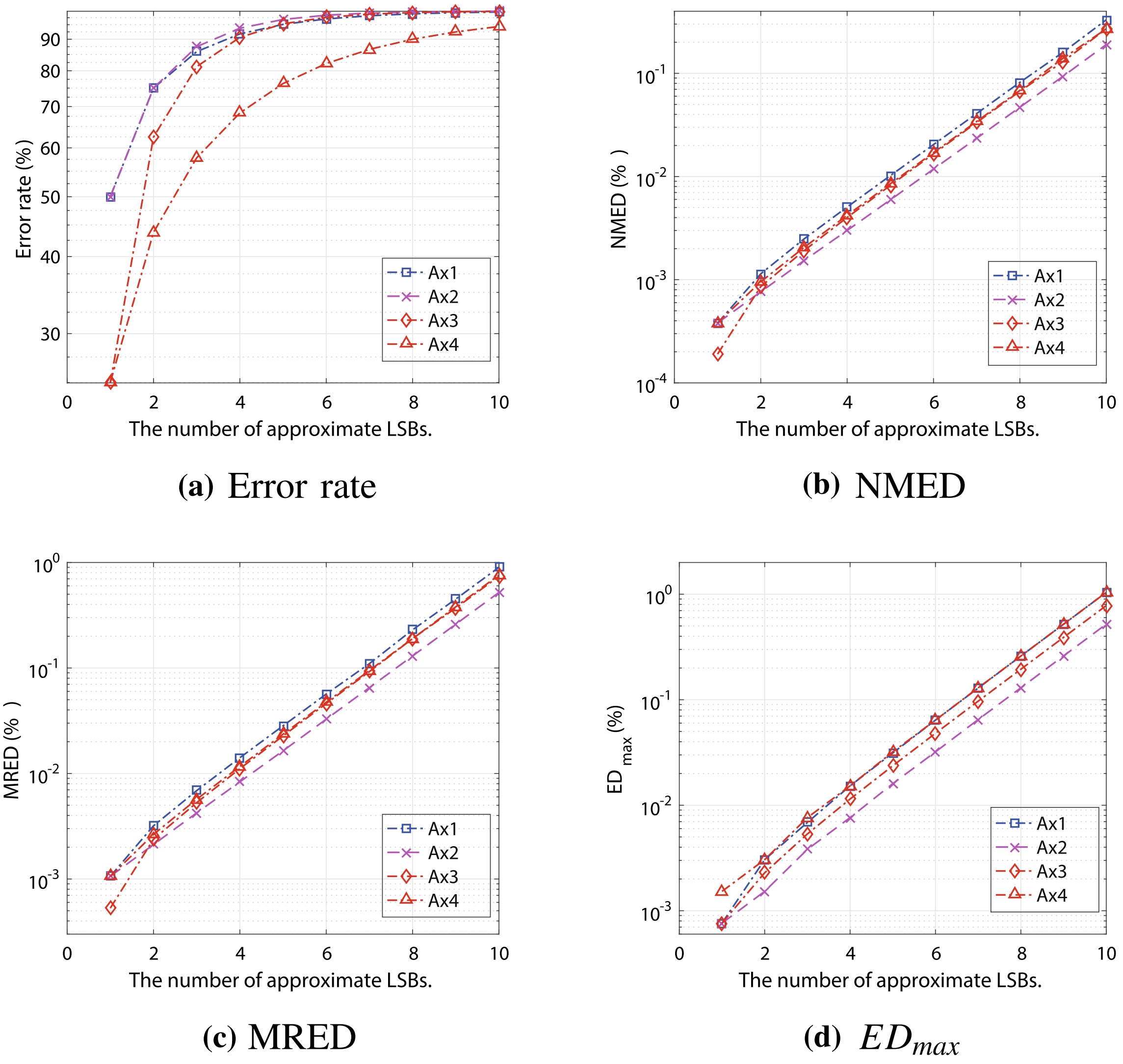
(a) Error rate

(b) NMED

(c) MRED

(d) ED_{max}

Fig. 5 The error characteristics of 8-bit adders with different number of approximate LSBs implemented by different approximate designs



(a) Error rate

(b) NMED

(c) MRED

(d) ED_{max}

Fig. 6 The error characteristics of 16-bit adders with different number of approximate LSBs implemented by different approximate designs

LSBs is larger than 4. The adders using Ax1 have the largest NMEDs and MREDs. These comparison results are slightly different from the ones obtained from Table 8 due to the dependencies between adjacent AxFAs in a multibit adder. Being consistent with Table 8, the adders using Ax1 and Ax4 have the largest ED_{max}s. The adders based on Ax3 show medium values in all the four error metrics.

5 Case study of MTJ interplay switching for image processing applications

5.1 Image sharpening

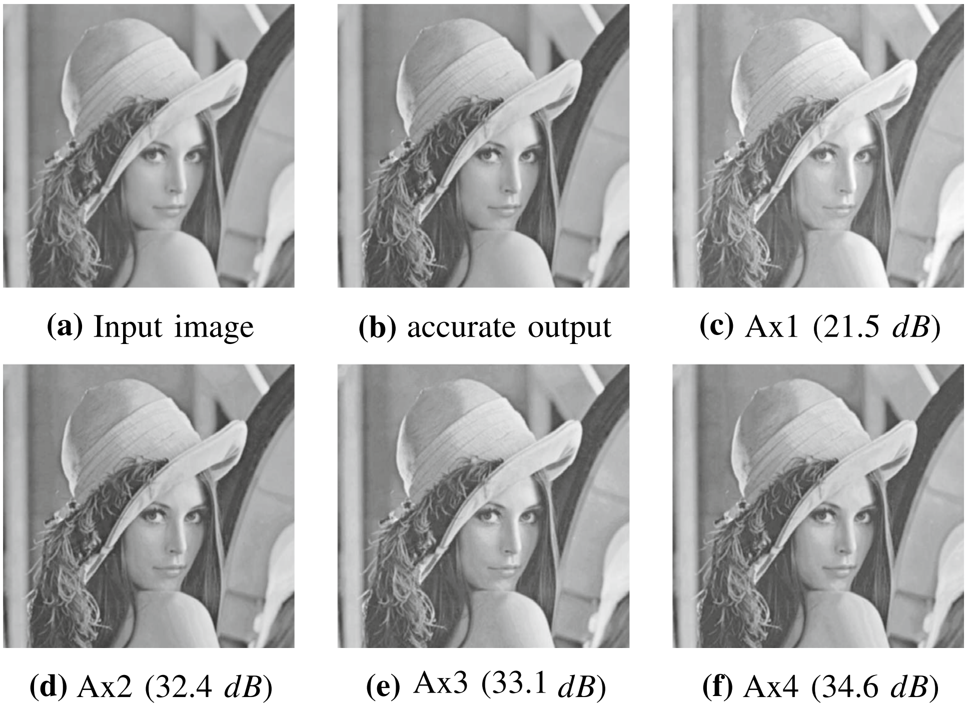
To assess the availability of the proposed approximate designs, the proposed AxFAs are evaluated in the image sharpening application. Figure 7 shows the image sharpening results by 16-bit approximate adders implemented by AxFAs with nine approximate LSBs, where the inputs are 512 × 512 pixels in 8-bit gray-scale. The peak signal-to-noise ratio (PSNR) shown below each image illustrates that the images sharpened using approximate 16-bit approximate adders consisting of nine Ax2, Ax3 and Ax4 have a similar quality

Table 8 The error characteristics of the AxFAs

Error metric	Ax1	Ax2	Ax3	Ax4
ER (%)	62.5	37.5	62.5	**25.0**
NMED (%)	29.2	**12.5**	20.8	16.7
MRED (%)	60.4	**25.0**	41.7	37.5
ED_{max} (%)	66.7	**33.3**	**33.3**	66.7

Bold values indicate the minimum error metric realized with different approximate adder implementation

Fig. 7 The image sharpening results using 16-bit approximate adders



(a) Input image **(b)** accurate output **(c)** Ax1 (21.5 *dB*)

(d) Ax2 (32.4 *dB*) **(e)** Ax3 (33.1 *dB*) **(f)** Ax4 (34.6 *dB*)

as the accurate result. The PSNRs for the images sharpened by approximate adders with different k values are shown in Table 9. It shows that the image sharpening results are of good quality when the number of approximate LSBs is less than 10 for Ax2, Ax3 and Ax4. However, the maximum number of approximate LSBs is 7 for Ax1 to achieve a good enough image sharpening result.The crossbar size of the Pj-AxMTJ based scheme is set to 256 × 256, the binary CIM scheme is set to 128 × 128. Matrix splitting is required if the crossbar size is smaller than the network layer size. The Pj-AxMTJ based scheme and binary CIM scheme can work at 200MHz. The accuracy simulation is performed in TensorFlow platform based on HSPICE-circuit extracted parameters.

5.2 Image similarity metric evaluation

To assess its effectiveness in image processing, the proposed VCMA based in-memory computing scheme is included in the image similarity metric evaluation, which is relevant to error-tolerant applications. In the process of two data comparison, if the write data is ‘1’ (‘0’), a high (low) level voltage is applied to the BL. No matter what state does VCMA-MTJ is, when a high level voltage is applied to the BL, the state of the VCMA-MTJ will flip to the opposite state in the current state (from AP to P or from P to AP). Therefore, if the storage data is same as (different from) the write data, the comparison result of tow data is ‘0’ (‘1’). By comparing the different pixels of the two images, the comparison results of the two images stored in VCMA-MTJ will be obtained.

There are two methods to detect the result of data comparison. First method uses SA to read every bit of all compared data. The ratio of the number of data calculated to the total amount of data is the similarity between the two sets of data. Although a precise similarity can be obtained from this method, it needs long time and great energy to sense the data. Another method which can avoid long reading time and effectively reduce energy is reading multi-bit data at the same time with one SA. However, this method can only get the similarity of this set of data is greater than or less than a certain threshold. The threshold can be adjusted from changing the reference array of sense amplifier. To the best of our knowledge, the proposed image similarity evaluation is the only PIM architecture that achieves in-memory image similarity evaluation, which is different from conventional image similarity evaluation based CNN/BNN acceleration (Angizi et al. 2019, 2020), but also can implement a full set of input Boolean in-memory logic functions as well as majority-based logic operations using its distinct computing methods.

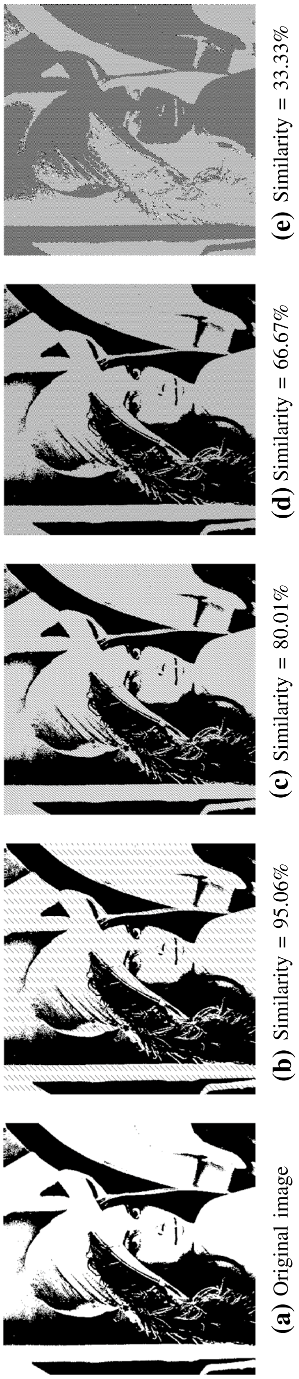
In image similarity evaluation method, multi cells located in an identical column can be selected by word line (WL) and sensed simultaneously to realize similarity evaluation. For instance, consider the data organization shown in Fig. 9a, where Array A and Array B operands correspond to two different pictures, respectively. Based on the previous discussion, the interplay is achieved with two times VCMA induced MTJ switching as the XOR function. Initial image processed into binary 0, 1 and stored into the array; the contrast image is converted to data as the array write operation control signal, the XOR results are restored into the array as Array C. The computational XOR results array can perform similarity evaluation function by setting WL to 1. As shown in Fig. 9b, reference voltage of sense amplifier is set at appropriate threshold. The number of date 1 in a column is judged by the readout data of the sensitive amplifier. When the number of date 1 in a column is more than the threshold value, the readout data of the amplifier is 1, and it is determined that these pixel points are mismatch.

Figure 8 illustrates image similarity evaluation using VCMA based MTJ interplay writing. The interplay is achieved with two times VCMA induced MTJ switching as the XOR function. The image demonstration is setup with 95.06%, 80.01%, 66.67%, 33.33% similarity with the original image. As shown in Fig. 10, the related circuit composed by writing and sensing circuits are implemented for image similarity evaluation. For example, suppose the resistance of reference array is less than R1 but greater than R2 (the resistance after four high-resistance (AP state VCMA-MTJ) and six low-resistance (P state of VCMA-MTJ) parallels is R1, the resistance after three high-resistance and seven low-resistance parallels is R2), if the data from reading 10 bits at the same time is ‘1’ (‘0’), t means that the similarity of the 10 bits is less than or equal to 60% (greater than or equal to 70%). Figure 11 shows the simulation waveform of image similarity evaluation circuits.

The simulation demonstrates the impact of different references on SA readout results. The data corresponding to SA1, SA2, SA3 and SA4 in the simulation diagram are 9P-1AP (9 MTJs in parallel state, 1 MTJ in anti-parallel state), 8P-2AP, 7P-3AP and 6P-4AP respectively. If the resistance of the reference source is set to three different criteria, *ref*1 (reference resistance greater than 9P-1AP and less than 8P-2AP), *ref*2 (reference resistance greater than 8P-2AP and less than 7P-3AP), and *ref*3 (reference resistance greater than 7P-3AP and less than 6P-4AP), three different results shown in the simulation diagram will be got. Under the standard of ref1, only the data corresponding to SA4 is similarity, but as the reference source resistance rises, the data corresponding to SA3 and SA2 gradually becomes similarity. The range of data included in the similarity is varied according to the setup of *ref* resistance.

Considering the energy consumption, if the reference bit-cells and data bit-cell are both with P state, the read energy is the largest. In this case, the energy consumption of reading 16-bit data at 16 times and 16-bit data at the same time is 222.1 fJ and 85.72 fJ respectively. A 61.4% energy reduction is achieved using VCMA mechanism interaction based XOR functions.

Table 10 shows the related works comparison, to the best of our knowledge, this is the first work that proposes the VCMA + STT/SOT + STT/SOT/P-VCMA PIM as an



(a) Original image **(b)** Similarity = 95.06% **(c)** Similarity = 80.01% **(d)** Similarity = 66.67% **(e)** Similarity = 33.33%

Fig. 8 Image similarity evaluation using VCMA based MTJ interplay writing, 95.06%, 80.01%, 66.67%, 33.33%

Table 9 PSNR values for images sharpened by 16-bit approximate adders with different values of k (dB)

AxFA	$k = 7$	$k = 8$	$k = 9$	$k = 10$
Ax1	33.0	26.9	21.5	16.0
Ax2	43.4	38.1	32.4	27.1
Ax3	44.8	39.1	33.1	29.0
Ax4	47.1	41.0	34.6	28.7

accelerator for a wide variety of tasks such as CNN/BNN acceleration and data encryption. Pj-AxMTJ is the only PIM architecture that achieves in-memory image similarity evaluation, but also can implement a full set of input Boolean in-memory logic functions as well as majority-based logic operations using its distinct computing methods.

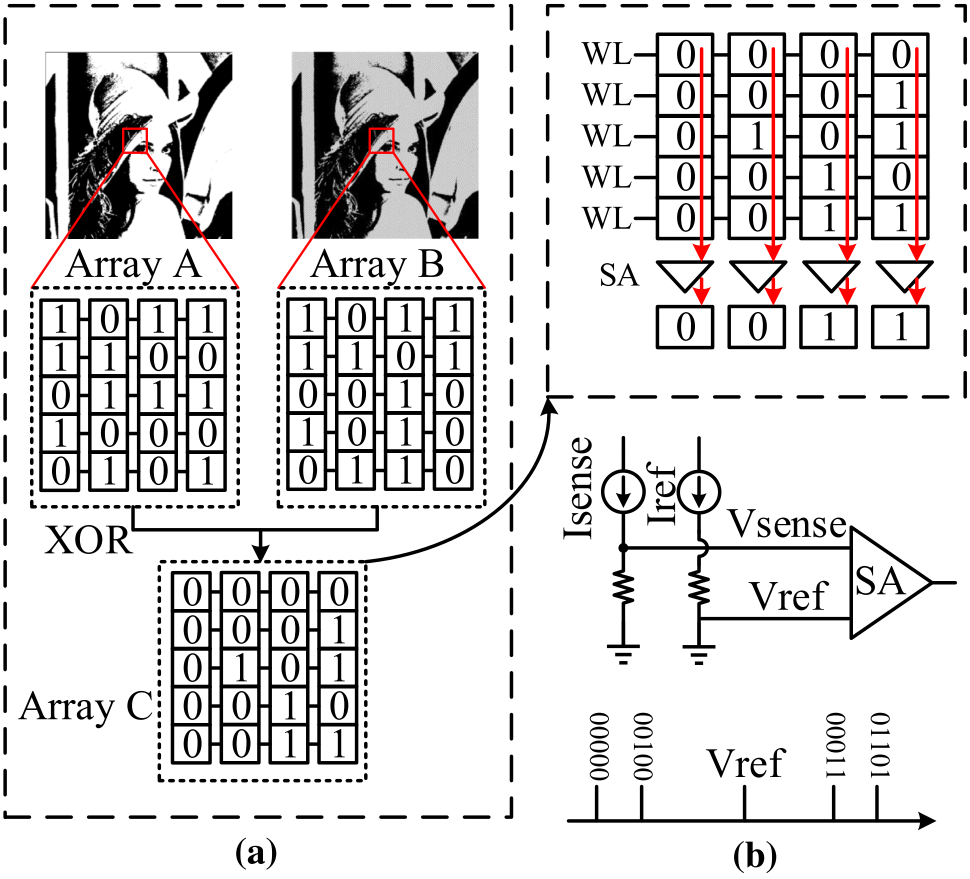
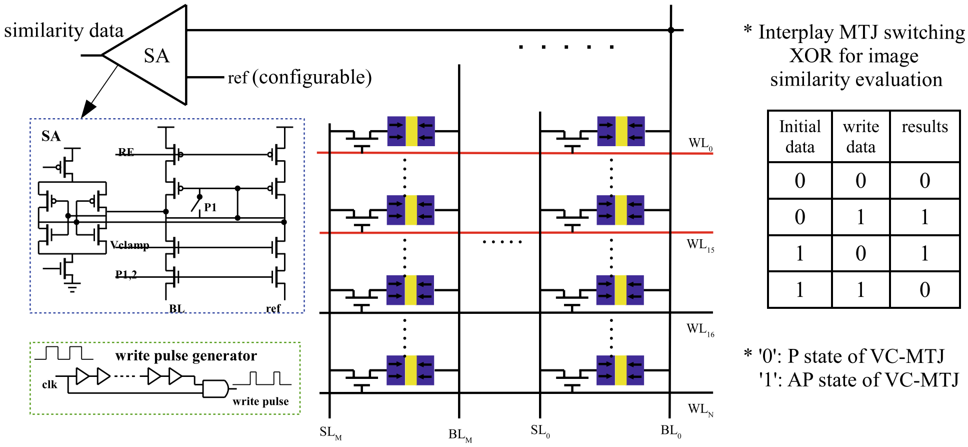


Fig. 9 **a** Mapping and computation of Image similarity evaluation. **b** Image similarity evaluation of voltage comparison between Vsenseand Vref



Initial data	write data	results
0	0	0
0	1	1
1	0	1
1	1	0

Fig. 10 Image similarity evaluation circuits using interplay VCMA based XOR logic operation

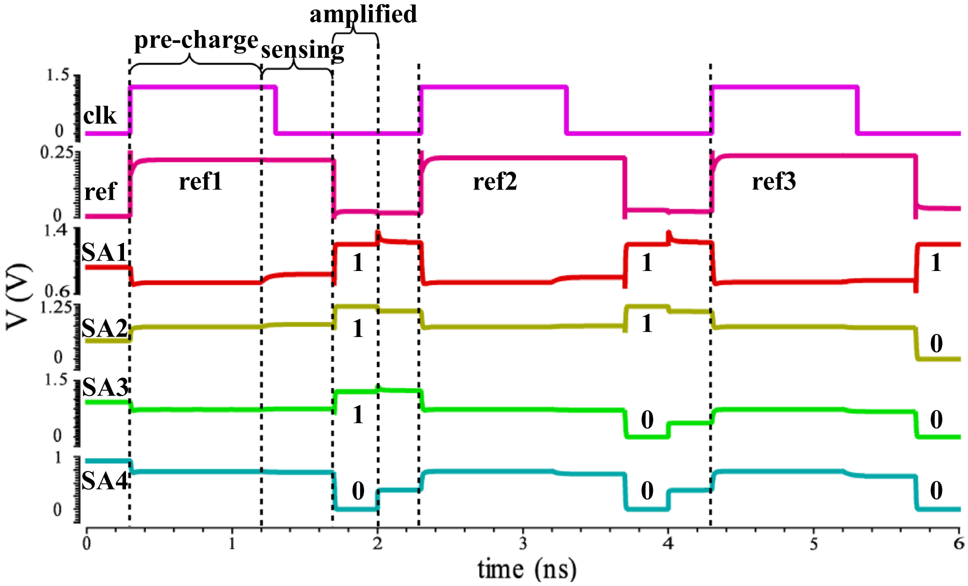


Fig. 11 The simulation waveform of image similarity evaluation circuits

Table 10 Related works comparison

Features	Pinatubo	RIMPA	STT-CiM	GraphS	MRIMA	Pj-AxMTJ
Technology	PCM/ReRAM	DWM	STT-MRAM	SOT-MRAM	STT-MRAM	VCMA+STT/SOT+STT/SOT/P-VCMA
Volatility	Non-volatile	Non-volatile	Non-volatile	Non-volatile	Non-volatile	Non-volatile
In-memory logic	AND/NAND/OR/NOR/XOR/XNOR	AND/NAND/OR /NOR/Maj-logic	AND/NAND/OR/NOR/XOR	AND/NAND/OR /NOR/XOR/XNOR /add/sub	AND/NAND/OR/NOR /XOR/XNOR/Maj-logic	AND/NAND/OR /NOR/XOR/XNOR /add/sub
In-memory convolver	N/A	N/A	N/A	Bit-wise adder/bit-wise convolver	Bit-wise adder/bit-wise convolver	Bit-wise adder /bit-wise convolver
Image recognitio	N/A	N/A	N/A	BNN/CNN	BNN/CNN	BNN/CNN/in-memory image similarity evaluation

6 Conclusion

In this paper, a novel nonvolatile process-in-memory strategy is proposed for approximate computing denoted as $Pj-AxMTJ$, which manifests in a different process-in-memory way using only non-volatile writing operations. By using joint magnetization switching mechanisms, such as precessional VCMA, STT-assisted precessional VCMA, as well as SOT erasing-STT programming, the bit-wise addition has been approximately realized in the bit-cells without sensing/reading operation and extra peripheral circuits. Simulation results demonstrate controllable approximate operations, and the perspective VCMA switching show a good power-delay performance. As case studies, image processing results were achieved as the accurate design, when the proposed AxFAs are used in the LSBs of a 16-bit adder. Interplay MTJ switching based in-memory XOR was implemented to evaluate image similarity. This interplay method can be further used in other bitwise operations especially within the low precision computational memory.

References

Amiri, P.K., Alzate, J.G., Cai, X.Q., Ebrahimi, F., Hu, Q., Wong, K., Grzes, C., Lee, H., Yu, G., Li, X., Akyol, M., Shao, Q., Katine, J.A., Langer, J., Ocker, B., Wang, K.L.: Electric-field-controlled magnetoelectric RAM: Progress, challenges, and scaling. IEEE Trans. Magn. **51**(11), 1–7 (2015)

Angizi, S., Sun, J., Zhang, W., Fan, D., Graphs: a graph processing accelerator leveraging SOT-MRAM. In: Design. Automation Test in Europe Conference Exhibition (DATE), pp. 378–383 (2019)

Angizi, S., He, Z., Awad, A., Fan, D.: Mrima: An MRAM-based in-memory accelerator. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **39**(5), 1123–1136 (2020)

Baek, S., Park, K., Kil, D., Jang, Y., Park, J., Lee, K., Park, B.: Complementary logic operation based on electric-field controlled spinorbit torques. Nat. Electron. **1**(7), 388–403 (2018)

Cai, H., Wang, Y., Naviner, L., Zhao, W.: Robust ultra-low power nonvolatile logic-in-memory circuits in FD-SOI technology. IEEE Trans. Circuits Syst. I Regul. Pap. **64**(4), 847–857 (2017)

Dong, X., et al.: Nvsim: A circuit-level performance, energy, and area model for emerging non-volatile memory. In: In emerging memory technologies, pp. 15–50. Springer, New York (2014)

Frustaci, F., Blaauw, D., Sylvester, D., Alioto, M.: Approximate SRAMs with dynamic energy-quality management. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **24**(6), 2128–2141 (2016)

Frustaci, F., Blaauw, D., Sylvester, D., Alioto, M.: Approximate SRAMs with dynamic energy-quality management, IEEE Trans. Very Large Scale Integr. VLSI Syst. **24**(6), 2128–2141 (2016)

Jain, S., Ranjan, A., Roy, K., Raghunathan, A.: Computing in memory with spin-transfer torque magnetic RAM. IEEE Trans. Very Large Scale Integr. VLSI Syst. **26**(3), 470–483 (2018)

Kanai, S., Yamanouchi, M., Ikeda, S., Nakatani, Y., Matsukura, F., Ohno, H.: Electric field-induced magnetization reversal in a perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. Appl. Phys. Lett. **101**(12), 122403 (2012)

Kanai, S., Nakatani, Y., Yamanouchi, M., Ikeda, S., Sato, H., Matsukura, F., Ohno, H.: Magnetization switching in a CoFeB/MgO magnetic tunnel junction by combining spin-transfer torque and electric field-effect. Appl. Phys. Lett. **104**(21), 212406 (2014)

Le Gallo, M., Sebastian, A., Mathis, R., Manica, M., Giefers, H., Tuma, T., Bekas, C., Curioni, A., Eleftheriou, E.: Mixed-precision in-memory computing. Nat. Electron. **1**(4), 246–253 (2018)

Li, S., Xu, C., Zou, Q., Zhao, J., Lu, Y., Xie, Y.: Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories, In: 2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC), pp. 1–6 (2016)

Liang, J., Han, J., Lombardi, F.: New metrics for the reliability of approximate and probabilistic adders. IEEE Trans. Comput. **62**(9), 1760–1771 (2013)

Locatelli, N., Vincent, A.F., Querlioz, D.: Use of magnetoresistive random-access memory as approximate memory for training neural networks. In: 2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS), pp. 553–556 (2018)

Maruyama, T., Shiota, Y., Nozaki, T., Ohta, K., Toda, N., Mizuguchi, M., Tulapurkar, A., Shinjo, T., Shiraishi, M., Mizukami, S., Ando, Y., Suzuki, Y.: Large voltage-induced magnetic anisotropy change in a few atomic layers of iron. Nat. Nanotechnol. **4**, 158–161 (2009)

Mittal, S.: A survey of techniques for approximate computing. ACM Comput. Surv. **48**(4), 62:1–62:33 (2016)

Monazzah, A.M.H., Shoushtari, M., Miremadi, S.G., Rahmani, A.M., Dutt, N.: QuARK: Quality-configurable approximate STT-MRAM cache by fine-grained tuning of reliability-energy knobs. In: 2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), pp. 1–6 (2017)

Natsui, M., Suzuki, D., Sakimura, N., Nebashi, R., Tsuji, Y., Morioka, A., Sugibayashi, T., Miura, S., Honjo, H., Kinoshita, K., Ikeda, S., Endoh, T., Ohno, H., Hanyu, T.: Nonvolatile logic-in-memory LSI using cycle-based power gating and its application to motion-vector prediction. IEEE J. Solid State Circuits **50**(2), 476–489 (2015)

Oboril, F., Shirvanian, A., Tahoori, M.: Fault tolerant approximate computing using emerging non-volatile spintronic memories. In: 2016 IEEE 34th VLSI Test Symposium (VTS) (2016)

Ranjan, A., Venkataramani, S., Fong, X., Roy, K., Raghunathan, A.: Approximate storage for energy efficient spintronic memories. In: 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), pp. 1–6 (2015)

Ranjan, A., Venkataramani, S., Pajouhi, Z., Venkatesan, R., Roy, K., Raghunathan, A.: STAxCache: An approximate, energy efficient STT-MRAM cache. In: Design, Automation Test in Europe Conference Exhibition (DATE), pp. 356–361 (2017)

Sampson, A., Nelson, J., Strauss, K., Ceze, L.: Approximate storage in solid-state memories. In: 2013 46th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp. 25–36 (2013)

Spinmodel Library (2015). http://www.spinlib.com/STT_PMA_MTJ.html. Accessed 1 Oct 2015

Teimoori, M.T., Hanif, M.A., Ejlali, A.,Shafique, M.: AdAM: Adaptive approximation management for the non-volatile memory hierarchies. In: 2018 Design, Automation Test in Europe Conference Exhibition (DATE), pp. 785–790 (2018)

Wang, W.-G., Li, M., Hageman, S., Chien, C.L.: Electric-field-assisted switching in magnetic tunnel junctions. Nat. Mater. **11**, 64–68 (2012)

Wang, S., Lee, H., Ebrahimi, F., Amiri, P.K., Wang, K.L., Gupta, P.: Comparative evaluation of spin-transfer-torque and magnetoelectric random access memory. IEEE J. Emerg. Sel. Topics Circuits Syst. **6**(2), 134–145 (2016)

Wang, L., Kang, W., Ebrahimi, F., Li, X., Huang, Y., Zhao, C., Wang, K.L., Zhao, W.: Voltage-controlled magnetic tunnel junctions for processing-in-memory implementation. IEEE Electron Device Lett. **39**(3), 440–443 (2018)

Wang, M., Cai, W., Zhu, D., Wang, Z., Kan, J., Zhao, Z., Cao, K., Wang, Z., Zhang, Y., Zhang, T., Park, C., Wang, J., Fert, A., Zhao, W.: Field-free switching of a perpendicular magnetic tunnel junction through the interplay of spinorbit and spin-transfer torques. Nat. Electron. **1**(4), 582–588 (2018a)

Wang, Z., Zhang, L., Wang, M., Wang, Z., Zhu, D., Zhang, Y., Zhao, W.: High-density nand-like spin transfer torque memory with spin orbit torque erase operation. IEEE Electron Device Lett. **39**(3), 343–346 (2018b)

Wang, Z., Zhou, H., Wang, M., Cai, W., Zhu, D., Klein, J., Zhao, W.: Proposal of toggle spin torques magnetic RAM for ultrafast computing. IEEE Electron Device Lett. **40**(5), 726–729 (2019)

Yamaga,Y., Deguchi,Y., Fukuyama,S., Takeuchi,K.: 5x reliability enhanced 40 nm TaOx approximate-ReRAM with domain-specific computing for real-time image recognition of IoT edge devices. In: 2018 IEEE Symposium on VLSI Technology, pp. 109–110 (2018)

Yesil, S., Akturk, I., Karpuzcu, U.R.: Toward dynamic precision scaling. IEEE Micro **38**(4), 30–39 (2018)

Zeinali, B., Karsinos, D., Moradi, F.: Progressive scaled STT-RAM for approximate computing in multimedia applications. IEEE Trans. Circuits Syst. II Express Briefs **65**(7), 938–942 (2018)

Zeinali, B., Karsinos, D., Moradi, F.: Progressive scaled STT-RAM for approximate computing in multimedia applications. IEEE Trans. Circuits Syst. II Express Briefs **65**(7), 938–942 (2018)

Zhang, H., Kang, W., Wang, L., Wang, K.L., Zhao, W.: Stateful reconfigurable logic via a single-voltage-gated spin hall-effect driven magnetic tunnel junction in a spintronic memory. IEEE Trans. Electron Devices **64**(10), 4295–4301 (2017)

Zhao, H., Xue, L., Chi, P., Zhao, J.: Approximate image storage with multi-level cell STT-MRAM main memory. In: IEEE/ACM International Conference on Computer-Aided Design (ICCAD) 2017, pp. 268–275 (2017)