Energy Efficient Data Search Design and Optimization Based on A Compact Ferroelectric FET Content Addressable Memory

Jiahao Cai[1], Mohsen Imani[2], Kai Ni[3], Grace Li Zhang[4], Bing Li[4], Ulf Schlichtmann[4], Cheng Zhuo[1] and Xunzhao Yin[1,5*]
[1]College of Information Science & Electronic Engineering, Zhejiang University, China
[2]Department of Computer Science, University of California Irvine, USA
[3]Department of Electrical & Microelectronic Engineering, Rochester Institute of Technology, USA
[4]Chair of Electronic Design Automation, Technology University of Munich, Germany
[5]Zhejiang Lab, China *Corresponding author email: xzyin1@zju.edu.cn

ABSTRACT

Content Addressable Memory (CAM) is widely used for associative search tasks in advanced machine learning models and data-intensive applications due to the highly parallel pattern matching capability. Most state-of-the-art CAM designs focus on reducing the CAM cell area by exploiting the nonvolatile memories (NVMs). There exists only little research on optimizing the design and energy efficiency of NVM based CAMs for practical deployment in edge devices and AI hardware. In this paper, we propose a general compact and energy efficient CAM design scheme that alleviates the design overhead by employing just one NVM device in the cell. We also propose an adaptive matchline (*ML*) precharge and discharge scheme that further optimizes the search energy by fully reducing the *ML* voltage swing. We consider Ferroelectric field effect transistors (FeFETs) as the representative NVM, and present a 2T-1FeFET CAM array including a sense amplifier implementing the proposed *ML* scheme. Evaluation results suggest that our proposed 2T-1FeFET CAM design achieves 6.64×/4.74×/9.14×/3.02× better energy efficiency compared with CMOS/ReRAM/STT-MRAM/2FeFET CAM arrays. Benchmarking results show that our approach provides 3.3×/2.1× energy-delay product improvement over the 2T-2R/2FeFET CAM in accelerating query processing applications.

1 INTRODUCTION

Content addressable memories (CAMs) are a promising type of computing-in-memory (CiM) hardware solutions [1, 2] that address the memory wall issues presented in Von Neumann machines. Due to the highly parallel search capabilities, CAMs exhibit great potentials for data-intensive applications nowadays, including machine learning, neuromorphic computing and lookup tables, etc. [2–8].

Conventional CMOS CAMs [9] suffer from high power consumption and low area density, thus researchers consider building compact and efficient CAM designs based on emerging non-volatile memory (NVM) devices, such as resistive RAM (ReRAM) [10], spin transfer torque magnetic RAM (STT-MRAM) [11] and Ferroelectric field effect transistor (FeFET) [12–14], etc. ReRAMs and STT-MRAMs featuring with merged variable resistor and non-volatile storage can encode their low resistance states (LRS)/high resistance states (HRS) as logic values '1'/'0', respectively, and thus can replace CMOS SRAM in building the CAM designs. The three-terminal FeFET devices can function as 1T non-volatile storage and switches rather than variable resistors due to their unique hysteresis *I-V* characteristics, high on-off current (i.e., I_{ON}-I_{OFF}) ratio and high OFF resistance, and thus have been promising to build compact and efficient CAM designs with less area overhead and energy consumption compared with CMOS CAM designs. These NVM based innovations are mainly dedicated to exploiting NVMs for compact CAM designs, hence achieve improved area footprints of CAM cells.

However, **the energy efficiency, as another critical metric for the practical usage of CAM designs in power-constrained edge devices and AI hardware, has not yet been fully optimized.** Moreover, multiple NVM devices are employed in aforementioned CAM designs, causing significant design overhead associated with the write schemes of NVMs. Specifically, ReRAM based CAM designs suffer from high search and write energies due to the low HRS/LRS resistance ratios and current-driven write mechanism. Since the STT-MRAM exhibits a limited tunnel magneto-resistance (TMR) ratio, STT-MRAM based CAM cells [11] require extra transistors to facilitate reliable write and search operations, incurring extra area overhead and degrading the energy efficiency. FeFET based CAM designs [12] can improve the energy and area efficiency compared with other CAM designs with its electric field driven write mechanism and compact design. However, further optimization on energy efficiency and design overhead of FeFET based CAMs still call for exploration.

To address the design overhead and energy-aware issues of NVM based CAMs, in this paper, **we propose a general compact and energy efficient CAM design scheme that employs the minimum number of NVM,** i.e., just one device as storage element for an efficient design and energy optimization. We consider FeFET as a representative NVM device, and propose a novel compact and energy efficient 2T-1FeFET based CAM design which utilizes just one FeFET to alleviate design overhead associated with the write scheme, and can be adopted to other NVMs. To further optimize the search energy overhead associated with MVM based CAM arrays, **we propose an adaptive matchline (*ML*) precharge and discharge scheme which is implemented by a threshold inverter quantization (TIQ) comparator-based sense amplifier (SA)**. The proposed *ML* scheme can terminate the *ML* precharge and discharge, thus achieving reduced voltage swing and improved energy efficiency. The structures and operations of our proposed 2T-1FeFET CAM design are discussed, and the operation principles and energy analysis of our proposed *ML* precharge and discharge scheme are illustrated. Evaluation results suggest that our proposed 2T-1FeFET CAM array using adaptive *ML* precharge and discharge scheme can achieve 6.64×/4.74×/9.14×/3.02× better energy efficiency than the CMOS/ReRAM/STT-MRAM/2FeFET CAM. Benchmarking on the query processing application demonstrates that our proposed design can achieve 2.5×/2.1× (4.0×/3.3×) speedup/energy-delay product (EDP) improvement over GPU based associative memory (AM) compared with 2FeFET(ReRAM) based approach.

DAC '22, July 10–14, 2022, San Francisco, CA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9142-9/22/07…$15.00
https://doi.org/10.1145/3489517.3530527

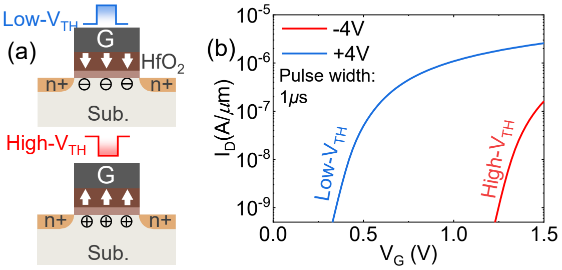

Figure 1: (a) FeFET polarization directions and channel conditions after memory write operations; (b) The FeFET I_D-V_G characteristics after positive/negative gate write voltages.

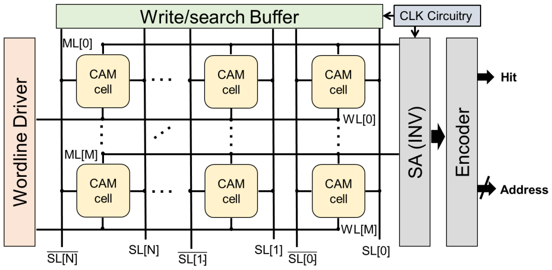

Figure 2: Schematic of an M×N CAM array.

2 PRELIMINARIES AND RELATED WORKS

2.1 FeFET Device and Model

FeFETs which integrate HfO_2 as the ferroelectric dielectric become a competitive candidate for embedded NVM as they are CMOS-compatible, energy-efficient, and compact [15]. As shown in Fig. 1(a), applying a positive/negative gate pulse will switch the ferroelectric polarization within the FeFET gate layer to the channel/gate metal direction, thus setting the FeFET to the low-V_{TH}/high-V_{TH} state, respectively. A gate voltage between the low-V_{TH} and high-V_{TH} can be applied to read the stored data (i.e., '1' and '0') through the drain current (i.e., I_{ON} and I_{OFF}). Compared with other NVM devices, i.e., ReRAM, which require large currents and thus significant power during the write, FeFETs exhibit superior write energy due to the electric field driven write scheme [16].

In this paper, we utilize the experimentally calibrated Preisach FeFET model [17] for the device and circuit simulations. The Preisach model considers the ferroelectric layer as a film composed of multiple domains with independent Q_{FE}-V_{FE} hysteresis loops. By combining the ferroelectric film response and the history tracking with non-saturated hysteresis loops, the ferroelectric model is developed, and integrated with an underlying MOSFET model for FeFET. The model has been calibrated with experiments [17]. Fig.1(b) shows the I_D-V_G curve of a FeFET with approximately 1V memory window.

2.2 Existing CAM Designs

Fig. 2 shows a NOR-type CAM array. A *ML* is shared by all cells within the word, and the searchlines (*SLs*) are used to distribute input bits across the array. The search operation starts with precharging the *MLs*, and then applying input to *SLs*. When a word match with input, the *ML* remains high, indicating a match. When a word mismatches with input, *ML* discharges to 0, indicating a mismatch. Below we review the CAMs that can be adopted in this array.

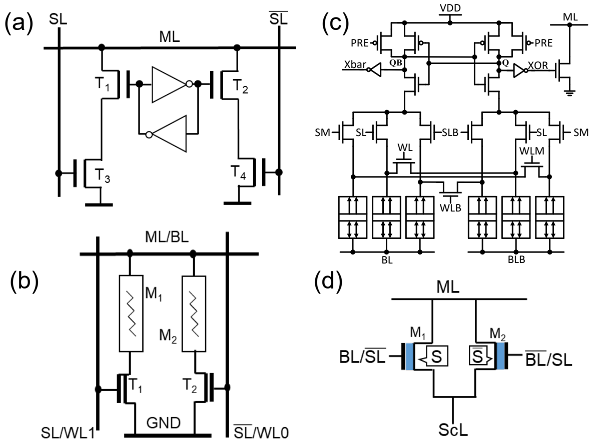

Figure 3: CAM designs based on (a) 10T CMOS; (b) 2T-2ReRAM; (c) 20T-6MTJ; (d) 2FeFET.

A conventional CMOS based 10T CAM design [18] is shown in Fig. 3(a). Fig. 3(b) demonstrates the most common 2T-2R CAM design which has a compact structure [10]. While the 2T-2R CAM consumes less area than a conventional CAM design, the write and search energies become a major concern due to the low variable resistance, low HRS/LRS ratios, current-driven write mechanism associated with large access transistors. A state-of-the-art STT-MRAM based CAM design as shown in Fig. 3(c) was presented in [11], exhibiting fast search performance and enhanced search functionality. However, a number of devices are required to address the degraded sensing margin caused by the limited TMR ratios of STT-MRAMs, thus heavily sacrificing the area and energy efficiency. FeFETs stand out among NVMs due to their high I_{ON}/I_{OFF} ratio, three-terminal structure, and low I_{OFF}. A 2FeFET based CAM has been proposed [12] as shown in Fig. 3(d), demonstrating improved performance, energy, and area efficiency over conventional CMOS based and other NVM based CAMs. Nevertheless, the energy efficiency of such CAM design can still be optimized, and the 2FeFET structure requires a sophisticated design for the voltage driven write scheme. Taking the 2FeFET based CAM design as an example, the precharge energy E_{pre} of an array is expressed as below:

$$E_{pre} = C_{ML} V_{DD} \Delta V \tag{1}$$

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

where V_{DD}, ΔV, C_{ML}, C_{PMOS}, C_{cell}, $C_{parasitic}$ and C_{drain} are the supply voltage, *ML* voltage swing, the capacitance associated with *ML*, the drain capacitance of the PMOS transistor precharging the *ML*, total capacitance of a CAM cell associated with the *ML*, the parasitic capacitance of each cell, and the drain capacitance of a transistor, respectively. C_{cell} consists of two drain capacitances. From Eq. (1) and (2) it can be seen that a potential optimization method to reduce the precharge energy is to reduce the number of transistors associated with the *ML*. Following this method, a novel energy-aware CAM design based on FeFET that associates one transistor with the *ML* has been proposed in [19]. However, such CAM design contains 2 FeFETs, resulting in extra design overhead for the write operation. Moreover, the design is optimized just at cell level, further optimization on the *ML* has not yet explored. Therefore, we propose a novel 2T-1FeFET CAM cell that associates only one FeFET device to *ML* to alleviate the design overhead regarding the write scheme shown in Fig. 1.

2.3 Energy-Aware ML Optimization Schemes

According to Eq. (1), besides the voltage scaling [20], another method to optimize the precharge energy is to reduce the voltage swing ΔV

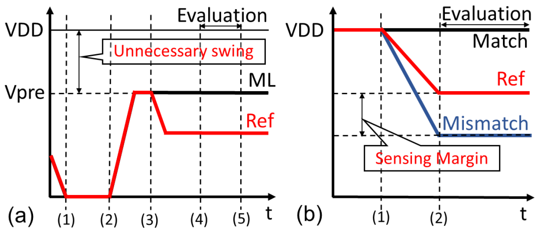

Figure 4: Energy-aware *ML* schemes: (a) early termination *ML* precharge scheme. (b) adaptive ML discharge scheme.

Table 1: Operations of 2T-1FeFET CAM Cell

V_{write}=4V V_{search}=1V	SL	$\overline{SL}$	WL	CL
Write '1'	0	0	V_{write}	V_{search}
Write '0'	0	0	$-V_{write}$	V_{search}
Search '1'	V_{search}	0	V_{read}	V_{search}
Search '0'	0	V_{search}	V_{read}	0

at array *ML*s. [21] proposed an early termination *ML* precharge scheme to eliminate the unnecessary *ML* precharge as shown in Fig. 4(a). On the contrary, [22] as shown in Fig. 4(b) proposed an adaptive *ML* discharge scheme to eliminate redundant *ML* discharge. Both improve the energy efficiency by reducing the *ML* voltage swing. That said, the *ML* voltage swing can still be further reduced regrading the existing schemes. Moreover, neither of the schemes is feasible for the 2T-2R CAM design in Fig. 3(b), as the *ML*s of the CAM array always discharge regardless of match/mismatch condition. In this sense, for the NVMs with limited HRS/LRS ratios, applying aforementioned optimization schemes to the CAM array will result in extra design overhead, making the optimization schemes less attractive. Therefore, it is necessary to propose a holistic co-design approach that considers both the compactness of the CAM cell and the feasibility of the *ML* optimization scheme when optimizing the energy efficiency of a CAM array.

In this paper, we propose a novel *ML* optimization scheme which can be combined with the proposed 2T-1FeFET CAM cell to fully leverage the advantages of both FeFET and the optimization scheme, and ultimately optimize the energy efficiency of the CAM array.

3 FEFET CAM DESIGN AND OPERATION

3.1 2T-1FeFET CAM Cell

Fig. 5(a) shows the structure of the proposed 2T-1FeFET CAM cell, which consists of one FeFET (M_0) and two NMOS transistors (T_1 and T_2). M_0 and T_1 are connected in series, forming a voltage divider, which is supplied by searchlines SL and $\overline{SL}$. As will be explained later, this voltage divider can realize an XOR operation between the search data and stored words, such that the internal node D voltage will be high only when mismatch happens, otherwise the D voltage will be low upon a match. Node D connects to the gate of an NMOS (i.e., T_2) which can implement the required XNOR operation in the CAM cell such that when mismatch happens the *ML* discharges.

The operations of the 2T-1FeFET CAM cell are explained in 5(b). When bit '1' is searched, the SL and $\overline{SL}$ will be high and low, respectively. Therefore, the node D voltage during search '1' is:

$$V_{D,Sr1} = \frac{V_{search}R_{M0}}{R_{M0} + R_{T1,Sr1}} \tag{3}$$

where V_{search} is the SL voltage, $R_{T1,Sr1}$ is the T_1 transistor resistance during search '1'; R_{M0} is the FeFET resistance, which could be either R_{low} or R_{high}, depending on the FeFET V_{TH} state. Therefore, when state '1' is stored (i.e., low-V_{TH} or R_{low}), node D voltage is

$$V_{D,Sr1St1} = \frac{V_{search}R_{low}}{R_{low} + R_{T1,Sr1}} \tag{4}$$

When state '0' is stored (i.e., high-V_{TH} or R_{high}), node D voltage is

$$V_{D,Sr1St0} = \frac{V_{search}R_{high}}{R_{high} + R_{T1,Sr1}} \tag{5}$$

Therefore, by choosing an appropriate bias for transistor T_1, its resistance $R_{T1,Sr1}$ can be set in between R_{low} and R_{high} and thus the corresponding $V_{D,Sr1St1}$ and $V_{D_{Sr1St0}}$ is below and above the V_{TH} of transistor T_2, respectively. In this way, the match and mismatch operations are realized.

Similarly, when bit '0' is searched, the voltages on the SL and $\overline{SL}$ are reversed. The node D voltage is:

$$V_{D,Sr0} = \frac{V_{search}R_{T1,Sr0}}{R_{M0} + R_{T1,Sr0}} \tag{6}$$

By setting the T_1 resistance, $R_{T1,Sr0}$ between R_{low} and R_{high}, the node D voltage for stored '1'

$$V_{D,Sr0St1} = \frac{V_{search}R_{T1,Sr0}}{R_{low} + R_{T1,Sr0}} \tag{7}$$

and node D voltage for stored '0'

$$V_{D,Sr0St0} = \frac{V_{search}R_{T1,Sr0}}{R_{high} + R_{T1,Sr0}} \tag{8}$$

will be above and below the V_{TH} of transistor T_2, respectively. It therefore realizes the correct search functionality for searching '0'. Fig. 5(c) shows the waveforms of the proposed 2T-1FeFET CAM cell which stores logic '1'. The node D voltage is only high when mismatch happens, validating the CAM operation. Our proposed 2T-1FeFET CAM cell associates only one transistor to the *ML*, resulting in *ML* capacitance reduction. The modified C_{ML} equation is:

$$C_{ML} \approx C_{PMOS} + N \times (C_{drain} + C_{parasitic}) \tag{9}$$

Therefore, 2T-1FeFET CAM cell decreases the precharge overhead of the *ML* by reducing *ML* capacitance compared with existing precharge-based CAMs. The FeFET device count reduction also alleviates the design overhead when considering the write operation.

Table 1 summarizes the write and search operations of the proposed CAM cell. Since the cell has only one FeFET as a storage element, write operation can be completed in only one step. To write a logic '1' or '0', V_{write} or $-V_{write}$ is applied to WL, respectively, while SL and $\overline{SL}$ are grounded. It is worth noting that irrespective of writing logic '1' or logic '0', V_{search}, or V_{DD}, is applied to CL and ground is applied to SL. Therefore, node D voltage is well defined at ground during the write. With these write configurations to the cell, desired data is written into the FeFET.

Based on these analyses, it can readily be seen that the proposed CAM cell is a general design that can be applied beyond the FeFET and to other NVMs as well, such as ReRAM and PCM, etc. This is because the voltage divider only requires the existence of two resistance states of NVM, not relying on any specific technology. Another good feature is that T_1 can act as the access transistor for the memory element during the write operations, especially for two-terminal resistive memories. In addition, it can solve the issue of limited ratio of some resistive memory (i.e., $R_{high}/R_{low} \approx 100$ in ReRAM and PCM) by amplifying through the action of transistor (i.e., T_2). By inducing a small voltage difference in node D (e.g., $\approx 0.5V$), a large ratio of *ML* discharge current (e.g., $\approx 10^4$) between the mismatch and match can be obtained. Therefore, it represents a general solution for compact, energy-efficient CAM design.

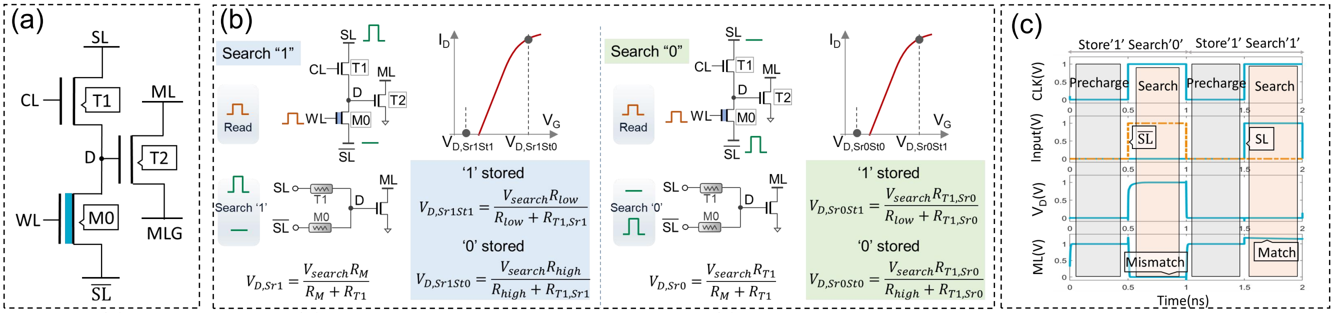
(a)
SL
CL
T1
ML
D
T2
WL
M0
MLG
SL
(b)
Search "1"
SL
CL
T1
ML
D
T2
WL
M0
SL
Read
I_D
V_G
V_D,Sr1St1
V_D,Sr1St0
SL
T1
ML
Search '1'
D
M0
SL
$V_{D,Sr1} = \frac{V_{search}R_M}{R_M + R_{T1}}$
'1' stored
$V_{D,Sr1St1} = \frac{V_{search}R_{low}}{R_{low} + R_{T1,Sr1}}$
'0' stored
$V_{D,Sr1St0} = \frac{V_{search}R_{high}}{R_{high} + R_{T1,Sr1}}$
Search "0"
Read
V_D,Sr0St0
V_D,Sr0St1
Search '0'
$V_{D,Sr0} = \frac{V_{search}R_{T1}}{R_M + R_{T1}}$
'1' stored
$V_{D,Sr0St1} = \frac{V_{search}R_{T1,Sr0}}{R_{low} + R_{T1,Sr0}}$
'0' stored
$V_{D,Sr0St0} = \frac{V_{search}R_{T1,Sr0}}{R_{high} + R_{T1,Sr0}}$
(c)
Store'1' Search'0'
Store'1' Search'1'
CLK(V)
Input(V)
V_D(V)
ML(V)
Precharge
Search
Precharge
Search
SL
SL
Mismatch
Match
0
0.5
1
1.5
2
Time(ns)

Figure 5: (a) Schematic and (b) operations of 2T-1FeFET CAM cell; (c) Transient waveforms of 2T-1FeFET CAM cell stored '1'.

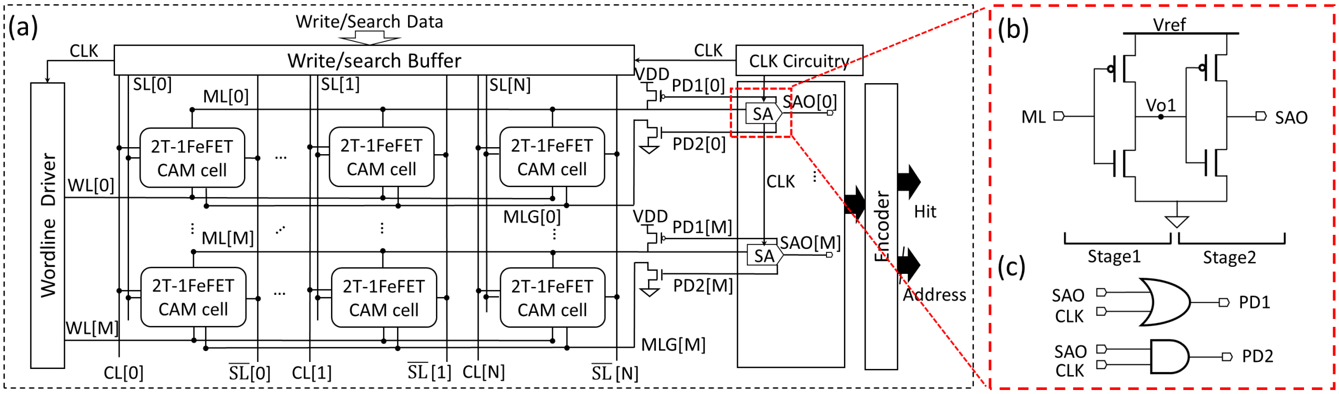

Figure 6: (a) Architecture of an M*N 2T-1FeFET CAM array; (b) TIQ comparator; (c) Output peripheral circuit.

3.2 2T-1FeFET CAM Array

Fig. 6(a) shows the proposed 2T-1FeFET CAM array. The array includes the CAM core of $M \times N$ size, the write/search buffer, the wordline driver, the SAs and peripheral circuitry. The searchlines (i.e., *SLs*) and control lines (i.e., *CLs*) are shared by the cells within the same column, while the wordlines (i.e., *WLs*) connect the cells within the same rows. The SA detects the *ML* voltage, and outputs the match/mismatch result, denoted as *SAO*. Fig. 6(b) and (c) demonstrate the SA, which contains a TIQ comparator [23] and two logic gates. The TIQ comparator where two inverters are cascaded is driven by a reference voltage V_{ref} to compare *ML* voltage V_{ML} with the threshold voltage of inverter. The *OR* and *AND* gates generate *PD*1 and *PD*2 control signals. When both *CLK* and *SAO* are low, *PD*1 will be low, and precharge *ML*. When both *CLK* and *SAO* are high, *PD*2 will be high, and pulldown *ML*. Note that the *MLG* node of the CAM cells within a word is grounded via a pulldown transistor which is gated by *PD*2 signal. The gate of the *ML* precharge PMOS transistor is controlled by *PD*1 signal.

For the write operation, we apply $V_{write}/2$ inhibition bias scheme to all *WLs* associated with unselected rows and $\overline{SL}$s associated with unselected columns to avoid write disturbance [24]. During the search, the *MLs* of the array are first precharged by *PD*1 signals, and then *PD*2s turn on the pulldown transistors, connecting the CAM cells to ground via *MLGs*. All *WLs* are activated and the $SL/\overline{SL}$s, *CLs* are set to the search voltages according to the input data as summarized in Table 1. Upon a mismatch, the discharging path is turned on, discharging the *ML* to ground via the mismatched CAM cells and the pulldown transistor. Upon a match, the *ML* stays at high as there exists no discharging path. The TIQ comparator compares V_{ML} with the threshold voltage, and generates output signal. Below we illustrate the detailed search operations using our proposed adaptive *ML* precharge and discharge scheme.

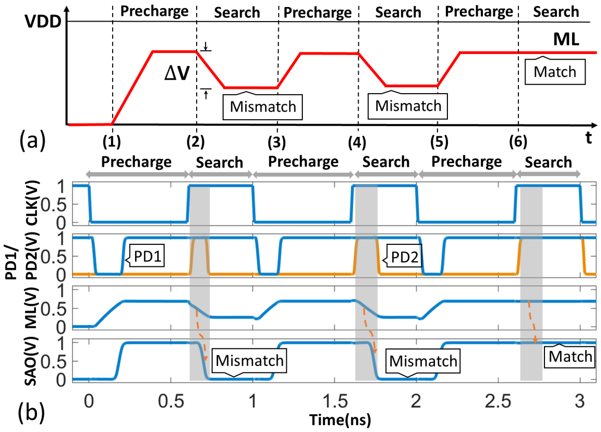

Figure 7: (a) Proposed adaptive *ML* precharge and discharge scheme and (b) corresponding simulation waveforms.

3.3 Adaptive ML Precharge & Discharge

Fig. 7(a) shows the proposed adaptive *ML* precharge and discharge scheme. It starts with *ML* precharge at timepoint (1). However, instead of precharging the *ML* to supply voltage, the precharge is terminated when the SA detects that V_{ML} exceeds a threshold voltage. The search operation then starts at timepoint (2). When a match occurs, *ML* stays at its voltage level. Upon a mismatch, *ML* is discharged. Similarly, SA will shut down the discharge path to terminate the discharge once V_{ML} falls below the threshold voltage. As such, the proposed scheme saves the precharge energy by reducing the voltage swing to small ΔV according to Eq. (1).

Fig. 7(b) shows the corresponding transient waveforms of the CAM array and the SA. During the precharge of search operation, the *ML* voltage V_{ML} is lower than the threshold voltage V_{th}, thus the TIQ comparator output *SAO* is low. The clock signal *CLK*, *PD*1

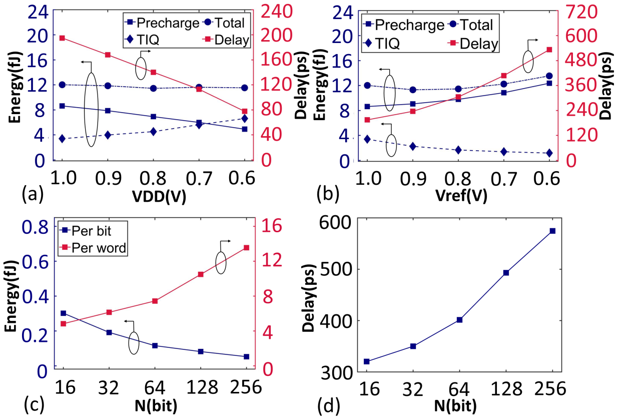

Figure 8: Energy and delay of proposed 2T-1FeFET CAM array with different (a) V_{DD} and (b) V_{ref}, respectively. (c) Search energy and (d) delay with different wordlength.

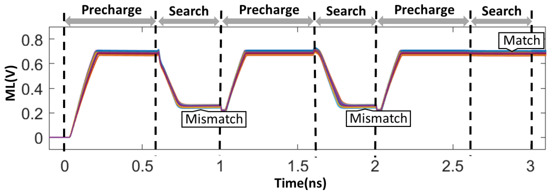

Figure 9: Transient waveforms of the 2T1-FeFET CAM array upon device variability.

and $PD2$ are all at low level, turning off the pulldown transistor, and turning on the precharge PMOS, respectively, the ML is precharged. Once V_{ML} is precharged to exceed V_{th}, SAO rises to high level, and drives $PD1$ to high level, thus terminating the precharge. When search starts, CLK and SAO are at high level, thus $PD2$ is driven to high level, turning on the pulldown NMOS transistor. Upon a match, ML remains at high, and SAO stays at high level, indicating a match. When a mismatch occurs, ML is discharged by a pulldown path from ML to ground. Once V_{ML} is discharged below V_{th}, SAO becomes 0, thus $PD2$ is driven to 0. As a result, the pull down transistor is turned off, and terminates the discharge.

4 EVALUATION

4.1 CAM Array Evaluation

Here we conduct the evaluations of our proposed 2T-1FeFET CAM array with SPECTRE using the Preisach FeFET model [17]. The write voltage for the 2T-1FeFET CAM is ±4V. The 45nm PTM model [25] is adopted for all MOSFETs with TT process corner at 25°C. We assume the minimum sized transistors in order to reduce power. Wiring parasitics for 45nm technology node are extracted from DESTINY [26]. The wordlength is 64. The search delay is measured for the worst case, where there is only one-bit mismatch.

Fig. 8(a) and (b) show the energy and delay of a word of the proposed 2T-1FeFET CAM array under V_{DD} and V_{ref} scaling, respectively. The search energy of the array mainly consists of two parts: (i) the precharge energy associated with the MLs, and (ii) SA energy consumption. The ML prechage energy is dependent on the ML associated capacitance and the voltage swing per Eq. (1), while the SA energy consumption is dominated by the TIQ comparator. As shown in Fig. 8(a), as V_{DD} scales down, the precharge strength is weakened, lowering down the upper bound of voltage swing. Then the reduced ML voltage swing results in the decreasing of precharge energy and the search delay. However, the reduced voltage swing implies that V_{ML} swings around the threshold voltage of TIQ comparator, and the near-threshold V_{ML} tends to keep the comparator transistors conducting current, resulting in higher static power consumption. Therefore, the change in total energy consumption is negligible. On the contrary, scaling V_{ref} down will slow down the response speed of TIQ comparator, causing longer precharge and discharge time. This results in larger voltage swing, thus larger search delay and precharge energy as shown in Fig. 8(b). Overall, the total energy change is negligible. Based on above analysis, scaling V_{ref} down reduces the SA energy consumption at the cost of increasing voltage swing. This can be partly addressed by scaling V_{DD} down to lower down the upper bound of the swing.

Table 2: Metric Comparison Summary of CAM Designs

Reference	[18]	[10]	[11]	[12]	This work
Technology	CMOS	ReRAM	STT-MRAM	FeFET	FeFET
Transistors/cell	10T	2T-2R	20T-6MTJ	2FeFET	2T-1FeFET
Cell size(μm^2)	3.3★	0.41†	18.05	0.15	0.36
Search delay	1.07ns	350.6ps	170ps	340.8ps	401.4ps
Energy [fJ/bit/search]	0.77 6.64X	0.55 4.74X	1.06 9.14X	0.35 3.02X	0.116 1X

★: The 10T CMOS CAM cell size is implemented using a standard 65 nm/1.2 V CMOS process.
†: The 2T-2R ReRAM based CAM cell is based on 90nm.

According to the voltage scaling analysis, we use the lowest operating V_{DD}(=0.6V) and V_{ref}(=0.6V). Fig. 8(c) and (d) shows the search energy and delay of the proposed 2T-1FeFET CAM array with varying wordlength. As the wordlength increases, the associated ML capacitance increases, slowing down the precharge/discharge speed, and resulting in an increase of search delay. The increasing capacitance also leads to larger precharge energy per word. However, increasing wordlength has negligible impact on the SA energy consumption, which is dependent on the TIQ comparator. In this sense, the search energy per bit is decreased as shown in Fig. 8(c).

Table 2 summarizes the metrics of 2T-1FeFET CAM and other CAMs, such as the technology, device count per cell, cell size, search delay and search energy per bit per search. The cell sizes are estimated based on a 2X2 2T-1FeFET CAM array layout. The cell size of the proposed 2T-1FeFET CAM is 10.9% of the conventional 10T CMOS CAM. Less area overhead of the proposed CAM leads to less parasitic capacitance associated with ML, resulting in less search energy and search delay. With the novel 1FeFET based CAM design and the proposed adaptive precharge and discharge scheme, our proposed 2T-1FeFET CAM is 6.64× and 2.67× more energy efficient and faster than 10T CMOS CAM design, respectively. Due to the scaling V_{DD} and V_{ref} for the reduced voltage swing and search energy, our approach is a bit slower than 2T-2R and 2FeFET CAMs, which is still acceptable, as our proposed CAM design achieves 4.74×/3.02× more energy efficiency than 2T-2R/2FeFET CAM designs. While the search delay of STT-MRAM CAM is 58% less than our proposed CAM, the cell size of the our proposed CAM cell is just 1.99% of that of the STT-MRAM CAM, thus incurring a huge density advantage that can compensate the slightly degraded performance. Our proposed 2T-1FeFET CAM is 9.14× more energy efficient than 20T-6MTJ CAM design. These results validate the efficiency of our 2T-1FeFET CAM array with adaptive ML precharge and discharge scheme for data-intensive search applications.

We also validate the robustness of our proposed 2T-1FeFET CAM design and the ML scheme. FeFET devices are assumed an experimental variability with σ=54mV for the low/high V_{TH} state [27],

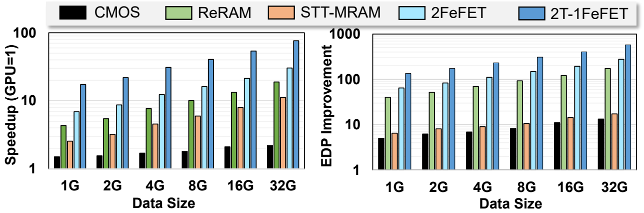

Figure 10: Query processing: performance and energy-delay product of CAM-based solutions compared to GPU.

and CMOS with 5% size variation. Fig. 9 shows the transient waveforms of *ML* with process variations included in the proposed CAM array during the search operations. The 200 Monte Carlo simulations in Fig. 9 show that the proposed 2T-1FeFET CAM array enables error-free search, suggesting the reliability and robustness of our proposed design and optimization methods.

4.2 CAM based AM Benchmarking

We further benchmark our proposed approach in associative memory based applications. Query processing executes a query over the stored data of file systems in a step-wise process [28]. It generally has two main steps: query filtering and query execution. The query filtering is a search operation that identifies the memory blocks that contain the required information of the query. In conventional processors, e.g., CPU, the query runs over a large amount of data stored in memory, resulting in significant data movement between memory and processing cores [28, 29]. The most common operation in many query processors is looking up a set of data that matches the input. A typical search query involves a brute-force search through a CAM storing the data. This is typically implemented in (i) word-by-word search and (ii) bit-by-bit search. A word-by-word search looks through every stored word in the CAM sequentially to find a match. In the worst case, it involves processing every element in the CAM. The bit-by-bit search scans one bit with the same index for multiple words at a time. This exact search operation supports several important queries: (1) check the specific data existence in the database, (2) get the number of rows that match certain criteria, and (3) find the existence of a pattern of data in the rows.

Fig. 10 compares the efficiency of different CAM architectures accelerating the query processing. All results are respective to the same query running on NVIDIA GTX 1080 GPU. For fairness, all evaluations are performed when all CAM-based solutions are providing the same chip area. In contrast to GPU that supports a limited amount of parallelism, CAMs support row-parallel searches, which enable accelerating multiple queries over the stored data in memory. Particularly, CAM-based search provides parallelism substantially over large-scale data that directly translates to faster and more efficient computation. The efficiency comes from CAM capability in addressing data movement issues by eliminating data transfer between off-chip memory and GPU cores. Evaluation on 1GB data shows that our 2T-1FeFET CAM provides 12.4× faster computation than GPU. This performance speedup increases to 38.2× and 54.5× when data size increases to 16GB and 32GB, respectively.

Comparing different CAM-based solutions, the computation efficiency depends on cell density and search efficiency. Due to the low cell density of CMOS and STT-MRAM, FeFET-based CAM solutions offer higher computation efficiency. Although ReRAM has higher density than CMOS and STT-MRAM based CAMs, it consumes more energy than FeFET-based CAMs. Due to the energy efficiency at array level, our CAM provides 2.5× and 2.1× (4.0× and 3.3×) faster and higher search EDP improvement over 2FeFET (ReRAM) CAM in accelerating query processing applications.

5 CONCLUSION

In this paper, we propose a general compact and energy efficient CAM design that employs the minimum number of NVM as storage element to alleviate the design overhead and optimize the search energy. We propose a novel 2T-1FeFET CAM design which utilizes just one FeFET and can be adoped to other NVMs. We then propose an adaptive *ML* precharge and discharge scheme implemented by a TIQ comparator based SA for further energy optimization. Evaluation results and query processing application benchmarking suggest that our proposed 2T-1FeFET CAM array with *ML* optimization scheme achieves better energy efficiency and performance when compared with other state-of-the-art CAM approaches.

ACKNOWLEDGMENTS

This work was partially supported by Zhejiang Provincial Key R&D program (2020C01052, 2022C01232), NSF (LQ21F040006, LD21F040003), NSFC (62104213, 92164203), Zhejiang Lab (2021MD0AB02).

REFERENCES

[1] C. Li *et al.*, "A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing," in *IEEE IEDM*, pp. 1–4, 2020.
[2] R. Karam *et al.*, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1311–1330, 2015.
[3] C. Chen *et al.*, "Pam: A piecewise-linearly-approximated floating-point multiplier with unbiasedness and configurability," *IEEE TC*, 2021.
[4] M. Imani *et al.*, "Searchd: A memory-centric hyperdimensional computing with stochastic training," *IEEE TCAD*, vol. 39, no. 10, pp. 2422–2433, 2019.
[5] X.-T. Nguyen *et al.*, "An efficient i/o architecture for ram-based content-addressable memory on fpga," *IEEE TCAS-II*, vol. 66, no. 3, pp. 472–476, 2018.
[6] X. S. Hu *et al.*, "In-memory computing with associative memories: a cross-layer perspective," in *2021 IEEE IEDM*, pp. 25–2, IEEE, 2021.
[7] X. Yin *et al.*, "Deep random forest with ferroelectric analog content addressable memory," *arXiv preprint arXiv:2110.02495*, 2021.
[8] X. Yin *et al.*, "An ultra-compact single fefet binary and multi-bit associative search engine," *arXiv preprint arXiv:2203.07948*, 2022.
[9] K. Pagiamtzis *et al.*, "Content-addressable memory (cam) circuits and architectures: A tutorial and survey," *JSSC*, vol. 41, no. 3, pp. 712–727, 2006.
[10] 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, no. 4, pp. 896–907, 2014.
[11] C. Wang *et al.*, "Design of magnetic non-volatile tcam with priority-decision in memory technology for high speed, low power, and high reliability," *IEEE TCAS-I*, vol. 67, no. 2, pp. 464–474, 2019.
[12] X. Yin *et al.*, "An ultra-dense 2fefet tcam design based on a multi-domain fefet model," *IEEE TCAS-II*, vol. 66, no. 9, pp. 1577–1581, 2018.
[13] X. Yin *et al.*, "Fecam: A universal compact digital and analog content addressable memory using ferroelectric," *IEEE TED*, vol. 67, no. 7, pp. 2785–2792, 2020.
[14] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, no. 11, pp. 521–529, 2019.
[15] A. I. Khan *et al.*, "The future of ferroelectric field-effect transistor technology," *Nature Electronics*, vol. 3, no. 10, pp. 588–597, 2020.
[16] S. Salahuddin *et al.*, "The era of hyper-scaling in electronics," *Nature Electronics*, vol. 1, no. 8, pp. 442–450, 2018.
[17] K. Ni *et al.*, "A circuit compatible accurate compact model for ferroelectric-fets," in *IEEE VLSI*, pp. 131–132, 2018.
[18] A. T. Do *et al.*, "Design of a power-efficient cam using automated background checking scheme for small match line swing," in *ESSCIRC*, pp. 209–212, IEEE, 2013.
[19] Y. Qian *et al.*, "Energy-aware designs of ferroelectric ternary content addressable memory," in *DATE*, pp. –, IEEE, 2021.
[20] S. Joshi *et al.*, "Multi-vdd design for content addressable memories (cam): A power-delay optimization analysis," *Journal of Low Power Electronics and Applications*, vol. 8, no. 3, p. 25, 2018.
[21] K. Lee *et al.*, "Low cost ternary content addressable memory based on early termination precharge scheme," in *ISCAS*, pp. 1–4, IEEE, 2019.
[22] W. Choi *et al.*, "Low cost ternary content addressable memory using adaptive matchline discharging scheme," in *ISCAS*, pp. 1–4, IEEE, 2018.
[23] S. Kumar *et al.*, "Design of a two-step low-power and high-speed cmos flash adc architecture," in *VDAT*, pp. 1–6, IEEE, 2020.
[24] K. Ni *et al.*, "Write disturb in ferroelectric fets and its implication for 1t-fefet and memory arrays," *IEEE EDL*, vol. 39, no. 11, pp. 1656–1659, 2018.
[25] R. Vattikonda *et al.*, "Modeling and minimization of pmos nbti effect for robust nanometer design," in *IEEE DAC*, pp. 1047–1052, 2006.
[26] M. Poremba *et al.*, "Destiny: A tool for modeling emerging 3d nvm and edram caches," in *DATE*, pp. 1543–1546, EDA Consortium, 2015.
[27] T. Soliman *et al.*, "Ultra-low power flexible precision fefet based analog in-memory computing," in *IEDM*, pp. 29–2, IEEE, 2020.
[28] F. Baig *et al.*, "Sparkgis: Resource aware efficient in-memory spatial query processing," in *ACM SIGSPATIAL GIS*, pp. 1–10, 2017.
[29] M. Imani *et al.*, "Nvquery: Efficient query processing in nonvolatile memory," *IEEE TCAD*, vol. 38, no. 4, pp. 628–639, 2018.