SCIENCE CHINA
Information Sciences

June 2021, Vol. 64 160402:1–160402:15
https://doi.org/10.1007/s11432-021-3220-0

• REVIEW •

Special Focus on Near-memory and In-memory Computing

A survey of in-spin transfer torque MRAM computing

Hao CAI[1†], Bo LIU[1†], Juntong CHEN[1], Lirida NAVINER[2], Yongliang ZHOU[1], Zhen WANG[3] & Jun YANG[1*]

[1]*National ASIC System Engineering Research Center, Southeast University, Nanjing* 210096*, China;*
[2]*Laboratoire Traitement et Communication de l'Information, Télécom Paris, Palaiseau* 91120*, France;*
[3]*Nanjing Prochip Electronic Technology Co., Ltd., Nanjing* 210000*, China*

Received 15 January 2021/Revised 5 March 2021/Accepted 17 March 2021/Published online 10 May 2021

Abstract In traditional von Neumann computing architectures, the essential transfer of data between the processor and memory hierarchies limits the computational efficiency of next-generation system-on-a-chip. The emerging in-memory computing (IMC) approach addresses this issue and facilitates the movement of significant data and rapid computations. Among the different memory types, intrinsic energy efficiency is demonstrated by in-magnetic random access memory (MRAM) computing with a low-power spintronic magnetic tunnel junction device and hybrid integration at an advanced complementary metal-oxide semiconductor node. This study reviews state-of-the-art techniques for managing IMC with an emphasis on spin-transfer torque-MRAM computing via design schemes at the bit-cell, circuit, and system levels. In addition, this study presents effective design techniques and potential challenges and demonstrates the existing limitations of in-MRAM computing and potential methods for overcoming these issues. This study also considers the design technology co-optimization from the IMC perspective.

Keywords spin-transfer torque-magnetoresistive random access memory, in-memory computing, magnetic tunnel junction, analog computing, nonvolatile memory, Boolean logic, neural network

Citation Cai H, Liu B, Chen J T, et al. A survey of in-spin transfer torque MRAM computing. Sci China Inf Sci, 2021, 64(6): 160402, https://doi.org/10.1007/s11432-021-3220-0

1 Introduction

The von Neumann computing architecture requires data transfer between processors and memory hierarchies, which limits the computational efficiency of next-generation data-centric system-on-a-chip (SoC) devices. The unique potential for emerging nonvolatile memories (NVMs) is to bring high-density memory arrays closer to processing elements (PEs) and creates high throughputs with low-latency access [1–3].

The in-memory computing (IMC) approach physically integrates processor-related computation and storage in a single chip [1–3]. The application of this approach can break through the memory/power wall, accelerate significant data transfer between PEs and memory subsystems, and produce highly efficient computations. Typically, this approach facilitates vector-matrix multiplication and executes it in parallel using analog computations, in which the input vectors can be activated with multiple rows. The dot product is obtained as the multiplication result of the input voltage and cell conductance, and the partial sum is calculated by the current column. Generally, an analog-to-digital converter (ADC) located at the edge of the memory array converts the partial sum to binary bits for post-processing purposes.

Extensive research has been conducted on realizing IMC with mature memories. Computing within static random access memory (SRAM) (possibly with modified bit-cells) was realized using a unitary mainstream complementary metal-oxide semiconductor (CMOS) process, e.g., with 28-nm sandwich-RAM [4] and 65-nm and 130-nm CMOS [2,5]. However, SRAM displays low density and is inherently volatile with significant standby leakage power consumption. Besides, both dynamic random access memory (DRAM)

[*] Corresponding author (email: dragon@seu.edu.cn)
[†] Cai H and Liu B have the same contribution to this work.

Table 1 Summary of main acronyms

Acronym	Definition	Acronym	Definition
STT	Spin transfer torque	NVM	Non-volatile memories
MRAM	Magnetoresistive random access memory	RRAM	Resistive random access memory
IMC	In-memory computing	PCM	Phase change memory
NMC	Near-memory computing	LIM	Logic-in-memory
MTJ	Magnetic tunnel junction	SRAM/DRAM	Static/dynamic random access memory
SOT	Spin orbit torque	MAC	Multiply-and-accumulate
VCMA	Voltage controlled magnetic anisotropy	CD	Critical dimension
VG-SHE	Voltage-gated spin hall effect	ECC	Error correction coding
TST	Toggle spin torques	PPA	Power-performance-area
TMR	Tunnel magnetoresistance ratio	SA	Sense amplifier
PMA	Perpendicular magnetic anisotropy	VSA/CSA	Voltage-type/current-type SA
PEs	Processing elements	ADC	Analog-to-digital converter
PTL	Pass-transistor-logic	FA	Full adder
WL/BL/SL	Word-line/bit-line/source-line	FDSOI	Fully depleted silicon-on-insulator
FinFET	Fin field-effect transistor	PUF	Physical unclonable function
PVT	Process-voltage-temperature	TRNG	True random number generator
FeFET	Ferroelectric field effect transistors	DNN/CNN/BNN	Deep/convolutional/binary neural network
SoC	System-on-chip	MeRAM	Magnetoelectric random access memory

and NOR-flash are characterized by properties of high density and low cost. Favorable performance was demonstrated by in-DRAM computing, e.g., Eyeriss [6] and in-NOR-flash neuromorphic computing [7]. However, the main disadvantages of these approaches include finite retention (IMC with DRAM) and low endurance (IMC with Flash). Accordingly, emerging NVMs with different information/bit physical storage mechanisms are better suited for power-performance-area constrained IMC platforms. Representative NVMs include resistive random access memory (RRAM), phase change memory (PCM), magnetic random access memory (MRAM), and ferroelectric field-effect transistor (FeFET)-based RAM. The non-volatility of these memories permits instant on/off switching and prevents the loss of stored data. Among the NVMs, MRAM is the only one expected to possess unlimited endurance; however, atomic motion in the storage material constrains the retention of both RRAM and PCM [8].

This study presents an overview of IMC, particularly in terms of foundry-available spin-transfer torque (STT)-MRAM. Table 1 lists the acronyms used in this paper, and the remaining manuscript is organized as follows: Section 2 introduces the novel paradigms besides memory storage, Section 3 reviews the state-of-the-art and design challenges of in-MRAM computing, Section 4 presents an in-MRAM computing perspective, and Section 5 concludes the paper.

2 Review of novel paradigms besides memory storage

As SoC technology becomes more data centric, energy dissipation and data throughput are improved by the use of novel memory approaches, including multistorage modes and IMC/near-memory computing (NMC) schemes. This paper hierarchically explores recent developments in this area from a bottom-up perspective from the bit-cell level to the circuit, architecture, and arithmetic levels (see Figure 1). Additionally, the study considers traditional (SRAM, DRAM, and Flash) and emerging NVMs (RRAM, PCM, and MRAM).

2.1 Bit-cell level attempt

As two of the most direct and effective approaches, multi-mode and IMC schemes can be realized by modification at the bit-cell level. Different bit-cell types and structures can be utilized to achieve the desired performance in specific scenarios based on memory device characteristics (see Figure 1(a)). Any optimization of bit-cell characteristics can significantly affect the performance of high-level memory circuits and subsystems.

Early approaches did not include a computational mode. A dual-mode NAND-flash that switches between multi-level program cell (MLC) and single level program cell (SLC) was proposed and evaluated in [9]. In comparison with the SLC mode, MLCs reduce the overhead bit-cell layout and enhance the

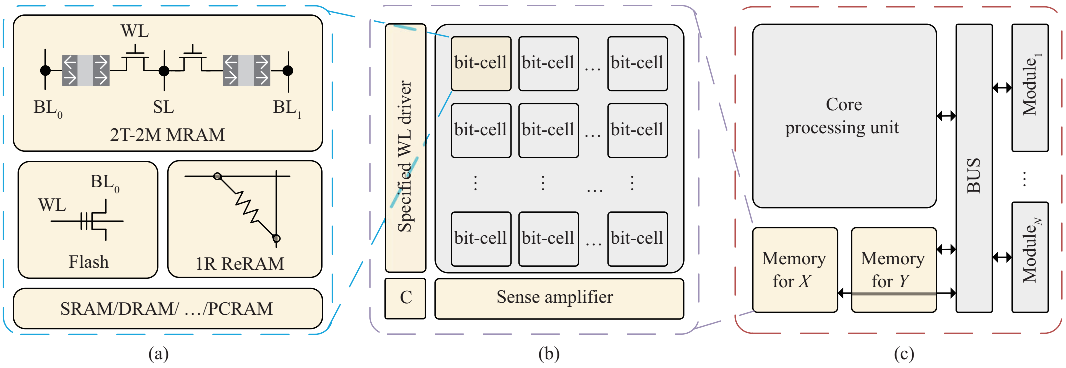

Figure 1 (Color online) (a) According to the device characteristics, different bit-cell types and structures can be implemented for desired performance in specific scenarios. (b) Commonly, circuit-level reconfiguration is realized using carefully designed peripheral circuits, including reading/writing drivers, sense amplifiers, and controllers. Basic functions are targeted, including storage and the computing unit. (c) Architecture-level reconfiguration is mainly focused on application (X or Y) and follows instructions from the core processing unit to execute assigned tasks.

density of NAND-flash. Notably, the programming performance can be reduced by the requirement for tight threshold voltage ($V_{\rm th}$) control. Conversely, the SLC mode is less sensitive to $V_{\rm th}$. This makes the MLC mode more suitable for scenarios of low program throughput, while the SLC mode is better suited for high-performance applications. An adjustable incremental step pulse and a self-boosting scheme were implemented in NAND-flash [9]; they exert tight control over the $V_{\rm th}$ of the cell and enable the memory to function with the high-density MLC and high-performance SLC modes. This design concept was further applied to RRAM [10, 11] and PCM [12] based on redesigned writing and peripheral circuit blocks.

Significant potential for using the IMC approach for emerging device-based NVMs was demonstrated by processing the in-FeFET cells with an intrinsic compact area and distinguishable multiple states [13, 14]. As a demonstration of PCM, dual-mode double-density PCM based on a novel stressing-mode storage scheme was studied in [15]. The threshold for RESET switching can be shifted by stressing a current on the memory cell, and the R-I curve can be used to store logic states. The independence of the two modes prevents interference between them and enables a double storage density. Additionally, PCM is applicable to IMC via physics crystallization or modified peripheral circuits [16, 17].

Notably, modifications of memory bit-cell and device parameters require fabrication support from foundries. Typically, customized bit-cells are expensive and require long periods of development/validation. Consequently, the optimal schemes for energy-efficient or high-performance memory subsystems utilize the standard features of the device and bit-cell and implement peripheral circuits.

2.2 Circuit-level implementation

With respect to the circuit-level implementation of storage blocks, memory subsystems can perform the storage function as well as the associated arithmetic and computing units. IMC and NMC were investigated, and the majority of them underwent silicon verification in SRAM [4, 18–22] and several NVMs, including RRAM [23–25], STT-MRAM [26–31], spin-orbit torque (SOT), and MRAM [32–34].

In addition to the modification of bit-cells for IMC, the peripheral circuit design method offers an effective approach for realizing novel IMC and NMC approaches. As shown in Figure 1(b), the reconfiguration or allocation of the reference generator, modified sense amplifier (SA), and specific driver/controller circuits can be reconfigured or allocated to realize basic Boolean logic and convolutional operations. Modified SA circuits can efficiently compute bitwise logic operations with data prestored in the memory. During analog signal processing, the complex reference structure shrinks the memory-sensing margin among different logic operations. As a solution, Ref. [18] proposed a common-mode insensitive small-offset voltage-type SA (VSA) in SRAM to enhance the yield of the computational output.

Pinatubo [35] and NV-logic-in-memory (LIM) [26] structures were demonstrated as the universal approaches for several Boolean logic (OR, AND, XOR, and INV operations) in-memory realizations, and these structures were applicable to NVM. As an effective processing-in-memory (PIM) scheme, a Pinatubo architecture enables the bulk bitwise efficient operations of two or more memory rows in resistive-cell-featured NVMs and supports one-step multirow operations with insignificant area overheads. Demonstrated with MRAM design environment, the NV-LIM scheme applied pass-transistor logic (PTL) for

Table 2 Survey of recent emerging memory based MAC operation

	CMOS	Memory	Bit-cell	Array	Tape-out	MAC	Speed-up	Energy	Application
DAC'16 [49]	N/A	RRAM	1T1R	512×512	No	DA-AD	1000×	N/A	Inference
ISCA'16 [50]	32-nm	RRAM	1T1R	128×128	No	DA-AD	14.8×	380 GOPS/W	Inference
Nature'18 [51]	90-nm	PCM	3T1C	512×512	No	DA-AD	N/A	119.7 TOPS/W	Inference+training
Nat.Elec.'18 [52]	N/A	RRAM	1T1R	128×64	Yes	DA-AD	N/A	17×	Inference
ISSCC'18 [53]	55-nm	RRAM	1T1R	512×256	Yes	AD	2×	N/A	Inference
ISSCC'19 [54]	55-nm	RRAM	1T1R	256×512	Yes	AD	1.3×	53.17 TOPS/W	Inference
VLSI'18 [55]	180/40-nm	RRAM	1T1R	2 Mb/4 Mb	Yes	AD	N/A	66.5 TOPS/W	Inference
NIPS'18 [56]	22-nm	STT-MRAM	1T1M	40 Mb	Yes	SRAM	N/A	9.9 TOPS/W	Inference
IEDM'18 [57]	130-nm	RRAM	2T2R	N/A	Yes	SA+logic	N/A	25 nJ/img/Minst	Inference
DAC'18 [58]	45-nm	SOT-MRAM	2T1M	N/A	No	IMC	4.3×	0.74 μJ/img/Minst	Inference
TVLSI'19 [59]	28-nm	SOT-MRAM	2T1M	128×256	No	IMC	12.3×	96.6 image/s/W	Inference
ISCAS'19 [60]	22-nm	STT-MRAM	1T1M	64×576	No	AD	70×	4.5×	Inference
ISSCC'20 [61]	130-nm	RRAM	1T1R	158.8 Kbit	Yes	DA-AD	N/A	78.4 TOPS/W	Inference

an 8-bit nonvolatile full adder (FA), NV-logic, NV-flip-flops, and approximate FA [26, 36]. Because of ultrafast PTL behavior and limited modification of peripheral circuits, this scheme achieves hardware-accelerated large-scale integration. However, the NV-LIM scheme is ineffective with large-scale NVM arrays owing to additional layout cost between SAs and bit-cells. We notice that further development of the Pinatubo and NV-LIM schemes is limited by their ability to perform only low-complexity Boolean logic operations.

2.3 Architecture-level exploration

The research in the fields of deep learning and neural-network-related applications has attracted considerable attention over the past decade. In terms of multisubsystem application scenarios, hybrid memory systems involving the collaboration of different memories have been proposed; these systems combine the benefits of different memory types, e.g., high-performance SRAMs and energy-efficient NVMs.

The capability of multimode reconfiguration memories in different applications has been reported in the literature [37–39]. Technologies including multilevel bit-cells, IMC/NMC, and approximate/stochastic computing have been investigated. A feasible approach for achieving an energy-efficient design is the scaling down of supply voltage ($V_{\rm dd}$), although memories are impacted by bit-cell access failures when biasing with low $V_{\rm dd}$.

To solve this problem, Ref. [40] proposed a hybrid 8T&6T-SRAM block and exploited the finite robustness of neural networks. Traditional 6T-SRAM can store the least significant bit with limited neural network impacts, whereas 8T-SRAM stores the most significant bit to guarantee the rightness of the bit-cell. Thus, the downscaling of $V_{\rm dd}$ can be achieved with a negligible loss in network performance. Furthermore, Ref. [41] studied an ultralow-power ECG monitor by separating the functions into transmission and record; this approach exploits the advantages of a free standby power NVM for information storage and recruits a high-performance volatile memory for the transmission task. A hybrid precision PCM system that benefits from both the accuracy of digital computing and energy layout area efficiency of the IMC mode was studied in [42]. The proliferation of big data requires both real-time processing and rapid transformation. Data security based on the stochastic behavior of MRAM is also important, such as the physical unclonable function and true random number generator [43–45]. Additionally, studies have been conducted on bioinspired computing for achieving low-energy, intelligent, and highly adaptable computing systems based on a device's stochastic behavior [46–48].

In terms of data-intensive workloads, the development of SoC to further improve energy efficiency is a growing research focus. Data flow is often optimized to increase on-chip data reuse, and the preferred approach is the in-memory-array digital multiply-and-accumulate (MAC) operation. However, the majority of the inputs/outputs are moved across MAC arrays and from global buffers. Table 2 [49–61] reviews parts of recently emerged embedded NVM-based MAC operations. Most of the publications were applied to inference applications and depend on ADC-DAC blocks for signal conversion. Owing to reliability and variability challenges, the silicon-verified results (tape-out work) were obtained using a mature CMOS processor. Finally, Ref. [56] achieved STT-MRAM codesigned with a PIM convolutional neural network (CNN) accelerator, which has been commercialized previously.

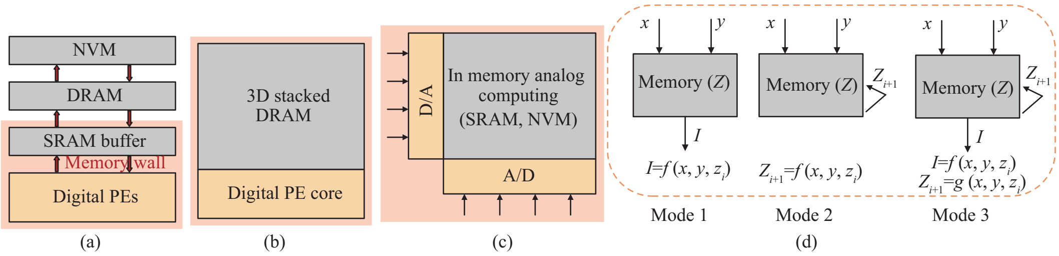

Figure 2 (Color online) Modified von-Neumann structure with in-memory computing. (a) PEs with memory hierarchies; (b) PEs with 3D stacked DRAM; (c) in-memory analog computing with SRAM and NVM; (d) three IMC modes.

2.4 Typical IMC modes

Recent demonstrations of all-digital deep neural network processors and accelerators with von Neumann architectures (Figure 2(a)) have shown that energy consumption and delays are dominated by the frequent movement of input data, weights, and intermediate data between the processor and memory. This has motivated the development of IMC in which in-place analog computing (see Figure 2(c)) exhibits substantial potential for addressing the energy requirements of computing and data movement. The above studies demonstrated that memory data movements and computations could be accelerated by both the multimode and IMC paradigms. Figure 2(d) presents the typical IMC modes, each of which exploits the advantages of the current storage state (Z_i) during computation. The modes can be simply classified as three different types:

- Mode 1, in which the computational result (output) is related to both input and storage states, and the output is generated directly.
- Mode 2 is similar to Mode 1, in which the computational result is stored in the storage unit.
- Mode 3, in which the computational result is related to the input and storage states. In this mode, simultaneous updates of the result outputs and the storage state occur.

As Mode 3 has similarities to the synaptic unit found in biological neural networks, it is more aligned with the development of artificial intelligence algorithms. Current applications based on Mode 3 are subject to significant constraints owing to the hardware resource limitation.

3 In-MRAM computing

MRAM has significant potential for use in next-generation low-power memory sub-systems because of its properties of high endurance (10^{15}), high density ($6F^2$), and efficient switching energy (0.1 pJ/bit writing) [8]. Several foundries have reported the data-access performance of MRAM macros to be stable and nonvolatile, which suggests good potential for in-MRAM computing. This section analyzes the electrical properties and bit-cells related to MTJ, and investigates the state-of-the-art computing schemes, challenges, and applications.

3.1 Implementation of MRAM hierarchy: an STT-MRAM example

Figure 3(a) demonstrates a typical single-access transistor MTJ (1T-1M)-MRAM bit-cell using a perpendicular MTJ device. Its characteristic is mainly determined by two ferromagnetic layers (CoFeB) and an oxide barrier (MgO). The magnetization direction of the reference layer is fixed, although it can be changed by the application of a sufficient current in the free layer. A lower resistance is achieved when the directions of the two ferromagnetic layers are parallel (R_P) rather than anti-parallel ($R_{\rm AP}$). The difference in resistance is represented as the tunnel magnetoresistance ratio (TMR) [62–64]:

$$\mathrm{TMR} = \frac{R_{\rm AP} - R_P}{R_P}. \tag{1}$$

When the bidirectional current is higher than the threshold current, the MTJ switches between the P and AP states (I_{c0}) [65]. I_{c0} is determined using (2) and (3). According to the STT mechanism, a bidirectional current I can change the MTJ between states when it is higher than the critical current I_{c0}:

$$I_{c0} = \alpha\frac{\gamma e}{\mu_B g}(\mu_0 M_s)H_K V = 2\alpha\frac{\gamma e}{\mu_B g}E, \tag{2}$$

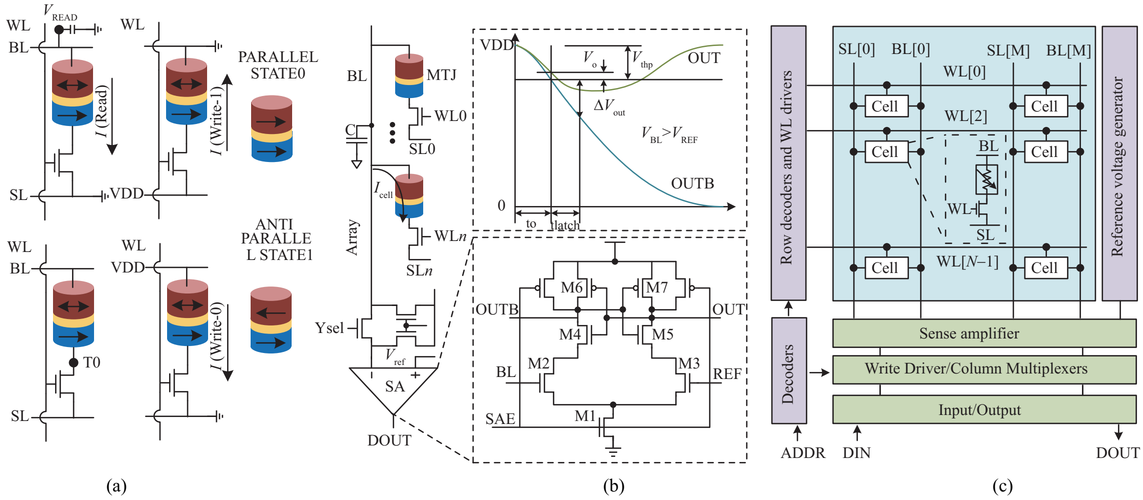

Figure 3 (Color online) MRAM bit-cell, main array, and peripheral blocks. (a) Bit-cell access bias; (b) latch-type voltage-mode sensing scheme and simulated transient waveform; (c) main STT-MRAM building blocks (1T-1M bit-cell).

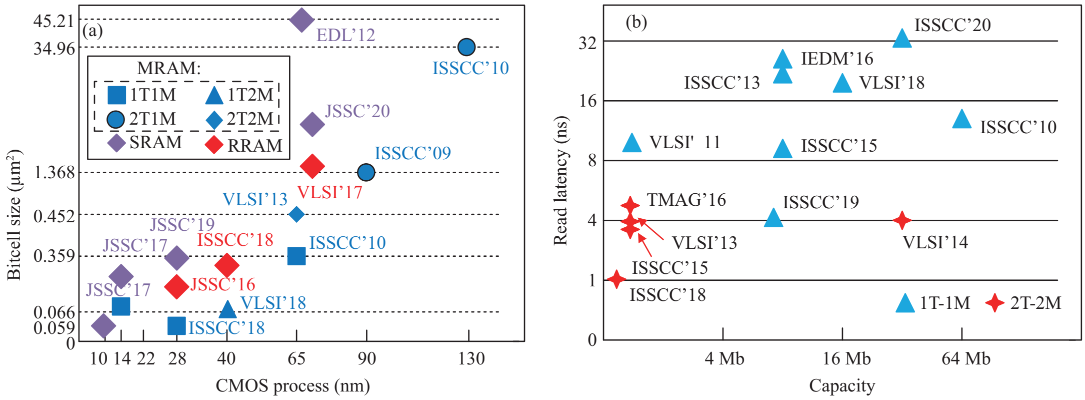

Figure 4 (Color online) Literature study of silicon-verified MRAM bit-cell structure. (a) The size of the bit-cells and their compatible CMOS process in MRAM, SRAM, and RRAM. (b) 1T-1M. The 2T-2M cell circuit is demonstrated as a suitable candidate for low-capacity and high-performance designs.

$$E = \frac{\mu_0 M_s \times H_k \times V}{2}, \tag{3}$$

where E is the energy barrier, α is the magnetic damping constant, γ is the gyromagnetic ratio, e is the elementary charge, μ_B is the Bohr magneton, μ_0 is the permeability of free space, M_s is the saturation magnetization, H_K is the effective anisotropy field, V is the volume of the free layer, and $g = \sqrt{\mathrm{TMR}(\mathrm{TMR}+2)/2(\mathrm{TMR}+1)}$ is the spin polarization efficiency factor.

Generally, the read operation differentiates the MTJ resistances by converting the resistance states into voltage differences using voltage sensing schemes [66, 67]. As presented in Figure 3(b), the bit-cell is read by enabling the word line (WL) and setting the source line (SL) to ground and bit line (BL) to V_{READ}. A small current (I_{cell}) is injected via the corresponding BL, and V_{BL} is determined by the effective resistance of the MTJ. Finally, the V_{BL} is compared with an intermediate reference voltage by an SA to detect the bit-cell data and ensure the availability of DOUT at the output of the SA.

Figure 3(c) depicts an overview of MRAM macro, which mainly comprises peripheral control circuits, sensing amplifiers, and MTJ-based core array circuits. The core array comprises bit-cells as $M \times N$, where M is the maximum bit number in a row, and N is the number of rows. The control units generate signals that inform other peripheral circuits to prepare for the incoming data access. A row decoder linked to a WL driver decodes the row address bits and selects a WL. Column multiplexers addressed by vertical BL and SL allow the write driver or SA to share a series of bits among multiple columns.

Figure 4 presents the literature studies of silicon-verified MRAM. To achieve diversified density, capacity, access latency, and energy consumption, the bit-cell structure of MRAM can be configured with 1T-1M, duplex 2T-2M, 2T-1M, and 4T-2M. An MRAM macro design (hybrid CMOS-MTJ integrated

Table 3 Survey of In-MRAM computing

	CMOS	Memory	Bit-cell	Operation	IMC with	SA type	Energy efficiency	Throughout
JSSC'15 [26]	90-nm	STT	4T2M	Read	Pass-transistor	CSA	48.3% improved	5×5 PE
TED'19 [68]	40-nm	STT	1T1M	Write	Bit-cell	CSA	237.3 fJ/bit	N/A
MEJ'18 [69]	45-nm	STT	1T1M	Read	Dual references	CSA	10 fJ/bit-wise	bit-wise
TVLSI'18 [70]	N/A	eNVM	1T1R	Read	References	VSA	38% improved	bit-wise
VLSI'18 [71]	45-nm	STT	1T1M	Read	References	CSA	N/A	bit-wise/DNN
DAC'18 [58]	45-nm	SOT	2T1M	Read	References	VSA	94 ×	N/A
Intermag'18 [72]	40-nm	STT	1T2M	Read	MTJ+references	CSA	N/A	N/A
TED'17 [73]	N/A	VCMA	1M	Write+read	MTJ	Crossbar	12 fJ/bit	N/A
DAC'19 [74]	45-nm	SOT	2T1M	Read	References	VSA	N/A	412.28 K/W
TVLSI'19 [75]	28-nm	STT	2T1M	Write/read	Threshold+references	CSA	68.5% saving vs. FPGA	235.1 image/s
TVLSI'19 [59]	28-nm	SOT	2T1M	Write+read	Bit-cell	CSA	1.41 W (CIFAR-10)	96.6-image/s/W
Tnano'19 [76]	40-nm	VG-SHE	1T1M	Write	Bit-cell	Crossbar	63.8 fJ/bit/FA	bit-wise
VLSI'20 [34]	22-nm	SOT	2T1M	Read	Analog IMC	N/A	N/A	N/A
ISSCC'20 [31]	22-nm	STT	1T1M	Read	Near-memory	CSA	0.23 pJ/bit read	42.67 GB/s read

circuit) that incorporated the electrical characteristics of an MTJ was constructed using a CMOS design kit. The size of the access transistor controls the bit-cell size, as the MTJ is normally located in the top metal layer. According to the literature study presented in Figure 4(b), the 2T-2M bit-cell was shown to be a suitable choice for a low-capacity and high-performance scenario. The lower write energy consumption and more compact structure of the 1T-1M counterpart make it more suitable for energy harvesting and high-density scenarios.

3.2 The state-of-the-art of in-MRAM computing

Numerous in-MRAM computing schemes have been demonstrated using on-chip arrays to realize fundamental Boolean logic operations (i.e., AND, OR, XOR, and FA) and complex arithmetical functions (i.e., neural networks). Table 3 [26, 31, 34, 58, 59, 68–76] shows the literature study on recent in-MRAM computing. Main approaches are executed using bit-cell modification, reference adaptation, PTL-enabled NMC, and in-memory analog computation schemes.

3.2.1 *Sensing-based scheme*

As listed in Table 3, a sensing-based IMC scheme was applied to STT, SOT, and voltage-controlled magnetic anisotropy (VCMA)-MTJ switching to realize bitwise operations based on the modified read/sense circuit; multiple rows in an MRAM array can be activated simultaneously. Additionally, bitwise logic operations can be performed by comparing the analog BL value (e.g., current or voltage amplitude) with reference bit-cells. However, the drawback is that the complex peripheral circuits must be included for bitwise operation, which increases both the energy consumption and the layout area. Additionally, the performance is highly sensitive to the CMOS-MTJ device variations; e.g., the limited TMR of MTJ leads to a small sensing margin and low $V_{\rm BL}$ swing, which deteriorates the computational accuracy and limits IMC stability. Recently, considerable research has been conducted on MRAM-sensing margin enhancement techniques [66, 67, 77] to benefit the performance optimization of sensing-based schemes.

Ref. [31] proposed a near-memory scheme, in which the output latch was reused in the output buffer and reconfigured as a D-flip-flop to realize shift and rotate operations. The sensing circuit was modified to facilitate the adaptation of NMC (see Figure 5). Silicon-verified results validated that the NMC scheme achieved a 33% reduction in logic area and 48.8% reduction in power consumption in a 22-nm STT-MRAM processor [31].

3.2.2 *Writing-based scheme*

An alternative solution to sensing-based IMC utilizes a writing (MTJ switching) procedure to perform simultaneous calculation and storage functions within the MRAM bit-cell [68, 76, 78, 79]. It can perform a logic function based on a series of write cycles, thereby alleviating the effect of limited TMR and wide distribution of cell resistance. However, high MTJ switching current and latency remain the limitations of a writing-based scheme.

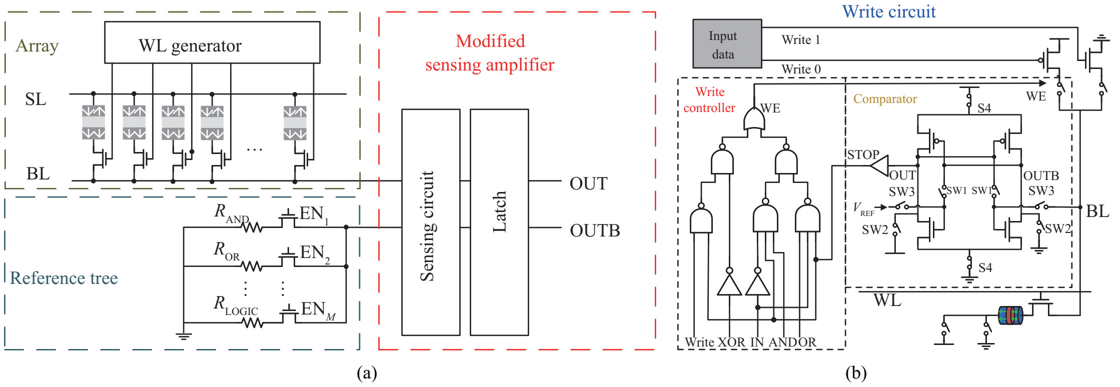

Figure 5 (Color online) (a) A common structure of analog computation based IMC. WL driver, reference tree, and modified sensing amplifier are included. (b) Writing-only IMC circuit, basic Boolean logic functions can be realized.

A write-operation-based IMC encodes the access transistor and MTJ control signals and then records the result into bit-cells during a write operation process [59, 68, 75]; e.g., the logic function $B_{i+1} = AC + \overline{A}B_i$ can be realized. Figure 5(b) presents the write-only IMC scheme. An enhanced self-write termination circuit in MRAM is proposed for the realization of logic functions based on a configurable write controller. Simulation results show a 59.07% energy reduction within a 20-ns write duration.

3.2.3 *Implication logic-based scheme*

Implication logic-based IMCs supply different voltage pulses to enable in-MRAM logic operations; the interconnection of memory bit-cells facilitates multiple parallel operations within the memory array. However, the main drawback with this scheme is that the realization of logic functions is at the expense of iteration, which shows low efficiency. To reduce MRAM energy consumption, a nonvolatile approximate FA was implemented [36]. In this design, input C_i can be eliminated in the SUM operation network (dashed rectangle) with Sum $= A \otimes B$, whereas the C_0 function maintains its accurate computation. According to the experimental results, this approximation originates from an insufficient MTJ write operation in which the B input is a floating state that follows a probability distribution related to V_{dd}. A 79.7% reduction in dynamic power and an acceptable computational accuracy is offered by the scaled V_{dd} (under 0.5 V). Simulation results were obtained using a 28-nm fully depleted silicon-on-insulator (FDSOI process) and a 40-nm MTJ compact model.

3.3 Challenge of in-MRAM computing

3.3.1 *Device-level stability and reliability*

Bi-directional MTJ writing leads to unbalanced switching latency and energy consumption. This latency can be calculated according to different regimes with Sun model ($I > I_{c0}$) and Neel-Brown model ($I < I_{c0}$) [80–82].

$$\tau = \tau_0 \exp\left(\frac{E}{k_B T}\left(1 - \frac{I}{I_{c0}}\right)\right) \quad (I > I_{c0}), \tag{4}$$

$$\frac{1}{\tau} = \left[\frac{2}{C + \ln(\frac{\pi^2 \xi}{4})}\right] \frac{\mu_B P_{\mathrm{ref}}}{e m_m (1 + P_{\mathrm{ref}} P_{\mathrm{free}})} \quad (I < I_{c0}), \tag{5}$$

where τ is the switching time, τ_0 is the attempt period, k_B is the Boltzmann constant, T is the temperature, $C \approx 0.577$ is the Euler's constant, $\xi = \frac{E}{k_B T}$ is the thermal stability factor, e is the elementary charge, m_m is the magnetization moment, and P is the tunneling spin polarizations. For the conventional one-pulse writing, applying the writing pulse with the same duration to all the bit-cells, the writing energy when switching MTJ from AP to P and P to AP can be calculated as

$$\begin{aligned} \mathrm{Energy}_{(P)} &= I_{(\mathrm{AP}\to P)} \tau V_{\mathrm{dd}} + I_P V_{\mathrm{dd}} (T_{\mathrm{pulse}} - \tau) \ \ \text{(When writing `}P\text{')}, \\ \mathrm{Energy}_{(\mathrm{AP})} &= I_{(P\to\mathrm{AP})} \tau V_{\mathrm{dd}} + I_P V_{\mathrm{dd}} (T_{\mathrm{pulse}} - \tau) \ \ \text{(When writing `AP')}, \end{aligned} \tag{6}$$

Table 4 Writing/reading failure mechanisms and key causes

Affect	Mechanisms	Major factors
Read	Decision fault	Process variations, limited TMR, low supply voltage
	Read disturb	Read and write share the same path, growing with technology scaling
	Incorrect read fault	Opposite temperature dependence resistance, parasitic effects, tiny SA sensing margin
	Retention failures	Intrinsic thermal instability, thermal noise
Write	Transition faults	Stochastic nature of write operation, thermal fluctuations
	Coupling faults	Neighboring cells switching
	Write polarization asymmetry	Higher P-AP switching current, varied writing time

where $\mathrm{Energy}_{(P)}$ and $\mathrm{Energy}_{(\mathrm{AP})}$ is the energy consumption during writing AP and P respectively, $I_{(\mathrm{AP}\to P)}$ and $I_{(P\to\mathrm{AP})}$ are the current of writing AP and P, V_{dd} is the supply voltage and T_{pulse} is the applied writing duration. Because of the redundant write time ($T_{\mathrm{pulse}}-\tau$), the extra energy consumption is inevitable for reliable MTJ writing operation.

The main issues in perpendicular MTJ-based MRAM include low magnetoresistance and limited TMR. The simultaneous activation of multiple rows by conventional sensing-based in-MRAM Boolean logic operations relies on the analog signal on BL to perform computation. An adequate difference between R_P and R_{AP} is required to achieve successful 1-bit in-MRAM Boolean logic operations. As the R_P-R_{AP} difference is highly sensitive to process-voltage-temperature (PVT) variation, a low resistance could result in a high column current density, causing a distinct IR-drop along the interconnecting wire and deteriorating the MRAM-sensing margin.

RRAM is a suitable candidate among NVM devices for nonvolatile logic because its high R-ratio facilitates multibit storage and provides a wide sensing margin. TMRs of over 600% have been demonstrated by in-plane anisotropy-based MRAM [83], whereas deficiencies in the reading/writing reliability have been shown by in-plane anisotropy. According to silicon-verified perpendicular MTJ-based STT-MRAM, the R_{AP} is approximately 10 kΩ, while the normal TMR could be around 200%. A 249% TMR using atom-thick W layers and double MgO/CoFeB interfaces was reported by [84], which could be beneficial for future in-MRAM computing.

Another important issue is that the design of hybrid CMOS magnetic circuits requires joint optimization to protect against PVT variations. Differences occur in the physical parameters of MTJs (from bit-cell to bit-cell) in terms of the resistance caused by process variations. Additionally, the R_P and R_{AP} have opposite temperature dependences. At high temperatures, there is a decrease in high resistance and a slight increase in a low resistance, which results in a higher TMR at high temperatures than at low temperatures. Consequently, the sensing margin between adjacent computing states is affected, and the number of operations is limited.

The downscaling of an MTJ device at a critical dimension (CD) of 40 nm produced considerable reductions in average switching time and critical switching current. An MTJ with a sub-40-nm CD enables faster switching and consumes less switching energy [85–87]. Therefore, writing-based in-MRAM computing has attracted considerable attention due to its low costs in terms of iterative write energy and latency.

3.3.2 *Reliability and variability at circuit and system levels*

Performance degradation occurs as a result of unreliable bit-cells a transistor reliability issues [88, 89]. Table 4 lists the writing/reading failure mechanisms and their key causes. An MRAM array may be affected by reading disturbance, which is the corruption of data under the effects of a significant read current across MTJ. However, due to the continuously diminishing MTJ switching current, the difference between the write and read currents is reduced. Therefore, a clamp transistor with BL is required to prevent read disturbance, although the sensing margin is deteriorated further by this setup.

Although the sensing margin can be optimized, several limitations of IMC must be addressed. In particular, a large number of activated WLs generate high currents along the BL (I_{BL}), which produces an inaccurate BL-clamping voltage. A high I_{BL} requires a large array area due to the wide metal lines required to support a high current density. To solve this, Ref. [31] implemented a BL-in/out multibit computing scheme using a single WL-on and input-aware multibit BL clamping.

The use of modified/customized sensing and peripheral circuits unavoidably impacts the layout area. Although computation efficiency is improved, analyses usually ignore the energy consumption of addi-

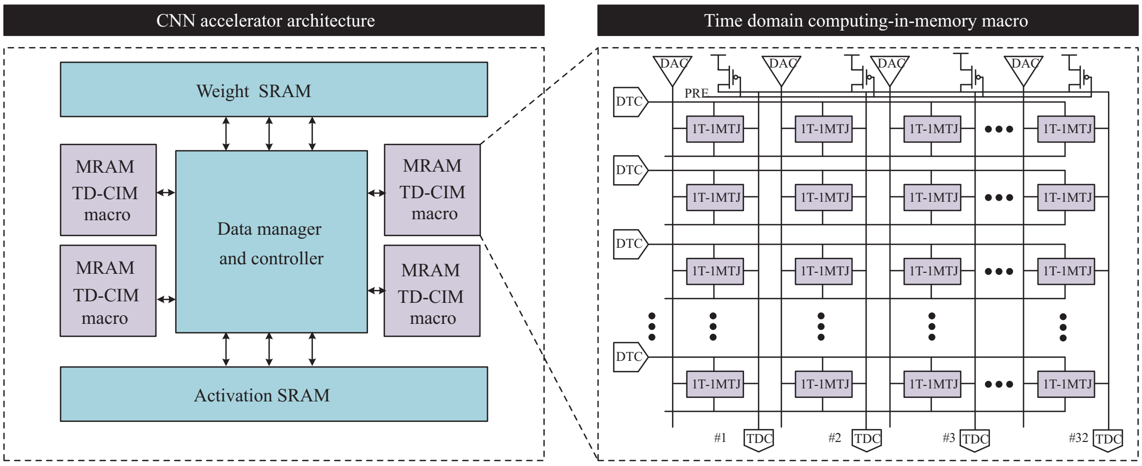

Figure 6 (Color online) Structure of a time-domain computing-in-memory macro. The architecture of the CNN accelerator with 4 CIM macros. Each CIM macro contains an MRAM array. A row-wise digital-time-converter is used to convert an activation into a time pulse signal. Both ADC and DAC are implemented at each column to provide MAC read-out and analog write-in. Similar to prior schemes, global SRAMs are used to store weight and input/output activation data before being fetched into CIM macro. A date manager is used to manage data sequencing and pre-/post-processing.

Table 5 Survey of MRAM based neural network

	CMOS	Tape-out	Memory	Capacity	CIM	Speed up	Energy saving	Area	Application
NIPS'18 [56]	22-nm	✓	STT	40 Mb	✓	N/A	9.9 TOPS/W	N/A	NLP
IEDM'18 [90]	45/28-nm	✓	STT	32 Kb	✓	N/A	82%	N/A	Computer vision
DAC'18 [58]	45-nm	×	SOT	N/A	×	4.3×	67%	N/A	Computer vision
TVLSI'19 [59]	28-nm	×	SOT	2048×256	✓	12.3×	60.8%	N/A	Computer vision
TVLSI'19 [75]	28-nm	×	SOT	N/A	✓	4.7×	29%	N/A	Computer vision
JETCAS'19 [29]	22-nm	×	STT	8 Mb	×	4.85×	83.5%	N/A	Computer vision
DAC'19 [91]	N/A	×	STT	N/A	×	N/A	79.4%	57%	Computer vision
DATE'19 [92]	28-nm	×	SOT	1024×512	✓	2.12×	14%	N/A	Computer vision
ASP-DAC'20 [93]	45-nm	×	SOT	256×512	✓	N/A	63%	7.9%	Computer vision
VLSI'20 [34]	22-nm	✓	SOT	×	✓	N/A	N/A	N/A	DNN

tional control units. To reduce the significant losses in area and energy, a multibit SA was proposed by [31].

3.3.3 *Energy efficiency challenge*

Analog computations based on Kirchhoff's current law could realize Boolean logic operation and memory access [58, 69–72, 74]. Typical analog signals include voltage amplitude, current density, and pulse duration, which are also used for memory access control. Figure 5 illustrates a typical in-MRAM computing scheme, including specified WL/column drivers, a modified SA, a reference tree, and a controller. During the calculation, multiple WLs are activated, and the accumulation results are obtained using the cross-coupled sensing circuit. Increased energy efficiency can be achieved through in-MRAM computation by utilizing the enormous mini-/sub-array bandwidth along with high computational throughputs.

Figure 6 presents an example of a time-domain in-MRAM computing macro. Previous IMC techniques had high quantization errors because ADC block was unable to track bit-cell PVT variations in MRAM arrays. CNN operations using 3 × 3 kernels, 2b-inputs, and 3b-weights utilize the MNIST handwritten digit recognition dataset with the LeNet-5 CNN. A total of 200 test images can be run through the two convolutional and three fully-connected layers. Our work achieves a classification error rate of 1% after the first two convolutional layers and 4% after all the five layers, which demonstrates the ability of the time-domain IMC architecture to compute convolutions. Furthermore, an energy efficiency improvement of 40%–70% was achieved via its parallel in-memory delay computations.

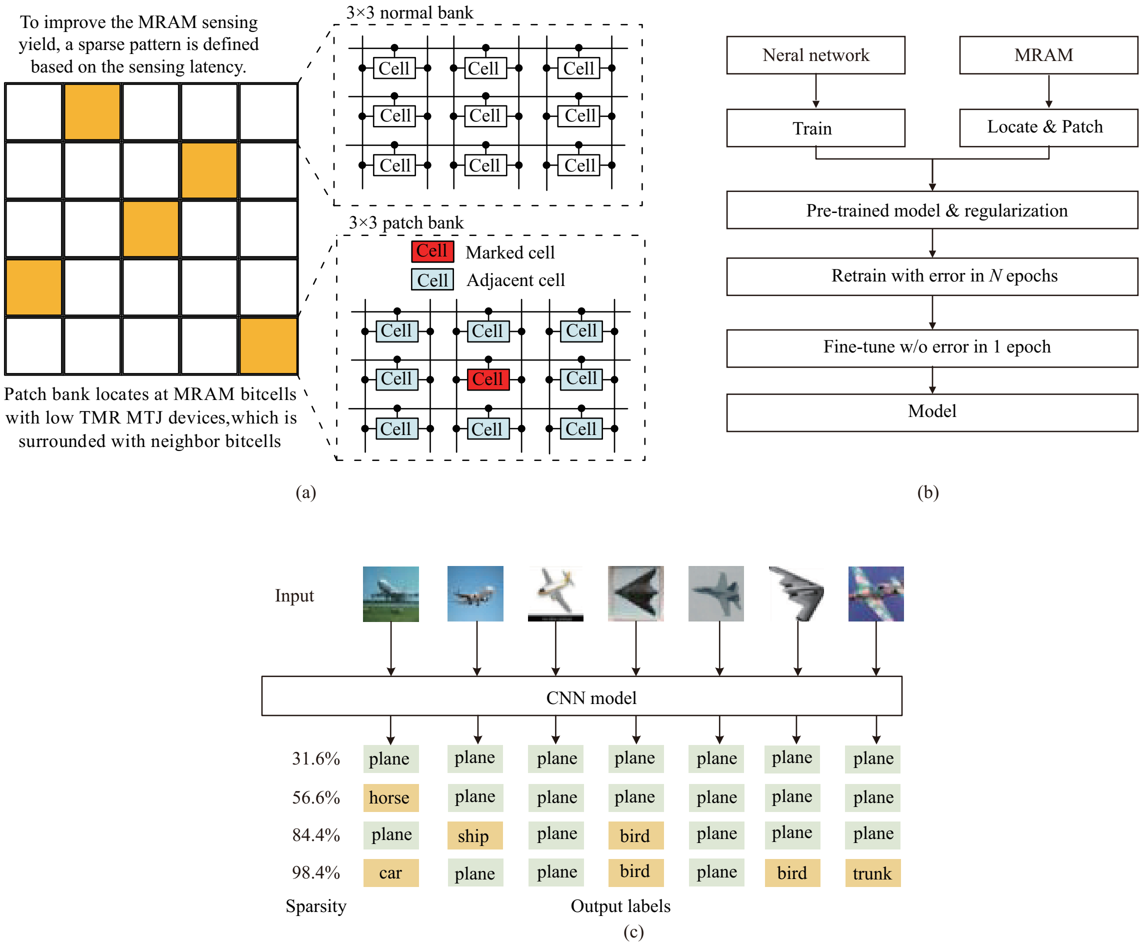
To improve the MRAM sensing yield, a sparse pattern is defined based on the sensing latency.
3×3 normal bank
Cell
3×3 patch bank
Cell Marked cell
Cell Adjacent cell
Patch bank locates at MRAM bitcells with low TMR MTJ devices,which is surrounded with neighbor bitcells
(a)
Neral network
MRAM
Train
Locate & Patch
Pre-trained model & regularization
Retrain with error in N epochs
Fine-tune w/o error in 1 epoch
Model
(b)
Input
CNN model
31.6%
56.6%
84.4%
98.4%
plane
horse
ship
bird
car
trunk
Sparsity
Output labels
(c)

Figure 7 (Color online) Sparse realization in unreliable MRAM for CNN [110]. (a) Patch bank structure of MRAM; (b) the flowchart of sparse realization in unreliable MRAM; (c) the visualization of a retrained model on CIFAR-10. The increased sparsity of in-MRAM can lead to the generation of incorrect labels during classification.

4 In-MRAM computing perspective

Table 5 [29, 34, 56, 58, 59, 75, 90–93] lists the recent neural network implementations. Several in-MRAM computing studies were implemented and verified using a 2x nm CMOS process. Currently, several foundries can customize, design, and fabricate STT-MRAMs at an advanced 2x nm CMOS node. The practical application of MRAMs in IoT scenarios requires fast read speeds and low power consumption [94] (of approximately 1 μA/MHz/b). Embedded STT-MRAMs with a capacity of up to 1 Gb were fabricated for industrial MCU/IoT applications based on a 28-nm FDSOI process, and endurance of 10^{10} was reported [95–97]. The planer FDSOI CMOS process provides tunable energy efficiency, e.g., forward/reverse body biasing, thereby alleviating the previous in-MRAM computational limitations. Conversely, the fin field-effect transistor (FinFET) guides MRAM into a 1x nm node [98–100]. The design space of MRAM arrays and peripheral circuits can be further extended using FDSOI and FinFET; therefore, there is significant potential for foundry-verified in-MRAM computing circuits and systems. The intrinsic stochastic behavior of in-MRAM neuromorphic computing could be an interesting research topic.

A physical mechanism was used to validate the novel efficient switching behavior of MTJ. For example, the performance of writing-based in-MRAM computing could be enhanced by emerging magnetoelectric RAM (MeRAM) with 5-fJ writing energy and 2-Gb/cm^2 bit density [101–103]. A SOT-MRAM with an accessing operational speed of approximately 100 MHz was fabricated in [104], which supports the prospect of a high-speed computing paradigm. Ref. [105] proposed a combination of SOT and VCMA switching with exchange bias; this ultralow power-switching method required a gate voltage of only 0.6 V. Based on these energy-efficient MTJ switching mechanisms, the design space of high-level in-MRAM computing can be explored using the SPICE behavioral model. Another potential solution for high-performance IMC is offered by the joint effect of STT- and SOT-induced switching, which provides high energy efficiency and speed [106, 107]. In [108], efficient in-memory approximate computing for image processing applications was realized using joint magnetization switching mechanisms, including precessional VCMA, STT-assisted precessional VCMA, and SOT-erasing STT programming.

The realizations of IMC at high levels show considerable potential. An 8-bit-string NAND-like spintronics memory structure was proposed that achieved 5× write energy reduction compared with STT-MRAM and a 25% reduction in density compared with SOT-MRAM [109]. Additionally, the co-optimization of design technology is important. Previously, row-wise memory access was achieved, although the proposed design obtains a binary neural network result over many bits, thereby amortizing the accessing cost. Figure 7 [110] illustrates our recent MRAM realization with array sparsity. The simulation results using representative dataset CIFAR-10 confirm that MRAM-sensing operation can be speedup to 6.4× with 84.46% sparsity. With a suitable sparsity selection, unreliable sensing issues can be solved by the proposed training and retraining phases. The MTJ-CMOS process and the co-optimization of memory/IMC circuits and systems are of considerable importance.

5 Conclusion

This paper presented an analysis of state-of-the-art in-MRAM computing with an emphasis on IMC approaches and their implementation with STT-MRAM. It identified the realization of energy-efficient memory access and the provision of efficient computational performance as key considerations. Potential circuit design schemes were reviewed by allocating STT-MRAM into bottom-up hierarchies, and energy efficiency was improved via multimode memory reconfiguration. SPICE-compatible simulations were used to implement and analyze several designs at the device, circuit, and system levels.

Besides the intrinsic energy efficiency obtained from spintronic devices, the key advantage of multimode MRAM reconfiguration is the ability of hierarchical design-space exploration, and the macro adaptation to various energy-constraint scenarios. Designer preference should influence the selection of multimode reconfiguration in MRAM and design strategies. However, to achieve high efficiency, customized MRAM design techniques require the following three improvements. First, the dimensions of the MTJ device (CD) must scale beyond 40 nm. This applies not only to STT but also to other advanced spintronic switching mechanisms, e.g., SOT, VCMA, and several interplay MTJ switching methods. Second, the development of MRAM/in-MRAM computing compatible with front- and back-end design tools remains at an early stage. Finally, in-MRAM computing with dimensional scaling is recommended for co-optimization with MTJ/CMOS processes and new computing architectures.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61904028).

References

1 Wong H S P, Salahuddin S. Memory leads the way to better computing. Nat Nanotech, 2015, 10: 191–194

2 Verma N, Jia H, Valavi H, et al. In-memory computing: advances and prospects. IEEE Solid-State Circ Mag, 2019, 11: 43–55

3 Yu S, Chen P Y. Emerging memory technologies: recent trends and prospects. IEEE Solid-State Circ Mag, 2016, 8: 43–56

4 Yang J, Kong Y Y, Wang Z, et al. Sandwich-RAM: an energy-efficient in-memory BWN architecture with pulse-width modulation. In: Proceedings of IEEE International Solid-State Circuits Conference, 2019. 394–396

5 Tang Y Q, Zhang J T, Verma N. Scaling up in-memory-computing classifiers via boosted feature subsets in banked architectures. IEEE Trans Circ Syst II, 2019, 66: 477–481

6 Chen Y H, Krishna T, Emer J S, et al. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J Solid-State Circ, 2017, 52: 127–138

7 Merrikh-Bayat F, Guo X, Klachko M, et al. High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays. IEEE Trans Neural Netw Learn Syst, 2018, 29: 4782–4790

8 Kent A D, Worledge D C. A new spin on magnetic memories. Nat Nanotech, 2015, 10: 187–191

9 Cho T, Lee Y T, Kim E C, et al. A dual-mode NAND flash memory: 1-Gb multilevel and high-performance 512-Mb single-level modes. IEEE J Solid-State Circ, 2001, 36: 1700–1706

10 Sheu S S, Chang M F, Lin K F, et al. A 4 Mb embedded SLC resistive-RAM macro with 7.2 ns read-write random-access time and 160 ns MLC-access capability. In: Proceedings of IEEE International Solid-State Circuits Conference, 2011. 200–202

11 Wang P Q, Ji Y, Hong C et al. SNrram: an efficient sparse neural network computation architecture based on resistive random-access memory. In: Proceedings of the 55th ACM/ESDA/IEEE Design Automation Conference (DAC), 2018. 1–6

12 Jiang L, Zhao B, Zhang Y T, et al. Improving write operations in MLC phase change memory. In: Proceedings of International Symposium on High-Performance Comp Architecture, 2012. 1–10

13 Ni K, Grisafe B, Chakraborty W, et al. In-memory computing primitive for sensor data fusion in 28 nm HKMG FeFET technology. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018

14 Yoon I, Khan A, Datta S, et al. A FerroFET-based in-memory processor for solving distributed and iterative optimizations via least-squares method. IEEE J Explor Solid-State Comput Dev Circ, 2019, 5: 132–141

15 Wu J Y, Lee M H, Khwa W S, et al. A double-density dual-mode phase change memory using a novel background storage scheme. In: Proceedings of IEEE Symposium on VLSI Technology, 2014

16 Cassinerio M, Ciocchini N, Ielmini D. Logic computation in phase change materials by threshold and memory switching. Adv Mater, 2013, 25: 5975–5980

17 Pozidis H, Papandreou N, Stanisavljevic M et al. Circuit and system-level aspects of phase change memory. IEEE Trans Circ Syst, 2021, 68: 844–850

18 Khwa W S, Chen J J, Li J F, et al. A 65 nm 4 Kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3 ns and 55.8TOPS/W fully parallel product-sum operation for binary DNN edge processors. In: Proceedings of IEEE International Solid-State Circuits Conference, 2018. 496–498

19 Valavi H, Ramadge P J, Nestler E, et al. A 64-Tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute. IEEE J Solid-State Circ, 2019, 54: 1789–1799

20 Su J W, Si X, Chou Y C, et al. A 28 nm 64 Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips. In: Proceedings of IEEE International Solid-State Circuits Conference, 2020. 240–242

21 Dong Q, Sinangil M, Erbagci B, et al. A 351TOPS/W and 372.4GOPS compute-in-memory SRAM macro in 7 nm FinFET CMOS for machine-learning applications. In: Proceedings of IEEE International Solid-State Circuits Conference, 2020. 242–244

22 Jia H Y, Valavi H, Tang Y Q, et al. A programmable heterogeneous microprocessor based on bit-scalable in-memory computing. IEEE J Solid-State Circ, 2020, 55: 2609–2621

23 Chen W H, Lin W J, Lai L Y, et al. A 16 Mb dual-mode ReRAM macro with sub-14 ns computing-in-memory and memory functions enabled by self-write termination scheme. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2017

24 Liu Q, Gao B, Yao P, et al. A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing. In: Proceedings of IEEE International Solid-State Circuits Conference, 2020. 500–502

25 Xue C X, Huang T Y, Liu J S, et al. A 22 nm 2 Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit MAC computing for tiny AI edge devices. In: Proceedings of IEEE International Solid-State Circuits Conference, 2020. 244–246

26 Natsui M, Suzuki D, Sakimura N, et al. Nonvolatile logic-in-memory LSI using cycle-based power gating and its application to motion-vector prediction. IEEE J Solid-State Circ, 2015, 50: 476–489

27 Hanyu T, Endoh T, Suzuki D, et al. Standby-power-free integrated circuits using MTJ-based VLSI computing. Proc IEEE, 2016, 104: 1844–1863

28 Wu M H, Hong M C, Chang C C, et al. Extremely compact integrate-and-fire STT-MRAM neuron: a pathway toward all-spin artificial deep neural network. In: Proceedings of IEEE Symposium on VLSI Technology, 2019. 34–35

29 Yoon I, Anwar M A, Joshi R V, et al. Hierarchical memory system with STT-MRAM and SRAM to support transfer and real-time reinforcement learning in autonomous drones. IEEE J Emerg Sel Top Circ Syst, 2019, 9: 485–497

30 Chang C C, Wu M H, Lin J W, et al. NV-BNN: an accurate deep convolutional neural network based on binary STT-MRAM for adaptive AI edge. In: Proceedings of the 56th Annual Design Automation Conference, 2019. 1–6

31 Chang T, Chiu Y, Lee C, et al. A 22 nm 1 Mb 1024b-read and near-memory-computing dual-mode STT-MRAM macro with 42.6 GB/s read bandwidth for security-aware mobile devices. In: Proceedings of IEEE International Solid-State Circuits Conference, 2020. 224–226

32 Yang L, Angizi S, Fan D L, et al. A flexible processing-in-memory accelerator for dynamic channel-adaptive deep neural networks. In: Proceedings of the 25th Asia and South Pacific Design Automation Conference (ASP-DAC), 2020. 313–318

33 Ostwal V, Zand R, DeMara R, et al. A novel compound synapse using probabilistic spin-orbit-torque switching for MTJ-based deep neural networks. IEEE J Explor Solid-State Comput Dev Circ, 2019, 5: 182–187

34 Doevenspeck J, Garello K, Verhoef B, et al. SOT-MRAM based analog in-memory computing for DNN inference. In: Proceedings of IEEE Symposium on VLSI Technology, 2020

35 Li S C, Xu C, Zou Q S, et al. Pinatubo: a processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories. In: Proceedings of the 53rd ACM/EDAC/IEEE Design Automation Conference (DAC), 2016. 1–6

36 Cai H, Wang Y, de Barros N L A, et al. Robust ultra-low power non-volatile logic-in-memory circuits in FD-SOI technology. IEEE Trans Circ Syst I, 2017, 64: 847–857

37 Indiveri G, Liu S C. Memory and information processing in neuromorphic systems. Proc IEEE, 2015, 103: 1379–1397

38 Liu B, Cai H, Wang Z, et al. A 22 nm, 10.8μW/15.1μW dual computing modes high power-performance-area efficiency domained background noise aware keyword- spotting processor. IEEE Trans Circ Syst I, 2020, 67: 4733–4746

39 Zhang Y Q, Xu L, Dong Q, et al. Recryptor: a reconfigurable cryptographic cortex-M0 processor with in-memory and near-memory computing for IoT security. IEEE J Solid-State Circ, 2018, 53: 995–1005

40 Srinivasan G, Wijesinghe P, Sarwar S S, et al. Significance driven hybrid 8T-6T SRAM for energy-efficient synaptic storage in artificial neural networks. In: Proceedings of Design, Automation and Test in Europe Conference and Exhibition (DATE), 2016. 151–156

41 Bortolotti D, Mangia M, Bartolini A, et al. An ultra-low power dual-mode ECG monitor for healthcare and wellness. In: Proceedings of Design, Automation and Test in Europe Conference and Exhibition (DATE), 2015. 1611–1616

42 Le Gallo M, Sebastian A, Mathis R, et al. Mixed-precision in-memory computing. Nat Electron, 2018, 1: 246–253

43 Chen Y S, Wang D Y, Hsin Y C, et al. On the hardware implementation of MRAM physically unclonable function. IEEE Trans Electron Dev, 2017, 64: 4492–4495

44 Yang K Y, Dong Q, Wang Z H, et al. A 28 nm integrated true random number generator harvesting entropy from MRAM. In: Proceedings of IEEE Symposium on VLSI Circuits, 2018. 171–172

45 Choi W, Lv Y, Kim J, et al. A magnetic tunnel junction based true random number generator with conditional perturb and real-time output probability tracking. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2014

46 Vincent A F, Larroque J, Locatelli N, et al. Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems. IEEE Trans Biomed Circ Syst, 2015, 9: 166–174

47 Grollier J, Querlioz D, Stiles M D. Spintronic nanodevices for bioinspired computing. Proc IEEE, 2016, 104: 2024–2039

48 Pedretti G, Bianchi S, Milo V, et al. Modeling-based design of brain-inspired spiking neural networks with RRAM learning synapses. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2014

49 Hu M, Strachan J P, Li Z Y, et al. Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication. In: Proceedings of the 53rd ACM/EDAC/IEEE Design Automation Conference

(DAC), 2016. 1–6

50 Shafiee A, Nag A, Muralimanohar N, et al. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In: Proceedings of the 43rd Annual International Symposium on Computer Architecture (ISCA), 2016. 14–26

51 Ambrogio S, Narayanan P, Tsai H, et al. Equivalent-accuracy accelerated neural-network training using analogue memory. Nature, 2018, 558: 60–67

52 Wang Z, Joshi S, Savel'ev S, et al. Fully memristive neural networks for pattern classification with unsupervised learning. Nat Electron, 2018, 1: 137–145

53 Chen W H, Li K X, Lin W Y, et al. A 65 nm 1 Mb nonvolatile computing-in-memory ReRAM macro with sub-16 ns multiply-and-accumulate for binary DNN AI edge processors. In: Proceedings of IEEE International Solid-State Circuits Conference, 2018. 494–496

54 Xue C X, Chen W H, Liu J S, et al. A 1 Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge processors. In: Proceedings of IEEE International Solid-State Circuits Conference, 2019. 388–390

55 Mochida R, Kouno K, Hayata Y, et al. A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture. In: Proceedings of IEEE Symposium on VLSI Technology, 2018. 175–176

56 Sun B H, Liu D, Yu L, et al. MRAM co-designed processing-in-memory CNN accelerator for mobile and IoT applications. 2018. ArXiv:1811.12179

57 Bocquet M, Hirztlin T, Klein J, et al. In-memory and error-immune differential RRAM implementation of binarized deep neural networks. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018

58 Angizi S, He Z Z, Rakin A S, et al. CMP-PIM: an energy-efficient comparator-based processing-in-memory neural network accelerator. In: Proceedings of the 55th Annual Design Automation Conference, 2018. 1–6

59 Chang L, Ma X, Wang Z H, et al. PXNOR-BNN: in/with spin-orbit torque MRAM preset-XNOR operation-based binary neural networks. IEEE Trans VLSI Syst, 2019, 27: 2668–2679

60 Patil A, Hua H C, Gonugondla S, et al. An MRAM-based deep in-memory architecture for deep neural networks. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2019. 1–6

61 Liu Q, Gao B, Yao P, et al. A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing. In: Proceedings of IEEE International Solid-State Circuits Conference, 2020. 500–502

62 Julliere M. Tunneling between ferromagnetic films. Phys Lett A, 1975, 54: 225–226

63 Wang M X, Cai W L, Cao K H, et al. Current-induced magnetization switching in atom-thick tungsten engineered perpendicular magnetic tunnel junctions with large tunnel magnetoresistance. Nat Commun, 2018, 9: 671

64 Wang Y, Cai H, Naviner L A B, et al. Compact model of dielectric breakdown in spin-transfer torque magnetic tunnel junction. IEEE Trans Electron Dev, 2016, 63: 1762–1767

65 Chappert C, Fert A, van Dau F N. The emergence of spin electronics in data storage. Nat Mater, 2007, 6: 813–823

66 Zhou Y L, Cai H, Xie L, et al. A self-timed voltage-mode sensing scheme with successive sensing and checking for STT-MRAM. IEEE Trans Circ Syst I, 2020, 67: 1602–1614

67 Zhou Y L, Cai H, Liu B, et al. MTJ-LRB: proposal of MTJ-based loop replica bitline as MRAM device-circuit interaction for PVT-robust sensing. IEEE Trans Circ Syst II, 2020, 67: 3352–3356

68 Zhang H, Kang W, Cao K H, et al. Spintronic processing unit in spin transfer torque magnetic random access memory. IEEE Trans Electron Dev, 2019, 66: 2017–2022

69 Zhang L Y, Deng E, Cai H, et al. A high-reliability and low-power computing-in-memory implementation within STT-MRAM. MicroElectron J, 2018, 81: 69–75

70 Jain S, Ranjan A, Roy K, et al. Computing in memory with spin-transfer torque magnetic RAM. IEEE Trans VLSI Syst, 2018, 26: 470–483

71 Dou C M, Chen W H, Xue C X, et al. Nonvolatile circuits-devices interaction for memory, logic and artificial intelligence. In: Proceedings of IEEE Symposium on VLSI Technology, 2018. 171–172

72 Pan Y, Ouyang P, Zhao Y, et al. A MLC STT-MRAM based computing in-memory architec-ture for binary neural. In: Proceedings of IEEE International Magnetics Conference (INTERMAG), 2018

73 Zhang H, Kang W, Wang L Z, et al. Stateful reconfigurable logic via a single-voltage-gated spin hall-effect driven magnetic tunnel junction in a spintronic memory. IEEE Trans Electron Dev, 2017, 64: 4295–4301

74 Angizi S, Sun J, Zhang W, et al. AlignS: a processing-in-memory accelerator for DNA short read alignment leveraging SOT-MRAM. In: Proceedings of the 56th ACM/IEEE Design Automation Conference (DAC), 2019. 1–6

75 Chang L, Ma X, Wang Z H, et al. DASM: data-streaming-based computing in nonvolatile memory architecture for embedded system. IEEE Trans VLSI Syst, 2019, 27: 2046–2059

76 Zhang H, Kang W, Wu B, et al. Spintronic processing unit within voltage-gated spin Hall effect MRAMs. IEEE Trans Nanotechnol, 2019, 18: 473–483

77 Cai H, Han M L, Zhou Y L, et al. Triple sensing current margin for maintainable MRAM Yield at sub-100% tunnel magnetoresistance ratio. IEEE Trans Magnetic, 2021, 57: 3400305

78 Cao K H, Cai W L, Liu Y Z, et al. In-memory direct processing based on nanoscale perpendicular magnetic tunnel junctions. Nanoscale, 2018, 10: 21225–21230

79 Mahmoudi H, Windbacher T, Sverdlov V, et al. Implication logic gates using spin-transfer-torque-operated magnetic tunnel junctions for intrinsic logic-in-memory. Solid-State Electron, 2013, 84: 191–197

80 Koch R H, Katine J A, Sun J Z. Time-resolved reversal of spin-transfer switching in a nanomagnet. Phys Rev Lett, 2004, 92: 088302

81 Worledge D C, Hu G, Abraham D W, et al. Spin torque switching of perpendicular Ta|CoFeB|MgO-based magnetic tunnel junctions. Appl Phys Lett, 2011, 98: 022501

82 Heindl R, Rippard W H, Russek S E, et al. Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions. J Appl Phys, 2011, 109: 073910

83 Ikeda S, Hayakawa J, Ashizawa Y, et al. Tunnel magnetoresistance of 604% at 300K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature. Appl Phys Letter, 2008, 93: 082508

84 Wang M X, Cai W L, Cao K H, et al. Current-induced magnetization switching in atom-thick tungsten engineered perpendicular magnetic tunnel junctions with large tunnel magnetoresistance. Nat Commun, 2018, 9: 671

85 Song J, Dixit H, Behin-Aein B, et al. Impact of process variability on write error rate and read disturbance in STT-MRAM devices. IEEE Trans Magn, 2020, 56: 1–11

86 Wang H T, Kang W, Zhang Y G, et al. Modeling and evaluation of sub-10-nm shape perpendicular magnetic anisotropy magnetic tunnel junctions. IEEE Trans Electron Dev, 2018, 65: 5537–5544

87 Iba Y, Takahashi A, Hatada A, et al. A highly scalable STT-MRAM fabricated by a novel technique for shrinking a magnetic tunnel junction with reducing processing damage. In: Proceedings of IEEE Symposium on VLSI Technology, 2014

88 Cai H, Wang Y, de Barros N L A, et al. Addressing failure and aging degradation in MRAM/MeRAM-on-FDSOI integration. IEEE Trans Circ Syst I, 2019, 66: 239–250

89 Lin I C, Law Y K, Xie Y. Mitigating BTI-induced degradation in STT-MRAM sensing schemes. IEEE Trans VLSI Syst, 2018, 26: 50–62

90 Xu N, Lu Y, Qi W Y, et al. STT-MRAM design technology co-optimization for hardware neural networks. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018

91 Li H T, Bhargav M, Whatmough P, et al. On-chip memory technology design space explorations for mobile deep neural network accelerators. In: Proceedings of the 56th Annual Design Automation Conference, 2019. 1–6

92 Chang L, Ma X, Wang Z, et al. CORN: in-buffer computing for binary neural network. In: Proceedings of Design, Automation and Test in Europe Conference and Exhibition (DATE), 2019. 384–389

93 Yang L, Angizi S, Fan D. A flexible processing-in-memory accelerator for dynamic channel-adaptive deep neural networks. In: Proceedings of ASP-DAC, 2020. 313–318

94 Chih Y, Shih Y, Lee C, et al. A 22 nm 32 Mb embedded STT-MRAM with 10 ns read speed, 1M cycle write endurance, 10 years retention at 150°C and high immunity to magnetic field interference. In: Proceedings of IEEE International Solid-State Circuits Conference, 2020. 222–224

95 Lee K, Bak J, Kim Y, et al. 1 Gbit high density embedded STT-MRAM in 28 nm FDSOI technology. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018

96 Lee Y, Song Y, Kim J, et al. Embedded STT-MRAM in 28-nm FDSOI logic process for industrial MCU/IoT application. In: Proceedings of IEEE Symposium on VLSI Technology, 2018. 181–182

97 Boujamaa E, Ali S, Wandji S, et al. A 14.7 Mb/mm^2 28 nm FDSOI STT-MRAM with current starved read path, 52 Ω/sigma offset voltage sense amplifier and fully trimmable CTAT reference. In: Proceedings of IEEE Symposium on VLSI Technology, 2020. 1–12

98 Wei L, Alzate J, Arslan U, et al. A 7 Mb STT-MRAM in 22FFL FinFET technology with 4 ns read sensing time at 0.9 V using write-verify-write scheme and offset-cancellation sensing technique. In: Proceedings of IEEE International Solid-State Circuits Conference, 2019. 214–216

99 Golonzka O, Alzate J, Arslan U, et al. MRAM as embedded non-volatile memory solution for 22FFL FinFET technology. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018

100 Huynh-Bao T, Veloso A, Sakhare S, et al. Process, circuit and system co-optimization of wafer level co-integrated FinFET with vertical nanosheet selector for STT-MRAM applications. In: Proceedings of the 56th Annual Design Automation Conference, 2019. 1–6

101 Wang K L, Lee H, Amiri P K. Magnetoelectric random access memory-based circuit design by using voltage-controlled magnetic anisotropy in magnetic tunnel junctions. IEEE Trans Nanotechnol, 2015, 14: 992–997

102 Noguchi H, Ikegami K, Abe K, et al. Novel voltage controlled MRAM (VCM) with fast read/write circuits for ultra large last level cache. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018

103 Li X, Lee A, Razavi S A, et al. Voltage-controlled magnetoelectric memory and logic devices. MRS Bull, 2018, 43: 970–977

104 Natsui M, Tamakoshi A, Honjo H, et al. Dual-port SOT-MRAM achieving 90-MHz read and 60-MHz write operations under field-assistance-free condition. IEEE J Solid-State Circ, 2021, 56: 1116–1128

105 Peng S Z, Lu J Q, Li W X, et al. Field-free switching of perpendicular magnetization through voltage-gated spin-orbit torque. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2019

106 Wang M X, Cai W L, Zhu D Q, et al. Field-free switching of a perpendicular magnetic tunnel junction through the interplay of spin-orbit and spin-transfer torques. Nat Electron, 2018, 1: 582–588

107 Grimaldi E, Krizakova V, Sala G, et al. Single-shot dynamics of spin-orbit torque and spin transfer torque switching in three-terminal magnetic tunnel junctions. Nat Nanotechnol, 2020, 15: 111–117

108 Cai H, Jiang H L, Zhou Y L, et al. Interplay bitwise operation in emerging MRAM for efficient in-memory computing. CCF Trans HPC, 2020, 2: 282–296

109 Wang Z H, Zhang L, Wang M X, et al. High-density NAND-like spin transfer torque memory with spin orbit torque erase operation. IEEE Electron Dev Lett, 2018, 39: 343–346

110 Cai H, Chen J T, Zhou Y L, et al. Sparse realization in unreliable spin-transfer-torque RAM for convolutional neural network. IEEE Trans Magn, 2021, 57: 1–5