Design of Ultracompact Content Addressable Memory Exploiting 1T-1MTJ Cell

Cheng Zhuo, *Senior Member, IEEE*, Zeyu Yang, Kai Ni, *Member, IEEE*, Mohsen Imani, *Member, IEEE*, Yuxuan Luo, *Member, IEEE*, Shaodi Wang, *Member, IEEE*, Deming Zhang, *Member, IEEE*, and Xunzhao Yin, *Member, IEEE*

***Abstract*—Content addressable memories (CAMs) are a promising category of computing-in-memory (CiM) elements that can perform highly parallel and efficient search operations for routers, pattern matching, and other data-intensive applications. Various magnetic tunnel junction (MTJ)-based CAM designs have been proposed to realize zero standby power and high-performance search. However, due to the relatively small tunnel magneto-resistance (TMR) ratio, MTJ-based CAMs require extra transistors and differential MTJ branches to distinguish between the parallel and anti-parallel resistance states, resulting in significant area and energy overhead. In this article, we propose a device-circuit co-design approach for an ultracompact CAM design by only exploiting a 1T-1MTJ structure in each cell. We propose a 2-step search scheme to enable the parallel in-memory search operation across the proposed CAM array and demonstrate the sufficient sensing margin of the array in a successful search operation. Evaluation results suggest that our proposed 1T-1MTJ-based CAM design improves 179×/301× area efficiency compared with the state-of-the-art 15T-4MTJ/20T-6MTJ CAM design. Application benchmarking on hyperdimensional computing (HDC) inference shows a 54.6×/12.8× speedup compared with GPU/20T-6MTJ CAM-based approaches.**

***Index Terms*—Computing-in-memory (CiM), content addressable memory (CAM), hyperdimensional computing (HDC), magnetic tunneling junction (MTJ).**

Manuscript received 31 March 2022; revised 5 July 2022; accepted 25 August 2022. Date of publication 6 September 2022; date of current version 21 April 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFE0126300; in part by the Zhejiang Provincial Key Research and Development Program under Grant 2022C01232; in part by NSF under Grant LQ21F040006 and Grant LD21F040003; in part by NSFC under Grant 62104213, Grant 92164203, and Grant 61901017; and in part by Zhejiang Lab under Grant 2021MD0AB02. This article was recommended by Associate Editor A. Gamatie. *(Corresponding author: Xunzhao Yin.)*

Cheng Zhuo and Zeyu Yang are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China.

Kai Ni is with the Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623 USA.

Mohsen Imani is with the Department of Computer Science, University of California at Irvine, Irvine, CA 92697 USA.

Yuxuan Luo is with the School of Micronanotelectronics, Zhejiang University, Hangzhou 310058, China.

Shaodi Wang is with Witmem Company Ltd., Beijing 100083, China.

Deming Zhang is with the Fert Beijing Research Institute, School of Electronics and Information Engineering, Beihang University, Beijing 100191, China.

Xunzhao Yin is with the College of ISEE, Zhejiang University, Zhejiang University, Hangzhou 310058, China, and also with the Zhejiang Lab, Zhejiang University, Hangzhou 310058, China (e-mail: xzyin1@zju.edu.cn).

Digital Object Identifier 10.1109/TCAD.2022.3204515

I. Introduction

As THE era of big data approaches, a growing number of data-intensive applications call for energy-efficient computing-in-memory (CiM) hardware with parallel data processing capabilities, especially the parallel search functionality, to overcome the so-called "memory wall" bottleneck in Von Neumann machines [1], [2]. As a promising CiM alternative, content addressable memory (CAM) addresses the memory wall issue by enabling the content addressing property and parallel search functions over the stored memory array given an input query, thus, has been utilized for pattern matching, IP routers and advanced machine learning models, etc. [3], [4], [5], [6], [7].

However, conventional CMOS-based CAMs suffer from high leakage power and low area density, which has become a major concern for date-intensive applications as the CMOS technology scales down [8]. To combat this issue, researchers are looking for device-level solutions to build compact and efficient CAM arrays. Recently, a number of efficient CAM designs have been proposed based on emerging nonvolatile memory (NVM) devices with nonvolatile storage, near-zero leakage power, high storage density, and high-speed state switching properties, such as magnetic tunnel junction (MTJ) [9], [10], [11], [12], [13], resistive RAM (ReRAM) [14], [15], [16], [17], and Ferroelectric FET (FeFET) [18], [19], [20], [21], [22], [23]. Among these CAM designs, MTJ-based CAMs consisting of MTJs and CMOS transistors stand out due to its superior endurance and data retention, high integration density, high access speed, and compatibility with CMOS technology [24].

Current MTJ-based CAM designs can be classified into two categories. The CAM designs from [25], [26], [27], [28], [29] have realized the search operations based on a voltage-dividing sensing scheme, thus are categorized as voltage dividing CAMs. While these designs consume a small number of transistors, the major problem of such approach is the low search reliability caused by the limited resistance ratio of MTJ. To enhance the reliability of the above designs, another group of designs called latch-based CAMs employ extra transistors to distinguish between the stored parallel (P) and anti-parallel (AP) resistance states of MTJs using a differential sensing scheme and positive feedback loop [30], [31], [32], [33]. Such a design methodology can enable a larger sensing margin, however, at the cost of significant area

overhead. Both types of MTJ-based CAM designs focus on the CAM cell structure innovations, and employ a number of transistors and MTJ pairs for complementary data storage to facilitate search functionality, incurring large area overhead and thus failing to leverage the compact and CMOS compatibility advantages of MTJ devices. Therefore, it is critical to propose a novel MTJ-based CAM design methodology that addresses the significant area overhead issue, while maintaining the high search reliability, high speed, and energy efficiency.

In this article, we propose a novel and general device-circuit co-design approach for compact CAM design, which leverages the small footprint and nonvolatility storage of NVMs with limited resistance ON/OFF ratios. The primary contributions of this article are as follows.

1) Instead of employing MTJ pairs for complementary storage and differential sensing, we propose an ultracompact MTJ CAM design by exploiting only a 1T-1MTJ structure in each CAM Cell, which is the mainstream cell structure for MTJ products.
2) A 2-step search scheme is then proposed to enable the parallel in-memory search operation, with reference columns and rows employed in the array to facilitate the search operation of the proposed CAM array.
3) To address the weak sensing margin caused by the limited resistance ratio of MTJ devices, we propose to use a voltage-based 2-stage sense amplifier (SA) to guarantee the functionality and the scalability of the proposed 1T-1MTJ-based CAM array.
4) We propose a segmentation scheme for the 1T-1MTJ CAM array to ensure the reliable search function and improve the search performance when executing the long-word query search.
5) Evaluation results at array level and application level demonstrate the efficiency of our proposed 1T-1MTJ-based CAM design over other state-of-the-art MTJ-based CAM designs.

The proposed 1T-1MTJ CAM design has been evaluated and compared with other state-of-the-art MTJ-based CAM designs in terms of performance, area and energy efficiency, etc. The evaluation results show that our proposed 1T-1MTJ CAM design has 179×/301× better area efficiency compared with the latest 15T-4MTJ/20T-6MTJ CAM design while remaining the energy efficiency. Moreover, the proposed CAM design has also been benchmarked in hyperdimensional computing (HDC)-based applications, suggesting 54.6×/12.8× speed up over GPU/20T-6MTJ CAM-based approaches.

The remainder of this article is organized as follows. Section II briefly introduces the spin transfer torque (STT)-based MTJ and current state-of-the-art MTJ-based CAM designs. Section III presents the proposed 1T-1MTJ CAM design, along with the write/search operation and further the segmentation scheme. Section IV validates the functionality of the proposed design. Section V presents the scalability, reliability analysis, and performance evaluations. Section VI further benchmarks the proposed approach in IIDC applications. Finally, Section VII concludes this article.

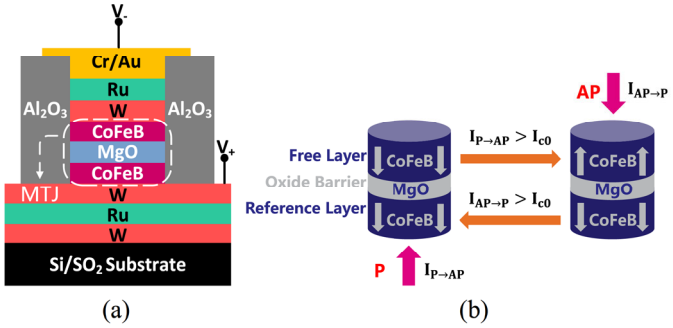


Fig. 1. (a) Structure of the MTJ device. (b) Illustration of STT write mechanism. To change from P to AP, the write current should be larger than the critical current I_{c0} and be flowed from pinned layer to free layer. For changing AP to P, the write current is reversed.

II. Preliminaries and Related Works

A. Spin Transfer Torque-Based Magnetic Tunnel Junction

Fig. 1(a) shows the structure of the MTJ with perpendicular magnetic anisotropy (PMA) we adopt in our design [34], which is composed of two ferromagnetic (FM) layers (e.g., CoFeB) and an ultrathin oxide barrier layer (e.g., MgO). The magnetization of one FM layer (i.e., pinned layer) is fixed in one direction, while that of the other FM layer (i.e., free layer) is reversible. The two resistance states of MTJ device are realized by aligning the relative magnetization orientation of the two FM layers. With P magnetization orientation or AP orientation of the two layers, the MTJ device can encode binary value "0" or "1" into the device states, i.e., low resistance R_{P} or high resistance R_{AP}, respectively. The resistance difference is denoted as tunnel magneto-resistance ratio (TMR $= (R_{\mathrm{AP}} - R_{\mathrm{P}})/R_{\mathrm{P}}$). This TMR ratio can reach 249% at room temperature [34]. Due to the relatively low critical current, high write and read speed, STT-based MTJ devices are considered as one of the most promising NVM technologies for building CiM elements. Fig. 1(b) illustrates the write mechanisms of the MTJ device [35]. It can be seen that the device can switch between R_{AP} and R_{P} depending on the amplitude and the direction of the applied current.

B. MTJ-Based CAM Designs

Since the conventional CMOS CAM design as shown in Fig. 2 suffers from large area overhead and high leakage current, thus, large standby power, NVMs have been extensively proposed to address aforementioned issues. By exploiting the nonvolatility, current drive write and read mechanism, long retention time and technology compatibility with CMOS, etc. properties, various MTJ-based CAM designs have been proposed to achieve high-speed search, compact design overhead, and close-to-zero standby power. Fig. 3 shows the schematics of two types of conventional MTJ-based CAM designs, where complementary data are stored in the MTJ device pairs. The MTJ device pairs have formed the basic cores of CAM cells based on voltage-dividing scheme [25], [26], [27], [28], [29] and latch structure [30], [31], [32], [33], respectively. The voltage-dividing-based 9T-2MTJ CAM design [29] as shown in Fig. 3(a) generates

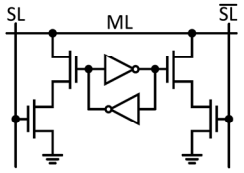


Fig. 2. Schematic of the conventional CMOS CAM design [36].

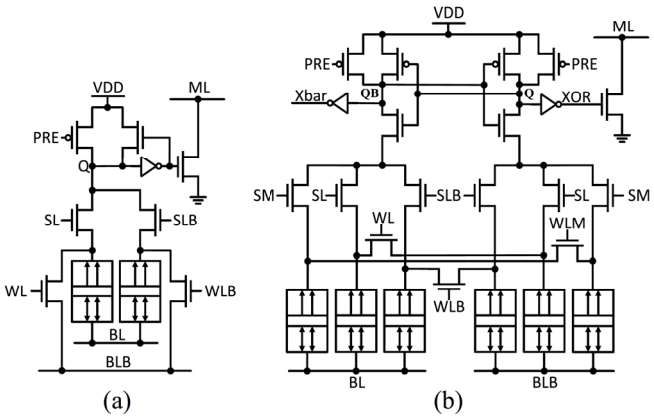
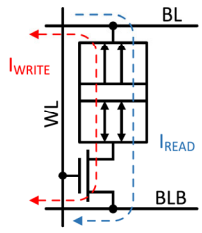


Fig. 3. Schematics of two types of conventional MTJ-based CAM designs. (a) Voltage-dividing-based 9T-2MTJ CAM cell [29]. (b) Latch-based 20T-6MTJ CAM cell [33].

the XNOR output (node Q) of input data and stored data by forming a voltage-dividing network between the pMOS and the connected MTJ, and then depending on the XNOR output, the access transistor associated with the matchline (ML) is activated to discharge the ML upon a mismatch, or turned off to keep the ML high voltage, indicating a match. Though the voltage-dividing-based CAM design consumes less number of transistors, it suffers from the low search function reliability due to the limited resistance ON/OFF ratio of MTJ devices. The latch-based 20T-6MTJ CAM design [33] as shown in Fig. 3(b) has demonstrated higher search reliability to the limited MTJ TMR and higher search energy efficiency by exploiting a differential sensing and positive feedback of two cross-coupled inverters. However, such designs suffer from significant area overhead.

Meanwhile, both types of the MTJ CAM designs employ MTJ devices in pairs to store complementary logic in the devices, which is a typical store and sense scheme to address the limited resistance ON/OFF ratios of MTJ devices. However, such storage and sensing scheme requires large number of transistors and consumes significant area overhead, thus, inhibiting the compact structure property of MTJs. Although, Matsunaga *et al.* [10] presented a 1T-1MTJ CAM design to achieve the high density, its search operation is performed in a bit-serial manner, which seriously affects the search speed.

In this work, instead of employing more than one MTJ device pair in each cell for the complementary storage and differential sensing mechanism, we propose to exploit just a 1T-1MTJ structure in each CAM cell and a voltage-based differential SA to build a novel MTJ-based CAM array for parallel search with ultracompact area overhead, and comparable energy efficiency and search function reliability with other state-of-the-art MTJ CAM arrays.



Operations	Write '1'	Write '0'	Read
WL	$V_{EN-WRITE}$	$V_{EN-WRITE}$	$V_{EN-READ}$
BL	V_{WRITE}	GND	I_{READ}
BLB	GND	V_{WRITE}	GND

(b)

Fig. 4. (a) 1T-1MTJ cell structure. (b) Biasing conditions for 1T-1MTJ operations.

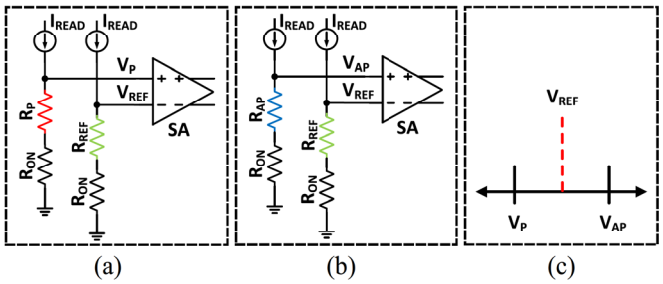


Fig. 5. Idea of voltage comparison between (a) V_P and V_{REF}, and (b) V_{AP} and V_{REF}. (c) Relationship between V_P, V_{AP}, and V_{REF}. The activated access transistor, the reference transistor, the MTJ in P state and the MTJ in AP state can be equivalent as R_{ON}, R_{REF}, R_P, and R_{AP}, respectively.

III. 1T-1MTJ-Based CAM Design

Here, we present the proposed ultracompact 1T-1MTJ-based CAM design. We first describe the basics of the 1T-1MTJ storage cell, and then discusses the operation principles of our proposed CAM design along with the general array structure. Finally, we conduct the segmented scheme for the 1T-1MTJ CAM array to ensure the search reliability for a long search query.

A. 1T-1MTJ Cell

Fig. 4(a) shows the 1T-1MTJ cell structure, which consists of an MTJ and a transistor. The cell is correspondingly controlled by bitlines (BL and BLB) and wordline (WL). The biasing conditions for cell write and read operations are presented in Fig. 4(b). During a write, the WL is set to $V_{EN-WRITE}$ to turn on the access transistor. To write "1" into the cell, V_{WRITE} and GND are applied to BL and BLB, respectively, to form a write current (Iwrite) exceeding the critical current as shown in Fig. 4(a). The BL and BLB bias are reversed for writing bit "0." Depending on the written data, the magnetization orientation of the free layer is switched between P and AP states, thus, achieving the write operation.

For the cell read operation, the WL is set to $V_{EN-READ}$ to turn on the access transistor, and a read current (I_{READ}) is applied on the BL, resulting in a read voltage at BL. The voltage at BL is then sensed and compared with a reference voltage (V_{REF}) generated by a reference cell, using a voltage-based SA. Fig. 5 shows the idea of voltage comparison for the cell read. Note that the equivalent reference resistance R_{REF} is realized by a reference transistor, which is biased to generate a reference voltage V_{REF} between the read voltages developed

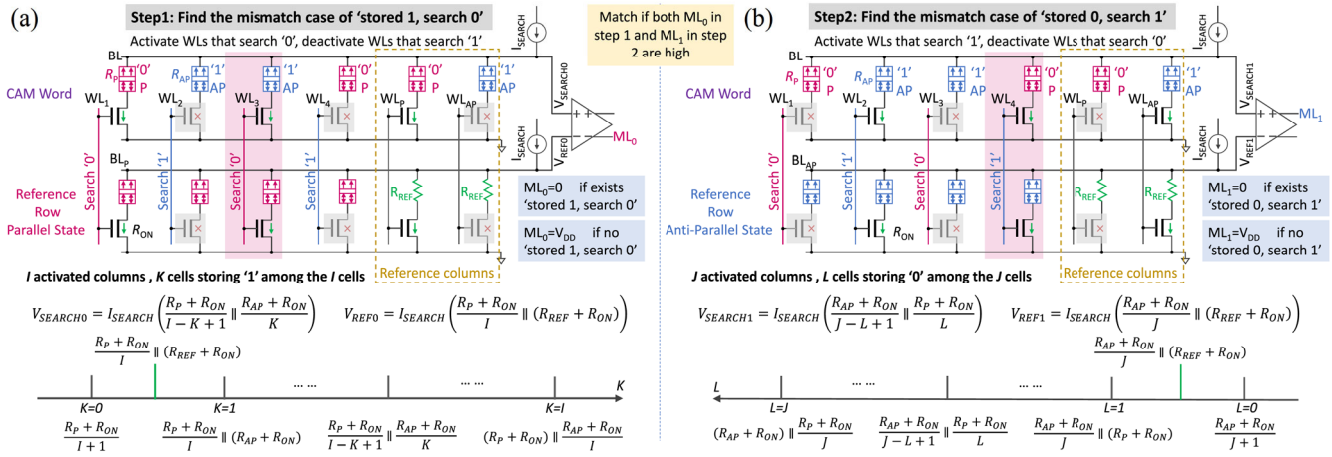


Fig. 6. Conceptual illustration of the 2-step search operation of the proposed 1T-1MTJ CAM array. (a) In step 1, the mismatch case of stored "1" search "0" is identified by utilizing the reference columns and the reference row storing P state. The inverting output (ML_0) of the sense amplifier (SA_0) indicates the search result of step 1. (b) In step 2, the mismatch case of stored "0" search "1" is searched by leveraging the reference columns and the reference row storing AP state. The noninverting output (ML_1) of another sense amplifier (SA_1) indicates the search result of step 2. Given the search results of the above two steps, the stored date matches the search information if both ML_0 in step 1 and ML_1 in step 2 are high.

on the BL when reading bit "0" and "1," i.e., V_P and V_{AP}, respectively. The voltages can be expressed as below

$$V_P = I_{READ}(R_P + R_{ON}) \tag{1}$$

$$V_{AP} = I_{READ}(R_{AP} + R_{ON}) \tag{2}$$

$$V_{REF} = I_{READ}(R_{REF} + R_{ON}) \tag{3}$$

$$V_P < V_{REF} < V_{AP} \tag{4}$$

where R_P, R_{AP}, and R_{ON} are the equivalent resistance of the MTJ in the P state, the MTJ in the AP state and the access transistor, respectively. When the MTJ is in P state, the read voltage at BL (V_P) is lower than V_{REF}, and the SA produces a low/high voltage at the noninverting/inverting output, indicating a "0." When the MTJ stores an AP state, the read voltage at BL (V_{AP}) is higher than V_{REF}, resulting in the fact that the SA produces a high/low voltage at its noninverting/intervting output, indicating a "1." Although the 1T-1MTJ cell structure is widely used in the typical STT-MRAM design for in-memory logic operation [37], [38], [39], it is rarely seen for CAM design due to the limited TMR ratio of the device. In this article, we propose a novel device-circuit co-design method that employs the 1T-1MTJ structure (or other NVM devices with limited resistance ON/OFF ratios) to realize a compact and reliable CAM array design.

B. Operation Principles of the Proposed 1T-1MTJ CAM

It can be seen from Fig. 5 that the resulted read voltage V_P and V_{AP} can be uniquely identified by biasing a reference voltage between the two voltages. Such voltage relationship still holds when comparing a group of parallel connected 1T-1MTJ cells and a group of the same number of parallel connected reference cells. Based on the above observation, we propose a 2-step search scheme for the proposed 1T-1MTJ CAM to enable the parallel in-memory search operations, as illustrated in Fig. 6. To find out whether the search information matches the stored data, it is only necessary to identify the two mismatch scenarios, namely store "1" search "0" and store "0" search "1." Therefore, if both scenarios can be ruled out, a match will be identified. The 2-step search scheme is thus implemented to search the two mismatch scenarios, respectively.

Fig. 6(a) shows the configuration of search step 1, where the mismatch case of stored "1" search "0" is identified. To enable a successful operation, a reference row where all the cells store "0" (i.e., P state) is included as the sensing reference. Two additional reference columns (i.e., one stored "0" and the other stored "1" in CAM words and two biasing cells in the reference rows) are added at the end to ensure the relationship between the well defined search voltage and the reference voltage. For step 1 of the search operation, the same search current (I_{SEARCH}) is applied on each BL, all the WLs that need to search "0" and the WL associated with reference column storing "0" (i.e., WL_P) are activated, enabling the corresponding memory cells. The rest of the WLs are deactivated, disabling the corresponding cells. Assuming there are I bit "0" in the search data, which corresponds to $I+1$ columns (i.e., I CAM columns and one reference column) activated during the first step search. If there are K cells storing "1" among the I cells and $I-K$ cells storing "0," the equivalent resistance of these $I+1$ cells in parallel is $[(R_P + R_{ON})/(I-K+1)]//[(R_{AP} + R_{ON})/K]$, where R_{ON} is the equivalent resistance of the enabled access transistor. As a result, the voltage on BL is

$$V_{SEARCH0} = I_{SEARCH}\left(\frac{R_P + R_{ON}}{I-K+1} // \frac{R_{AP} + R_{ON}}{K}\right). \tag{5}$$

As shown in Fig. 6(a), there are I activated cells storing "0" (i.e., P state) and an extra activated 2T cell structure (i.e., the biasing cell) effectively defining $R_{REF} + R_{ON}$ in the reference row. The equivalent resistance of these $I+1$ cells in parallel is $[(R_P + R_{ON})/I]//(R_{REF} + R_{ON})$, thus, the voltage on BL_P is

$$V_{REF0} = I_{SEARCH}\left(\frac{R_P + R_{ON}}{I} // (R_{REF} + R_{ON})\right). \tag{6}$$

$V_{SEARCH0}$ and V_{REF0} are further used as the noninverting and inverting inputs of the SA for searching "0" (SA_0), whose inverting output (ML_0) indicates the search result of step 1.

Since the resistance of R_{REF} is between the R_P and R_{AP}, only when $K = 0$ (i.e., the cells activated in step 1 all store matched "0"), the $V_{SEARCH0}$ is lower than V_{REF0}, resulting in the ML_0 remaining high. Otherwise, $V_{SEARCH0}$ is greater than V_{REF0} and ML_0 becomes low.

In the step 2, as shown in Fig. 6(b), the mismatch scenario stored "0" search "1" is searched. And another additional reference row where all the cells store "1" (i.e., AP state) is used as the sensing reference. During step 2 of the search operation, all the WLs that need to search "1" (assume J bit "1" in the search data) and the WL associated with the reference column storing "1" (i.e., WL_{AP}) are activated, thus, enabling the corresponding memory cells. At the same time, the rest of the WLs disable the corresponding memory cells. Assume that there are L cells storing "0" among the J cells, and the remaining $J-L$ cells store "1." The equivalent resistance of these $J+1$ cells in parallel is $[(R_{AP}+R_{ON})/(J-L+1)//[(R_P+R_{ON})/L]$. As a result, the voltage on BL is

$$V_{SEARCH1} = I_{SEARCH}\left(\frac{R_{AP}+R_{ON}}{J-L+1}//\frac{R_P+R_{ON}}{L}\right). \quad (7)$$

Meanwhile, the reference row contains J activated cells storing "1" (i.e., AP state) and an extra activated 2T cell structure defining R_{REF}+R_{ON} as shown in Fig. 6(b). The equivalent resistance of these $J+1$ cells in parallel is $[(R_{AP}+R_{ON})/J]//(R_{REF}+R_{ON})$. As a result, the voltage on BL_{AP} is

$$V_{REF1} = I_{SEARCH}\left(\frac{R_{AP}+R_{ON}}{J}//(R_{REF}+R_{ON})\right) \quad (8)$$

$V_{SEARCH1}$ and V_{REF1} are further used as the noninverting and inverting inputs of the SA for searching "1" (SA_1), whose noninverting output (ML_1) indicates the search result of step 2. Only when $L = 0$ (i.e., the cells activated in step 2 all store matched "1,") the $V_{SEARCH1}$ is greater than V_{REF1}, leading to the ML_1 remaining high. Otherwise, $V_{SEARCH1}$ is lower than V_{REF1} and the ML_1 is low. Given the search results of the above two steps, only when ML_0 is high in step 1 and ML_1 is high in step 2, then it indicates that the stored data matches the search information, otherwise the mismatch happens.

Fig. 7 gives two concrete instances of the 2-step search operation when a CAM word "1100" is stored. As illustrated in Fig. 7(a), When the matched data "1100" is searched, the ML_0 is high because $V_{SEARCH0}$ is smaller than V_{REF0} in step 1, and then the ML_1 is high because $V_{SEARCH1}$ is greater than V_{REF1} in step 2, it means that the stored word matches the search query. However, when the mismatched data "0110" is searched in Fig. 7(b), the ML_0 is low because $V_{SEARCH0}$ is greater than V_{REF0} in step 1, and the ML_1 is also low because $V_{SEARCH1}$ is smaller than V_{REF1} in step 2, thus a mismatch scenario is detected.

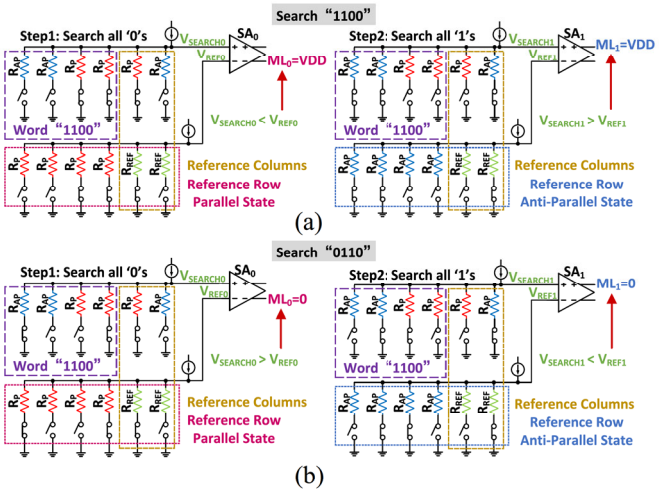


Fig. 7. Two concrete instances of the proposed 2-step search scheme: (a) Search "1100" with respect to the stored word "1100", the result is a match; (b) Search "0110" with respect to the stored word "1100," the result is a mismatch. For simplicity, we use R_P and R_{AP} here to represent the state of MTJs for storing "0" and "1," respectively, and a physics-based STT-MTJ compact model [40] is used during the simulations.

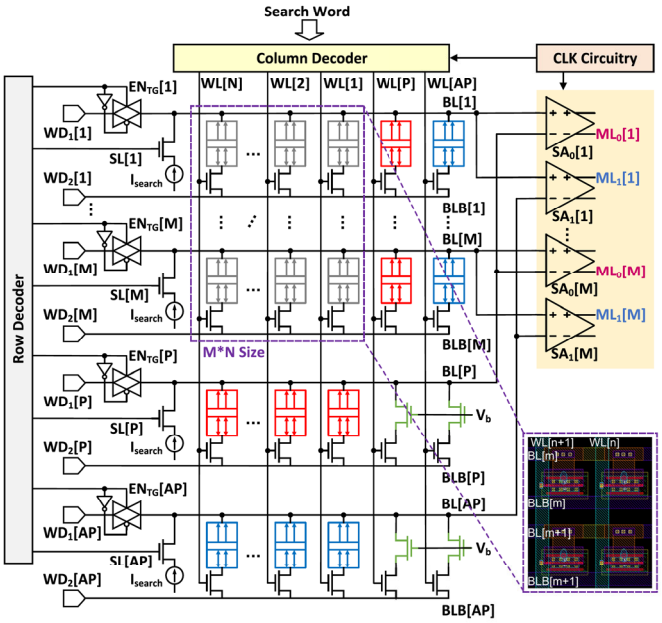


Fig. 8. Schematic of the 1T-1MTJ CAM array with $M \times N$ size. The cells in gray are used to store data; the cells in red store the P state, which forms the reference row and column for searching "0"; the cells in blue store the AP state, which forms the reference row and column for searching "1". The four reference transistors in green are biased by V_b to obtain an equivalent resistance R_{REF} between R_P and R_{AP}. The layout of a 2×2 array is sketched in the purple box.

C. 1T-1MTJ CAM Array Design

Based on the proposed operation principles of 1T-1MTJ CAM, we describe the CAM array design here. Fig. 8 shows the schematic of the proposed 1T-1MTJ-based CAM array design, along with a 2×2 layout of the 1T-1MTJ CAM array. The array mainly consists of four parts: 1) the CAM array core of size $M \times N$ that stores the data words; 2) the reference rows and columns that generate reference voltages; 3) the 2-stage SAs; and 4) the peripheral circuits including decoders, current sources, write drivers (WDs), and transmission gates. Unlike most of prior MTJ-based CAM designs that generate the one-bit search output within each complex CAM cell and sense the ML through the access transistor of each cell as shown in Fig. 3, our proposed CAM array design directly associates the BLs of all 1T-1MTJ cells in each row with the inputs of SAs, and determines the search output *via* the row-wise voltage

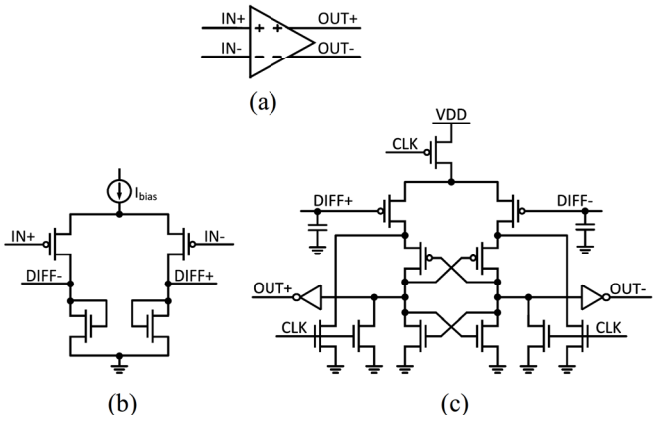


Fig. 9. Schematic of the 2-stage SA. (a) Symbol. (b) First-stage differential preamplifier. (c) Second-stage dynamic latched voltage comparator.

comparisons. WLs are vertically placed across the array, and used to enable the write and the 2-step search operations. A word is written into a row by two WDs through a transmission gate, and the current source of each row is used to generate the search voltage on BL by its search current.

The reference rows and columns are used to generate the reference voltages for the comparisons during the search operation as illustrated in Fig. 6. The reference row in red (blue) is associated with BL_P (BL_{AP}) and all the cells in the reference row store P state (AP state). The two reference columns associated with WL_P and WL_{AP} are shown in red and blue as well, representing that P state and AP state are stored, respectively. Additionally, four extra biasing cells that each contains a biasing transistor and an access transistor are located at the intersection of reference rows and reference columns, as shown in Fig. 8. Specifically, two biasing cells are associated with WL_P, and in parallel with two reference rows, while the other two biasing cells are associated with WL_{AP}, and in parallel with the reference rows as well. As illustrated in Fig. 5, the biasing cells are used to ensure that the reference voltages of the reference rows is distinct from the read voltages of other word rows during the search, which has been elucidated in details in Section III-B. It is worth noting that the above 2-step search scheme can be done with only two biasing cells located at the intersections of the reference row and column in red for searching "0" and the reference row and column in blue for searching "1," respectively. Nonetheless, in this case, when all bits in the search query are "1s," since only WL_P is activated in search step 1, BL_{AP} associated with the reference row in blue is charged to high voltage. In the search step 2, BL_{AP} then discharges through the reference row in blue, consuming extra energy consumption and search time. A similar situation occurs when all bits in the search query are "0's", and BL_P grows to high voltage in the search step 2 and then discharges in the step 1 of next search. To avoid these situations, four biasing cells are used.

We adopt two 2-stage SAs (i.e., SA_0 and SA_1) at the end of each CAM word to compare the read voltages and the reference voltages from the CAM array core. The SA is composed of a differential preamplifier and a dynamic latched comparator, as shown in Fig. 9. The BL of each row is used as one input to both SAs. Another input of SA_0 comes from the BL of the reference row storing P state, i.e., BL_P, while the input of SA_1 comes from BL_{AP}, the AP reference row Bitline. The inverting output ML_0 of SA_0 in step 1 and the noninverting output ML_1 of SA_1 in step 2 indicate whether a match or a mismatch occurs. Fig. 9(c) shows the schematic of the dynamic latched voltage comparator in the SA [38]. When the clock signal (CLK) is high (i.e., precharge phase), the two outputs of the comparator, i.e., V_{out+} and V_{out-} are precharged to high voltage level. When CLK is low (i.e., search phase), the difference between the two inputs V_{diff+} and V_{diff-} triggers the positive feedback of the internal cross-coupled inverters and then generates the comparison decision.

It should be pointed out that the dynamic CLK in the latched comparator can introduce kick-back noise to the BLs if the comparator is directly connected with the CAM array. As it can been seen from Fig. 8, the BLs of the reference rows, i.e., BL_P and BL_{AP} are associated with all the SA_0s and all the SA_1s of data rows, respectively, while the BL of each data row is associated with just two SAs, thus incurring unbalanced capacitance load associated with the differential inputs of SAs. As the capacitance load on the BLs are unbalanced, the effects of the kick-back noise to the BLs are also unbalanced, and thus bring in significant differential error. In order to suppress this error, a differential preamplifier as shown in Fig. 9(b) is inserted in between the BLs and the latched comparator. The preamplifier can suppress the kick-back noise by its open-loop gain. Besides, it provides isolation for the second-stage latched comparator, thus ensuring the functionality of SA and the scalability of the proposed 1T-1MTJ CAM array.

D. Write Operations

The write operation of the proposed 1T-1MTJ CAM array is performed word-wise, and divided into two steps by enabling the row decoders and WDs. During the write operation, the sourcelines (SLs) are set to ground to deactivate the transistors associated with the search current sources as shown in Fig. 8. The unselected rows are deactivated by disabling the corresponding transmission gates and WDs. According to Fig. 4(b), to write "1," the BL and BLB of the selected cells are driven to V_{WRITE} and GND, respectively, and the corresponding WLs of selected cells are enabled to generate the write current path. Subsequently, the voltage across the 1T-1MTJ cells of selected columns can switch the MTJ state to R_{AP}. To write "0," the selected cells are applied with similar voltage conditions per Fig. 4(b) to switch the MTJs of the cells to R_P, while the unselected columns are deactivated by setting the corresponding WLs to ground to avoid write disturbance.

E. Proposed Segmented Design

It can be seen from the expressions of search voltages and reference voltages (5)–(8) that as the length of a CAM word increases, it becomes harder for SAs to distinguish between the match and the worst 1-bit mismatch conditions, thus resulting in less reliable search functionality. Therefore, we propose to employ a segmentation scheme for the 1T-1MTJ CAM array to ensure the reliable search function and improve the search

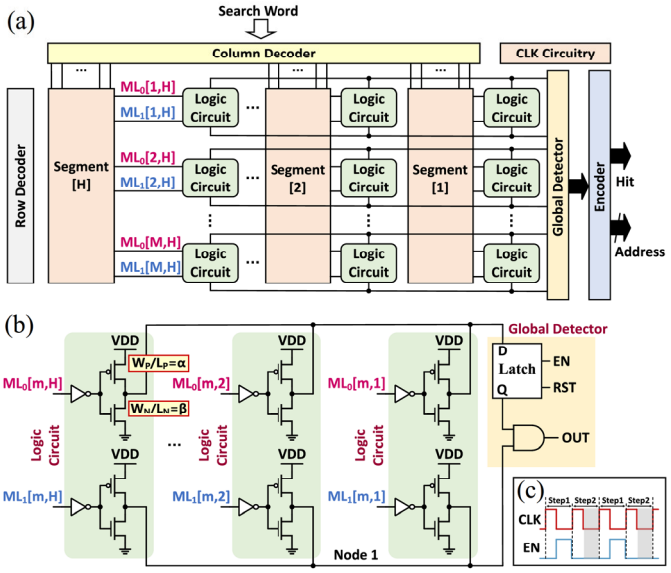


Fig. 10. (a) Schematic of the segmented 1T-1MTJ CAM array. (b) Structures of the logic circuit and the global detector for generating the overall search result. (c) Diagram of the CLK and the enable signal of the latch.

performance as depicted in Fig. 10(a). The CAM array is segmented into several subarray, each containing the reference rows, reference columns and SAs to generate the segmented search results simultaneously. Then the overall search result is aggregated by the logic circuits and the global detector to determine whether the stored word matches the search query. Fig. 10(b) illustrates the structures of the logic circuit and the global detector. We consider using the ganged CMOS-based AND circuit in the logic circuits [41]. The outputs of the two sets of skewed inverters are connected together to generate the search results for steps 1 and 2, respectively. Design parameters α and β as defined in Fig. 10(b) determine the behavior of the circuit, and the AND operation is implemented when α is sufficiently smaller than β, where the skewed inverter has a strong pull-down, thus, the output of the ganged AND circuit is pulled up to high level only when all the pull-down nMOS transistors are turned off (i.e., all ML_0s/ML_1s are high). Since the search result of step 1 is generated one cycle earlier than that of step 2, the output is D-latched, and then combined with the step 2 output for generating the overall search result as shown in Fig. 10(b) and (c).

IV. Functional Verification

To verify the functionality of our proposed CAM design, a physics-based STT-MTJ compact model [40] and a 45-nm predictive technology model (PTM) model [42] have been adopted in SPICE simulations. The critical parameters of the MTJ devices are summarized in Table I. The supply voltage is 1.1 V and the search current is 25 μA. We assumed the minimum sized transistors for 1T-1MTJ cells to increase the sensing margin corresponding to the BL voltages and achieve a higher search speed and compact cell area. The functionality of the proposed 1T-1MTJ CAM design including the write and search operations has been validated.

TABLE I
Summary of the Critical Parameters and Their Default Values of the MTJ Devices in the Simulations

Parameter	Description	Value
D	Diameter of MTJ	40nm
R_P	resistance in P state	1.84kΩ
R_{AP}	resistance in AP state	4.60kΩ
TMR(0)	TMR ratio with zero V_{bias}	150%
T_{free}	Free layer thickness	1.3nm
T_{oxide}	Oxide barrier thickness	0.75nm
R•A	Resistance*Area product	$5\Omega{\cdot}\mu m^2$
ΔTMR	TMR ratio variation	3%
ΔT_{free}	Variation of free layer thickness	3%
ΔT_{oxide}	Variation of oxide barrier thickness	3%
V_{DD}	Voltage supply	1.1V

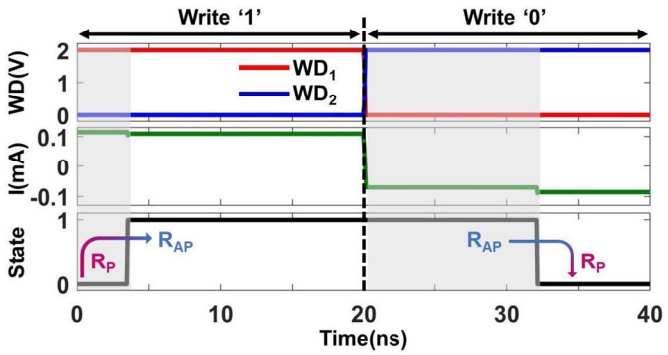


Fig. 11. Transient simulation waveforms of the proposed 1T-1MTJ CAM cell for writing "0" and "1."

A. Verification of Write Operation

Fig. 11 shows the transient simulation waveforms of the proposed 1T-1MTJ CAM cell for writing "1" and "0" successively. The MTJ of the cell is initialized in P state. The EN_{TG} and WL are activated to enable the transmission gate and the access transistor of the cell, respectively. To write "1," the WD_1 and WD_2 are set to be 2 V and GND, respectively, resulting in the write current higher than the critical current of MTJ. After the switching duration, the MTJ is flipped to AP state successfully. Subsequently, the WD_1 and WD_2 are reversed for writing "0." As a result, the MTJ is switched from AP to P state. Since a high write current leads to fast writing on the premise that the MTJ is not broken down [40], the required switching duration to write "1" is shorter than the duration required to write "0," as shown in Fig. 11.

B. Verification of Search Operation

We validate the proposed 2-step search scheme that enables the parallel search operation. Fig. 12 demonstrate the transient simulation waveforms of four 4-bit CAM words upon a search query 1010 as an example. The CAM rows are storing "1010," "1011," "0010," and "0011," respectively. The 2-step search operation is completed in two clock periods. As illustrated in Section III-C, the two outputs of the 2-stage SA are precharged to high voltage level when CLK is high, then the search and output generation are followed when CLK is low. For each word row, the inverting output (ML_0 in red) of SA_0

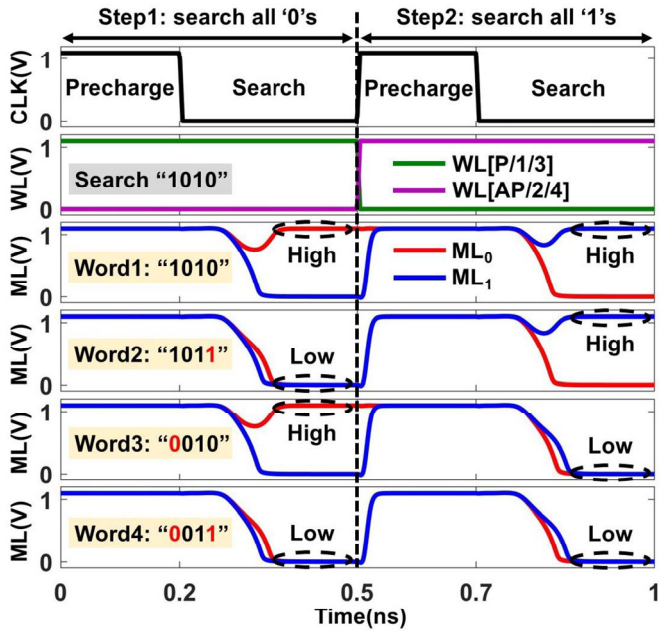


Fig. 12. Transient simulation waveforms of the proposed 2-step search scheme implemented in the proposed 1T-1MTJ CAM array when four stored CAM words are "1010," "1011," "0010," and "0011" upon a search query "1010." Only the ML corresponding to the word storing "1010" keeps high during the entire search operation.

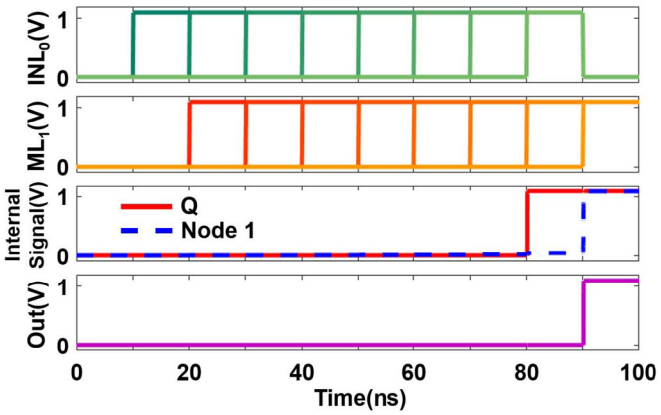


Fig. 13. Transient simulation waveforms of the proposed segmented structure with eight segments.

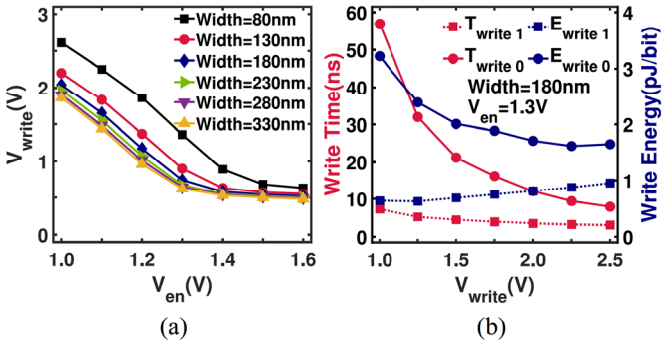


Fig. 14. (a) Required minimum write voltage (V_{write}) on the WD for write operation versus the enabled voltage (V_{en}) on WL and EN_{TG} with different transistor widths of transmission gate where TMR = 150%. (b) Write time and energy per bit of the proposed 1T-1MTJ CAM array versus different write voltages on the WD when V_{en} = 1.3 V and width = 180 nm. Note that the width marked in the figure is the nMOS width of transmission gate, and the corresponding pMOS width is double sized.

indicates the search result of step 1 while the noninverting output (ML_1 in blue) of SA_1 indicates the search result of step 2. As shown in Fig. 12, in search step 1, to find the mismatch case of "stored 1, search 0," the WL[P] of reference column and the WLs corresponding to all "0"s in the search query (i.e., WL[1] and WL[3] when searching "1010") are enabled, and ML_0 corresponding to "1011" and ML_0 corresponding to "0011" become low when CLK is low, indicating a mismatch for searching all "0s," in step 1. In the search step 2, to find the other mismatch case of "stored 0, search 1," the rest of the WLs, i.e., the WL[AP] of reference column, WL[2] and WL[4] are enabled. During the search phase, ML_1 corresponding to "0010" and ML_1 corresponding to "0011" become low, indicating a mismatch for searching all "1s." Combining the ML outputs of the two search steps, only the word storing "1010" matches with the input query, while the mismatch occurs to other three words during the search.

Fig. 13 demonstrates the function of the proposed segmented structure in Section III-E with the design parameters α and β set to 65 nm and 1.3 μm, respectively. Once the search step outputs (i.e., ML_0 and ML_1) of each segment grow to high voltage, the outputs (i.e., D and Node 1 in Fig. 10) of two ganged CMOS-based AND circuits become high, respectively. The latch maintains the ganged circuit output D at its output Q until the other ganged circuit output Node 1 is generated, and the output of the global detector is generated.

V. Evaluation

After validates the functionality of the proposed 1T1MTJ CAM design, the scalability, reliability analysis, and performance evaluations are considered. The reference columns and rows are also included in the evaluation.

A. Evaluations of Write Speed and Write Energy

As explained in Section III-D, the write operation of the proposed 1T-1MTJ CAM array is performed by enabling the transmission gate, two WDs and the corresponding WLs of selected cells. As a result, the enable voltage (V_{en}) on WL and EN_{TG}, the transistor widths of the transmission gate and the required minimum write voltage (V_{write}) on the WD affect the write efficiency. Fig. 14(a) shows the required minimum V_{write} on the WD versus the V_{en} on WL and EN_{TG} with different transistor widths of transmission gate, the TMR ratio is 150%. It can be seen that, with the increase of the V_{en}, the required minimum V_{write} decreases gradually. Besides, a larger transistor width of transmission also requires a lower minimum V_{write} at the expense of additional area overload. To optimize the area overhead while avoiding the breakdown of the transistors, the nMOS (pMOS) width of the transmission gate and V_{en} are set to be 180 nm (360 nm) and 1.3 V, respectively. With this configuration, we explore the write speed and energy consumption per bit of the proposed 1T-1MTJ CAM array with varying V_{write}. Fig. 14(b) shows that a higher V_{write}

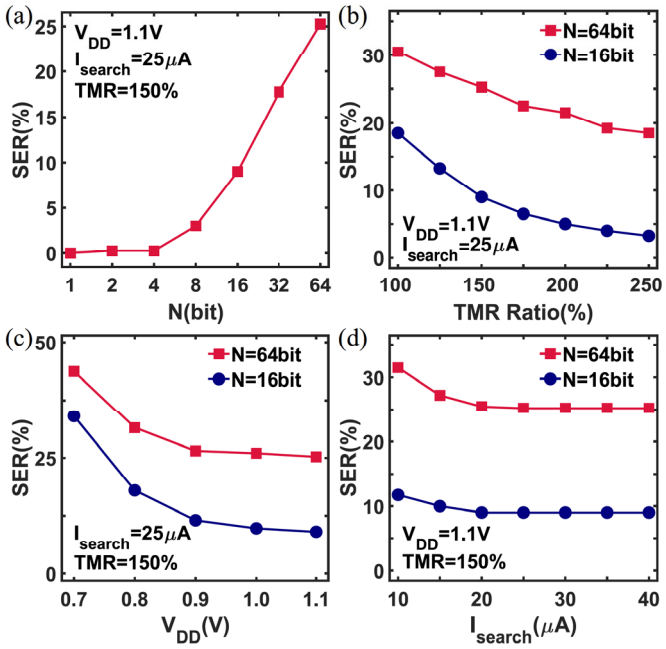


Fig. 15. SER of the proposed 1T-1MTJ CAM array versus (a) wordlength N, (b) TMR ratio, (c) supply voltage, and (d) search current by performing the Monte-Carlo simulations for the worst case where there is only a 1-bit mismatch.

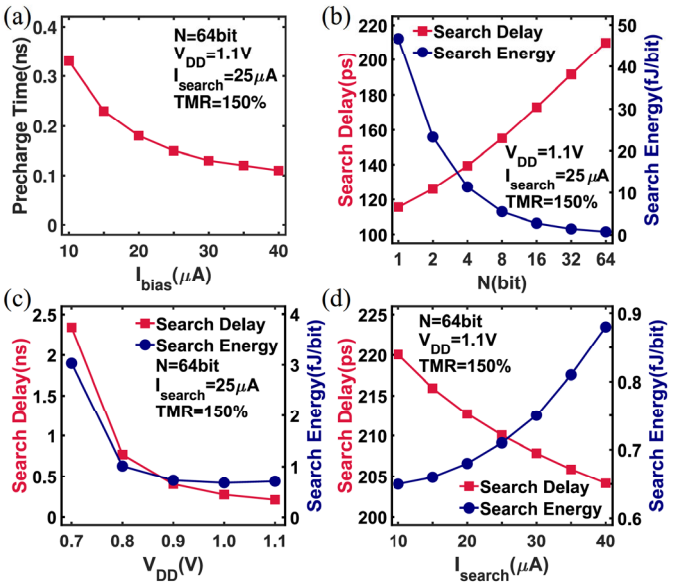


Fig. 16. (a) Minimum required precharge time for the search operation versus the bias current in the first-stage differential preamplifier. Search delay and energy per bit of the proposed 1T-1MTJ CAM array versus, (b) wordlength N, (c) supply voltage, and (d) search current.

consumes a shorter write time. Moreover, since the write currents for writing "1" and writing "0" are different, the write energy consumptions are different as well. To optimize the write energy and speed, we assumed the V_{write} to be 2 V in our evaluations, with an average write energy per bit of 1.26 pJ/bit, and a write time of 20 ns as shown in Fig. 11.

B. Reliability Analysis of 2-Step Search Scheme

To evaluate the function reliability of the proposed 2-step search scheme, we have performed the Monte-Carlo simulations for the worst case where there is only a 1-bit mismatch cell. We consider 5% process variation on threshold voltage of transistors as in [37], [43], [44]. As for the variation of MTJ, We follow the prior works [30], [45], [46] to choose the MTJ variation parameters, including free layer thickness, oxide barrier thickness, and TMR ratio for our design analysis. We assume the same values of the variation parameters, e.g., 3% process variations on free layer thickness, oxide barrier thickness, and TMR ratio of MTJ as in the 15T-4MTJ and 20T-6MTJ designs from [32], [33].

Fig. 15(a) shows the search-error-rate (SER) of the N-bit word CAM array employing the proposed 1T-1MTJ CAM design with the 2-step search scheme. When increasing the wordlength N from 1 to 64, the SER increases from 0% to 25.3%. That said, the search reliability of the array can be improved by leveraging the proposed segmentation design in Section III-E. We also investigate the relationship between the SER and TMR ratio as illustrated in Fig. 15(b). It can be seen that the SER decreases greatly when the TMR increases from 100% to 250%,[1] which is reasonable as the TMR directly affects the distinguishability of 1T-1MTJ CAM cells between P and AP states. The result indicates that larger TMR ratio is desired for high-reliable search function. Moreover, the differential discharging currents flowing through the two branches of the second-stage voltage comparator of the SA affects the comparison accuracy, the search function reliability can also be improved by increasing the supply voltage to enlarge the current difference of the differential currents as shown in Fig. 15(c). Fig. 15(d) illustrates the SER with different search current. If the search current is small, it is difficult for SAs to accurately compare the output voltages, thus causing additional errors. However, the impact of the search current on SER is negligible when the search current exceeds 25 μA.

[1]TMR ratio can reach 249% at room temperature [34].

C. Evaluations of Search Delay and Search Energy

Here, we evaluate the search speed and energy of the proposed N-bit word CAM array. As illustrated in Section III-C, the 2-stage SA is adopted at the end of the CAM word to compare the read voltages and the reference voltages from the CAM array core, and the response speed of the first-stage differential preamplifier correlates with the settle time of the input signals that are fed into the second-stage voltage comparator. The input settle time of the voltage comparator determines the minimum required precharge time before the voltage comparison, thus affecting the total search time of the array. Therefore, as the bias current I_{bias} of the differential preamplifier grows, the bandwidth of the differential preamplifier increases and the settle time decreases, resulting in less precharge time, and thus less cycle time per search. Fig. 16(a) shows the precharge time required versus the bias current of the preamplifier. Fig. 16(b) illustrates the relationship between the performance and energy of the proposed array and wordlength N. The search delay is measured for

TABLE II
PERFORMANCE COMPARISONS

Reference	[29]	[31]	[32]	[33]	This work
Transistors/ cell	9T-2MTJ	10T-4MTJ	15T-4MTJ	20T-6MTJ	1T-1MTJ
Cell area† (μm^2)	6.84	8.28	10.76	18.05	0.06
SER (144-bit)	30.0%	18.5%	2.7%	2.7%	9.0%
Search delay (ns)	0.20	1.28	0.17	0.17	0.17
Search energy (fJ)	5381.28	730.08	24.48	151.92	391.86
Search energy (fJ/bit)	37.37	5.07	0.17	1.06	2.72
Write energy (pJ/bit)	0.55	5.79	1.59	2.38	1.26

†: The 1T-1MTJ CAM cell area is estimated based on [47].
Note: Given the 144-bit search length, the performance of all these CAM cells was evaluated using the physics-based STT-MTJ compact model for a fair comparison [40]. The energy number of the proposed 1T-1MTJ CAM design includes the reference rows, reference columns and SAs and the CAM cells.

the worst case where there is only a 1-bit mismatch. As seen, the search delay increases with the increase of the wordlength N, which is due to the decreasing voltage difference on BLs. That said, even for a 64-bit word, the search delay is less than 0.22 ns. However, as for the search energy per bit, on one hand, the search voltage on BL decreases as the number of parallel cells increases, resulting in less power consumption within the array core. On the other hand, the search time does not increase significantly with the increase of the wordlength, thus the energy consumption of the SAs does not grow significantly. As a result, the overall search energy changes slightly with increasing N, and thus the search energy per bit significantly decreases. We also investigate the impact of the supply voltage on the performance, as depicted in Fig. 16(c). It can be seen that increasing the supply voltage can enlarge the differential current difference between the two branches of the second-stage dynamic latched voltage comparator in the SA, thus reducing the search delay and the search energy. Fig. 16(d) shows the search delay and energy with varying search current. As the search current increases, the voltage difference between the read voltage and the reference voltage increases, resulting in faster voltage comparison of the SA during the search phase, thus less delay, though at the expense of search energy.

D. Performance Comparisons

Above analysis shed light on the tradeoff design of the proposed 1T-1MTJ CAM array with respect to SER, energy, and delay versus design considerations, such as TMR, search current, and wordlength N, etc. Table II summarizes the performance metrics of the proposed 1T-1MTJ CAM design and other prior CAM designs, including device count per cell, cell area, SER, search delay, search energy, and write energy per bit. For a fair comparison, we adopt the same physics-based STT-MTJ compact model from [32], [33] and assume a wordlength of 144bit for all designs. The metric values of other MTJ-based CAM designs are extracted from [32], [33]. The cell area of the proposed 1T-1MTJ CAM is estimated based on [47], which is reduced greatly when compared to the previous MTJ-based designs. While previous MTJ-based CAM designs not only employ MTJ devices in pairs to store complementary logic in the devices, but also consume a large number of transistors in each cell to improve the reliability of in-cell search output, our proposed CAM design directly associates the BLs of all 1T-1MTJ cells in each row as the inputs of SAs, and determines the search output via the row-wise voltage comparisons. Therefore, our proposed design can fully leverage the compactness of MTJ devices. Although the proposed CAM design requires extra reference rows, reference columns, and SAs besides the array cells, the extra area overhead is negligible upon long-word query search. Meanwhile, our proposed CAM design achieves 1.2×/7.5× less search delay than 9T-2MTJ/10T-4MTJ CAM, and comparable search delay to the state-of-the-art 15T-4MTJ/20T-6MTJ CAM designs. Although the 15T-4MTJ/20T-6MTJ CAM designs consume less energy consumption than the proposed CAM design, their area overheads are much more than our ultracompact design. Due to the smaller number of transistors and MTJs in the write path, our approach achieves 4.6×/1.3×/1.9× more write energy efficiency than 10T-4MTJ/15T-4MTJ/20T-6MTJ CAM designs. While the 9T-2MTJ CAM achieves less write energy consumption than the proposed CAM, the SER of our 1T-1MTJ CAM is 72% less than that of the 9T-2MTJ CAM. These evaluation results suggest the area, energy efficiency, and performance improvement of our 1T-1MTJ CAM design, which again proves the advantages of our proposed device-circuit co-design method.

VI. APPLICATION BENCHMARKING

We employ our proposed 1T-1MTJ-based CAM design in the context of hyperdimensional computing (HDC) hardware for fast and parallel inference. HDC is motivated by the understanding that the human brain operates on *high-dimensional* representations of data originated from the large size of brain circuits [48]. It thereby models the human memory using points of a high-dimensional space, that is, with *hypervectors*. HDC performs learning tasks using the following steps [49]: 1) encode data into high-dimensional vectors using a well-defined set of mathematics performed over pregenerated random *base hypervectors* (e.g., the definition of alphabets for text or colors for image data); 2) during training, HDC superimposes the encoded signals to create a composite representation of a phenomenon of interest known as a "class hypervector;" and 3) in inference, the nearest neighbor search identifies an appropriate class for the encoded query hypervector.

As it has been shown by prior work, associative search is the most expensive operation of HDC that takes a majority of the inference cost [50]. We leverage our 1T-1MTJ CAM to enable the similarity search required between an encoded query hypervector and multiple pretrained class hypervectors. Ideally, HDC needs to support the nearest Hamming distance search. However, this search often requires costly SA and analog circuits to enable current or voltage-based competition between different CAM rows [51]. In contrast, we exploit our CAM capability in supporting exact search to allow the nearest

TABLE III
LIST OF DATASETS AND THE IMPACT OF SEGMENTATION ON HDC CLASSIFICATION ACCURACY

Dataset	Baseline Accuracy	Accuracy with Segmentation 4	8	16
Activity Recognition (HAR) [52]	97.8%	97.8%	97.1%	95.2%
Voice Recognition (ISOLET) [53]	96.9%	96.9%	96.3%	93.9%
Physical Monitoring (PAMAP) [54]	92.1%	92.1%	91.0%	88.3%
Face Detection (FACE) [55]	94.0%	94.0%	93.1%	91.2%
Context Recognition (EXTRA) [53]	78.3%	78.3%	72.7%	71.1%

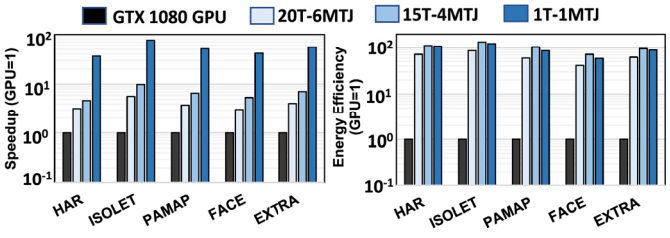


Fig. 17. Performance and energy efficiency of CAM-based solutions in accelerating HDC (normalized to GPU).

search functionality. Our design supports exact search operation on small CAM partitions. Then, for each CAM row, it counts the number of partitions that have been exactly matched with a query. A class (i.e., CAM row) with the highest number of matched partitions will be selected as the most similar class. In the following, we explore the accuracy and efficiency of our CAM accelerating HDC over five popular and large-scale classification problems, listed in Table III.

Quality of Learning: Table III shows the classification accuracy of 1T-1MTJ CAM when the segment length varies from 4 to 16. The results are reported compared to an exact platform (i.e., GPU) that supports the nearest search to perform inference. The segment size plays an important role in HDC classification accuracy, as our CAM exploits segmentation to approximate the nearest search using exact partial matches. Our evaluation shows that CAM with a smaller segment length provides a higher quality of learning. Increasing the segment length has two drawbacks: 1) increases our approximation in modeling the nearest search and 2) increases the search error rate in each segment. Our evaluation shows that CAM with a segment length of four provides the same accuracy as exact hardware. Increasing the segment size to 8 and 16 reduces the classification accuracy by 0.8% and 2.3%, respectively.

HDC Efficiency: Fig. 17 shows the energy consumption and performance of our 1T-1MTJ CAM compared to two state-of-the-art MTJ-based CAMs: 1) 20T-6MTJ and 2) 15T-4MTJ. All results are normalized to the performance and energy efficiency of the Nvidia 1080 GTX GPU platform. All CAM-based solutions are used when providing the same chip area. Our evaluation shows that all CAM-based solutions provide significant improvement compared to HDC running on the GPU platform. This higher efficiency comes from: 1) capability of CAM-based solutions to enable row-parallel search operation and 2) addressing data movement issue between memory and computing units.

Fig. 17 also compares our 1T-1MTJ CAM with prior CAM designs in accelerating HDC associative search. Due to the high density of our CAM cell in the same area, our solution provides substantially higher parallelism. This results in 54.6× (12.8× and 7.7×) speedup of our 1T-1MTJ compared to GPU (20T-6MTJ and 15T-4MTJ CAM) in accelerating HDC. In terms of energy consumption, all CAMs require the same number of operations to perform the task, thus providing comparable energy efficiency. Our results indicate a slightly higher energy efficiency of 15T-4MTJ as this CAM has the lowest search energy/bit.

VII. CONCLUSION

Most of MTJ-based CAM designs have been proposed by exploiting MTJ device pairs and differential sensing techniques for reliable search functionality, thus incurring significant area overhead. In this article, we propose a novel device-circuit co-design approach for building compact CAM designs that leverage the small footprint and nonvolatile storage of NVMs with limited resistance ON/OFF ratios without sacrificing the area efficiency. We use MTJ as a proxy, and propose a novel 1T-1MTJ-based CAM design that utilizes our proposed 2-step search scheme to realize improved area overhead, energy efficiency, and performance. The write and search operations have been validated, as well as the search reliability and scalability. The array and application level evaluations suggest that our proposed 1T-1MTJ CAM design is superior to other MTJ-based designs in terms of area, energy efficiency, and performance. Our proposed device-circuit co-design method has paved a promising way for enabling a simple NVM storage structure for ultracompact and efficient CAM designs, and further other CiM innovations.

REFERENCES

[1] B. Li, B. Yan, and H. Li, "An overview of in-memory processing with emerging non-volatile memory for data-intensive applications," in *Proc. Great Lakes Symp. VLSI*, 2019, pp. 381–386.
[2] X. S. Hu *et al.*, "In-memory computing with associative memories: A cross-layer perspective," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2021, pp. 25–32.
[3] R. Karam, R. Puri, S. Ghosh, and S. Bhunia, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proc. IEEE*, vol. 103, no. 8, pp. 1311–1330, Aug. 2015.
[4] Y.-J. Chang, "A high-performance and energy-efficient TCAM design for IP-address lookup," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 56, no. 6, pp. 479–483, Jun. 2009.
[5] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nat. Electron.*, vol. 2, no. 11, pp. 521–529, 2019.
[6] X. Yin *et al.*, "Deep random forest with ferroelectric analog content addressable memory," 2021, *arXiv:2110.02495*.
[7] X. Yin *et al.*, "An ultra-compact single FeFET binary and multi-bit associative search engine," 2022, *arXiv:2203.07948*.
[8] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: A tutorial and survey," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, Mar. 2006.
[9] W. Xu, T. Zhang, and Y. Chen, "Design of spin-torque transfer magnetoresistive RAM and CAM/TCAM with high sensing and search speed," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 1, pp. 66–74, Jan. 2009.
[10] S. Matsunaga *et al.*, "Design and fabrication of a one-transistor/one-resistor nonvolatile binary content-addressable memory using perpendicular magnetic tunnel junction devices with a fine-grained power-gating scheme," *Jpn J. Appl. Phys.*, vol. 50, no. 6R, 2011, Art. no. 063004.

[11] M. K. Gupta and M. Hasan, "Robust high speed ternary magnetic content addressable memory," *IEEE Trans. Electron Devices*, vol. 62, no. 4, pp. 1163–1169, Apr. 2015.

[12] D. Cho, K. Kim, and C. Yoo, "Variation-tolerant non-volatile ternary content addressable memory with magnetic tunnel junction," *J. Semicond. Technol. Sci.*, vol. 17, no. 3, pp. 458–464, 2017.

[13] R. Govindaraj and S. Ghosh, "Design and analysis of STTRAM-based ternary content addressable memory cell," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 4, pp. 1–22, 2017.

[14] J. Li *et al.*, "1Mb 0.41 μm 2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing," in *Proc. Symp. VLSI Technol.*, 2013, pp. C104–C105.

[15] M.-F. Chang *et al.*, "A ReRAM-based 4T2R nonvolatile TCAM using RC-filtered stress-decoupled scheme for frequent-OFF instant-ON search engines used in IoT and big-data processing," *IEEE J. Solid-State Circuits*, vol. 51, no. 11, pp. 2786–2798, Nov. 2016.

[16] C.-C. Lin *et al.*, "7.4 a 256b-wordlength ReRAM-based TCAM with 1ns search-time and 14 improvement in wordlength-energyefficiency-density product using 2.5 T1R cell," in *Proc. IEEE ISSCC*, 2016, pp. 136–137.

[17] M.-F. Chang *et al.*, "A 3T1R nonvolatile TCAM using MLC ReRAM for frequent-off instant-on filters in IoT and big-data processing," *IEEE J. Solid-State Circuits*, vol. 52, no. 6, pp. 1664–1679, Jun. 2017.

[18] X. Yin, X. Chen, M. Niemier, and X. S. Hu, "Ferroelectric FETs-based nonvolatile logic-in-memory circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 1, pp. 159–172, Jan. 2019.

[19] X. Yin, K. Ni, D. Reis, S. Datta, M. Niemier, and X. S. Hu, "An ultra-dense 2FeFET TCAM design based on a multi-domain FeFET model," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 9, pp. 1577–1581, Sep. 2019.

[20] J. Cai *et al.*, "Energy efficient data search design and optimization based on a compact ferroelectric FET content addressable memory," in *Proc. 59th ACM/IEEE Design Autom. Conf.*, 2022, pp. 751–756.

[21] X. Yin *et al.*, "Ferroelectric ternary content addressable memories for energy efficient associative search," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, early access, Aug. 9, 2022, doi: 10.1109/TCAD.2022.3197694.

[22] C. Li *et al.*, "A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2020, pp. 29–33.

[23] X. Yin *et al.*, "FeCAM: A universal compact digital and analog content addressable memory using ferroelectric," *IEEE Trans. Electron Devices*, vol. 67, no. 7, pp. 2785–2792, Jul. 2020.

[24] X. Fong, Y. Kim, R. Venkatesan, S. H. Choday, A. Raghunathan, and K. Roy, "Spin-transfer torque memories: Devices, circuits, and systems," *Proc. IEEE*, vol. 104, no. 7, pp. 1449–1488, Jul. 2016.

[25] S. Matsunaga, A. Mochizuki, T. Endoh, H. Ohno, and T. Hanyu, "Design of an energy-efficient 2T-2MTJ nonvolatile TCAM based on a parallel-serial-combined search scheme," *IEICE Electron. Exp.*, vol. 11, no. 3, 2014, Art. no. 20131006.

[26] S. Matsunaga *et al.*, "Fully parallel 6T-2MTJ nonvolatile TCAM with single-transistor-based self match-line discharge control," in *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, 2011, pp. 298–299.

[27] S. Matsunaga *et al.*, "A 3.14 um 2 4T-2MTJ-cell fully parallel TCAM based on nonvolatile logic-in-memory architecture," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, 2012, pp. 44–45.

[28] S. Matsunaga *et al.*, "Complementary 5T-4MTJ nonvolatile TCAM cell circuit with phase-selective parallel writing scheme," *IEICE Electron. Exp.*, vol. 11, no. 10, 2014, Art. no. 20140297.

[29] S. Matsunaga, A. Katsumata, M. Natsui, T. Endoh, H. Ohno, and T. Hanyu, "Design of a nine-transistor/two-magnetic-tunnel-junction-cell-based low-energy nonvolatile ternary content-addressable memory," *Jpn. J. Appl. Phys.*, vol. 51, no. 2S, 2012, Art. no. 02BM06.

[30] W. Zhao *et al.*, "Synchronous non-volatile logic gate design based on resistive switching memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 2, pp. 443–454, Feb. 2014.

[31] B. Song, T. Na, J. P. Kim, S. H. Kang, and S.-O. Jung, "A 10T-4MTJ nonvolatile ternary CAM cell for reliable search operation and a compact area," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 64, no. 6, pp. 700–704, Jun. 2017.

[32] C. Wang, D. Zhang, L. Zeng, E. Deng, J. Chen, and W. Zhao, "A novel MTJ-based non-volatile ternary content-addressable memory for high-speed, low-power, and high-reliable search operation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 4, pp. 1454–1464, Apr. 2019.

[33] C. Wang, D. Zhang, L. Zeng, and W. Zhao, "Design of magnetic non-volatile TCAM with priority-decision in memory technology for high speed, low power, and high reliability," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 2, pp. 464–474, Feb. 2020.

[34] M. Wang *et al.*, "Current-induced magnetization switching in atom-thick tungsten engineered perpendicular magnetic tunnel junctions with large tunnel magnetoresistance," *Nat. Commun.*, vol. 9, no. 1, pp. 1–7, 2018.

[35] H. Cai, Y. Wang, L. A. D. B. Naviner, and W. Zhao, "Robust ultra-low power non-volatile logic-in-memory circuits in FD-SOI technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 4, pp. 847–857, Apr. 2017.

[36] A. T. Do, C. Yin, K. S. Yeo, and T. T.-H. Kim, "Design of a power-efficient CAM using automated background checking scheme for small match line swing," in *Proc. IEEE ESSCIRC (ESSCIRC)*, 2013, pp. 209–212.

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

[38] Z. He, S. Angizi, and D. Fan, "Exploring STT-MRAM based in-memory computing paradigm with application of image edge extraction," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, 2017, pp. 439–446.

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

[40] Y. Wang *et al.*, "Compact model of dielectric breakdown in spin-transfer torque magnetic tunnel junction," *IEEE Trans. Electron Devices*, vol. 63, no. 4, pp. 1762–1767, Apr. 2016.

[41] M. Imani *et al.*, "SearcHD: A memory-centric hyperdimensional computing with stochastic training," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 10, pp. 2422–2433, Oct. 2020.

[42] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of pMOS NBTI effect for robust nanometer design," in *Proc. IEEE DAC*, 2006, pp. 1047–1052.

[43] K.-W. Kwon, X. Fong, P. Wijesinghe, P. Panda, and K. Roy, "High-density and robust STT-MRAM array through device/circuit/architecture interactions," *IEEE Trans. Nanotechnol.*, vol. 14, no. 6, pp. 1024–1034, Nov. 2015.

[44] X. Fong, S. H. Choday, and K. Roy, "Bit-cell level optimization for non-volatile memories using magnetic tunnel junctions and spin-transfer torque switching," *IEEE Trans. Nanotechnol.*, vol. 11, no. 1, pp. 172–181, Jan. 2012.

[45] Y. Wang, Y. Zhang, E. Deng, J.-O. Klein, L. A. Naviner, and W. Zhao, "Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses," *Microelectron. Rel.*, vol. 54, nos. 9–10, pp. 1774–1778, 2014.

[46] W. Kang, L. Zhang, J.-O. Klein, Y. Zhang, D. Ravelosona, and W. Zhao, "Reconfigurable codesign of STT-MRAM under process variations in deeply scaled technology," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1769–1777, 2015.

[47] S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nat. Electron.*, vol. 1, no. 8, pp. 442–450, 2018.

[48] P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cogn. Comput.*, vol. 1, no. 2, pp. 139–159, 2009.

[49] M. Imani, J. Morris, J. Messerly, H. Shu, Y. Deng, and T. Rosing, "BRIC: Locality-based encoding for energy-efficient brain-inspired hyperdimensional computing," in *Proc. 56th Annu. Design Autom. Conf.*, 2019, pp. 1–6.

[50] M. Imani *et al.*, "Revisiting HyperDimensional learning for FPGA and low-power architectures," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, 2021, pp. 221–234.

[51] M. Imani, A. Rahimi, D. Kong, T. Rosing, and J. M. Rabaey, "Exploring hyperdimensional associative memory," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, 2017, pp. 445–456.

[52] D. Anguita *et al.*, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Proc. AAL*, 2012, pp. 216–223.

[53] "UCI machine learning repository." Accessed: May 27, 2021. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/ISOLET.

[54] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. IEEE 16th Int. Symp. Wearable Comput. (ISWC)*, 2012, pp. 108–109.

[55] A. Angelova, Y. Abu-Mostafam, and P. Perona, "Pruning training sets for learning of object categories," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 1–9.

Cheng Zhuo (Senior Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Zhejiang University, Hangzhou, China, in 2005 and 2007, respectively, and the Ph.D. degree in computer science and engineering from the University of Michigan, Ann Arbor, MI, USA, in 2010.

He is currently a Professor with the College of Information Science and Electronic Engineering, Zhejiang University. He has published over 100 technical papers. His current research interests include computing in memory, deep learning, hardware accelerator, and general VLSI EDA areas.

Prof. Zhuo received Four Best Paper Nominations in DAC'16, CSTIC'18, ICCAD'20, and VTS'21. He also received 2012 ACM/SIGDA Technical Leadership Award, and 2017 JSPS Invitation Fellowship. He has served on the technical program and organization committees of many international conferences and as an Associate Editor for IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, *ACM Transactions on Design Automation of Electronic Systems*, and *Elsevier Integration*. He is a Fellow of IET.

Zeyu Yang received the B.S. degree in electronic science and technology from Zhejiang University, Hangzhou, China, in 2021, where he is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering.

His current research interests include design and optimization of circuits, and architectures for in-memory computing.

Kai Ni (Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2011, and the Ph.D. degree of electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2016, by working on characterization, modeling, and reliability of III-V MOSFETs.

Since then, he became a Postdoctoral Associate with the University of Notre Dame, Notre Dame, IN, USA, working on ferroelectric devices for nonvolatile memory and novel computing paradigms. He is currently an Assistant Professor with the Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY, USA. He has 80 publications in top journals and conference proceedings, including nature electronics, IEDM, VLSI symposium, IRPS, and EDL. His current interests lie in nanoelectronic devices empowering unconventional computing, AI accelerator, and 3-D memory technology.

Mohsen Imani (Member, IEEE) received the B.Sc. and M.S. degrees from the School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, in 2011 and 2014, respectively, and the Ph.D. degree from the Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA, USA, in 2020.

He is currently an Assistant Professor with the University of California at Irvine, Irvine, CA, USA. His current research interests include brain-inspired computing, approximation computing, and processing in-memory.

Yuxuan Luo (Member, IEEE) received the B.Eng. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2014, and the Ph.D. degree from the National University of Singapore, Singapore, in 2018.

From 2018 to 2020, he worked as a Research Fellow with the National University of Singapore. He is currently a Faculty with the School of Micro-Nanoelectronics, Zhejiang University, Hangzhou, China. His current research interest is on sensing circuits and systems.

Dr. Luo was a recipient of the 2018 the Institution of Engineers, Singapore Prestigious Engineering Award. He serves as a Technical Program Committee Member of the IEEE International Conference on Integrated Circuits, Technologies, and Applications, a Subtrack Chair of 2022 IEEE Asia Pacific Conference on Circuits and Systems, and a Guest Editor of *Frontiers in Nanotechnology*.

Shaodi Wang (Member, IEEE) received the B.S. degree from Peking University, Beijing, China, in 2011, and the Ph.D. degree in electrical engineering from UCLA, Los Angeles, CA, USA, in 2017.

He founded Witmem Company Ltd., Whitman, MA, USA, in 2017, where he is currently serves as the CEO. He and Witmem are dedicated to developing computing-in-memory technology. He has published over 20 journal and conference papers and applied over 50 patents.

Deming Zhang (Member, IEEE) received the B.S. and Ph.D. degrees in electronic and information engineering from Beihang University, Beijing, China, in 2011 and 2017, respectively.

He is currently an Associated Professor with the School of Integrated Circuit Science and Engineering, Beihang University. He has authored/coauthored more than 50 scientific papers. His research interests include spintronic devices modeling and IC design, i.e., memory, logic and computing-in-memory.

Xunzhao Yin (Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2013, and the Ph.D. degree in computer science and engineering from the University of Notre Dame, Notre Dame, IN, USA, in 2019.

He is an Assistant Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He has published top journals and conference papers, including *Nature Electronics*, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS, IEEE TRANSACTIONS ON ELECTRON DEVICES, DAC, IEDM, and Symposium on VLSI. His research interests include emerging circuit/architecture designs and novel computing paradigms with both CMOS and emerging technologies.

Mr. Yin has received the best paper award nomination of ICCAD 2020, VLSI Test Symposium 2021, and DATE2022, Zhejiang University 2020 Top 10 academic research advancements nomination. He serves as an Associate Editor of *ACM SIGDA E-Newsletter*.