Energy-Aware Designs of Ferroelectric Ternary Content Addressable Memory

Yu Qian[1], Zhenhao Fan[1], Haoran Wang[1], Chao Li[1], Mohsen Imani[2], Kai Ni[3], Grace Li Zhang[4], Bing Li[4], Ulf Schlichtmann[4], Cheng Zhuo[1*], and Xunzhao Yin[1*]
[1]Zhejiang University, Hangzhou, China, [2]University of California, Irvine, USA
[3]Rochester Institute of Technology, USA, [4]Technical University of Munich, Germany
* Corresponding authors (e-mail: xyin1@zju.edu.cn, czhuo@zju.edu.cn)

Abstract*—Ternary content addressable memories (TCAMs) are a special form of computing-in-memory (CiM) circuits that aim to address the so-called memory wall issues by merging the parallel search function with memory blocks. Due to the content addressing nature, TCAMs have been widely utilized for search intensive tasks in low-power, data analytic applications, such as IP routers, associative memories, and learning models. While most state-of-the-art TCAM designs focus on improving the TCAM density by harnessing compact nonvolatile memories (NVMs), little efforts have been spent on reducing and optimizing the energy consumption of the NVM based TCAM. In this paper, by exploiting the Ferroelectric FET (FeFET) as a representative NVM, we propose two compact and energy-aware designs of ferroelectric TCAMs for low power applications. We first introduce a novel 2FeFET based XOR-like gate structure that can also be adopted to other NVMs, and then leverage the structure to propose two TCAM designs that achieve high energy efficiency by either reducing the associated precharge overhead (2FeFET-1T cell), or eliminating the precharge phase typically required by TCAMs (2FeFET-2T cell). We evaluate and compare the designs w.r.t area, search energy and delay at array level with other existing designs, and benchmark the proposed TCAM designs in an associative memory based GPU architecture. The results suggest that the proposed 2FeFET-1T/2FeFET-2T TCAM design consumes 3.03X/8.08X less search energy than the conventional 16T CMOS TCAM, while the proposed design cell area is only 32.1%/39.3% of the latter. Compared with the state-of-the-art 2FeFET only TCAM array, our proposed designs still achieve 1.79X and 4.79X search energy reduction, respectively. Moreover, our proposed designs can achieve, on average, 45.2%/51.5% energy saving compared with the conventional GPU based architecture at the application level.

I. Introduction

In the era of Big Data, a variety of data-intensive applications call for efficient and parallel data analytic operations to replace the sequential, time, and energy consuming operations in conventional digital machines, specifically the search function [1]. Ternary content addressable memory (TCAM), which supports parallel searches over the stored memory array given an input vector, is a potential solution to address the processor-memory bottleneck challenges [2]. Due to the content addressing and their fully parallel property, TCAMs have been applied in many areas, e.g., neuromorphic computing, IP routers and in-memory data processing, etc. [3]–[9].

While the conventional TCAM design based on standard CMOS technology has been proposed [3], it suffers from large area overhead and leakage as CMOS technology scales down to the physical limit. Emerging non-volatile memories (NVMs) such as the two-terminal resistive RAM (ReRAM) [10] and the three-terminal Ferroelectric FET (FeFET) [11] can encode logic values '0'/'1' using their high/low resistance state, thus are exploited to build more compact TCAM designs [12]–[14]. Highly promising as NVM based TCAM designs are, these designs focus more on reducing the TCAM cell size with small NVMs. The potential of combining both the NVMs and the energy-aware design schemes for TCAMs to improve energy efficiency has yet to be explored.

In this paper, we exploit FeFET as a representative NVM, and propose a novel 2FeFET based structure, which performs an in-memory XOR-like function and can be generally adapted to other NVMs. Based on the above structure, we propose two compact and energy efficient TCAM designs, shedding light on two potential design schemes to improve the energy efficiency and performance of TCAMs. Specifically, the first 2FeFET-1T TCAM design improves the energy efficiency and performance by reducing the effective precharge/discharge capacitance associated with the TCAM matchline, while the second 2FeFET-2T TCAM design further reduces the search energy by fully eliminating the precharge phase prior to every search operation. The structures, operations, and energy/performance analysis of the proposed TCAMs are discussed. The area, search energy, and delay of the proposed designs are 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 two design schemes. We also examine the energy efficiency of the proposed FeFET TCAM designs in the context of an associative memory application that leverages TCAM arrays as fast search elements integrated within a conventional GPU architecture. Our evaluation results show that the proposed 2FeFET-1T/2FeFET-2T TCAM consumes only 32.1%/39.3% of the cell size of a CMOS TCAM design, and offers 3.03X/8.08X and 1.79X/4.79X better energy efficiency than the CMOS TCAM and the state-of-the-art 2FeFET TCAM design, respectively. At the application level, the proposed 2FeFET-1T and 2FeFET-2T TCAM designs can offer 41.6% and 47.9% (45.2% and 51.5%) more energy saving than the CMOS TCAM design (exact GPU), respectively.

Sec. II introduces the FeFET basics and existing TCAM works. Sec. III introduces the proposed FeFET TCAM designs, along with the energy and performance discussions. Evaluation results are presented in Sec. IV. Sec. V concludes.

II. Background

In this section, we first review the FeFET device and model [15], and then describe the CAM basics as well as existing TCAMs based on CMOS, ReRAM and FeFET, respectively.

A. The 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 1nm thickness, greatly outperforming its perovskite ferroelectric counterparts [16]. Ferroelectric memory exhibits

978-3-9819263-5-4/DATE21/©2021 EDAA

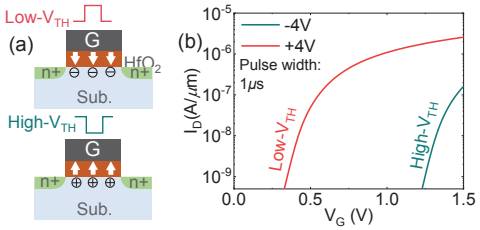


Fig. 1. (a) FeFET polarization directions and channel conditions after memory write operations; (b) The FeFET I_D-V_G characteristics after positive/negative gate write voltages.

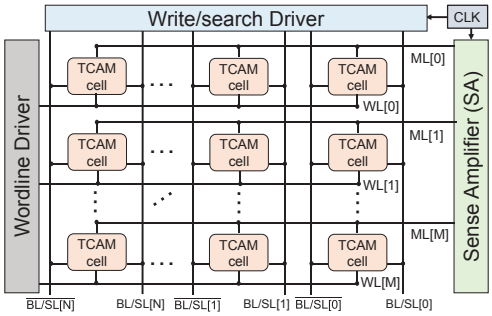


Fig. 2. Schematic of an M×N TCAM array.

superior write energy efficiency because the polarization switching process is driven by electric field, while the write process in other types of NVMs requires a high conduction current [17], thus consuming a high write power. When integrating a ferroelectric film as the gate insulator in a MOSFET, a FeFET is obtained. By applying a positive/negative gate pulse, the ferroelectric polarization will switch to the direction of 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) [18], a Preisach model [15], and a comprehensive Monte Carlo model [19]. In this work, we are utilizing the experimentally calibrated Preisach compact model for FeFET due to its computational efficiency and accuracy [15]. 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 [15]. The final ferroelectric model can be accomplished by including the ferroelectric history tracking algorithm and non-saturated hysteresis loops, explained in [15]. Integrating such a ferroelectric model with a MOSFET model, such as the standard BSIM, a FeFET model is obtained. Such a model has been calibrated with experimental results [15]. Fig.1(b) shows the I_D-V_G characteristic after memory write with ±4V gate voltages. A memory window of approximately 1V 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 driving the search lines with input data. MLs on which all cells match with the input will remain high, indicating a match. MLs that have at least one cell mismatching with the input, will discharge. The SAs sense the ML states of their associated words.

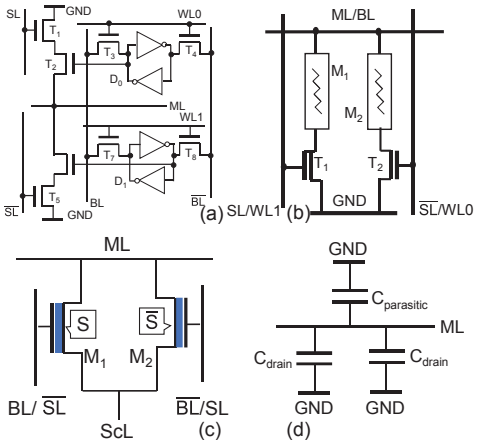


Fig. 3. TCAM designs: **(a)** 16T CMOS; **(b)** 2T-2R ReRAM; **(c)** 2FeFET; **(d)** Equivalent schematic of (a, b, c) TCAM cells.

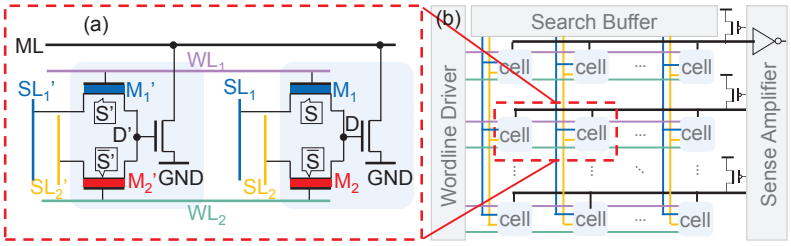
Numerous NVM based TCAM designs adopt the parallel search function with smaller cell size than the conventional 16T CMOS TCAM (Fig. 3(a)), due to the NVM properties. Fig. 3 shows the most commonly used TCAM designs. The 16T CMOS TCAM [3] is volatile, and stores binary states with 2 SRAMs, while the 2T-2R ReRAM based [12] and 2FeFET based [13], [17] TCAMs encode binary states into the NVMs with much less transistors, thus smaller cell sizes. It can be seen from Fig. 2 and Fig. 3 that the TCAM designs aforementioned are precharge-based, requiring a precharge phase prior to every search. The search energy of TCAM arrays consist of the precharge energy and the leakage energy:

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

where E_{pre} represents the energy precharged to the matchline 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 matchlines. The equivalent schematic of a TCAM cell in a NOR-type matchline connection is depicted in Fig. 3(d), formulating C_{ML} as below:

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

where C_{PMOS}, C_{cell}, $C_{parasitic}$ and C_{drain} are the drain capacitance of the PMOS transistor that precharges the matchline, total capacitance of a TCAM cell associated with the matchline, the parasitic capacitance of each cell, and the drain capacitance of a transistor, respectively. Each TCAM cell connects the matchline with two transistors, thus C_{cell} consists of two drain capacitances, which can be extracted from PTM technology [20], and $C_{parasitic}$ can be extracted from DESTINY [21]. 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 matchlines given an array size. As such, one of our proposed designs introduces a general



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

Mode	WL_1	WL_2	SL_1	SL_2	S	$\overline{S}$	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)** The write and search operations.

design scheme to improve the energy efficiency by reducing the number of transistors associated with the matchline.

Another design scheme focuses on eliminating the energy consuming precharge phase. A CMOS based precharge-free TCAM design was proposed to eliminate the precharge phase [22]. However, the TCAM consumes 14 transistors, leading to large area overhead. Though the TCAM performs two searches within one clock cycle, the relatively large search delay prevents the design from practical usage in fast and small devices. To overcome the above challenges, we propose a compact TCAM design that leverages FeFETs and a precharge-free design scheme to achieve better area and energy efficiency.

III. Energy-Aware TCAM Designs

Here, we propose two energy efficient FeFET based TCAM designs by either (i) reducing the associated matchline capacitance, or (ii) eliminating the precharge phase. We first describe the general structures and operations of the proposed designs, and then conduct the array analysis to explain how the two proposed designs can enable energy saving and performance improvement compared with previous TCAMs.

A. 2FeFET-1T TCAM leveraging ML load reduction

Fig. 4(a) shows the design of our proposed TCAM, consisting of a 2FeFET structure and a NMOS transistor. The drains of the FeFETs connect to the gate of the NMOS as node D, while the sources of the FeFETs connect to the searchlines SL_1 and SL_2, respectively. The gates of the FeFETs are controlled by two wordlines WL_1 and WL_2, and the matchline ML senses the drain of the NMOS. When placed in an array, the searchlines are shared by the cells vertically, and the wordlines connect the cells horizontally.

The 2FeFET structure in the proposed cell forms the core for constructing efficient TCAM designs. When stored values S=1 (M_1 in low V_{TH} state) and $\overline{S}$=0 (M_2 in high V_{TH} state), then D=SL_1. When S=0 and $\overline{S}$=1, D=SL_2. Such transmission gate feature enables the structure as an XOR-like gate:

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

This structure can be adopted to other NVM devices by overcome their limited ON/OFF resistance ratio, and combined with TCAM design schemes discussed in this paper, as long as Eq. 3 can be readily realized.

Fig. 4(c) summarizes the write and search operations of the proposed TCAM cell. According to the write scheme discussed in Sec. II-A, input data is written into the FeFETs as S and $\overline{S}$. To write a logic '0', -V_w and V_w are applied to WL_1 and WL_2, respectively, and 0 is applied to the searchlines. The gate-source voltages of M_1 and M_2 are -4V and 4V, switching the polarization of FeFETs and setting S and $\overline{S}$ to '0' and '1', respectively. Similarly, logic '1' can be written into the cell by applying V_w and -V_w to WL_1 and WL_2, respectively. To write a 'don't care' state 'X', -V_w is applied to both wordlines. In the proposed TCAM array in Fig. 4(b), the wordlines associated with unselected words are set to either $V_w/2$ or -$V_w/2$ to inhibit write disturbance [23]. With such configurations for write and search operation, the node D voltage will be low enough for the match case, cutting off the NMOS transistor while turning ON the NMOS transistor for the mismatch case.

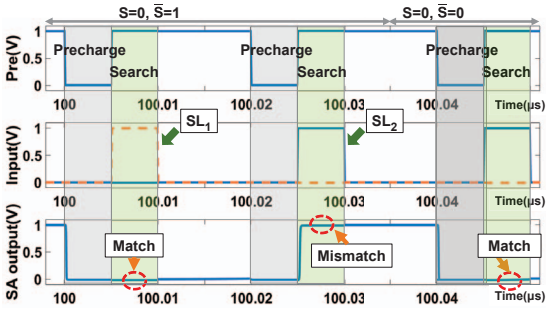


Fig. 5. Transient waveforms of a 2FeFET-1T TCAM.

During the search operation, the matchlines of the array are precharged to a high level. The wordlines are activated, and the search voltage V_s (1V) is applied to the searchlines according to the input data as summarized in Fig. 4(c). The matchline state ML is controlled by the output of the 2FeFET structure core, thus can be formulated as below:

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

The truth table of the search operation in Fig. 4(c) is consistent with Eq. 4, validating the functionality of our proposed TCAM design. Fig. 5 demonstrates the transient waveforms of the proposed 2FeFET-1T TCAM cell.

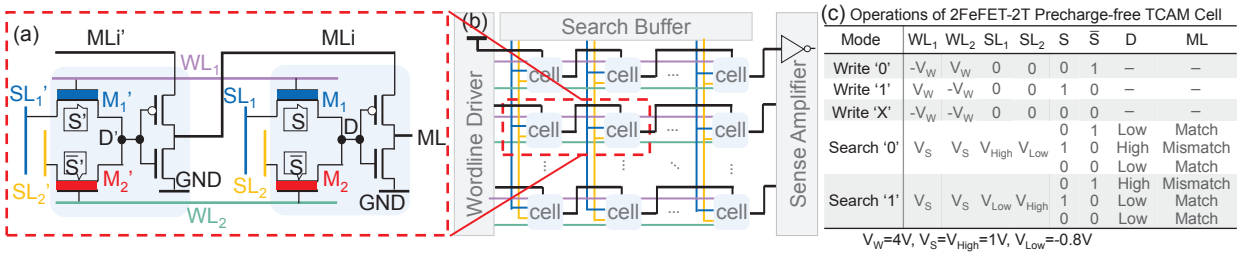
Compared with existing precharge-based TCAM designs, our proposed 2FeFET-1T TCAM cell leverages ML capacitance reduction scheme by associating only one transistor of each cell to the matchline, resulting in less precharge energy:

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

The search delay of a precharge-based TCAM array is determined by the effective RC constant as below:

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

where R_{eff} represents the effective resistance between the matchline and ground upon a mismatch search. Since the ML load reduction scheme associates one transistor of each cell to the matchline, it can be expected from Eq. (2, 5 and 6) that the proposed 2FeFET-1T TCAM array has much less search delay than the existing TCAM designs.



(c) Operations of 2FeFET-2T Precharge-free TCAM Cell

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_{High}	V_{Low}	0	1	Low	Match
					1	0	High	Mismatch
					0	0	Low	Match
Search '1'	V_S	V_S	V_{Low}	V_{High}	0	1	High	Mismatch
					1	0	Low	Match
					0	0	Low	Match

V_W=4V, V_S=V_{High}=1V, V_{Low}=-0.8V

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

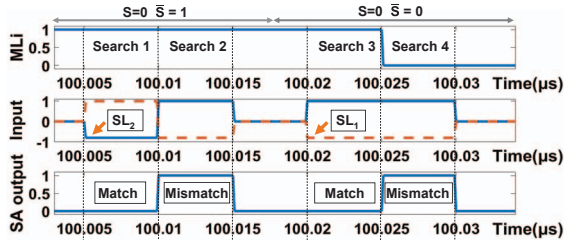


Fig. 7. Transient waveforms of a 2FeFET-2T precharge-free TCAM.

B. 2FeFET-2T TCAM leveraging precharge-free scheme

Compared with the ML load reduction scheme for precharge-based TCAMs, a precharge-free scheme that eliminates the energy consuming precharge phase of the TCAMs can further improve the energy efficiency. We propose a 2FeFET-2T precharge-free TCAM design that combines both the advantages of NVMs and the precharge-free scheme.

Fig. 6 shows the structure of our proposed 2FeFET-2T TCAM cell, along with its array schematic. The design consists of the 2FeFET structure and an inverter supplied by the matchline of the previous cell MLi. When organized in an 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 write and search operations. The write scheme of the 2FeFET-2T TCAM design is the same as that of the proposed 2FeFET-1T TRCAM design. The $V_{w/2}$ scheme is applied to inhibit write disturbance. Fig. 6(c) summarizes the search operation of the TCAM cell. The truth table is valid only when the previous matchline voltage level MLi is high. The ML state is thus formulated as below:

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

Note that in a case where 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 given the internal node D is at 0. This is because a PMOS pass transistor can only pass a weak ground. When MLi falls below the PMOS threshold voltage $V_{TH,P}$, the absolute gate-source voltage of PMOS falls below the threshold, thus turning off the device, and leaving ML at around $V_{TH,P}$. Such incomplete ML swing can be continuously degraded along with the array word, resulting in a function failure. Therefore, in order to achieve full ML swing, V_{Low}=-0.8V is applied to the 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.

Unlike the default precharge phase in the precharge-based TCAM designs, the proposed 2FeFET-2T TCAM design does not require the precharge phase before every search, as the

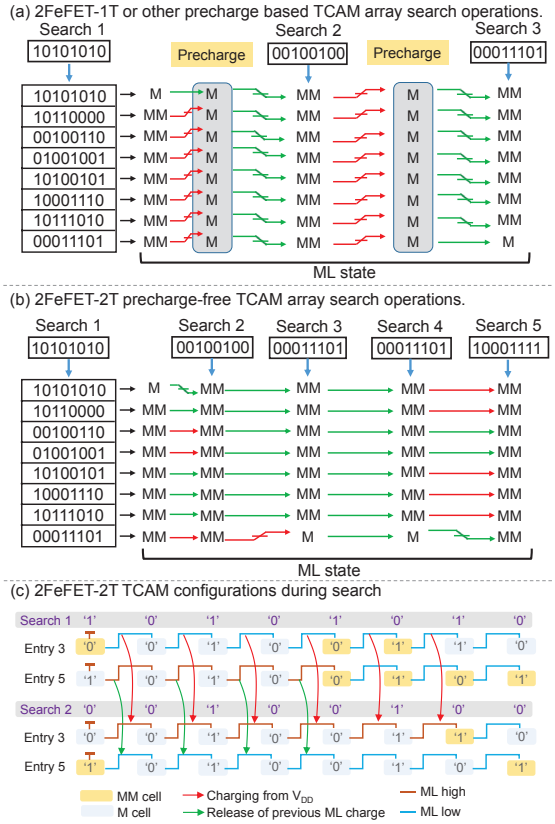


Fig. 8. Functionality of TCAM arrays with an example illustration. M represents a match state of an entry, and MM represents a mismatch state of an entry. The red arrows indicate a matchline charging from the voltage supply to the entry, while the green arrows mean no matchline charging from the supply to the entry, but rather possible matchline discharge.

matchline state of each TCAM cell is determined by both the cell evaluation output D and the matchline state of the previous cell MLi. The matchline state transition of a cell only depends on the state of previous evaluation: A mismatch in the first search discharges the MLi to ground, and a match in the second search charges the MLi again. If consecutive searches result in the same match/mismatch state to the cell, then MLi remains without transitions. Here we discuss the only situation where the matchlines along with a TCAM word will be charged. One cell C_i of an array word will be charged upon a search only when the following conditions are satisfied: (i) 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; (ii) the matchlines of all previous $i-1$ cells are all at match state upon the search, thus forming a charging path from voltage supply to the MLi of C_i.

Fig. 8 shows an example illustration 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 across the TCAM array, and we assume that a TCAM array will have at most one entry matching with the search input data. It can be seen from Fig. 8(a) that regardless of the ML state of an entry in the previous search, precharge phase is inevitable in the precharge-based TCAM designs. However, as shown in Fig. 8(b), the matchline charging situation of the 2FeFET-2T precharge-free TCAM array is different. Take the entry 3 and 5 storing '00100110' and '10100101' as an example, Fig. 8(c) shows the detailed matchline state transitions and charging for search 1 and search 2. Search 1 results in a mismatch for both entries. Since the first cell of entry 3 is at mismatch state, all cell states within entry 3 are mismatch. Since the first 4 cells within entry 5 are at match state, their associated matchlines are at high level regardless of the mismatch state of the entry. Upon Search 2, the first cell of entry 3 transitions from mismatch to match, resulting in a match state for the first consecutive 6 cells. 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, thus resulting in the discharge of the matchlines associated with the first 4 cells within the entry. Similar analysis can be conducted to other entries through all searches, and the entry state transitions as well as the matchline charging situation are visualized as shown in Fig. 8(b). It can be seen that much less matchline charging is needed by the 2FeFET-2T precharge-free TCAM array for four consecutive searches (from Search 2 to Search 5), thus causing significantly less energy consumption per search compared with 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 increase.

Since the proposed 2FeFET-2T TCAM array adopts the NAND-type matchline connection for precharge-free design scheme, the delay in a worst case is expected to be large. Assuming that initially the cells within a word are all at the same match/mismatch condition, the worst case of a search is when only the first cell has a matchline state transition. The SA that senses the last cell then needs to wait until the matchline state transition at the first cell propagates through the entire word, causing a large search delay. We will quantitatively compare the energy and performance of the proposed TCAMs against existing approaches in the next section.

IV. Evaluation

In this section, we compare the search energy and delay of the two proposed TCAMs with 16T CMOS, 2T-2R ReRAM, 2FeFET and 14T CMOS [22] based TCAMs to prove the benefits of the proposed two design schemes with FeFETs.

A. TCAM array evaluation

We conduct the evaluation simulations of our proposed FeFET based TCAM arrays (Fig. 4(b) and Fig. 6(b)) through SPECTRE, and compare the results with aforementioned TCAM designs. The 45nm PTM model is adopted for all MOSFET devices with minimized sizes. For the 2T-2R ReRAM based TCAM, we assume a LRS/HRS of 20kΩ/2MΩ [24]. Wiring parasitics are extracted from DESTINY [21].

Table I summarizes TCAM metrics including the transistor count per cell, estimated cell size, search style, search delay and search energy per bit per search of different TCAM arrays. The 2X2 layouts of the proposed TCAM arrays have been sketched, and the cell sizes have been estimated based on the layout[1]. For the precharge-based TCAM arrays, the matchline precharge occurs before every search, while for the precharge-free TCAM arrays, the matchline precharge only occurs when the conditions of the cell matchline state discussed in Sec. III-B are satisfied. Therefore, the search energy of the proposed 2FeFET-2T precharge-free TCAM is highly dependent on the search patterns. Here we evaluate the 2FeFET-2T precharge-free TCAM based on the randomly chosen search patterns in Fig. 8(b) as an example. For the precharge-based TCAM arrays, we measure the search delay in a worst case, where there is only one-bit mismatch. The delay of the proposed precharge-free TCAM array is measured in the worst case as discussed in Sec. III-B.

TABLE I
Metric Comparison Summary of TCAM Designs

Reference	[3]	[12]	[13]	[22]	Fig. 4	Fig. 6
Technology	CMOS	ReRAM	FeFET	CMOS	FeFET	FeFET
Transistors/ cell	16T	2T-2R	2FeFET	14T	2FeFET -1T	2FeFET -2T
Cell size (μm^2)	1.12⋆	0.41†	0.15	8.9	0.36	0.44
Search style‡	P	P	P	PF	P	PF
Search delay	582.4ps	350.6ps	340.8ps	∼20ns	252.8ps	1.43ns
Energy [fJ/bit/search]	0.59 8.08X	0.55 7.53X	0.35 4.79X	0.18 2.48X	0.195 2.67X	0.073 1X

⋆: The 16T CMOS TCAM cell size is projected based on MOSIS Scalable CMOS design rules for 45 nm, i.e., SCMOS DEEP rules [25] and the "push rule" SRAM scaling trend [26], i.e., $124F^2$ at 65 nm and $171F^2$ at 45 nm.
†: The 2T-2R ReRAM based TCAM cell is based on 90nm.
‡: P denotes precharge-based, PF denotes precharge-free.

The cell sizes of the proposed 2FeFET-1T TCAM and 2FeFET-2T TCAM are 32.1% and 39.3% of that of a conventional 16T CMOS TCAM, respectively. Less area overhead of the proposed TCAMs leads to less parasitic capacitance associated with the matchlines at array level, resulting in less precharge energy compared with the CMOS TCAM array. The search energy and delay numbers in Table I show that our proposed 2FeFET-1T TCAM is 3.02X/1.79X more energy efficient and 2.3X/1.35X faster than the 16T CMOS TCAM/2FeFET TCAM, respectively. The aforementioned results validate the efficiency of the ML load reduction scheme which associates only one transistor to the matchline, in addition to the area efficiency gained by FeFETs. Our proposed 2FeFET-2T precharge-free TCAM is 8.08X/4.79X 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.

B. Benchmarking of TCAM based associative memory

We benchmark our proposed TCAMs in a TCAM enhanced GPU architecture for applications with the temporal locality. Fig. 9 shows a GPU floating-point unit (FPU) cores enhanced with the proposed TCAMs. The TCAMs and the associated memory blocks are used as associative memories next to each FPU to store frequent arithmetic operations and corresponding results. During the execution, the GPU cores search the input data on a TCAM block, parallel with the first pipeline stage of the FPU. Upon a match, a corresponding pre-stored result will be fetched as the output, while the FPU pipeline will

[1]The 2X2 layouts of the proposed TCAM arrays have been omitted due to the page limit, the layout drawing and the area estimation follow the same principle from [13], [24].

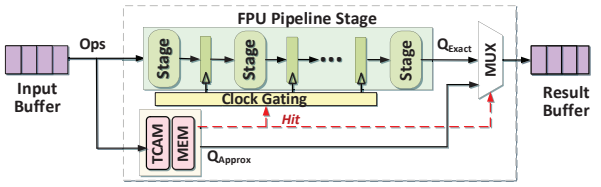


Fig. 9. GPU cores enhanced with TCAM block.

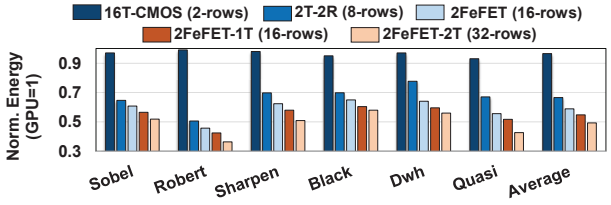


Fig. 10. Normalized energy consumption of enhanced GPU integrating different TCAM cells.

be clock-gated. With low power TCAM designs, this method reduces repeated computations by enabling the computation reuse, thus significantly reducing the energy consumption of processors in low power applications [27], [28].

This method has the following challenges: (i) we need to pay the cost of TCAM search, regardless of hit or miss in the TCAM. (ii) To provide a high TCAM hit rate, it is desirable to use a larger TCAM. However, the large power consumption of the TCAM block often limits the TCAM size to a small number of rows. We compare the proposed low power TCAM designs with the TCAMs based on CMOS, ReRAM, and 2FeFET in this GPU architecture. The TCAM based associative memories are employed in an Nvidia Kepler architecture GeForce GTX Titan. TCAM is implemented with the FPU within each of the cores in the GPUs, including adder (ADD), multiplier (MUL), and multiply accumulator (MAC). We evaluate the efficiency of our enhanced architecture over a wide range of signal processing applications and the Caltech 101 dataset [29]. Fig. 10 shows the normalized energy consumption of the enhanced GPU with the proposed TCAMs and other TCAMs[2]. For each TCAM, we report the results for the number of rows that provide maximum efficiency. TCAMs with high energy consumption, e.g., 16T-CMOS, provide maximum efficiency using a small number of rows. This is because the TCAM search energy overhead dominates the gain coming from a higher hit rate. Our evaluation shows that the proposed 2FeFET-1T/2FeFET-2T TCAM can save, on average, 45.2%/51.5% the GPU energy consumption as compared to the conventional GPU. Compared with the 2FeFET TCAM enhanced GPU approach, our proposed designs can still achieve 4.9%/10.6% more energy saving. The higher efficiency results from the lower power consumption of the proposed TCAMs, which enable a larger TCAM block with a higher hit rate, thus reducing the active time of GPU cores.

V. Conclusion

In this paper, we propose a novel FeFET based structure that exploits NVMs, and two compact TCAM designs indicating two potential design methods for energy efficiency improvements of TCAMs. We present the structures and our proposed 2FeFET-1T and 2FeFET-2T precharge-free TCAMs, and analyze their energy efficiency advantages over existing works. We then evaluate the proposed TCAM designs, and the results indicate that our proposed two design methods with FeFETs can achieve promising area, energy efficiency and performance metrics w.r.t. to existing TCAM designs. Benchmarking of a GPU based architecture further emphasized the energy efficiency of our proposed TCAM designs.

[2]14T CMOS TCAM is not included due to larger search delay than the GPU cycle.

Acknowledgment

This work was partially supported by Zhejiang Provincial NSF with Grant No. LQ21F040006 and LD21F040003, and Semiconductor Research Corporation (SRC) Task No. 2988.001.

References

[1] C. Zhuo *et al.*, "Noise-aware DVFS for efficient transitions on battery-powered iot devices," *TCAD*, vol. 39, pp. 1498–1510, 2020.
[2] D. Gao *et al.*, "Eva-cim: A system-level performance and energy evaluation framework for computing-in-memory architectures," *TCAD*, vol. 39, pp. 5011–5024, 2020.
[3] K. Pagiamtzis *et al.*, "Content-addressable memory circuits and architectures: A tutorial and survey," *JSSC*, vol. 41, pp. 712–727, 2006.
[4] T. Kohonen, *Associative memory: A system-theoretical approach*. Springer Science & Business Media, 2012, vol. 17.
[5] Y.-J. Chang, "A high-performance and energy-efficient tcam design for ip-address lookup," *IEEE TCAS-II*, vol. 56, pp. 479–483, 2009.
[6] X. Yin *et al.*, "Fecam: A universal compact digital and analog content addressable memory using ferroelectric," *IEEE TED*, vol. 67, pp. 2785–2792, 2020.
[7] S. Zhang *et al.*, "Aging-aware lifetime enhancement for memristor-based neuromorphic computing," in *IEEE DATE*, 2019, pp. 1751–1756.
[8] A. F. Laguna *et al.*, "Seed-and-vote based in-memory accelerator for dna read mapping," in *IEEE ICCAD*, 2020, pp. 1–9.
[9] Y. Zhu *et al.*, "Statistical training for neuromorphic computing using memristor-based crossbars considering process variations and noise," in *IEEE DATE*, 2020, pp. 1590–1593.
[10] H.-S. P. Wong *et al.*, "Metal–oxide rram," *Proceedings of the IEEE*, vol. 100, pp. 1951–1970, 2012.
[11] K. Ni *et al.*, "Critical role of interlayer in hf 0.5 zr 0.5 o 2 ferroelectric fet nonvolatile memory performance," *IEEE TED*, vol. 65, pp. 2461–2469, 2018.
[12] J. Li *et al.*, "1 mb 0.41 μm^2 2t-2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing," *JSSC*, vol. 49, pp. 896–907, 2014.
[13] X. Yin *et al.*, "An ultra-dense 2fefet tcam design based on a multi-domain fefet model," *IEEE TCAS-II*, vol. 66, pp. 1577–1581, 2018.
[14] C. Li *et al.*, "A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing," in *IEEE IEDM*, 2020, pp. 1–4.
[15] K. Ni *et al.*, "A circuit compatible accurate compact model for ferroelectric-fets," in *IEEE VLSI*, 2018, pp. 131–132.
[16] T. Böscke *et al.*, "Ferroelectricity in hafnium oxide: Cmos compatible ferroelectric field effect transistors," in *IEEE IEDM*, 2011, pp. 24–5.
[17] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, pp. 521–529, 2019.
[18] A. Aziz *et al.*, "Physics-based circuit-compatible spice model for ferroelectric transistors," *EDL*, vol. 37, pp. 805–808, 2016.
[19] S. Deng *et al.*, "A comprehensive model for ferroelectric fet capturing the key behaviors: Scalability, variation, stochasticity, and accumulation," in *IEEE VLSI*, 2020, pp. T1–T2.
[20] R. Vattikonda *et al.*, "Modeling and minimization of pmos nbti effect for robust nanometer design," in *IEEE DAC*, 2006, pp. 1047–1052.
[21] M. Poremba *et al.*, "Destiny: A tool for modeling emerging 3d nvm and edram caches," in *DATE*. EDA Consortium, 2015, pp. 1543–1546.
[22] T. V. Mahendra *et al.*, "Energy-efficient precharge-free ternary content addressable memory (tcam) for high search rate applications," *TCAS-I*, 2020.
[23] S. Mueller *et al.*, "From mfm capacitors toward ferroelectric transistors: Endurance and disturb characteristics of hfo2-based fefet devices," *TED*, vol. 60, pp. 4199–4205, 2013.
[24] X. Yin *et al.*, "Design and benchmarking of ferroelectric fet based tcam," in *DATE*. EDAA, 2017, pp. 1448–1453.
[25] D. Rules, "Mosis scalable cmos (scmos)."
[26] S. G. Narendra *et al.*, "Through the looking glass? the 2015 edition: Trends in solid-state circuits from isscc," *SSC Magazine*, vol. 7, pp. 14–24, 2015.
[27] M. Imani *et al.*, "Resistive configurable associative memory for approximate computing," in *DATE*. IEEE, 2016, pp. 1327–1332.
[28] ——, "Exploring hyperdimensional associative memory," in *HPCA*. IEEE, 2017, pp. 445–456.
[29] http://www.vision.caltech.edu/Image_Datasets/Caltech1.