Compact and Efficient CAM Architecture through Combinatorial Encoding and Self-Terminating Searching for In-Memory-Searching Accelerator

Weikai Xu[1], Jin Luo[1], Qianqian Huang[1,2,3*], and Ru Huang[1,2,3*]
[1]School of Integrated Circuits, Peking University, Beijing 100871, China; [2]Beijing Advanced Innovation Center for Integrated Circuits, Beijing 100871, China; [3]Chinese Institute for Brain Research, Beijing 102206, China.
*Corresponding author: ruhuang@pku.edu.cn; hqq@pku.edu.cn

***Abstract*—Content addressable memory (CAM) has triggered a lot of attention for data-intensive applications due to highly parallel pattern searching capability. Most state-of-the-art works focus on reducing hardware cost of CAM by exploiting various emerging non-volatile memory (NVM) technologies. However, the existing CAM designs still mainly follow the conventional encoding scheme which requires two complementary storage nodes and search signals for each bit of entry and query respectively, along with separate precharging and evaluation phases for bit-vector searching, limiting the further improvement of area- and energy-efficiency. In this work, a compact and efficient CAM architecture is proposed through two techniques: (1) a combinatorial encoding scheme for CAM by encoding entry/query states with permutations and combinations of multiple storage nodes as a group, which can significantly improve the encoding efficiency and thus greatly reduce the hardware implementation cost of CAM compared with conventional encoding scheme; (2) an one-step self-terminating searching scheme for CAM by detecting matching condition during precharging phase and terminating precharging once a match is detected, which can further reduce the search delay and energy. The experiments and evaluations of the proposed CAM architecture with co-optimization of combinatorial encoding and self-terminating searching are carried out based on ferroelectric FET (FeFET), which can reduce the area-energy-delay product (AEDP) by 1182× over the conventional CMOS-based CAM in data searching tasks, showing its great potential for area- and energy-efficient in-memory-searching accelerator.**

I. Introduction

Content addressable memory (CAM), which can perform massively parallel in-memory-searching (IMS) operations by comparing each input query with all entries stored in the CAM array, has been widely used for large-scale search-based applications, such as network routing, text mining, machine learning, natural language processing and bioinformatics, et al [1-7] (Fig. 1a).

Generally, the features in datasets of large-scale search problems are represented by multi-dimensional vectors, and each dimension has various number of states [4]. The numerous features are encoded as Booleans strings stored in the CAM array, enabling parallel IMS operations in $O(1)$ time [7] (Fig. 1b). The bitwise XNOR-logic comparison between query and entry in CAM is carried out, and the whole feature is encoded by bitwise expansion of Booleans "0/1" [6-14]. Generally, based on the conventional bitwise complementary encoding scheme, each single bit of stored entry and input query are typically encoded with complementary form of two storage nodes and two search signals, respectively (Fig. 1c). It suffers from the following challenges: (1) the CAM requires twice the number of storage nodes to encode the same amount of states compared to the typical memory, leading to large area cost due to the inadequate utilization of various combinations of storage nodes with the inefficient encoding scheme; (2) large energy consumption of search lines (SLs) induced by the increased number of SLs; (3) large energy consumption and search delay of match lines (MLs) induced by the increased capacitance of MLs.

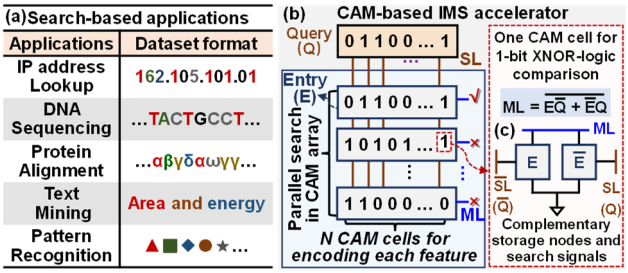

Fig. 1 (a) Various search-based applications which can be accelerated by content addressable memory (CAM). (b) The CAM-based in-memory-searching (IMS) accelerator, where the features in dataset are encoded and stored in CAM array for parallel IMS operations. (c) The simplified circuit model of CAM cell for 1-bit XNOR-logic comparison operation.

On the other hand, to implement the parallel comparison of bit-vectors, existing CAMs generally employ conventional precharge-high ML searching scheme, consisting of two phases [6-14]. In the first phase, all SLs are reset to ground for precharging MLs to high voltage. Then, in the second phase, SLs are precharged according to the query for evaluation of MLs, which will maintain high voltage or be discharged to low voltage in match or mismatch conditions. However, the conventional searching scheme with separate ML precharging and evaluation phases suffers from several challenges: (1) large energy consumption of SLs and MLs induced by the frequent charging/discharging processes with full-swing voltage for continuous searches; (2) large search delay in ML evaluation phase limited by the 1-bit mismatch case with the slowest discharging speed.

In recent years, most state-of-the-art CAM designs focus on reducing the hardware cost of CAM by introducing emerging non-volatile memory (NVM) technologies [7-13], such as resistive random-access memory (RRAM), phase change memory (PCM) and ferroelectric FET (FeFET) et al., while they still follow the conventional encoding and searching schemes. Very recently, there is study that implements CAM cell with single FeFET, while it requires two search operations and suffers from the significant decrease in the sense margin [15]. Since it involves time-multiplexing the storage element, the current 1FeFET-based CAM architecture will have double search delay penalty.

In this work, a **C**ombinatorial **E**ncoding and self-**T**erminating **S**earching based **CAM** architecture (CETS-CAM) is proposed, which can address the challenges mentioned above. By utilizing the proposed combinatorial encoding scheme for entry storage, the encoding efficiency is largely improved which saves the required storage nodes of CAM, resulting in lower area cost and energy consumption.

DAC '24, June 23–27, 2024, San Francisco, CA, USA

ACM 979-8-4007-0601-1/24/06
https://doi.org/10.1145/3649329.3655666

Moreover, we merge the precharging and evaluation phases into one-step self-terminating ML charging process, further reducing the search energy and delay of CAM. The main contributions of this work are summarized as follows.

- A combinatorial encoding scheme for CAM is proposed, which encodes entry/query states with permutations and combinations of multiple storage nodes as a group. It can improve the encoding efficiency from 50% of conventional encoding scheme to the approaching 100% which is the theoretical maximum value. Therefore, the proposed encoding scheme can save the area overhead of storage nodes, SLs and MLs, and thus reduce the search energy and delay of CAM.
- A self-terminating searching scheme for CAM is proposed, which detects matching condition with one-step precharging phase and self-terminates ML precharging once a match is detected. It further reduces the search delay and energy by eliminating the evaluation delay and reducing charging/discharging frequency of SLs along with the voltage swing of MLs.
- Experiments and evaluations of the proposed CETS-CAM are carried out based on emerging FeFET device with low write energy. The results show that the proposed CETS-CAM architecture can achieve 11×/3.4× improvement in speed/energy-efficiency and reduce the area-energy-delay product (AEDP) by 61.9× compared with conventional CAM architecture. Moreover, the 1FeFET-based CETS-CAM can further reduce the AEDP by 1182× compared with conventional CMOS-based CAM, showing its great potential for area- and energy-efficient in-memory-searching accelerator.

II. Background

In this section, we review the existing CAM designs based on different memory technologies and introduce the conventional encoding and searching schemes.

A. Existing CAM Designs

Fig. 2a shows the general architecture of CAM array. Each row of CAM cells of one entry shares one ML for row-wise match-detection. Each column of CAM cells shares the complementary SLs/$\overline{\text{SLs}}$ for query input, enabling comparing a given query with all entries stored in CAM array with high parallelism. The search results of match/mismatch are generally detected by the sense amplifier of each ML [6].

Fig. 2b shows the conventional CMOS-based CAM cell for basic binary search, which consists of 6 transistors (SRAM) for entry bit storage and extra 4 transistors for bitwise comparison [6], suffering from high hardware cost and energy consumption. Recently, in order to improve the storage density and capacity of CAM, various emerging NVMs with high density have been applied for CAM design with reduced hardware overhead, such as 2T2R-based [8-9] and 2FeFET-based [12] CAM cells (Fig. 2cd). Among the emerging NVMs, HfO_2-based three-terminal FeFET has advantages of CMOS compatibility, high I_{ON}/I_{OFF} ratio and low write energy, which is considered as a promising device candidate for CAM design [12-13].

B. Conventional Bitwise Complementary Encoding Scheme

In CAM architecture, every input search query is compared with massive entries parallelly, where the entries of practical application need to be encoded and stored in memory devices, and the comparison operation is carried out within the CAM array. In conventional encoding scheme, the entry and

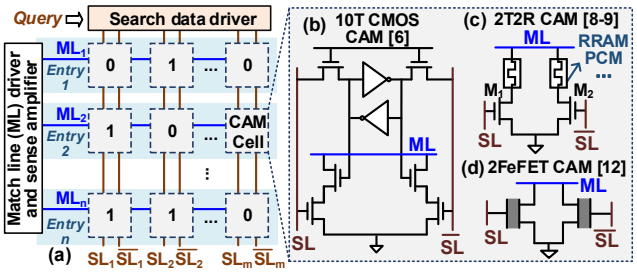

Fig. 2 (a) The architecture of a general n×m CAM array. (b) 10T CMOS CAM cell [6]. (c) 2T2R CAM cell [8-9]. (d) 2FeFET CAM cell [12].

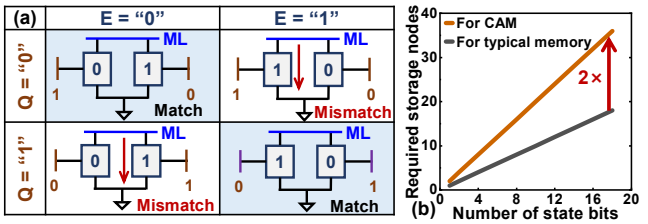

Fig. 3 (a) The truth table of CAM cell based on conventional encoding scheme. (b) The required number of storage nodes of CAM is twice as large as that in typical memory for encoding the same number of state bits.

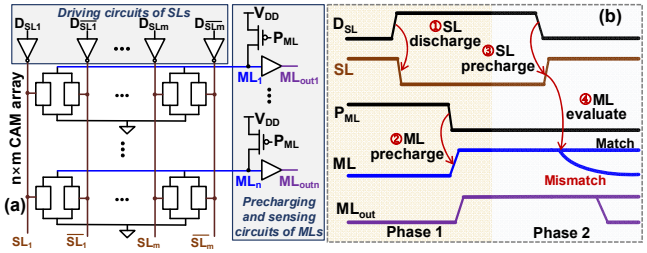

Fig. 4 (a) The schematic of CAM array and (b) timing diagram of signal transition of conventional precharge-high searching scheme with separate precharging and evaluation phases of match line (ML).

query vectors are bitwise encoded, and each bit of entry and query is represented by complementary storage nodes (E/$\overline{\text{E}}$) and search signals (Q/$\overline{\text{Q}}$) for 1-bit ("0"/"1" state) XNOR-logic comparison operation (Fig. 3a). For CMOS-based CAM, the complementary E and $\overline{\text{E}}$ are stored in the inherent two complementary storage nodes of SRAM cell [6]. Following the conventional encoding scheme, various NVM-based CAMs use two independent storage elements with complementary states for E and $\overline{\text{E}}$, such as R_{high}/R_{low} of resistive memories and high/low threshold voltage states (V_H/V_L) of FeFET devices [9, 12].

As shown in the truth table (Fig. 3a), the comparison results are reflected in the conduction status of the complementary discharge path. Only when the query matches the entry, can both discharge paths of ML be cut off, resulting in an extremely low ML current (I_{ML}). Otherwise, the I_{ML} is large due to one discharge path being conductive, indicating the mismatch condition. Based on the 1-bit comparison of CAM cell, the multi-dimensional feature can be encoded with bitwise expansion [6]. Therefore, to encode *k*-bit states, 2*k* storage nodes are needed which is twice as large as that in typical memory (Fig. 3b).

C. Conventional Precharge-High ML Searching Scheme

To implement the search operation of bit-vectors, the conventional CAM architecture generally employs precharge-high ML searching scheme, which consists of separate ML precharging and evaluation phases in one search operation. Fig. 4a shows the schematic of CAM array with driving circuits of SLs and precharging and sensing circuits of MLs based on the conventional searching scheme.

As shown in the Fig. 4b, in phase one, SLs are discharged

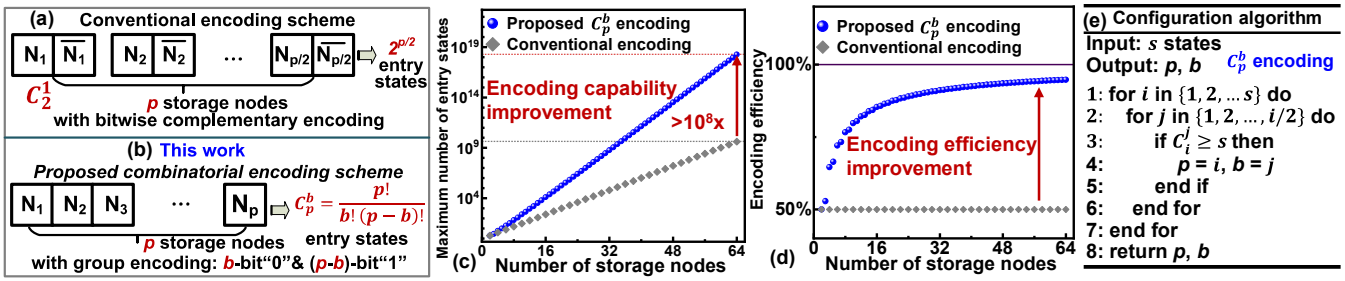

Fig. 5 (a) Conventional bitwise encoding scheme. (b) Proposed combinatorial encoding scheme. Comparison of (c) encoding capability and (d) encoding efficiency between conventional and proposed encoding scheme. (e) Configuration algorithm of the proposed encoding scheme for encoding s states.

to ground by SL driving circuits, and then MLs are precharged to high voltage (i.e., V_{DD}) by ML precharging circuits. In phase two, SLs are precharged to high voltage or ground according to the input search query for evaluation of MLs. Only when all CAM cells of one ML are matched with the input, can the ML remain high voltage, indicating a match condition. Otherwise, the ML will discharge due to the significant discharging current of mismatched cell, which will be amplified to full swing by sensing circuits for detectin. Generally, only one entry will match the query per search in the practical applications, and most mismatched MLs will discharge to low voltage and need to be precharged for next search operation [6].

III. Proposed CAM Architecture with Combinatorial Encoding and Self-Terminating Searching

A. Combinatorial Encoding Scheme for CAM

1) ***Challenges of Conventional Encoding Scheme***: In conventional bitwise complementary encoding scheme, two independent storage devices that could originally encode 4 states ("00, 01, 10, 11"), can only be used to encode 2 states ("01, 10") for representing 1-bit entry. Therefore, p storage nodes can only represent $2^{p/2}$ states of entry (Fig. 5a), suffering from the following challenges.

First, it reduces the encoding efficiency of CAM which is defined as the number of entry bits that a certain amount of storage nodes can encode, resulting in the waste of storage elements. Second, it leads to an increased number of SLs, which will enlarge the dynamic energy consumption of SLs. Third, it leads to the increased capacitance of MLs due to the increased storage devices and the wire length in each ML, resulting in elevated dynamic energy consumption of MLs and search delay in ML evaluation phase. Therefore, the conventional bitwise complementary encoding scheme limits the further improvement of CAM density, energy-efficiency and search speed.

2) ***Proposed Combinatorial Encoding Scheme***: In order to maximize the state utilization of independent storage nodes, and thus optimize the AEDP of CAM, a combinatorial encoding scheme for CAM is proposed. As shown in Fig. 5b, by encoding entry/query states with permutations and combinations of multiple storage nodes as a group, multi-bit XNOR-logic comparison can be implemented. With b nodes/SLs of a group of p nodes/SLs set to "0"/"1", and the others set to "1"/"0", there are a total of C_p^b different combinations, which can represent C_p^b states of entry/query, called as C_p^b encoding in this work. Actually, the conventional encoding scheme can be considered as a specific case of the proposed C_p^b encoding, i.e., C_2^1 (Fig. 5a).

With p storage nodes, the maximum number of encodable entry states based on proposed C_p^b encoding can be expressed as follows:

$$\max(C_p^b) = \begin{cases} C_p^{p/2} = \dfrac{p!}{(p/2)!\,(p/2)!} & when\ p\ is\ even \\ C_p^{(p-1)/2} = \dfrac{p!}{((p-1)/2)!\,((p+1)/2)!} & when\ p\ is\ odd \end{cases} \quad (1)$$

While the number based on conventional encoding scheme is $2^{p/2}$ and p must be even. As shown in Fig. 5c, the maximum value of C_p^b (i.e., $C_p^{p/2}$ or $C_p^{(p-1)/2}$) is larger than $2^{p/2}$ when p is larger than 2, and the difference grows increasingly with the increase of p, i.e., the number of storage nodes, indicating the encoding capability improvement of proposed encoding scheme. Therefore, it leads to the improvement of encoding efficiency, which will gradually approach 100% as p increases in the proposed encoding scheme (Fig. 5d), and thus addresses the challenges in conventional encoding scheme.

Fig. 5e shows the configuration algorithm of the proposed combinatorial encoding scheme for encoding s states, where the value of s depends on the practical problems. The minimum values of p and b can be determined to satisfy the condition that C_p^b can cover s, which will lead to the minimum number of total storage nodes and precharge-high SLs, respectively, for further optimizing the AEDP of CAM.

3) ***Proposed CAM Architecture based on FeFET***: Due to the excellent properties of emerging FeFET memory device as mentioned, the FeFET is adopted as the storage elements for experiments and evaluations of the proposed CAM architecture in this work. The FeFET-based CAM array is shown in Fig. 6a, where the threshold voltage of FeFET can be programmed to V_H/V_L state by applying relatively large programming voltages on SLs for ferroelectric polarization switching, which is used for entry storage of "0"/"1" (Fig. 6b). Benefiting from the three-terminal transistor feature of FeFET device, the search operation is implemented by applying non-destructive read-voltage to SLs with of $0V/V_S$ of relatively small amplitude to represent the input search query bit of "0"/"1", which can eliminate the extra transistors for comparison operation of CAM.

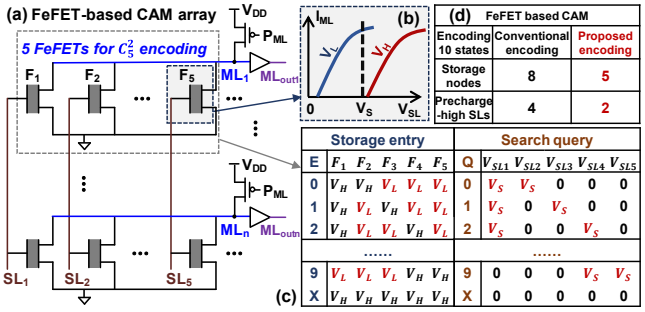

Fig. 6 (a) FeFET-based CAM array where the encoding and driving circuits of SLs are not drawn which are consistent with the previous description. (b) The FeFET device with programmable threshold voltage states for entry storage. (c) The C_5^2 encoding based on FeFET. (d) The comparison of hardware cost between conventional and proposed encoding schemes.

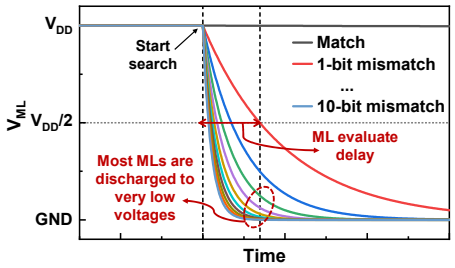

Fig. 7 The ML evaluation process in conventional searching scheme, resulting in high dynamic energy consumption due to most discharged MLs.

Taking the example of encoding 10 numeric characters (i.e., 10 states), according to the proposed configuration algorithm, $p=5$ and $b=2$ is determined with C_5^2 (i.e., 10) encoding, which means five FeFETs (F_1-F_5) are needed and encoded as a group, and two FeFETs in the group are set to "0"(V_H) and the other three FeFETs are set to "1"(V_L) (Fig. 6c). More FeFET devices can be connected in parallel in a row to store the whole entry. In search phase, search voltage is applied to SLs (V_{SL1}-V_{SL5}) according to the input query state. The number of precharge-high SLs (V_S) is equal to b. Only when search query matches the stored entry, can ML maintain high voltage since the group of FeFET are all turned off. Otherwise, ML will be discharged to ground as long as at least one FeFET in the group is turned on, indicating the mismatch condition.

Furthermore, by programming a group of all FeFETs to V_H, it can represent entry state of "X" that can match with all queries, which can be used for wildcard-search-based applications. Similarly, by setting all SLs to 0V, it can represent query "X" that will match all entries, which can be used for fragment-search-based applications.

As shown in Fig. 6d, in order to encode the same 10 states, conventional bitwise encoding scheme requires 8 storage FeFETs to cover the states (2^4>10), along with 4 precharge-high SLs. It clearly shows that the proposed combinatorial encoding scheme can save hardware overhead of storage elements and wires, and thus reduce the dynamic energy consumption for precharging MLs/SLs. Moreover, benefiting from the reduced number of SLs and precharge-high SLs which are equal to p and b respectively of proposed C_p^b encoding, the hardware cost and energy consumption of SLs can be further reduced.

B. Self-Terminating Searching Scheme for CAM

1) ***Challenges of Conventional Searching Scheme***: In conventional precharge-high ML searching scheme, the separate ML precharging and evaluation phases result in large search energy and delay penalty.

On the one hand, it leads to large dynamic energy consumption of SLs and MLs. First, all SLs need to be discharged to ground in phase one, and half of them will be charged to V_{DD} in phase two, regardless of the consecutive search queries. It results in significant dynamic power in SLs, much of which is unnecessary, e.g., the extreme scenario of searching consecutive identical queries. Second, in the ML evaluation process, only one matched ML remains high voltage, all mismatched MLs discharge with various rates, depending on the discharging current correlated to the number of mismatched bits (Fig. 7). To ensure that the 1-bit mismatched ML with the slowest discharge rate can discharge to $V_{DD}/2$, most of mismatched MLs are discharged to very low voltages, resulting in high dynamic power in MLs. On the other hand, it also leads to large search delay which will be

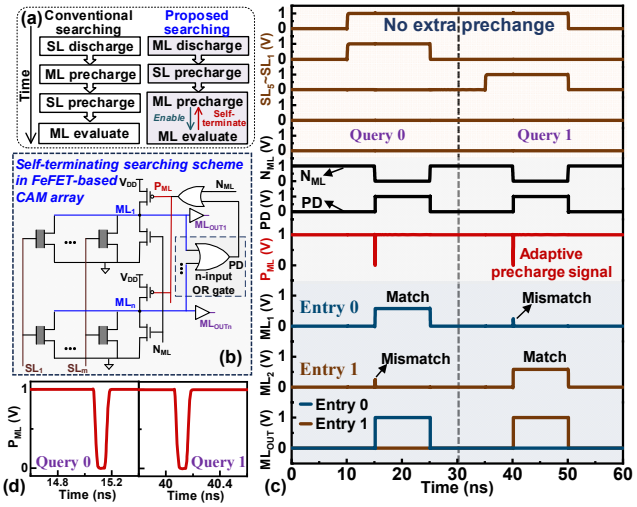

Fig. 8 (a) The diagram and (b) hardware implementation of proposed self-terminating searching scheme based on FeFET. The experimental results of (c) searching two different quires ("0" and "1") in sequence and (d) the adaptive precharge signals of the two search operations.

further increased as the length of entry increases due to the increased ML capacitance [9], limiting the search speed of large-scale search problems in CAM-based in-memory-searching accelerator.

2) ***Proposed Self-Terminating Encoding Scheme***: As mentioned above, the conventional precharge-high ML searching scheme consists of separate precharging phase and evaluation phase with frequently charging/discharging of SLs and MLs, suffering from high delay and energy consumption. In this work, a self-terminating searching scheme incorporating both precharging and evaluation phases is proposed (Fig. 8ab), which uses combined OR gate to detect matching condition with one-step precharging ML and can adaptively self-terminate the ML precharging with feedback connection, leading to higher speed and energy efficiency.

The experimental results are shown in Fig. 8c, which take entry/query of "0" and "1" with C_5^2 encoding as an example. Initially, by setting control signal (N_{ML}) high, all MLs are discharged to ground along with low feedback signal (PD) for disabling precharging phase. And then, SLs are precharged to corresponding voltages based on the search query. In comparison phase, N_{ML} is set to low voltage and P_{ML} changes to low voltage accordingly, which turns on the pullup transistors to precharge the MLs. Once the query is completely matched with one entry, a row of FeFETs are turned off, and the corresponding ML will be precharged to relatively high voltage (larger than $V_{DD}/2$), indicating that the search operation is complete, which will promptly force the PD to high voltage through multi-input OR logic and adaptively disable the precharge signal P_{ML} (Fig. 7d). Only the matched ML will maintain high, and the mismatched MLs will be discharged to ground by the discharging paths of mismatched cells. It should be noted that ML_{OUT} signals of mismatched MLs are always low since the voltage of MLs never exceeds $V_{DD}/2$ in the entire search process.

Therefore, the proposed one-step self-terminating searching scheme can reduce the search latency and energy consumption by integrating precharge-evaluation phase and self-terminating ML precharging. Moreover, different from the conventional searching scheme, the proposed self-terminating searching scheme eliminates the need to discharge SLs for precharging MLs per search operation, and precharge SLs with consecutively same query-bit input (Fig. 8c). The

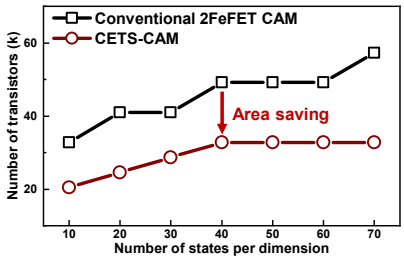

Fig. 9 The required number of transistors as the number of states per dimension increases of 2FeFET-based CAM and proposed 1FeFET-based CETS-CAM, indicating area saving of proposed CETS-CAM architecture.

precharge activity factor (α) of SL can thus be reduced, which can further reduce the dynamic energy of SL.

IV. Experimental Results

A. Experiment Setup

The experiments and evaluations of proposed CETS-CAM are carried out based on typical emerging FeFET device with circuit simulation in HSPICE. The BSIM model of 45nm technology node [16] is used for all MOSFETs, and Preisach ferroelectric switching model [17] is used for FeFETs. It should be noted that besides FeFET, the CETS-CAM architecture can also be implemented based on other storage technologies, such as RRAM, PCM and embedded DRAM. The key performance metrics including area, search energy, search delay and system AEDP are evaluated and compared with conventional CAM designs with the same technology node. Furthermore, the reliability and robustness of the proposed encoding and searching schemes are also investigated. The evaluated CAM array consists of 64 entries with 64 dimensions, and each dimension has various number of states for extensive evaluations.

B. Main Results

1) ***Area Cost***: Fig. 9 shows the comparison of required number of transistors between the reported state-of-the-art 2FeFET-based CAM with conventional encoding scheme [12] and the proposed 1FeFET-based CETS-CAM. Benefiting from the proposed combinatorial encoding scheme with higher encoding capability and efficiency, the required number of storage nodes is reduced, and thus the CETS-CAM exhibits lower area cost compared with conventional CAM designs. The improvement is more significant and tends towards 2× as the number of states per dimension increases.

2) ***Search Energy***: Fig. 10 shows the search energy comparison between 2FeFET-based CAM and the proposed 1FeFET-based CETS-CAM. The search energy mainly consists of dynamic energy of MLs and SLs. The ML energy consumption (E_{ML}) is dependent on the capacitance of ML (C_{ML}) and the voltage swing of ML (V_{swing}), which can be expressed as follows [2]:

$$E_{ML} = C_{ML} V_{DD} V_{swing} \quad (2)$$

where C_{ML} consists of capacitance of transistors connected to the ML and the parasitic wire capacitance is extracted according to [18]. The SL energy consumption (E_{SL}) can be expressed as [2]:

$$E_{SL} = C_{SL} V_{DD} V_S \alpha \quad (3)$$

where α is the precharge activity factor as mentioned above.

As shown in Fig. 10a, the E_{ML} of CAM array of the 1FeFET-based CETS-CAM is lower than that of the 2FeFET-based CAM. Due to the reduced transistors and wire length, the C_{ML} becomes smaller, leading to a 44% reduction of E_{ML} with 70 states. Moreover, the 1FeFET-based CETS-CAM

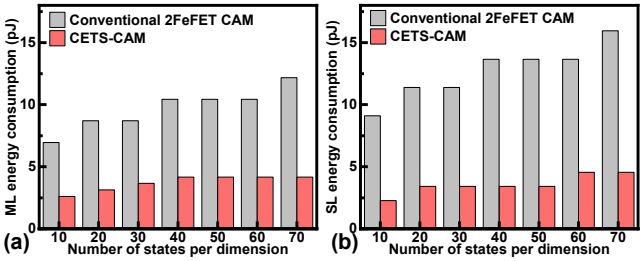

Fig. 10 The comparison of (a) ML energy and (b) SL energy between 2FeFET-based CAM and proposed 1FeFET-based CETS-CAM.

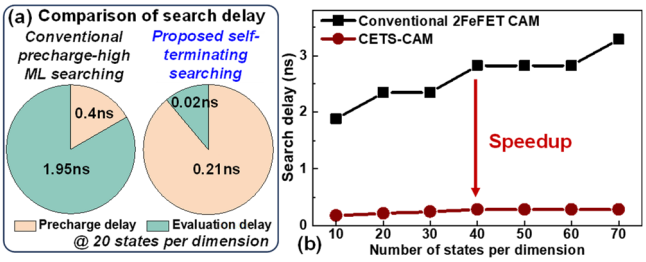

Fig. 11 (a) The composition of search delay for conventional searching scheme and proposed searching scheme. (b) the comparison of search delay between 2FeFET-based CAM and proposed 1FeFET-based CETS-CAM.

achieves ML evaluation during precharging based on the proposed self-terminating searching scheme, in which the precharging circuits are turned off when the matched ML is precharged to more than $V_{DD}/2$. As a result, the V_{swing} of ML is reduced, leading to a further 42% reduction of E_{ML} with 70 states.

As shown in Fig. 10b, the E_{SL} of CAM array shows the same tendency as the E_{ML}. Benefiting from the proposed encoding scheme, the number of precharged-high SLs is reduced, leading to a 47% reduction of E_{SL} per search operation with 70 states. On the other hand, in conventional searching scheme, all SLs are discharged low for precharging MLs and then precharged to the appropriate search-data values. Therefore, the value of α in Eq. (3) in conventional searching scheme is 1. While in proposed one-step self-terminating searching scheme, there is no need to precharge all SLs to ground before precharging ML, and the value of α depends on the search data. During each search operation, the voltages of SLs are either high or low, and thus the value of α is below 1, resulting in further 50% reduction of E_{SL} in average.

3) ***Search Delay***: The search delay of CAM array primarily includes the delay of SL and ML. The precharge and discharge delay of SL is mainly determined by the driving circuits of SL and the delay of ML is closely related to the CAM design. In conventional searching scheme, the delay of ML consists of precharge delay (t_{MLpre}) and evaluate delay (t_{MLeva}), which are defined as the risetime from $0.1V_{DD}$ to $0.9V_{DD}$ through the precharge transistor and the fall time from V_{DD} to $0.5V_{DD}$ as the 1-bit mismatched case [6]. They can be estimated as follows based on RC model [6]:

$$t_{MLpre} = 2.2 R_{EQpre} C_{ML} \quad (4)$$

$$t_{MLeva} = 0.69 R_{ML} C_{ML} \quad (5)$$

where R_{EQpre} is the equivalent resistance of the row shared precharge transistor, and R_{ML} is the largest pulldown resistance of ML among mismatch conditions. The t_{MLeva} is larger than t_{MLpre} and dominates the search delay in conventional searching scheme, as shown in Fig. 11a. While in proposed one-step self-terminating searching scheme, the evaluation is carried out simultaneously during precharging phase and self-terminating the precharge once a match is detected. The search delay is mainly determined by the t_{MLpre}, and the t_{MLeva} is the delay of adaptively feedback and self-

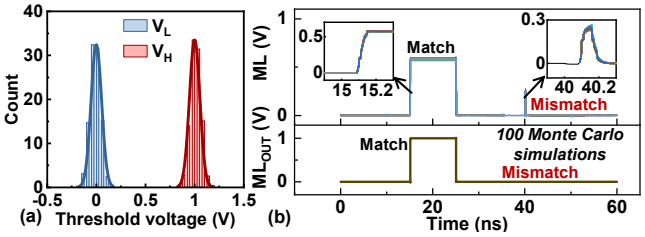

Fig. 12 (a) The distribution of high/low V_{TH} states of 100 FeFET devices. (b)Transient waveforms of 100 Monte Carlo simulations for two consequent search operations of CETS-CAM when considering device variation.

terminating circuit (Fig. 11a).

Fig. 11b shows the comparison of search delay between 2FeFET-based CAM and the proposed 1FeFET-based CETS-CAM, where the proposed CETS-CAM shows significantly reduced search delay compared with conventional 2FeFET CAM, especially for the increased number of states. Due to the reduced transistors and wire length, the C_{ML} is smaller than that of conventional encoding scheme, leading to the lower search delay according to Eq. (4)(5). Moreover, in proposed self-terminating searching scheme, the search delay is determined by the precharging speed of the matched ML in CAM array, rather than the discharging speed in the evaluate phase of conventional searching scheme, which is limited by the 1-bit mismatch case. As a result, the search speed of CETS-CAM when further adopting the proposed one-step self-terminating searching scheme is 11.7× faster than that of conventional-search-based CAM.

4) ***Robustness***: Additionally, considering the variation in practical circuits, the robustness of our proposed CETS-CAM architecture with combinatorial encoding and self-terminating searching schemes is validated. It is presumed that FeFET devices exhibit experimental variability with standard deviation of 54mV for the V_H/V_L state (Fig. 12a) [19], along with 5% size variation of CMOS devices in peripheral circuits of CAM array. The 100 Monte Carlo simulations are carried out, and the transient waveforms for two consequent search operations with match and mismatch conditions are shown in Fig. 12b. It shows that the 1FeFET-based CETS-CAM enables search operations without errors when considering the variation, indicating the reliability and robustness of our proposed CETS-CAM architecture.

C. CAM-based Protein Sequence Detecting Benchmarking

Based on the proposed CETS-CAM, the protein sequence detecting for read mapping which is a large-scale search problem in bioinformatics [4], is implemented and benchmarked. The proteins are composed of chains of amino acids with 20 states, which can be encoded by C_6^3 encoding scheme of CETS-CAM. While 10 storage nodes are needed for encoding one dimension with 20 states by using conventional bitwise encoding scheme. Table. I shows the performance comparison, where the energy and delay of SL/ML driving circuit, ML sense amplifiers, and the additional circuits of proposed searching scheme are further considered. The throughput (searches/s) and throughput/Watt (searches/J) are used to evaluate the delay and energy of different CAM designs, with the same 64×64 array size. As shown in Table. I, benefiting from the proposed compact and efficient CAM architecture, the 1FeFET-based CETS-CAM achieves 11×/3.4× improvement in speed/energy-efficiency and reduces the AEDP by 61.9× compared with conventional 2FeFET-based CAM. Moreover, by further utilizing the advantages of FeFETs, the 1FeFET-based CETS-CAM can reduce the search AEDP by 1182×/134.3× over CMOS-based/2T2R-based CAM, indicating the great potential for large-scale search applications.

Table I. Comparison of protein sequence alignment based on the proposed 1FeFET-based CETS-CAM or other reported advanced CAM designs.

CAM type	CMOS-CAM [6]	2T2R-CAM [9]	2FeFET-CAM [12]	CETS-CAM
Throughput (Searches/s)	24.8M	42.1M	42.6M	470.6M
Throughput / Watt (Searches/J)	21G	31.1G	50G	168.1G
AEDP reduction	1×	8.8×	19.1×	1182×

V. Conclusion

In this work, a novel CETS-CAM with co-optimization of reconfigurable combinatorial encoding and self-terminating searching is proposed and demonstrated with the lowest search AEDP. Based on the proposed encoding scheme, the encoding efficiency of CETS-CAM is significantly improved, and thus the area cost of storage elements is reduced. Furthermore, by merging the precharging and evaluation phases into one-step charging-detecting process, the search EDP can be further reduced. This work provides a general CAM architecture with higher area- and energy-efficiency compared with state-of-the-art CAM designs, showing its great potential for in-memory-searching system.

Acknowledgements

This work was supported by National Key R&D Program of China (2018YFB2202801), NSFC (61927901, 62374009), 111 Project (B18001).

Reference

[1] R. Karam et al., "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proceedings of the IEEE*, 2015.
[2] J. Cai et al., "Energy Efficient Data Search Design and Optimization Based on A Compact Ferroelectric FET Content Addressable Memory," in *DAC*, 2022.
[3] S. Liu et al., "A fast read alignment method based on seed-and-vote for next generation sequencing," *BMC Bioinformatics*, 2016.
[4] L. Liu., "A Reconfigurable FeFET Content Addressable Memory for Multi-State Hamming Distance," , *TCAS-I*, 2023.
[5] P. Tseng et al., "In-Memory Approximate Computing Architecture Based on 3D-NAND Flash Memories," in *VLSI*, 2022.
[6] K. Pagiamtzis et al, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," *JSSC*, 2006.
[7] X. S. Hu et al., "In-memory computing with associative memories: A cross-layer perspective," in *IEDM*, 2021.
[8] Y. Li et al., "Monolithic 3D Integration of Logic, Memory and Computing-In-Memory for One-Shot Learning," in *IEDM*, 2021.
[9] J. Li et al., "1 Mb 0.41 μm² 2T-2R Cell Nonvolatile TCAM With Two-Bit Encoding and Clocked Self-Referenced Sensing," *JSSC*, 2014.
[10] M. Chang et al., "Designs of Emerging Memory Based Non-Volatile TCAM for Internet-of-Things (IoT) and Big-Data Processing: A 5T2R Universal Cell," *ISCAS*, 2016.
[11] X. Yin et al., "Design and Benchmarking of Ferroelectric FET based TCAM," in *DATE*, 2017.
[12] K. Ni et al., "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, 2019.
[13] X. Yin et al., "An Ultra-Dense 2FeFET TCAM Design Based on a Multi-Domain FeFET Model," *TCAS-II*, 2019.
[14] H. Yang et al., "An Ultra-High-Density and Energy-Efficient Content Addressable Memory Design Based on 3D-NAND Flash," *SCIENCE CHINA Information Sciences,* 2022.
[15] X. Yin et al., "An Ultracompact Single-Ferroelectric Field-Effect Transistor Binary and Multibit Associative Search Engine," *Advanced Intelligent Systems*, 2023.
[16] W. Zhao et al, "New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration," *IEEE TED*, 2006.
[17] K. Ni et al., "A circuit compatible accurate compact model for ferroelectric-fets," in *VLSI*, 2018.
[18] H. Bhardwaj et al., "Power optimization using current-mode signalling technique for IoT applications," *Measurement: Sensors*, 2022.
[19] T. Soliman et al., "Ultra-low power flexible precision fefet based analog in-memory computing,"in *IEDM*, 2020.