Ferroelectric Ternary Content Addressable Memories for Energy-Efficient Associative Search

Xunzhao Yin, *Member, IEEE*, Yu Qian, Mohsen Imani, *Member, IEEE*, Kai Ni, *Member, IEEE*, Chao Li, Grace Li Zhang, *Member, IEEE*, Bing Li, *Senior Member, IEEE*, Ulf Schlichtmann, *Senior Member, IEEE*, and Cheng Zhuo, *Senior Member, IEEE*

***Abstract*—A fast and efficient search function across the database has been a core component for a number of data-intensive tasks in machine learning, IoT applications, and inference. However, the conventional digital machines implementing the search functionality with repetitive arithmetic operations suffer from the energy efficiency and performance degradation due to the significant data transfer between the storage and processing units in the Von Neumann architecture. Ternary content addressable memories (TCAMs) are an essential hardware form of computing-in-memory (CiM) designs that aim to overcome the data transfer bottlenecks by implementing the parallel associative search function within the memory blocks. While most state-of-the-art TCAM designs focus on improving the information density by harnessing compact nonvolatile memories (NVMs), little efforts have been spent on optimizing the energy efficiency of the NVM-based TCAM. In this article, by exploiting the ferroelectric FET (FeFET) as a representative NVM, we propose an NOR-type 2FeFET-1T and an NAND-type 2FeFET-2T TCAM designs that enable highly energy-efficient associative search by reducing the associated precharge overheads. We then propose a hybrid ferroelectric NAND-NOR (HFNN) TCAM design to further improve the energy efficiency. An HFNN-based segmented architecture is proposed to reduce the search delay and energy by search operation pipeline. Evaluation results suggest that the proposed 2FeFET-1T, 2FeFET-2T and HFNN TCAM design consume 3.03×, 8.08×, and 226.92× less search energy than the conventional 16T complementary metal oxide semiconductor (CMOS) TCAM, respectively. Application benchmarking shows that our proposed 2FeFET-1T/2FeFET-2T/HFNN TCAM can save, on average, 45.2%/50.6%/57.5% the GPU energy consumption as compared to the conventional GPU.**

***Index Terms*—Associative memory, computing-in-memory (CiM), ferroelectric FET (FeFET), ternary content addressable memory (TCAM).**

Manuscript received 29 March 2022; revised 8 June 2022; accepted 2 August 2022. Date of publication 9 August 2022; date of current version 21 March 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 and Grant 92164203; and in part by Zhejiang Lab under Grant 2021MD0AB02. This article was recommended by Associate Editor A. Gamatie. *(Corresponding author: Cheng Zhuo.)*

Xunzhao Yin is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China, and also with Zhejiang Lab, Hangzhou 310058, China.

Yu Qian, Chao Li, and Cheng Zhuo are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China (e-mail: czhuo@zju.edu.cn).

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

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

Grace Li Zhang, Bing Li, and Ulf Schlichtmann are with the Chair of Electronic Design Automation, Technology University of Munich, 80333 Munich, Germany.

Digital Object Identifier 10.1109/TCAD.2022.3197694

I. Introduction

In THE era of Big Data, the development of data-intensive IoT applications call for efficient and parallel data analytic operations to replace the sequential, time- and energy-consuming operations in the conventional digital processors [1], [2]. In addition to the common arithmetic functions, one such essential function is the search function [1], [3]. For example, to implement cognitive learning, memory augmented computing system needs efficient searching capabilities to identify the class of the new query [1], [3].

However, due to the emergence of large-scale data of computing models and the so-called memory wall issues, the processors implementing the search function are still slow, energy consuming, and prohibitively expensive. Computing-in-memory (CiM) where frequent computations are integrated within the memory blocks to reduce the data transfer between processors and memory, represents a promising solution to improve the energy efficiency of the hardware [4], [5]. As the essential hardware kernel of CiM, ternary content addressable memory (TCAM), which supports parallel searches over the stored memory array given an input vector, is a potential solution to meet the demands of AMs and address the processor-memory bottleneck present in conventional digital machines [4], [6]. Due to the content addressing and fully parallel property, TCAMs have been applied in many areas, e.g., neuromorphic computing, memory-augmented neural network, IP routers, in-memory data processing, etc. [1], [7], [8], [9], [10], [11], [12].

While the conventional TCAM designs based on the standard complementary metal oxide semiconductor (CMOS) technology have been proposed [13], they suffer from large area overhead and leakage as the CMOS technology scales down to the physical limit. Emerging nonvolatile memories (NVMs), such as the two-terminal resistive RAM (ReRAM) [14] and the three-terminal ferroelectric FET

(FeFET) [15] can encode logic values "0"/"1" using their high/low resistance state, and thus have been exploited to build more compact TCAM designs [16], [17]. Highly promising as NVM-based TCAMs are, these designs focus more on reducing the TCAM cell size with small NVMs. The potential of combining both the NVMs and the energy-efficient design schemes for TCAMs to improve energy efficiency has yet to be explored.

In this article, we exploit FeFET as a representative NVM, and propose a novel FeFET-based TCAM design to achieve ultrahigh energy efficiency. Based on a proposed 2FeFET structure to perform a in-memory XOR-like function, we first propose two compact and energy-efficient TCAM designs, i.e., NOR-type and NAND-type, shedding light on two potential design schemes to improve the energy efficiency and performance of TCAMs. Specifically, the 2FeFET-1T NOR-type TCAM design improves the energy efficiency and performance by reducing the effective precharge/discharge capacitance associated with the TCAM matchline, while the 2FeFET-2T NAND-type TCAM design further reduces the search energy by fully eliminating the precharge phase prior to every search operation. While these two designs achieve higher energy efficiency than prior works, the NOR-type TCAM design still consumes significant precharge energy for each search, and the NAND-type TCAM design is the slowest due to the series transistors along the discharge path.

Both proposed designs still need to improve the energy efficiency to support the in-memory information processing tasks. Therefore, we propose a hybrid ferroelectric NAND-NOR (HFNN) TCAM design by leveraging the advantages of both TCAM design styles (NOR-type and NAND-type) to achieve ultralow energy consumption with acceptable search delay. HFNN design adopts the 2FeFET-1T NOR-type TCAM and a 2FeFET-1T NAND-type TCAM modified from the 2FeFET-2T design in the array. It saves the energy by reducing the number of activated TCAM rows via the NAND-type matchline structure, and keeps the search speed by employing the NOR-type matchline structure. We further propose an HFNN-based segmented architecture to pipeline the search operation, which significantly reduces the energy and delay.

The structures, operations, and energy/performance analysis of the proposed TCAMs are discussed and evaluated at the array level and compared with alternative TCAM designs based on CMOS, ReRAM, and FeFET, respectively, to demonstrate the benefits of combining NVMs with the proposed TCAM design schemes. We also examine the energy efficiency of the proposed FeFET TCAM designs in the context of augmented GPU architecture. Evaluation results show that the proposed 2FeFET-1T/2FeFET-2T/HFNN TCAM design offers 3.03×/8.08×/226× and 1.79×/4.79×/134.62× better energy efficiency than the CMOS TCAM and the state-of-the-art 2FeFET TCAM design, respectively. The proposed 2FeFET-1T/2FeFET-2T/HFNN TCAM used as an associative search engine in the enhanced GPU architecture can save, on average, 45.2%/50.6%/57.5% energy consumption as compared to the conventional GPU. Compared with the 2FeFET TCAM enhanced GPU approach, our proposed designs can still achieve 4.9%/10.4%/16.5% more energy saving.

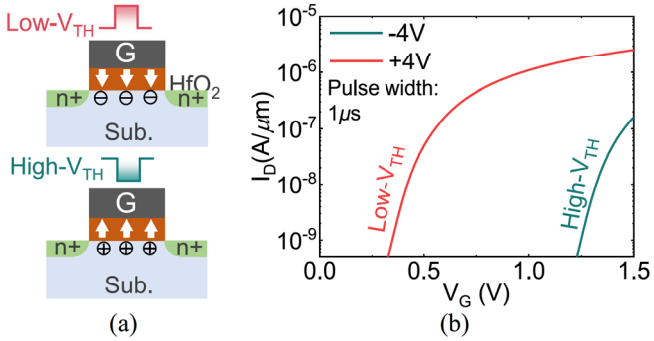


Fig. 1. (a) FeFET polarization directions and channel conditions after memory write operations. (b) FeFET I_D–V_G characteristics after positive/negative gate write voltages.

II. Background

In this section, we first review the basics of the FeFET device and model [18]. Then, we describe the TCAM basics, including the existing TCAMs based on CMOS, ReRAM, and FeFET, and derive the energy equations associated with the matchline.

A. FeFET Basics

FeFET has received renewed interest recently ever since the discovery of ferroelectricity in doped HfO_2. Such a material is CMOS-compatible and maintains its ferroelectricity even down to 1-nm thickness, greatly outperforming its perovskite ferroelectric counterparts [19]. Prior works [20], [21] also show that FeFET is compatible with the advanced 7 nm/14 nm CMOS technology. Ferroelectric memory exhibits superior write energy efficiency because the polarization switching process is driven by an electric field, while the write process in other types of NVMs requires a high conduction current [1], thus consuming a high write power. When integrating a ferroelectric film as the gate insulator in an MOSFET, an FeFET is obtained. By applying a positive/negative gate pulse, the ferroelectric polarization will be switched to point at the channel/gate, thus attracting electrons/holes in the channel and setting the device in low-V_{TH}/high-V_{TH} state, respectively, as shown in Fig. 1(a).

Several models have been proposed for FeFET, including a model based on hysteric negative capacitance FET (NCFET) [22], a Preisach model [18], and a comprehensive Monte Carlo model [23]. In this work, we are utilizing the experimentally calibrated Preisach compact model for FeFET due to its computational efficiency and accuracy [18]. In this model, a ferroelectric film is considered to contain a large amount of independent domains, where each domain has its own switching coercive field. Thus, the overall ferroelectric response is obtained by adding up all the domain responses, which can be well approximated by a hyperbolic tangent function [18]. The final ferroelectric model can be accomplished by, including the ferroelectric history tracking algorithm and nonsaturated hysteresis loops, explained in [18]. Integrating such a ferroelectric model with an MOSFET model, such as the standard BSIM, an FeFET model is obtained. Such a model has been calibrated with experimental results [18]. Fig. 1(b)

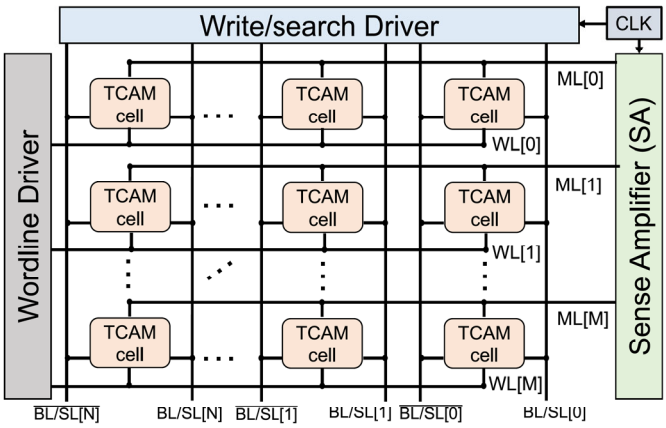


Fig. 2. Schematic of an $M \times N$ TCAM array.

shows the I_D–V_G characteristic after memory write with ±4 V gate voltages. A memory window of approximately 1 V is obtained.

B. Existing TCAM Designs

Fig. 2 shows the schematic of a TCAM array consisting of M words, with each word containing N cells placed horizontally. The cells within one word share a matchline (ML) in a NOR-type connection, which is sensed by the sense amplifier (SA). The cells within one column are associated with the same bit/search line pairs. A typical NOR-type precharge-based TCAM operation starts with precharging the MLs to match state (high voltage level), and then drives the search lines with the input data. The MLs on which all cells match with the input remain high, indicating a match, while those with at least one cell mismatching with the input, will discharge, indicating a mismatch. The SAs sense the ML states of their associated words.

Numerous NVM-based TCAM designs consume smaller cell sizes than the conventional 16T CMOS TCAM [Fig. 3(a)], due to the compact NVM structures. Fig. 3 shows the most commonly used TCAM designs. The 16T CMOS TCAM [13] is volatile, and stores ternary states ("0," "1," and wildcard "X") with two SRAMs, while the 2T-2R ReRAM-based [16] and 2FeFET-based [1], [24] TCAMs encode ternary states into the NVMs with much less transistors, thus smaller cell sizes. It can be seen from Figs. 2 and 3 that the TCAM designs aforementioned are precharge-based, requiring a precharge phase prior to every search. The search energy of TCAM arrays consists of the precharge energy and the leakage energy

$$E \approx E_{\text{pre}} + E_{\text{leak}} = C_{\text{ML}} V_{DD}^2 + E_{\text{leak}} \tag{1}$$

where E_{pre} represents the energy precharged to the ML during the precharge phase, E_{leak} represents the leakage energy from supply to ground, and C_{ML} is denoted as the capacitance associated with the array ML. The equivalent schematic of a TCAM cell in a NOR-type ML connection is depicted in Fig. 3(d), formulating C_{ML} as follows:

$$\begin{aligned} C_{\text{ML}} &\approx C_{\text{pMOS}} + N \times \left(C_{\text{cell}} + C_{\text{parasitic}}\right) \\ &= C_{\text{pMOS}} + N \times \left(2C_{\text{drain}} + C_{\text{parasitic}}\right) \end{aligned} \tag{2}$$

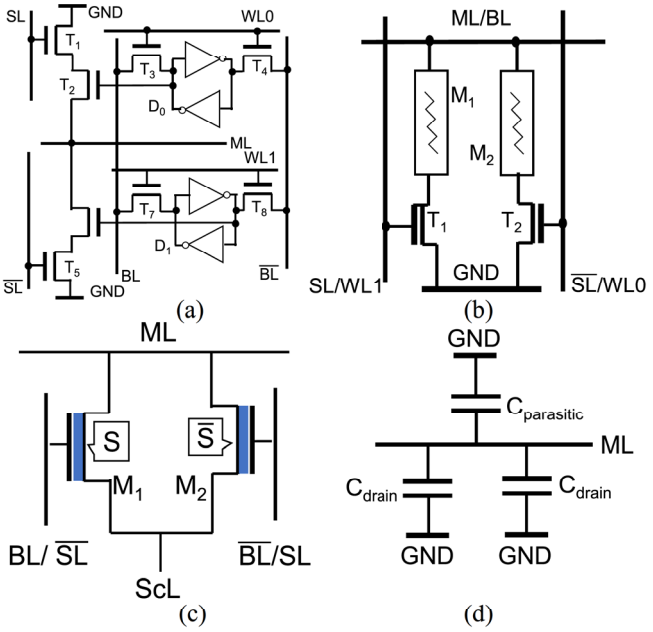
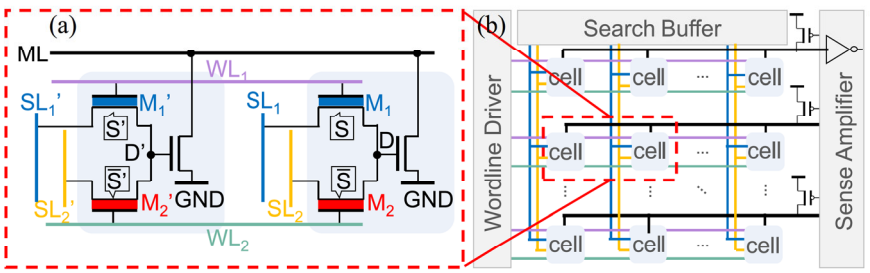


Fig. 3. TCAM designs: (a) 16T CMOS; (b) 2T-2R ReRAM; (c) 2FeFET; (d) Equivalent schematic of (a, b, c) TCAM cells.

where C_{pMOS}, C_{cell}, $C_{\text{parasitic}}$, and C_{drain} are the associated capacitance of the pMOS transistor that precharges the ML, the total capacitance of a TCAM cell associated with the ML, the parasitic capacitance of each cell, and the drain capacitance of a transistor, respectively. Each TCAM cell connects the ML with two transistors, thus C_{cell} consists of two drain capacitances, which can be extracted from the PTM technology [25], and $C_{\text{parasitic}}$ can be extracted from DESTINY [26]. N is the number of columns in the array. As E_{pre} dominates the total search energy, the search energy of an array is mainly dependent on the number of transistors per cell associated with the ML given an array size. To improve the energy efficiency, we therefore propose a general design scheme for the NOR-type TCAM which decreases the ML capacitance load by reducing the number of transistors associated with the ML.

The other proposed design scheme focuses on eliminating the precharge phase instead. Mahendra *et al.* [27] reported a 14T CMOS-based precharge-free TCAM design that eliminates the precharge phase at the cost of a large area and delay overheads. However, the large area size and search delay limits its practical usage in fast and small devices. We therefore propose to combine the precharge-free design with FeFETs-based NAND-type TCAM to build a compact and area, energy-efficient TCAM design.

To take advantages of both precharge-based NOR-type TCAM and precharge-free NAND-type TCAM designs, the hybrid NAND-NOR TCAM design scheme, which contains both NAND-type and NOR-type TCAM cells within the array, represents a promising solution to achieve a balance between the search delay and energy for TCAMs [28], [29], [30]. Upon the search operation, the NAND-type TCAM cells are first activated, and used to filter the words whose NAND-type TCAM cells match with the corresponding input bits. The NOR-type TCAM cells of the filtered words are then activated to start



(c) Operations of 2FeFET-1T TCAM Cell V_W=4V, V_S=1V

Mode	WL_1	WL_2	SL_1	SL_2	S	$\bar{S}$	D	ML
Write '0'	$-V_W$	V_W	0	0	0	1	–	–
Write '1'	V_W	$-V_W$	0	0	1	0	–	–
Write 'X'	$-V_W$	$-V_W$	0	0	0	0	–	–
Search '0'	V_S	V_S	V_S	0	0	1	Low	Match
					1	0	High	Mismatch
					0	0	Low	Match
Search '1'	V_S	V_S	0	V_S	0	1	High	Mismatch
					1	0	Low	Match
					0	0	Low	Match

Fig. 4. Schematic of (a) 2FeFET-1T TCAM cells; (b) TCAM array (precharge pMOS and the SA (inverter) included); (c) write and search operations.

precharge-based search operation. Such a design scheme significantly reduces the number of activated NOR-type cells, thus reduces the precharge energy while maintaining acceptable search delay. To fully leverage the emerging devices, such as FeFETs and the hybrid TCAM design scheme, we further propose HFNN TCAM to achieve ultralow energy consumption, which will be discussed in Section IV.

III. Energy-Aware TCAM Designs

In this section, we propose two novels FeFET TCAM designs with improved energy efficiency by either 1) reducing the associated ML capacitance or 2) eliminating the precharge phase [31]. We first review the basic structures and operations of the proposed TCAM designs, and then discuss how they improve the energy efficiency and performance at the array level.

A. 2FeFET-1T TCAM Leveraging Matchline Load Reduction

Fig. 4(a) shows the schematic of our proposed 2FeFET-1T NOR-type TCAM design, which consists of a 2FeFET structure and an nMOS discharge transistor. The wordlines WL_1 and WL_2 control the FeFET gates, and FeFET sources are connected to the searchlines SL_1 and SL_2, respectively. The gate node D of nMOS is connected to the drains of the FeFETs, and the matchline ML connects with the drain of nMOS. As shown in Fig. 4(b), each column shares the same searchlines, and each row shares the same wordlines.

The 2FeFET structure in the proposed cell performs the core function for an efficient TCAM design. Without loss of generality, assume the stored value $S = 1$(M_1 with low V_{TH} state) and $\bar{S} = 0$(M_2 with high V_{TH} state), then $D = SL_1$. Similarly, when $S = 0$ and $\bar{S} = 1$, $D = SL_2$. Fig. 4(c) shows the truth table of the structure performing such XOR-like function

$$D = SL_1 \times S + SL_2 \times \bar{S}. \tag{3}$$

Once (3) can be readily realized, this structure is applicable to other NVM devices and combined with the proposed TCAM designs schemes here as long as the limited ON/OFF resistance ratio of NVMs can be addressed.

Fig. 4(c) summarizes the write and search operations of the proposed 2FeFET-1T TCAM cell. Input bits are written into the FeFETs as states S and $\bar{S}$ (normally the FeFETs stores opposite values except the "do not care" state "X"). To write a logic "0" to the cell, $-V_w$ and V_w are applied to WL_1 and WL_2 to switch the polarization of FeFETs, respectively, and

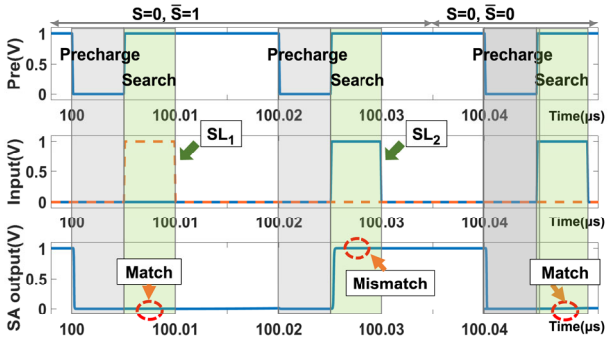


Fig. 5. Transient waveforms of a 2FeFET-1T TCAM.

0 is applied to the searchlines. Similarly, logic "1" can be written by applying V_w and $-V_w$ to WL_1 and WL_2, respectively. $-V_w$ is applied to both wordlines to write a "do not care" state "X." During the write, the wordlines associated with unselected words are set to either $V_w/2$ or $-V_w/2$ to inhibit write disturbance [32]. As such, the voltage of node D will be high/low enough to turn on/off the nMOS for the mismatch/match case, respectively.

The matchlines of the array are precharged to a high level when search operation starts. Meanwhile, $0/V_s$ (1 V) is applied to the searchlines/wordlines to set node D to 0. Then the wordlines and the searchlines are applied with V_s (1 V)/0 according to the input bits as summarized in Fig. 4(c). As a result, the matchline state ML can be formulated as below according to the 2FeFET structure function

$$ML = \bar{D} = \overline{SL_1 \times S + SL_2 \times \bar{S}}. \tag{4}$$

The transient waveforms shown in Fig. 5 validate function of 2FeFET-1T TCAM cell.

Compared with existing precharge-based TCAM designs, our proposed 2FeFET-1T TCAM cell applies the matchline capacitance reduction scheme by reducing the transistor number associated to ML to only 1, therefore consumes less precharge energy

$$C_{ML} \approx C_{pMOS} + N \times (C_{drain} + C_{parasitic}). \tag{5}$$

The search delay of a precharge-based TCAM array is dependent on the effective RC constant as follows:

$$\tau \approx C_{ML} \times R_{eff} \tag{6}$$

where R_{eff} represents the effective resistance between the matchline and ground upon a mismatch search.

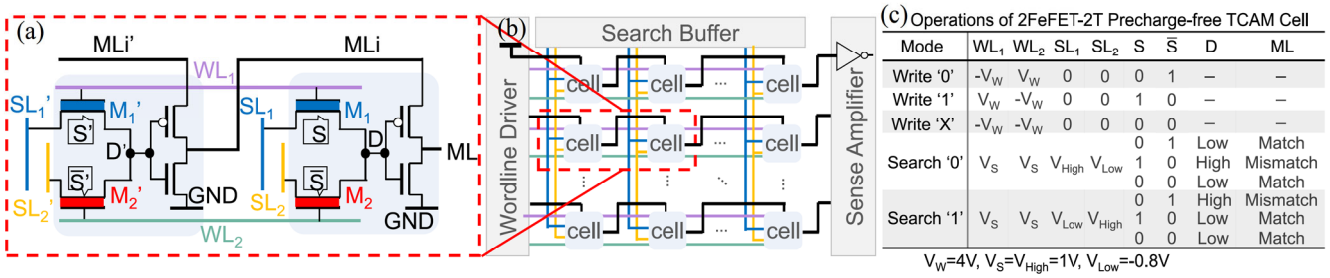


Fig. 6. Schematic of (a) 2FeFET-2T TCAM cell; (b) 2FeFET-2T TCAM array; (c) write and search operation summary.

Equations (2), (5), and (6) imply that the proposed 2FeFET-1T TCAM array can achieve much less search delay than the existing TCAM designs due to the reduction of the matchline load.

B. 2FeFET-2T TCAM Leveraging Precharge-Free Scheme

While the matchline load reduction scheme saves much energy for precharge-based TCAMs, our proposed 2FeFET-2T NAND-type precharge-free TCAM design can further improve the energy efficiency by eliminating the energy-consuming precharge phase. Fig. 6(a) and (b) shows the proposed 2FeFET-2T TCAM cell and the array schematic, respectively. The cell consists of the 2FeFET structure and an inverter supplied by the matchline of the previous cell *MLi*. In the 2FeFET-2T TCAM array, the matchline structure adopts the NAND-type connection, where the matchline of the previous cell connects to the supply rail of the inverter in the current cell.

Fig. 6(c) summarizes the write and search operations. The write scheme as well as the $V_{w/2}$ inihibition scheme of the 2FeFET-2T TCAM design are the same as that of the 2FeFET-1T TCAM design. The truth table is valid only when the previous matchline voltage level *MLi* is high. The matchline state ML is thus formulated as follows:

$$ML = MLi \times \overline{D}. \tag{7}$$

Note that when the previous cell is a mismatch, and *MLi* transitions from high level to ground, the ML of the current cell cannot completely follow the decrease of *MLi* when the internal node *D* is at 0. This is because a pMOS transistor can only pass a weak ground from the source to drain. When *MLi* falls below the pMOS threshold voltage $V_{TH,P}$, the pMOS is turned off, leaving ML at around $V_{TH,P}$. Such incomplete ML swing can be continuously degraded along the array word, resulting in a function failure. To achieve full ML swing, $V_{\text{Low}} = -0.8$ V is applied to the cell searchlines, lowering down node *D* to below $-V_{TH,P}$. The transient waveforms of the proposed 2FeFET-2T TCAM cell shown in Fig. 7 validates the proposed TCAM design.

Comparing to the precharge-based TCAM designs, the proposed 2FeFET-2T TCAM eliminates precharge in most cases, as the matchline state ML of each TCAM cell is determined by both the cell internal output *D* and the matchline state of the previous cell *MLi*. In such design, consecutive searches are performed, and a reset phase discharging the node

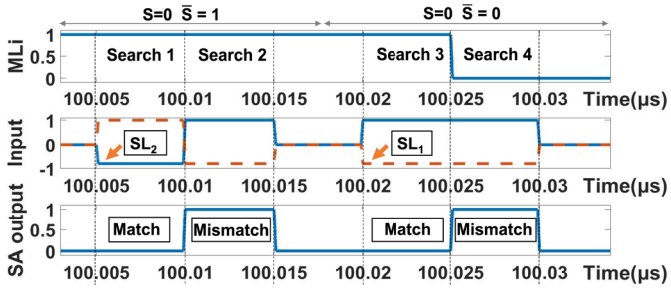


Fig. 7. Transient waveforms of a 2FeFET-2T precharge-free TCAM.

D is not required. Note that the matchline state transition of the previous cell only depends on the state of last evaluation: A mismatch in the last search discharges the *MLi* to ground, and a match in the current search charges the *MLi* again. If consecutive searches result in the same match/mismatch state to the previous cell, *MLi* remains without transitions. The only situation where the matchlines along with a TCAM word will be charged is illustrated below. Cell C_i will be precharged upon a search only when the following conditions are satisfied: 1) the *MLi* of C_i transitions from mismatch state (Low) to match state (High) upon the search, turning on the pMOS for charging ML in C_i and 2) the matchlines of all previous $i-1$ cells are all at match state, thus forming a charging path from voltage supply to the *MLi* of C_i. Such strict conditions greatly reduce the occurrence of charging, therefore consuming much less energy consumption.

Fig. 8 illustrates an example of how the matchlines of the two proposed FeFET-based TCAM arrays are charged in consecutive searches. Without loss of generality, randomly chosen patterns are searched, and we assume that a TCAM array has at most one-word entry matching with the search input. Fig. 8(a) shows that the precharge phase in the precharge-based TCAM designs is inevitable regardless of the matchline state of the last search. However, as shown in Fig. 8(b), 2FeFET-2T precharge-free TCAM array can avoid precharge in most cases.

Take the entry 3 and 5 storing "00 100 110" and "10 100 101" as an example, Fig. 8(c) shows the detailed matchline state transitions and charging situations for searchs 1 and 2. Search 1 results in a mismatch for both entries. Entry 3 has a mismatch at the first cell, therefore all the following cells within entry 3 are mismatch. Since the first 4 cells of entry 5 are at match state, their matchline states are high. In Search 2, the first cell of entry 3 transitions from

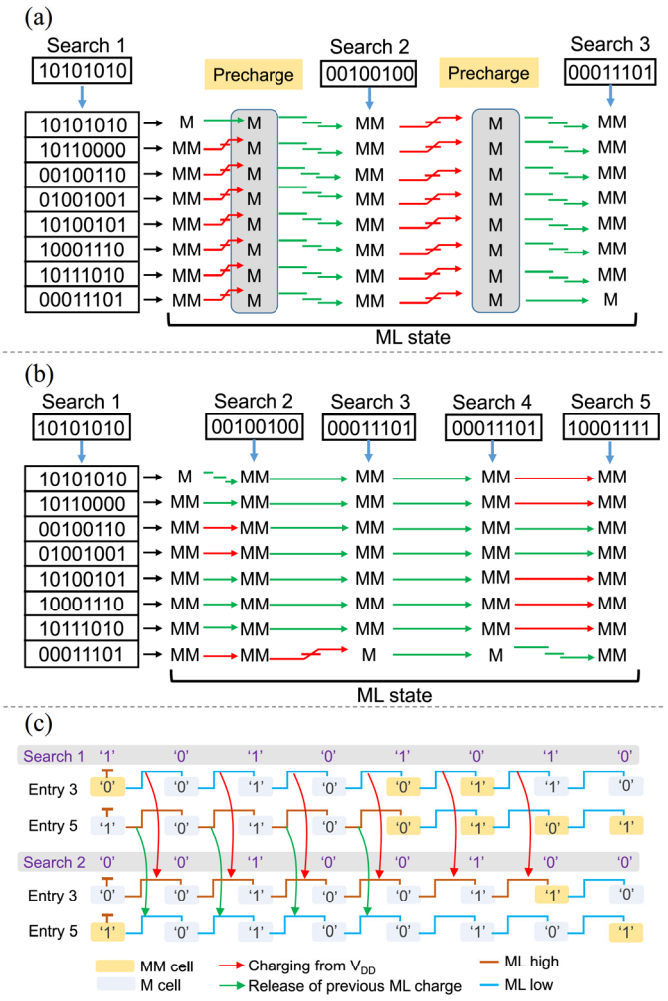


Fig. 8. Functionality of TCAM arrays with an example illustration. *M* represents a match state of an entry, and *MM* represents a mismatch state of an entry. The red arrows indicate a matchline charging from the voltage supply to the entry, while the green arrows mean no matchline charging from the supply to the entry, but rather possible matchline discharge. (a) 2FeFET-1T or other precharge based TCAM array search operations. (b) 2FeFET-2T precharge-free TCAM array search operations. (c) 2FeFET-2T TCAM configurations during search.

mismatch to match, thus the first 6 cells are at match state. The matchlines of the first 6 cells are then charged by the supply, though the entry is still at mismatch state. The first cell of entry 5 transitions from match to mismatch, resulting in the discharge of the matchlines associated with the first 4 cells within the entry. Similar analysis can be applied to other entries, the entry state transitions and the matchline charging situations are visualized in Fig. 8(b), where a red arrow indicates a matchline charging in Fig. 8(a) and (b). It can be seen that much less matchline charging is needed by the 2FeFET-2T precharge-free TCAM array in four consecutive searches (from Search 2 to Search 5), consuming much less search energy than precharge-based TCAM arrays. Moreover, as the number of searches increases, the energy saving of the proposed precharge-free TCAM array over the precharge-based TCAM array will further grow.

Since the proposed 2FeFET-2T TCAM array adopts the NAND-type matchline connection for precharge-free design scheme, the search delay will increase in general. Assuming that initially the cells within a 2FeFET-2T TCAM word are all match or mismatch, the worst case of the word occurs when only the first cell has a matchline state transition. In this case, the SA sensing the last cell output needs to wait until the matchline state transition at the first cell propagates through the entire word, resulting in a large search delay. The energy and performance of the proposed TCAMs against existing approaches are evaluated in Section VI.

IV. Hybrid Ferroelectric NAND-NOR Array

The proposed 2FeFET-1T TCAM in Fig. 4 associates just one transistor with the NOR-type matchline structure, thus resulting in a faster search operation and reduced precharge energy compared with previous designs in [1] and [24]. However, the precharge phase before search still consumes energy. The proposed 2FeFET-2T TCAM in Fig. 6 further improves the energy efficiency by eliminating the precharge energy consumed per search through the serially connected NAND-type matchline structure, despite at the cost of large search delay. To leverage the advantages of both proposed TCAM design styles, here we propose to combine the NOR-type and NAND-type matchline structures, and thus propose an HFNN TCAM array, which achieves ultralow search energy while maintaining acceptable latency.

Fig. 9 shows the proposed HFNN architecture, which consists of an $m \times n$ main array containing the proposed 2FeFET-1T NOR TCAM cells, 2FeFET-1T NAND TCAM cells modified from 2FeFET-2T NAND design [Fig. 9(c) and (d)] and peripheral circuits [Fig. 9(b)]. The main HFNN TCAM array has the data TCAM array associated with m matchlines and a replica associated with the replicated matchline. The replica consists of k series nMOS transistors in the replica NAND-type part and $n-k$ parallel nMOS transistors in the replica NOR-type part, as shown in Fig. 9(a). Each HFNN row consists of k NAND-type TCAM cells and $n-k$ NOR-type TCAM cells. The NAND-type TCAM cells are connected with the NOR-type TCAM cells via NAND and NOR precharge circuitry, and the searchline driver controls the searchlines associated with NOR-type TCAM array. To avoid negative search voltage during the search operation of 2FeFET-2T NAND TCAM, a modified NAND TCAM structure as shown in Fig. 9(c) is adopted. The TCAM replica row contains the same number of NAND and NOR TCAM cells as the main HFNN array, and is used to evaluate the search timing associated with the NAND-type TCAM array.

As a combination of the proposed TCAM design styles, the HFNN array utilizes the advantages of both NAND-type and NOR-type matchline designs. In the HFNN array, only several bits (k) are used for NAND cells and most bits ($n-k$) are used for NOR cells. Fig. 10 shows the activations in NAND-type MLs and NOR-type MLs. In phase 1, all NAND-type parts are activated to perform the search operation with the first k bits of input. In phase 2, the rows whose all NAND TCAM cells are matched activate the corresponding NOR TCAM cells within the same rows, i.e., the corresponding NOR-type MLs are precharged. In phase 3, the activated NOR-type TCAM rows perform search operation and generate the outputs.

In the HFNN array, the NOR-type ML is only precharged when the NAND-type ML corresponding to the same row is

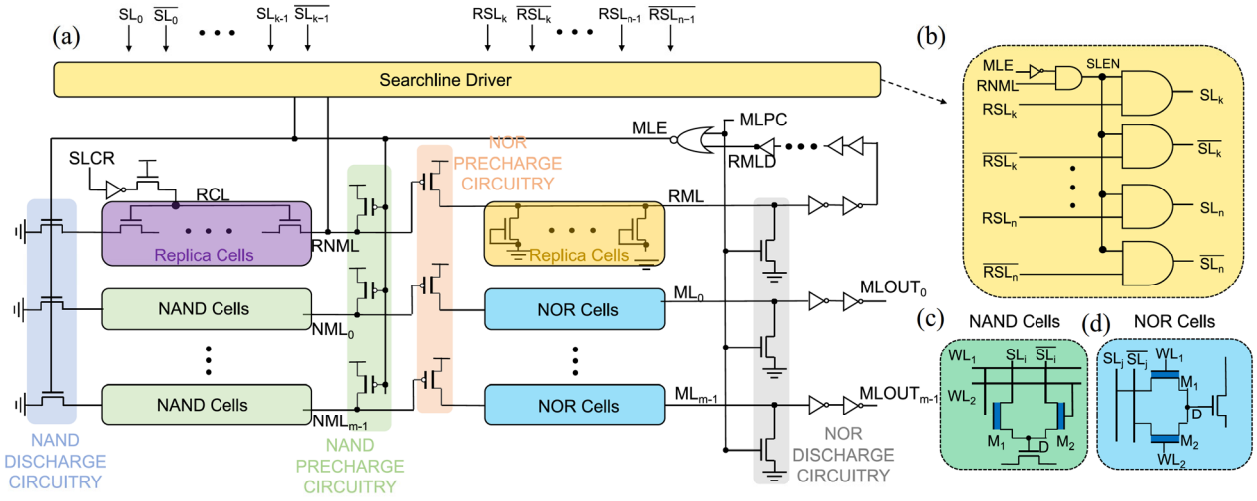


Fig. 9. Schematic of (a) HFNN array; (b) Searchline driver; (c) NAND TCAM cell; (d) NOR TCAM cell.

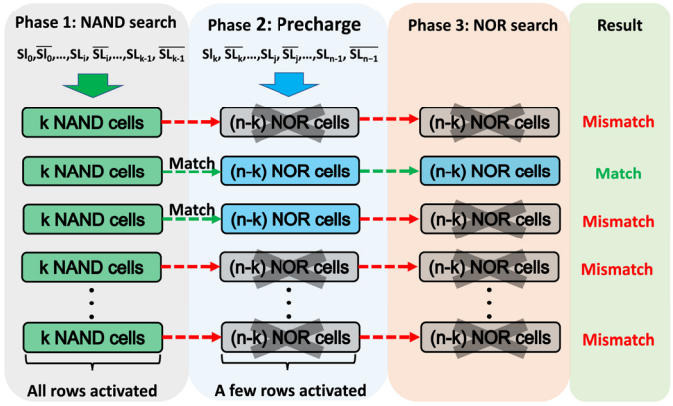


Fig. 10. Activation in NAND-type and NOR-type MLs of HFNN.

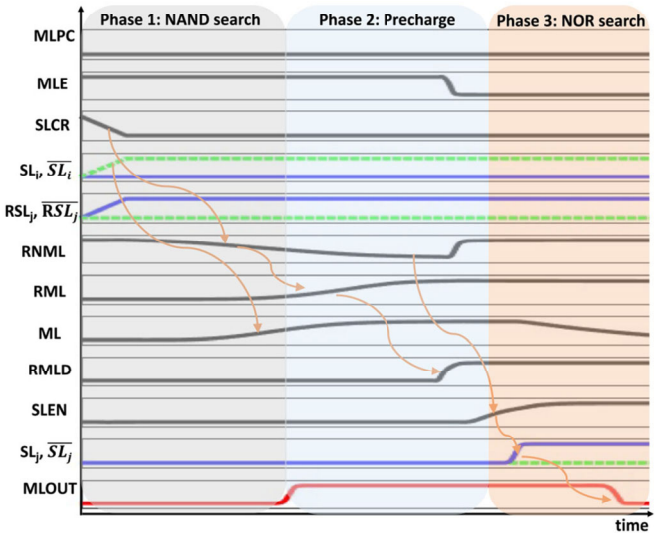


Fig. 11. Waveforms of HFNN array.

discharged to 0, indicating that the NAND TCAM cells match with the first k input bits. Assume that each TCAM cell has the probability of $(1/2)$ to match with the input, then the precharge probability of the NOR-type ML is

$$P_{\text{precharge}} = \left(\frac{1}{2}\right)^k \tag{8}$$

and the average energy of the array can be calculated as follows:

$$E_{\text{average}} = E_{\text{rep}_{\text{nand}}} + E_{\text{rep}_{\text{nor}}} + \frac{E_{\text{nand}} + E_{\text{nor}}}{2^k} \tag{9}$$

$$E_{\text{average/bit}} = \frac{E_{\text{average}}}{m \times n} \tag{10}$$

where E_{average} is the average energy consumption of the HFNN array. $E_{\text{rep}_{\text{nand}}}$ and $E_{\text{rep}_{\text{nor}}}$ are the energy consumed by the NAND TCAM cells and NOR TCAM cells within the replica row, respectively. E_{nand} and E_{nor} are the search energy consumption of the NAND and NOR TCAM cells in the main array, respectively. $E_{\text{average/bit}}$ is the average search energy per bit per search. According to (9), the HFNN saves the energy consumption of NOR TCAM part by reducing the number of activated NOR-type MLs that need to be precharged from m to $(m/2^k)$. The energy consumption associated with each part of HFNN can be derived as follows:

$$E_{\text{rep}_{\text{nand}}} = E_{\text{nand}} = ((2k+1) \times C_{\text{drain}} + C_{\text{pMOS}}) \times V_{DD}^2 \tag{11}$$

$$E_{\text{rep}_{\text{nor}}} = E_{\text{nor}} = ((n-k) \times C_{\text{drain}} + C_{\text{pMOS}}) \times V_{DD}^2 \tag{12}$$

$$E_{\text{average/bit}} = \frac{(n+1+k) \times C_{\text{drain}} + 2C_{\text{pMOS}}}{m \times n} \tag{13}$$

$$\times \frac{V_{DD}^2 \times (1 + \frac{1}{2^k})}{m \times n}. \tag{14}$$

Note that here we consider the worst precharge case, i.e., each NAND cell is precharged.

The overall search operation of an HFNN array can be divided into three phases, and the corresponding waveforms are illustrated in Fig. 11. In phase 1, the discharge control signal *MLPC* in Fig. 9 is set to "1" to activate the NOR discharge circuitry, and reset the MLs to ground, as shown in Fig. 12(a). As the replicated matchline *RML* is discharged to

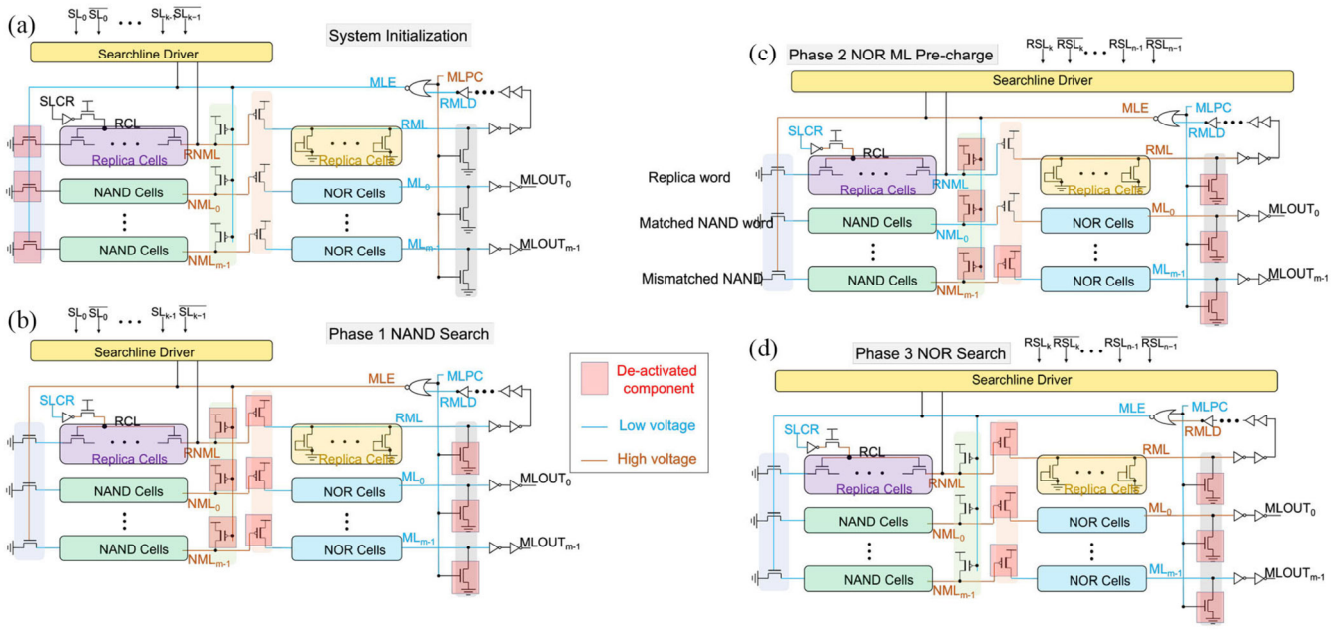


Fig. 12. Signal states during the three phase operation of the HFNN array: (a) Initialization; (b) Phase 1 NAND search; (c) Phase 2 NOR cells ML precharge; (d) Phase 3 NOR search.

ground, the other input of the NOR gate *RMLD* is set to "0." Meanwhile, the NOR gate output *MLE* is "0," and activates the NAND precharge circuitry, thus charging the replicated NAND matchline (*RNML*), NAND-type MLs (*NML*s) to high voltage level, as shown in Fig. 12(a). Upon the search operation of NAND-type part, *MLPC* is set to "0" to deactivate the NOR discharge circuitry, and thus *MLE* turns to "1," activating the NAND discharge circuitry, as shown in Fig. 12(b). The searchlines ($SL_i/\overline{SL_i}$) associated with the NAND TCAM cells and raw searchlines ($RSL_j/\overline{RSL_j}$) associated with the NOR TCAM cells are driven with search voltages according to the input data, respectively. Meanwhile, the searchline control signal (*SLCR*) associated with the replicated NAND TCAM row is set to "0," thus turning on all the transistors within the NAND replica, as shown in Fig. 12(b). *RNML* starts to discharge through the nMOS transistor series. As the discharge path contains the same number of transistors as other NAND TCAM rows, *RNML* has the same discharge delay as the matched NAND-type rows, thus is used as an indicator of the NAND search operation.

In phase 2, as *RNML* is discharged to "0," the NAND search operation finishes, and the NOR precharge circuitry is activated as shown in Fig. 12(c). The replicated NOR matchline (*RML*) is precharged to "1," and then *RMLD* turns to "1." The buffer chain associated with *RMLD* generates a delay to ensure that the NOR-type MLs associated with the matched NAND-type rows are fully precharged to "1." Then *MLE* becomes "0" again, turning on the NAND precharge circuitry, as shown in Fig. 12(d). *RNML*s and *RML*s are precharged to "1," and the NOR precharge circuitry is deactivated, disconnecting all MLs from the voltage supply. Meanwhile, *SLEN* in the searchline driver as shown in Fig. 9(b) becomes "1," thus passing the raw searchline signals RSL_js to the corresponding searchlines SL_js associated with the NOR TCAM cells. As shown in Fig. 12(d), in phase 3, the NOR-type TCAM rows corresponding to the matched NAND-type rows perform search operation, and the matched NOR TCAM row will generate a match output, indicating a fully matched word with the input.

It can be seen that HFNN array can significantly reduce the number of activated MLs by using the NAND-type TCAM rows to filter out the mismatched rows, thus saving the precharge energy. The architecture consumes precharge energy associated with all NAND-type MLs and a few NOR-type MLs, thus the energy consumption of HFNN is a bit larger than that of an NAND-type TCAM array considering a single row in the worst case. However, at the array level, the increasing precharge energy consumption of NAND part is far less than the precharge energy reduction in NOR part, which will be explained later in Section VI. The delay of HFNN is a bit larger than that of an NOR-type TCAM array, but much less than that of a NAND-type TCAM array.

V. HFNN-Based Segmented Architecture

While HFNN performs the search operation with much faster speed than NAND-type TCAM array, its delay as well as the energy consumption increase as the wordlength grows. To accommodate more bits for the search operation, we further propose HFNN-based segmented architecture as well as its pipeline operations for large-scale data search tasks. Fig. 13 shows the segmentation architecture. It consists of the data encoder, HFNN segments, and the latch circuitry between the segments. The data encoder drives the searchlines to HFNN segments per input data. Each segment contains a proposed HFNN array which performs the search operation as discussed in Section IV. The latch circuitry consists of a 3-input NAND gate, a latch nMOS transistor, and a buffer. Without loss of generality, the gate of the nMOS transistor, G_l, is driven by

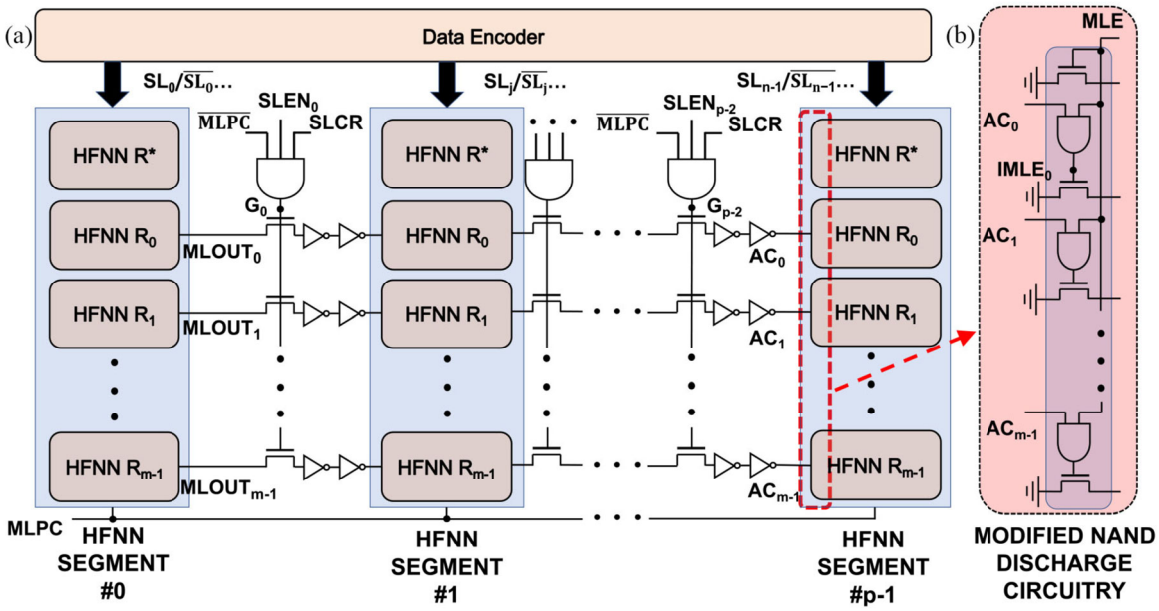


Fig. 13. Schematic of (a) main part of HFNN-based segmented architecture where HFNN R* represents the replica row and HFNN Rm represents the *m*th row of HFNN array; (b) Modified NAND discharge circuity.

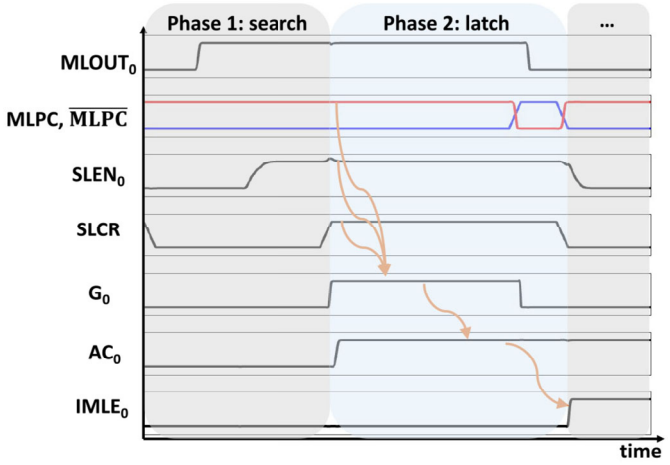


Fig. 14. Operation waveforms of HFNN-based segmented architecture.

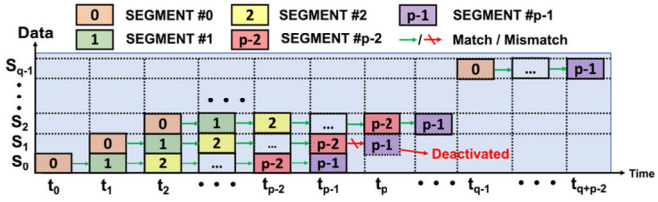


Fig. 15. Pipeline operation diagram of HFNN-based segmented architecture.

the NAND gate as follows:

$$G_l = \overline{MLPC} \cdot SLCR \cdot SLEN_l \tag{15}$$

where $\overline{MLPC}$ and $SLCR$ are the global control signals to enable the search, and $SLEN_l$ is the signal of the searchline driver in the lth HFNN segment, respectively, as shown in Fig. 9. The output signal of the latch circuitry corresponding to the ith row AC_i is the input of the NAND discharge circuitry corresponding to the ith row in the HFNN as shown in Fig. 13(b). For each HFNN segment except the first one, the NAND discharge circuitry in the HFNN segment is modified as shown in Fig. 13(b), the nMOS discharge transistor of the NAND-type TCAM array is driven by MLE and AC_i

$$IMLE_i = AC_i \cdot MLE. \tag{16}$$

Fig. 14 shows the simulated waveforms of an HFNN-based segmented architecture containing p HFNN segments. In phase 1, the HFNN segments search the input data across the stored data, and generate the outputs. Only the matching row outputs of the HFNN segment *MLOUT*'s become "1," otherwise "0." In phase 2, *SLCR* is set to "1," thus the gate voltage of the nMOS in the latch circuitry becomes "1" according to (15). The outputs of the HFNN segment are passed through the nMOS and the buffer, driving the outputs of latch circuitry *AC*'s to the value of *MLOUT*'s. Then, the rise of the global signal *MLPC* turns off the nMOS, latches the outputs of HFNN segments and resets all segments. In the next round of search, *MLPC* is set to "0," and MLE is driven to "1." If one row's latched output of the previous HFNN segment is "1," then the gate of the discharge nMOS within the modified NAND discharge circuitry *IMLE* equals to the control signal *MLE* according to (16), and the corresponding row within the HFNN segment is activated for the next search.

With the proposed HFNN-based segmented architecture, the search operation across large-scale data words can be pipelined. Fig. 15 shows an example of one HFNN row in the segmented architecture. S_0, S_1, ..., S_{q-1}, etc., are the input words with large wordlength. Each search word is divided to p subwords that have the same wordlength as an HFNN segment, thus each subword can be loaded into the HFNN segment for the search operation. During t_0, the subword of S_0 corresponding to HFNN segment #0 is searched, and if the output

TABLE I
METRIC COMPARISON SUMMARY OF TCAM DESIGNS

Reference	[13]	[16]	[17]	[24]	[27]	[34]	[35]	Fig. 4	Fig. 6	Fig. 9
CAM type	TCAM	TCAM	TCAM	TCAM	TCAM	BCAM*	BCAM*	TCAM	TCAM	TCAM
Technology	CMOS	ReRAM	ReRAM	FeFET	CMOS	CMOS	CMOS	FeFET	FeFET	FeFET
Transistors/ cell	16T	2T-2R	3T-1R	2FeFET	14T	10T	10T	2FeFET -1T	2FeFET -2T	HFNN
Node	$45nm$	$90nm$	$90nm$	$45nm$	$45nm$	$45nm$	$45nm$	$45nm$	$45nm$	$45nm$
Search style†	P	P	P	P	PF	PF	PF	P	PF	P
Supply[V]	1	1.2	1	1	1	1	1	1	1	1
Search delay [ns]	0.58	0.35	0.96	0.34	~20	1.25	-	0.25	1.43	1.23
EfS [fJ/bit/search]	0.59 226.92X	0.55 211.54X	0.51 205.43X	0.35 134.62X	0.18 63.23X	2.1 807.69X	0.66 2.54X	0.195 75.00X	0.073 28.08X	0.0026 1X
EfS_N [fJ/bit/search]	0.59 226.92X	0.23 88.46X	0.26 100.00X	0.35 134.62X	0.18 63.23X	2.1 807.69X	0.66 2.54X	0.195 75.00X	0.073 28.08X	0.0026 1X

⋆: BCAM denotes binary content addressable memory, storing binary states ('0'/'1').
†: P denotes precharge-based, PF denotes precharge-free.

is a match, then the subword of S_0 corresponding to HFNN segment #1 is searched during t_1, otherwise the search operation with this word terminates. Meanwhile, the subword of S_1 corresponding to HFNN segment #0 is loaded and searched. Following the above pipeline, the search words are loaded into the segmented architecture in time series, and ultimately a word search operation can be realized using the same delay as an HFNN segment. Such a pipeline search scheme significantly enables a fast search for large scale data words, and saves significant energy by terminating the search operation at the early stage of pipeline. It therefore imposes huge potential for low power, high speed, and data-intensive applications.

VI. EVALUATION

In this section, we evaluate and compare the search energy and delay of the proposed TCAM arrays with 16T CMOS, 2T-2R ReRAM, 14T CMOS [27], and the state-of-art 2FeFET-based TCAMs to validate the benefits of the proposed two design schemes with FeFETs and the hybrid design scheme.

A. TCAM Array Evaluation

Our proposed FeFET-based TCAM arrays [Figs. 4(b), 6(b), and 9(a)] are evaluated through SPECTRE simulations at 27 °C, and compared against existing state-of-the-art TCAM designs. The 45-nm PTM model with TT process corner is adopted for MOSFET devices with minimized sizes. LRS/HRS of 20 kΩ/2 MΩ is assumed for 2T-2R ReRAM-based TCAM. Wiring parasitics are extracted from DESTINY [26].

Table I summarizes the performance metrics of 64 × 64 TCAM arrays, including the CAM type, transistor count, technology, search style, search delay, search energy per bit per search (EfS), and normalized search energy per bit per search (EfS_N) defined in [33], which is for legitimate comparison

$$EfS_N = EfS \times \left(\frac{45\text{ nm}}{\text{Node}}\right) \times \left(\frac{1}{V_{DD}}\right)^2 \tag{17}$$

where the energy is normalized to 45 nm/1.0 V per (17).

Since the proposed 2FeFET-2T precharge-free TCAM array only needs precharge upon the matchline conditions discussed in Section III-B, the corresponding search energy is highly dependent on the search patterns. We assumed the randomly chosen search patterns in Fig. 8(b) for evaluations. For the precharge-based TCAM arrays, the worst case where only one-bit mismatch occurs is measured for the search delay. The worst case of the proposed precharge-free TCAM array as discussed in Section III-B is measured for the search delay.

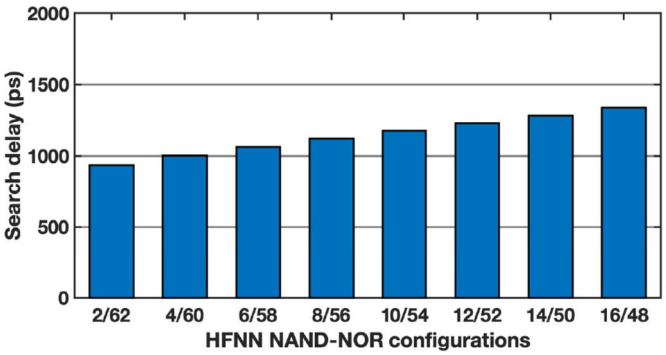


Fig. 16. Delay of HFNN with different NAND-NOR configurations (x/y denotes that each row consists of x NAND cells and y NOR cells).

The proposed 2FeFET-1T TCAM and 2FeFET-2T TCAM consume 32.1% and 39.3% of the cell size of a conventional 16T CMOS TCAM, respectively. Less cell area overhead of the proposed TCAMs leads to less parasitic capacitance associated with the matchlines at the array level, resulting in less precharge energy compared with the CMOS TCAM array. The search energy and delay in Table I show that our proposed 2FeFET-1T TCAM is 3.02×/1.79× more energy efficient and 2.30×/1.35× faster than the 16T CMOS TCAM/2FeFET TCAM, respectively. The results validate the efficiency of the matchline load reduction scheme which associates only one transistor to the matchline, in addition to the area saving obtained by FeFETs. Our proposed 2FeFET-2T precharge-free TCAM is 8.08×/4.79× more energy efficient than the 16T CMOS/2FeFET TCAM, at the expense of larger search delay due to the NAND-type matchline structure. The energy efficiency again proves the advantages of the precharge-free scheme with FeFETs. The maximum operating frequencies of these 64 × 64 TCAM arrays are further evaluated. The working cycle of NOR-type 2FeFET-1T TCAM consists of a precharge phase consuming 0.27 ns and a search phase consuming 0.25 ns, respectively, while the NAND-type 2FeFET-2T TCAM only performs a search with 1.43-ns latency due to its precharge-free scheme. The corresponding operating frequency of 2FeFET-1T/2FeFET-2T TCAM can therefore reach around 1.92 GHZ/699 MHz.

As discussed in Section IV, the numbers of NAND and NOR TCAM cells within an HFNN row, i.e., k and $(n-k)$, highly affect the energy and delay of the whole HFNN array. In the 64×64 HFNN array, different NAND-NOR configurations, i.e., $(k/(n-k))$, are evaluated in terms of delay and energy consumption, which are shown in Figs. 16 and 17,[1] respectively. The delay of the HFNN array is measured in the worst case where only one cell within the NOR-type TCAM row is mismatched. The matching probability of each TCAM cell is assumed to be (1/2). It can be seen that as the number of NAND TCAM cells increases, the delay of the entire array rises due to the increased number of series

[1]Fig. 17 reports the energy breakdown of an HFNN array that includes the activated NOR-type matchlines.

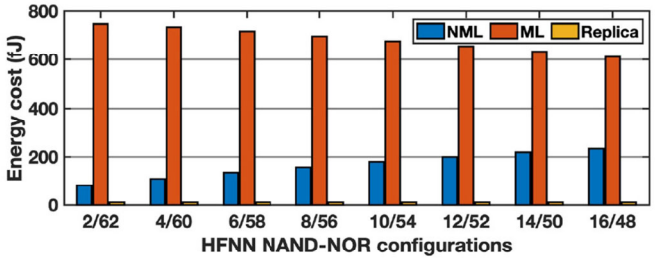


Fig. 17. Energy of HFNN with different NAND-NOR configurations (*x*/*y* denotes that each row consists of *x* NAND cells and *y* NOR cells).

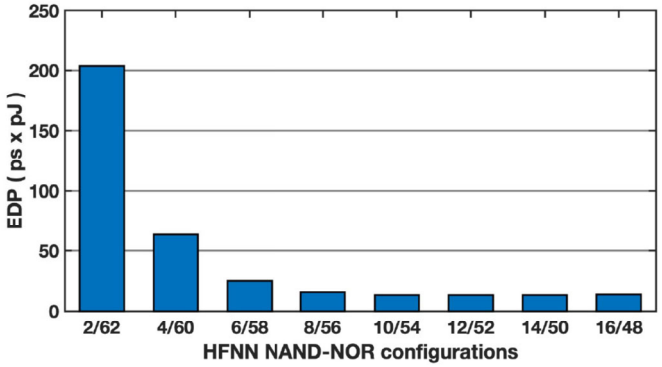


Fig. 18. EDP of HFNN with different NAND-NOR configurations (*x*/*y* denotes that each row consists of *x* NAND cells and *y* NOR cells).

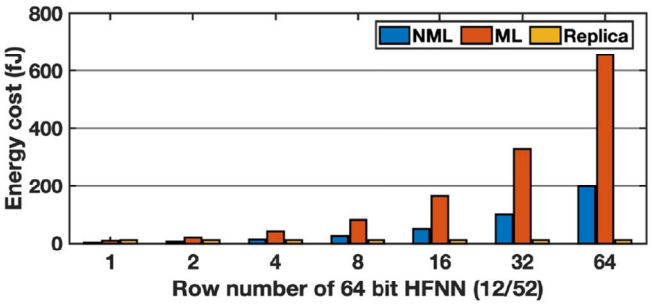


Fig. 19. Energy of HFNN (12/52) with different row numbers (12/52 denotes that each row consists of 12 NAND cells and 52 NOR cells).

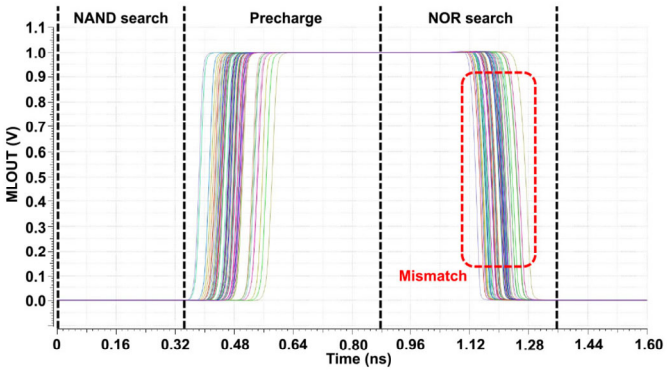


Fig. 20. 60 Monte Carlo simulation waveforms of an HFNN array (12/52) assuming a standard deviation of 45 mV for FeFET V_{TH} [36], and 10% for the transistor size [37].

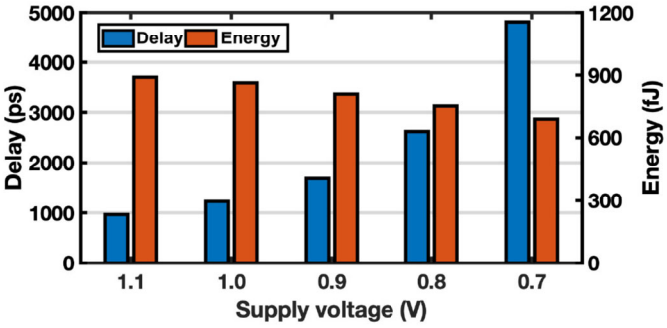


Fig. 21. Delay and energy of HFNN (12/52) under different supply voltages.

transistors along the discharge path in the NAND-type TCAM part. Fig. 17 shows the search energy consumption for the matchlines within the HFNN array. It mainly consists of the energy consumption by the NAND-type matchline (*NML*) and the NOR-type matchline. As the number of NAND TCAM cells increases, more series transistors of the NAND-type TCAM array are precharged, resulting in the increasing energy consumption associated with *NML*. On the contrary, the decrease of NOR TCAM cells results in the reduction of energy associated with the NOR-type matchline. The energy consumption by the replica row is negligible. According to (9) and (13), though increasing k leads to more *NML* precharge energy and larger search delay, the number of activated NOR-type matchlines is highly reduced, resulting in significant overall energy saving. Considering the tradeoff between the delay and the overall energy, Fig. 18 shows the energy-delay product (EDP) of the HFNN array. It can be seen that the HFNN with configuration, i.e., HFNN (12/52) achieves the optimal EDP. The working cycle of the HFNN (12/52) array is composed of a reset phase (set MLPC to "1" to discharge all NOR matchlines) taking 0.22 ns, and a search phase consuming 1.23-ns latency, respectively. The maximum operating frequency of the HFNN array with such configuration is thus 619 MHz.

Fig. 19 shows the search energy consumption breakdown of the HFNN array (12/52) with varying numbers of rows. It can be seen from Fig. 19 that the respective energy consumption of NAND-type and NOR-type matchlines of HFNN array linearly grows as the row number increases, resulting in linearly growing energy. Table I also compares the HFNN array with existing TCAM designs. The average search energy and delay metrics suggest that our proposed HFNN array can achieve 63.23×/28.08× more energy efficiency and 16.26×/1.16× faster search speed than 14T CMOS /2FeFET-2T TCAM, respectively. HFNN array is also 226.92×/134.62× more energy efficient than 16T CMOS TCAM/2FeFET TCAM, respectively. The PVT conditions of the proposed HFNN array are also validated. Fig. 20 shows the Monte Carlo simulation results of the HFNN array (12/52) with 60 runs, where standard deviations of 45 mV for the FeFET V_{TH} [36] and 10% for the transistor size [37] are applied, respectively. Fig. 21 shows the delay and energy metrics of HFNN (12/52) under scaling supply voltages. It can be seen that the energy decreases as the supply voltage scales down, which is consistent with (1). However, lower V_{DD} consumes more time to charge the associated load capacitance and switch the array output state,

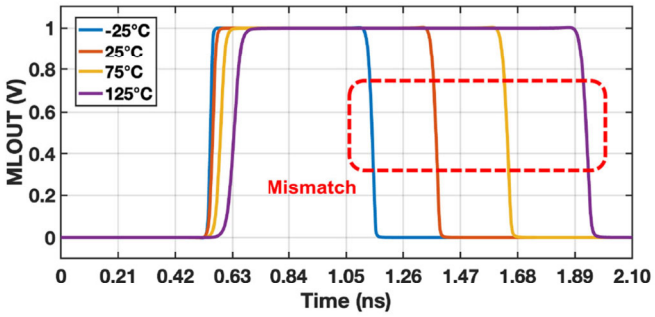


Fig. 22. Output waveforms of HFNN array (12/52) at different temperatures.

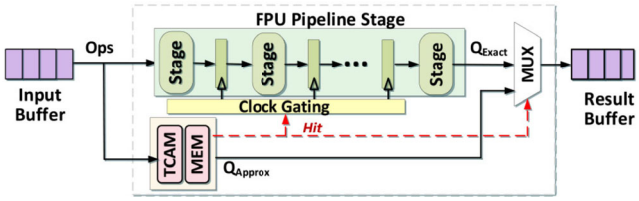


Fig. 23. GPU cores enhanced with TCAM block.

thus resulting in increasing search delay. Higher supply voltage leads to reduced search time. Fig. 22 shows the impact of varying temperatures on the performance of HFNN array (12/52). While higher temperature shifts the rise/fall timing of HFNN array output and increases the latency, HFNN array still functions correctly and thus is robust to the temperature.

B. Benchmarking of TCAM-Based Associative Memory

To benchmark the TCAM designs at the application level, we use a modified version of Multi2sim [38], a cycle accurate CPU-GPU simulator. The kernel code is modified to implement the enhanced GPU architecture and enable runtime simulation. To show the generality of our approach, we apply TCAMs to Nvidia Kepler GeForce GTX Titan. TCAMs are implemented next to the FPU within each of the cores in the GPUs. As Fig. 23 shows, TCAM is implemented for the GPU FPU operations which make up the majority of computation within the tested applications: adder (ADD), multiplier (MUL), and multiply accumulator (MAC). Our benchmarks include a wide range of signal processing applications and the Caltech 101 dataset [39].

The TCAMs and the associated memory are located close to the first stage of FPUs, storing frequent arithmetic operations and corresponding results, respectively. During the computation task, the enhanced GPU architecture searches the input data across the TCAM array, and meanwhile executes the first stage of the FPUs. Upon a TCAM match, a prestored result corresponding to the matched entry will be fetched as the output, while the FPU pipeline will be clock-gated and the subsequent FPU stages will not be executed. With low-power TCAM designs, such architecture reduces repeated and expensive FPU computations by exploiting the computation reuse, thus significantly reducing the energy consumption of computation cores for low-power applications [40], [41].

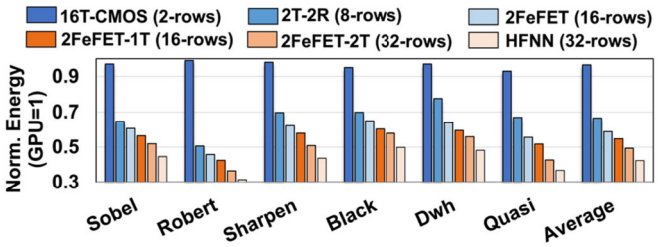


Fig. 24. Normalized energy consumption of enhanced GPU integrating different TCAM arrays.

Said associative memory-enhanced architecture faces the following challenges: 1) the energy and performance cost of TCAM search is inevitable and 2) a larger TCAM array typically has a higher hit rate which is desirable for the data-intensive applications. However, a TCAM array with a large number of rows consumes high power consumption, thus degrading the benefits of higher computation reuse in the enhanced GPU architecture. Fig. 24 shows the normalized energy consumption of the architecture enhanced with the proposed TCAMs and other TCAMs.[2] We show the TCAM array results with the number of rows that maximizes the energy efficiency. 16T CMOS TCAMs achieve maximum efficiency using a small number of rows, as the TCAM search energy overhead dominates the gain coming from a higher hit rate. Results show that the proposed 2FeFET-1T/2FeFET-2T/HFNN TCAM can save, on average, 45.2%/50.6%/57.5% energy consumption compared to the conventional GPU. Compared with the state-of-the-art 2FeFET TCAM approach, our proposed designs can still achieve 4.9%/10.4%/16.5% more energy saving. The high-efficiency results from the low power of the proposed TCAMs, which allow a large TCAM array with a high hit rate, thus reducing the expensive executions of GPU cores.

VII. Conclusion

TCAM is a hardware kernel of CiM design that can perform parallel in-memory search operations across the memory, thus is a promising candidate for information processing in MANN, IP routers, and advanced machine learning models. In this article, we propose a novel FeFET-based TCAM design that exploits NVMs and three potential design methods for energy efficiency improvements. We first present the proposed 2FeFET-1T NOR-type and 2FeFET-2T precharge-free NAND-type TCAMs, analyze their energy efficiency over existing approaches. We then propose HFNN array that leverages the advantages of both NAND-type and NOR-type TCAM designs to realize ultra energy-efficient search operation. An HFNN-based segmented architecture is presented for large scale array search. We evaluate and benchmark the proposed TCAM designs, and the results indicate that our proposed TCAM design methods with FeFETs can achieve promising energy efficiency and performance with respect to existing TCAM designs.

[2]Excluding 14T CMOS TCAM due to larger search delay than GPU cycle.

REFERENCES

[1] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nat. Electron.*, vol. 2, no. 11, pp. 521–529, 2019.

[2] C. Chen, W. Qian, M. Imani, X. Yin, and C. Zhuo, "PAM: A piecewise-linearly-approximated floating-point multiplier with unbiasedness and configurability," *IEEE Trans. Comput.*, early access, Dec. 1, 2021, doi: 10.1109/TC.2021.3131850.

[3] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs)," *IEEE Trans. Biomed. circuits Syst.*, vol. 12, no. 1, pp. 106–122, Feb. 2018.

[4] X. S. Hu *et al.*, "In-memory computing with associative memories: A cross-layer perspective," in *Proc. IEEE Int. Electron Devices Meeting*, 2021, pp. 25–32.

[5] T. Kohonen, *Associative Memory: A System-Theoretical Approach*, vol. 17. New York, NY, USA: Springer, 2012.

[6] D. Gao, D. Reis, X. S. Hu, and C. Zhuo, "EVA-CIM: A system-level performance and energy evaluation framework for computing-in-memory architectures," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 12, pp. 5011–5024, Dec. 2020.

[7] C.-K. Liu *et al.*, "COSIME: FeFET based associative memory for in-memory cosine similarity search," 2022, *arXiv:2207.12188*.

[8] X. Yin *et al.*, "An ultra-compact single FeFET binary and multi-bit associative search engine," 2022, *arXiv:2203.07948*.

[9] X. Yin *et al.*, "Deep random forest with ferroelectric analog content addressable memory," 2021, *arXiv:2110.02495*.

[10] 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.

[11] A. F. Laguna, H. Gamaarachchi, X. Yin, M. Niemier, S. Parameswaran, and X. S. Hu, "Seed-and-vote based in-memory accelerator for DNA read mapping," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, 2020, pp. 1–9.

[12] 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*, 2020, pp. 29–3.

[13] 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.

[14] H.-S. P. Wong *et al.*, "Metal–oxide RRAM," *Proc. IEEE*, vol. 100, no. 6, pp. 1951–1970, Jun. 2012.

[15] K. Ni *et al.*, "Critical role of interlayer in $Hf_{0.5}$ $Zr_{0.5}$ O_2 ferroelectric FET nonvolatile memory performance," *IEEE Trans. Electron Devices*, vol. 65, no. 6, pp. 2461–2469, Jun. 2018.

[16] 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. IEEE Symp. VLSI Technol.*, 2013, pp. C104–C105.

[17] M.-F. Chang *et al.*, "A 3T1R nonvolatile TCAM using MLC ReRAM with sub-1ns search time," in *Proc. Int. IEEE Solid-State Circuits Conf. (ISSCC)*, 2015, pp. 1–3.

[18] K. Ni, M. Jerry, J. A. Smith, and S. Datta, "A circuit compatible accurate compact model for ferroelectric-FETs," in *Proc. IEEE Symp. VLSI Technol.*, 2018, pp. 131–132.

[19] T. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, "Ferroelectricity in Hafnium oxide: CMOS compatible ferroelectric field effect transistors," in *Proc. Int. Electron Devices Meeting*, 2011, pp. 24–5.

[20] S. Thomann *et al.*, "On the reliability of in-memory computing: Impact of temperature on ferroelectric TCAM," in *Proc. IEEE 39th VLSI Test Symp. (VTS)*, 2021, pp. 1–6.

[21] G. Paim *et al.*, "On the resiliency of NCFET circuits against voltage over-scaling," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 4, pp. 1481–1492, Apr. 2021.

[22] A. Aziz, S. Ghosh, S. Datta, and S. K. Gupta, "Physics-based circuit-compatible SPICE model for ferroelectric transistors," *IEEE Electron Device Lett.*, vol. 37, no. 6, pp. 805–808, Jun. 2016.

[23] S. Deng *et al.*, "A comprehensive model for ferroelectric FET capturing the key behaviors: Scalability, variation, stochasticity, and accumulation," in *Proc. IEEE Symp. VLSI Technol.*, 2020, pp. 1–2.

[24] 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.

[25] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of pMOS NBTI effect for robust nanometer design," in *Proc. 43rd ACM/IEEE Design Autom. Conf.*, 2006, pp. 1047–1052.

[26] M. Poremba, S. Mittal, D. Li, J. S. Vetter, and Y. Xie, "Destiny: A tool for modeling emerging 3D NVM and EDRAM caches," in *Proc. IEEE Design Autom. Test Europe Conf. Exhibit.*, 2015, pp. 1543–1546.

[27] T. V. Mahendra, S. W. Hussain, S. Mishra, and A. Dandapat, "Energy-efficient precharge-free ternary content addressable memory (TCAM) for high search rate applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 7, pp. 2345–2357, Jul. 2020.

[28] B.-D. Yang and L.-S. Kim, "A low-power CAM using pulsed NAND-NOR match-line and charge-recycling search-line driver," *IEEE J. Solid-State Circuits*, vol. 40, no. 8, pp. 1736–1744, Aug. 2005.

[29] S. Choi, K. Sohn, and H.-J. Yoo, "A 0.7-fJ/bit/search 2.2-ns search time hybrid-type TCAM architecture," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 254–260, Jan. 2005.

[30] S.-H. Yang, Y.-J. Huang, and J.-F. Li, "A low-power ternary content addressable memory with Pai-sigma matchlines," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 10, pp. 1909–1913, Oct. 2012.

[31] Y. Qian *et al.*, "Energy-aware designs of ferroelectric ternary content addressable memory," in *Proc. IEEE Design Autom. Test Europe Conf. Exhibit. (DATE)*, 2021, pp. 1090–1095.

[32] S. Mueller *et al.*, "From MFM capacitors toward ferroelectric transistors: Endurance and disturb characteristics of HfO_2-based FeFET devices," *IEEE Trans. Electron Devices*, vol. 60, no. 12, pp. 4199–4205, Dec. 2013.

[33] P.-T. Huang and W. Hwang, "A 65 nm 0.165 fJ/bit/search 256×144 TCAM macro design for IPv6 lookup tables," *IEEE J. Solid-State Circuits*, vol. 46, no. 2, pp. 507–519, Feb. 2011.

[34] T. V. Mahendra, S. Mishra, and A. Dandapat, "Self-controlled high-performance precharge-free content-addressable memory," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 8, pp. 2388–2392, Aug. 2017.

[35] T. V. Mahendra, S. W. Hussain, S. Mishra, and A. Dandapat, "Low discharge precharge free matchline structure for energy-efficient search using CAM," *Integration*, vol. 69, pp. 31–39, Nov. 2019.

[36] T. Soliman *et al.*, "Ultra-low power flexible precision fefet based analog in-memory computing," in *Proc. IEEE Int. Electron Devices Meeting*, 2020, pp. 1–4.

[37] S. N. Mozaffari and A. Afzali-Kusha, "Statistical model for subthreshold current considering process variations," in *Proc. 2nd Asia Symp. Qual. Electron. Design (ASQED)*, 2010, pp. 356–360.

[38] X. Gong, R. Ubal, and D. Kaeli, "Multi2Sim kepler: A detailed architectural GPU simulator," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, 2017, pp. 269–278.

[39] F.-F. Li, F. Rob, and P. Pietro, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 178.

[40] 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.

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

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. His research interests include emerging circuit/architecture designs and novel computing paradigms with both CMOS and emerging technologies.

Dr. Yin has received the Best Paper Award Nomination of ICCAD 2020, the DATE2022 and VLSI Test Symposium 2021, and the Bronze Medal of Student Research Competition at ICCAD 2016. He serves as the Associate Editor for ACM SIGDA E-Newsletter.

Yu Qian received the B.S. degree in microelectronic science and engineering from Zhejiang University, Hangzhou, China, in 2022, where he is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering.

His current research interests include computing-in-memory circuit/architecture designs.

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 with 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.

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 in 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 non-volatile memory and novel computing paradigms. He is currently an Assistant Professor of Electrical and Microelectronic Engineering with the 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.

Chao Li is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang Unversity, Hangzhou, China.

His research interests include design and optimization of in-memory computing architectures.

Grace Li Zhang (Member, IEEE) received the Dr.-Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2018.

She is currently a Postdoctoral Researcher pursuing Habilitation with the Chair of Electronic Design Automation, TUM, where she leads the research team on heterogeneous computing. Her research interests include neural networks and neuromorphic computing, computer architectures, and machine learning for EDA.

Ms. Zhang has served/is serving on the technical committee of several conferences, including ICCAD, DATE, ASP-DAC, and GLSVLSI.

Bing Li (Senior Member, IEEE) received the Dr.-Ing. and Habilitation degrees from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2018, respectively.

He is currently a Researcher with the Chair of Electronic Design Automation, TUM. His research interests include high-performance and low-power design of integrated circuits and systems.

Dr. Li has served on the Technical Program Committees of several conferences, including DAC, ICCAD, DATE, and ASP-DAC.

Ulf Schlichtmann (Senior Member, IEEE) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and information technology from the Technical University of Munich (TUM), Munich, Germany, in 1990 and 1995, respectively.

He is a Professor and the Head of the Chair of Electronic Design Automation with TUM. He joined TUM in 2003, following ten years in industry. His current research interests include computer-aided design of electronic circuits and systems, with an emphasis on designing reliable and robust systems. Increasingly, he focuses on emerging technologies, such as lab-on-chip and photonics.

Cheng Zhuo (Senior Member, IEEE) received the B.S. and M.S. degrees 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 Electronic Engineering, Zhejiang University. He has published over 100 technical papers. His research interests include low-power optimization, 3-D integration, hardware acceleration, power, and signal integrity.

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