

SEE-MCAM: Scalable Multi-bit FeFET Content Addressable Memories for Energy Efficient Associative Search

Shengxi Shou^{1,3}, Che-Kai Liu², Sanggeon Yun³, Zishen Wan², Kai Ni⁴, Mohsen Imani³, X. Sharon Hu⁵,
Jianyi Yang⁶, Cheng Zhuo^{6*}, Xunzhao Yin^{1,7*}

¹College of Information Science and Electronic Engineering, Zhejiang University, China

²School of Electrical and Computer Engineering, Georgia Institute of Technology, USA

³Department of Information and Computer Science, University of California Irvine, USA

⁴Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, USA

⁵Department of Computer Science and Engineering, University of Notre Dame, USA

⁶School of Micro-Nano Electronics, Zhejiang University, China

⁷Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province, China

*Corresponding authors, email: {czhuo, xzyin1}@zju.edu.cn

Abstract—Artificial intelligence has made remarkable advancements in recent years, leading to the development of algorithms and models capable of handling ever-increasing amounts of data. The computational demands of these algorithms necessitate circuit and architecture designs that go beyond the von-Neumann paradigm. Content addressable memories (CAMs), which implement parallel associative search functionality within memory blocks to overcome the memory wall bottleneck, have proven to be effective for data-intensive tasks. While current CAM designs have achieved higher storage density and energy efficiency than their CMOS-based counterparts by leveraging emerging non-volatile memories (NVM), most of these implementations are limited to binary storage cells. In this work, we propose SEE-MCAM, scalable and compact *multi-bit CAM (MCAM)* designs that utilize the three-terminal ferroelectric FET (FeFET) as the proxy. By exploiting the multi-level-cell characteristics of FeFETs, our proposed SEE-MCAM designs enable multi-bit associative search functions and achieve better energy efficiency and performance than existing FeFET-based CAM designs. We validated the functionality of our proposed designs by achieving 3 bits per cell CAM functionality, resulting in $3\times$ improvement in storage density. The area per bit of the proposed SEE-MCAM cell is 8% of the conventional CMOS CAM. We thoroughly investigated the scalability and robustness of the proposed design. Evaluation results suggest that the proposed 2FeFET-1T SEE-MCAM achieves $9.8\times$ more energy efficiency and $1.6\times$ less search latency compared to the CMOS CAM, respectively. When compared to existing MCAM designs, the proposed SEE-MCAM can achieve $8.7\times$ and $4.9\times$ more energy efficiency than ReRAM-based and FeFET-based MCAMs, respectively. Benchmarking results show that our approach provides up to 3 orders of magnitude improvement in speedup and energy efficiency over a GPU implementation in accelerating a novel quantized hyperdimensional computing (HDC) application.

I. INTRODUCTION

In the era of artificial intelligence (AI), the exponential growth of data generated by machine learning applications, edge devices, and data centers has created significant demands on the efficiency of the underlying hardware. Such hardware needs to support high-performance and data-intensive applications. One crucial operation in these algorithms is the

data query operation, which involves searching for a vector among a large number of data vectors stored in a library. This operation is integral to various machine learning and neuromorphic models, such as hyperdimensional computing (HDC) [1], [2], few-shot learning [3], reinforcement learning [4], bioinformatics [5], and robotics [6], [7].

However, traditional Von-Neumann hardware faces a challenge known as the memory wall problem when handling these data-intensive workloads. The memory wall problem arises due to the substantial data movement required between the memory and computing units, resulting in significant data transfer overhead. This overhead dominates the total cost of data query operations, ultimately limiting the overall efficiency of the system.

To address this challenge and enhance efficiency, there is a strong demand for hardware solutions that support parallel associative search (i.e., data query operations) within the memory, effectively eliminating data transfer overhead. In-memory computing (IMC) has emerged as an alternative architectural paradigm that combines computational and storage units, offering promising solutions to overcome the memory wall challenge specifically for data search operations. Content addressable memory (CAM) is a key primitive of IMC, embedding parallel search functionality within memory blocks and enabling fast associative search. CAM has gained significant adoption as an associative memory (AM) to accelerate the inference phase of novel machine learning tasks mentioned earlier.

Conventional CMOS-based 16T CAM arrays [8] suffer from drawbacks such as high leakage and area overhead due to the energy-consuming 6T static random access memory (SRAM) design. Recent research focuses on leveraging emerging non-volatile memories (NVMs) such as ferroelectric field-effect transistors (FeFETs) [9], [10], resistive random access memory (ReRAM) [11], [12], [13], and spin-transfer torque magnetic RAM (STT-MRAM) [14] to develop more compact and efficient CAM designs. By storing binary logic values

inside NVM devices and performing bit-wise XOR logic operations between the query and the stored data, NVM-based binary CAM (BCAM) and ternary CAM (TCAM) designs have proven to be more compact and energy-efficient than their CMOS counterparts. These NVM-based CAM designs have been extensively studied under various data-intensive workloads [1], [3], [4], [5].

However, these NVM-based CAM designs have been limited to exploiting the single-level cell (SLC) property of NVM, hindering further improvements in CAM density. To overcome this limitation and enhance CAM density, leveraging the *multi-level cell (MLC)* property of NVMs for *multi-bit CAMs (MCAMs)* has become an appealing direction of research. Some approaches include a 6T-2R MCAM utilizing the MLC property of ReRAM devices, but it requires additional transistors and exhibits energy consumption due to analog inverters and current-based sensing [15], [16]. FeCAM in [17], achieves MCAM functionality using only two FeFET devices, but introduces high precharge energy associated with the CAM matchline (ML). A 2FeFET-1T CAM design [18] eliminates analog inverters but requires separate sensing circuitry for NOR-type and NAND-type ML structures, and is vulnerable to device variations. Additionally, a distance function based on FeFET conductance within the MCAM array [19] faces challenges related to FeFET variations.

To fully exploit the potential of NVM-based CAMs for accelerating data-intensive workloads in-memory, it is crucial to design CAMs that effectively address the aforementioned drawbacks of existing works. Such a CAM design could lead to significantly higher CAM density and improved performance and energy efficiency, while incurring minimal penalties in terms of robustness.

In this work, we propose SEE-MCAM, which leverages FeFETs as proxy NVMs, to design scalable NOR-type and NAND-type MCAMs for energy-efficient associative search. Using an experimentally calibrated FeFET Preisach model [9], we incorporate a 2FeFET multi-bit input binary output (MIBO) XOR logic structure into our proposed CAM cells, allowing for storage of multi-bit values and implementation of the Boolean XOR logic operation. Such structure effectively controls the access transistors associated with the CAM array matchline (ML), reducing or eliminating the precharge energy typically associated with the ML and ensuring robustness against device variations. Additionally, our proposed CAM array can be programmed by employing write inhibition schemes and directly applying multi-bit search operations to the FeFET source/drain [20], [21]. Building upon the 2FeFET MIBO XOR logic structure, we propose NOR-type 2FeFET-1T and NAND-type 2FeFET-2T SEE-MCAM designs that support multi-bit and parallel search functions. The NOR-type SEE-MCAM reduces ML-associated capacitance, resulting in energy savings, while the NAND-type SEE-MCAM eliminates the precharge phase.

We discuss and evaluate the structures, operations, simulation validation, and energy/performance analysis of the proposed SEE-MCAM designs at the array level. To demon-

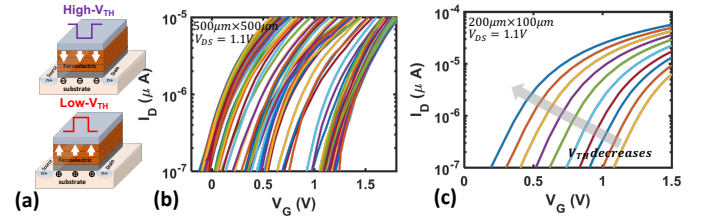


Fig. 1. (a) FeFET write schemes. (b) $I_D - V_G$ characteristics with MLC V_{TH} states of a fabricated FeFET. (c) Simulated $I_D - V_G$ characteristics with more than 3-bit V_{TH} states based on Preisach FeFET model.

strate the advantages of utilizing FeFETs' MLC property in conjunction with SEE-MCAM design schemes, we compare our designs with existing MCAM designs based on ReRAM and FeFET, respectively. Additionally, we investigate the scalability and robustness of SEE-MCAM arrays against device variations. We have demonstrated that SEE-MCAM achieves up to 3 bits per cell CAM density. With a simpler cell structure and sensing circuitry, the SEE-MCAM array significantly improves energy efficiency compared to prior works [15], [18], [22], [23], [10]. Specifically, SEE-MCAM demonstrates an area per bit efficiency of 8% of CMOS CAM and achieves $8.7\times$ and $4.9\times$ higher energy efficiency than ReRAM and FeFET-based MCAMs, respectively, while maintaining sufficient robustness against device variations. Furthermore, benchmarking results of a novel quantized HDC inference task using the SEE-MCAM array indicate a potential improvement of up to 3 orders of magnitude compared to conventional GPU-based approaches.

II. BACKGROUND

A. FeFET Basics

FeFETs based on HfO_2 [24] have emerged as highly competitive candidates due to their intrinsic CMOS structure, high I_{ON}/I_{OFF} ratio, low OFF current, excellent scalability, CMOS compatibility, and superior write energy efficiency. FeFETs are fabricated by integrating a ferroelectric layer in the gate stack of a metal-oxide-semiconductor field-effect transistor (MOSFET), where HfO_2 serves as the ferroelectric material (as shown in Figure 1(a)). Moreover, Figure 1(a) illustrates a FeFET that can store high- V_{TH} and low- V_{TH} states, respectively. By applying a positive (negative) voltage pulse to the gate terminal, the polarization of the ferroelectric layer will be switched towards the channel, programming the FeFET into the low- V_{TH} (high- V_{TH}) state. Figure 1(b) and (c) show the $I_D - V_G$ characteristics of the fabricated and simulated FeFET devices with different write pulses, respectively. FeFETs have been successfully deployed in various scenarios, including CAMs [23], frequency multipliers [25], crossbars [26], field-programmable gate arrays (FPGAs) [27], and oscillators [28], among others.

B. Existing Binary CAMs and Ternary CAMs

Various CAM designs have been proposed based on canonical CMOS and different NVM devices. Depending on the

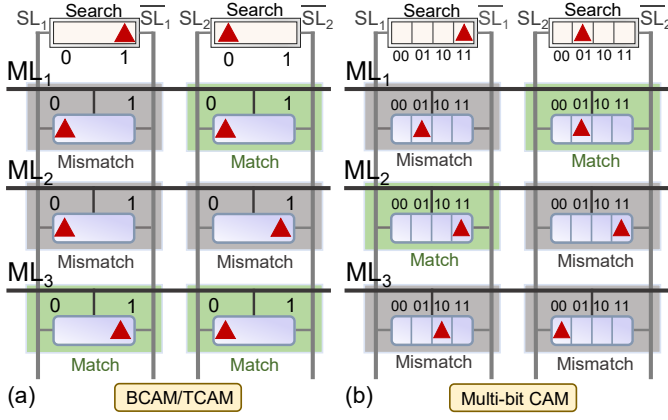


Fig. 2. (a) In a single bit CAM, a '0' or '1' is stored and searched in parallel, while (b) in a MCAM, multiple level values can be stored and searched in parallel.

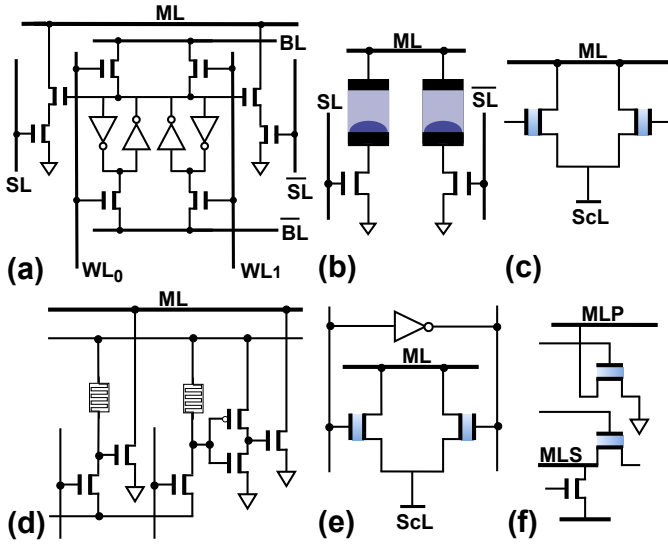


Fig. 3. Existing BCAM/TCAM designs based on (a) 16T CMOS [8], (b) 2T-2R [13], and (c) 2FeFET [10]. Existing MCAM designs based on (d) 6T-2ReRAM [15], (e) 2FeFET [17], and (f) 2FeFET-1T [18].

type of data stored in a CAM cell, CAMs can be categorized into binary CAM (BCAM), ternary CAM (TCAM), multi-bit CAM (MCAM), and analog CAM (ACAM) [29]. Most of the existing CAM designs are BCAM, with binary values stored in cell, implementing bit-wise XOR logic. TCAM can store an additional "don't care" bit besides binary values, which serves as a wildcard. Figure 2 illustrates the BCAM and TCAM storing a single bit. Figure 3 summarizes some representative BCAM and TCAM designs. Figure 3(a) presents the traditional 16T CMOS-based CAM, which consumes significant energy and area overhead. ReRAM-based CAMs such as the 2T-2R TCAM [13] (Figure 3(b)) and 3T-1R CAM [12] have been proposed and fabricated for memory-intensive tasks [30]. Though these design consume much less area overhead than CMOS, the low HRS/LRS (high/low resistance state) ratio, current driven write mechanisms and two-terminal NVM struc-

ture of ReRAM devices necessitate extra selectors and write facilitation circuitry, thus resulting in high energy consumption [10]. Recently, FeFET emerges as a promising device due to its high I_{ON}/I_{OFF} ratio, low I_{OFF} and three-terminal structure. A number of FeFET-based CAMs have been proposed for energy-efficient data-intensive computing tasks [31], [10], [23], [32], [33]. Figure 3(c) depicts a typical 2FeFET CAM cell [34]. That said, these FeFET designs only exploit the binary NVM property of FeFETs. As shown in Figure 1, the potential of FeFETs remains unexplored.

C. MCAM Concepts and Related Works

Above NVM-based BCAMs and TCAMs are limited to exploiting the SLC characteristic of NVMs, thus hindering from further CAM density improvement. Recent works explore the possibilities of exploiting the MLC properties of NVMs to construct MCAM designs to boost the CAM density. Unlike the single-bit CAM design shown in Figure 2(a), MCAMs store multi-bit values, and a multi-bit input query is applied for a search. Only when all multi-bit values in the query are identical to a stored entry, a match can be detected. [15] and [16] presented a 6T-2R MCAM (Figure 3(d)) that utilizes the MLC property of ReRAM devices, albeit at the cost of four additional transistors compared to the conventional 2T-2R TCAM design [13]. Moreover, their design incorporates an analog inverter and a current-based sensing mechanism, resulting in significant energy consumption. [17] proposed FeCAM, which employs only two FeFET devices to achieve MCAM functionality, shown in Figure 3(d). However, FeCAM associates the two FeFETs' drain capacitance with the CAM matchline (ML), introducing high precharge energy. Another approach, the 2FeFET-1T CAM introduced in [18], eliminates the need for analog inverters by dividing the two FeFETs into separate NOR-type and NAND-type ML branches, shown in Figure 3(f). Unfortunately, the two ML branches require different sensing circuitry, and the NAND-type branch is vulnerable to device variations. [19] and [35] propose to realize a novel distance function within an MCAM cell and the cell structure is the same as that of 2FeFET TCAM in Figure 3(c). However, such a distance function relies on accurate FeFET conductance in the linear and saturation region, which makes it vulnerable to FeFET variations. Our proposed SEE-MCAM designs aim to exploit the potential of FeFET devices, achieving higher CAM density, improved performance and energy efficiency than the above works, while maintaining the robustness.

III. PROPOSED SEE-MCAM DESIGNS

In this section, we propose two SEE-MCAM designs with multi-bit functionality and improved energy efficiency by either (i) reducing the NOR-type ML capacitance, or (ii) eliminating the precharge phase in NAND-type ML. We first introduce the 2FeFET structure implementing the key MIBO XOR logic, and then discuss the SEE-MCAM designs.

A. 2FeFET Structure for Multi-Bit Input Binary Output

Figure 4 shows the schematic and an example 2-bit operation principle of the 2FeFET structure to implement MIBO

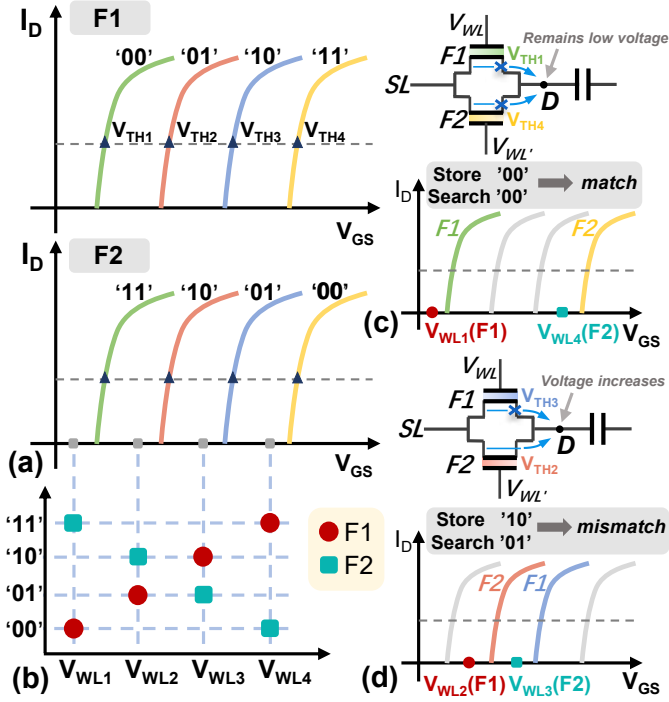


Fig. 4. A 2bit case of 2FeFET MIBO: (a) F1 and F2 are written to store 4 levels of threshold voltage states, respectively. (b) 4 sets of V_{WL} applied at the gates of F1 and F2 correspond to search 4 different values, respectively. (c) A match case where the stored value is '00', and the input value is '00'. The binary output D is at low. (d) A mismatch case where the stored value is '10', and the input value is '01'. The binary output D is at high.

XOR logic, which is the key function in a CAM cell. The 2FeFETs are connected in parallel, forming a push-pull structure. By programming multiple threshold voltages, i.e., V_{TH1} , V_{TH2} , V_{TH3} , V_{TH4} , into the 2FeFETs F1 and F2 to encode the stored multi-bit value as shown in Figure 2(a), and then applying pre-defined search voltages, i.e., V_{WL1} , V_{WL2} , V_{WL3} , V_{WL4} , corresponding to the search query values shown in Figure 4(b), a binary XOR output (match or mismatch) between the stored value and the applied query input can be generated. During the operation, The sourceline (SL) is pulled up to high level.

Figure 4(c) and (d) demonstrate a match and a mismatch cases, respectively. Consider a 2FeFET cell storing '00' as shown in Figure 4(c), F1 stores V_{TH1} and F2 stores V_{TH4} . Then an input value '00' is searched by applying corresponding V_{WL} set per Figure 4(b), i.e., applying $V_{WL1} + V_{SL}$ to the gate of F1, and $V_{WL4} + V_{SL}$ to the gate of F2, respectively. In this way, since both FeFETs are turned off, the output node D is not charged by SL , and remains at low level, indicating a match case. Similarly, Figure 4(d) demonstrates a 2FeFET cell storing '10', where F1 stores V_{TH3} and F2 stores V_{TH2} . An input query '01' is searched by applying $V_{WL2} + V_{SL}$ to F1 and $V_{WL3} + V_{SL}$ to F2, respectively. In this case, F1 is turned off but F2 turns on. The output node D will then be charged to high level, indicating a mismatch case. It can be seen that the 2FeFET structure implements a MIBO XOR logic, only

TABLE I
OPERATION SUMMARY OF A 3-BIT SEE-MCAM.

| D ML | Stored '000' | Stored '001' | Stored '010' | Stored '011' | Stored '100' | Stored '101' | Stored '110' | Stored '111' |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Search '000' | L M | H MM | H MM | H MM | H MM | H MM | H MM | H MM |
| Search '001' | H MM | L M | H MM | H MM | H MM | H MM | H MM | H MM |
| Search '010' | H MM | H MM | L M | H MM | H MM | H MM | H MM | H MM |
| Search '011' | H MM | H MM | H MM | L M | H MM | H MM | H MM | H MM |
| Search '100' | H MM | H MM | H MM | H MM | L M | H MM | H MM | H MM |
| Search '101' | H MM | H MM | H MM | H MM | H MM | L M | H MM | H MM |
| Search '110' | H MM | H MM | H MM | H MM | H MM | H MM | L M | H MM |
| Search '111' | H MM | H MM | H MM | H MM | H MM | H MM | H MM | L M |

For a given stored and input value pair. H/L indicates the high/low voltage level of the node D . M/MM indicates a match or mismatch state of ML.

when the input query is identical to the stored value, the output is at low and high otherwise. This MIBO principle is scalable depending on the number of distinguishable states a FeFET can store, and 3-bit per cell CAM function will be discussed in this section. This 2FeFET structure forms the basic of our proposed SEE-MCAM designs.

B. 2FeFET-1T SEE-MCAM

By embedding the 2FeFET structure in Sec. III-A into the CAM cell as shown in Figure 5(a), the proposed NOR type 2FeFET-1T SEE-MCAM is realized. Figure 5(a) shows the schematic of 2FeFET-1T SEE-MCAM design. A NMOS access transistor separates the 2FeFET structure from ML. The wordlines $WL1$ and $WL2$ shared by each column control the gates of 2FeFET structure, and the SL shared by a row connects the sources of the 2FeFET structures within the word. The MIBO node D is connected to the gate of NMOS. The sense amplifier (SA) of the array adopts a threshold inverter quantization (TIQ) comparator [32].

During the write, the SL s associated with unselected words are applied with write inhibition scheme [20], [21]. The SL s of selected words as well as the wordlines of selected columns are applied with write pulses to program the FeFETs into desired threshold voltage state per Figure 1. Table I summarizes the search operations of the 2FeFET-1T SEE-MCAM cell. The MLs of the array are precharged to a high level. Due to the NOR type ML connection, only when all input values are identical to the stored values of the cells within a word, the corresponding word ML can remain at high, otherwise drops down to a low level. Figure 5(b) shows the transient waveforms that validate the 3-bit MCAM function of NOR type SEE-MCAM.

Compared with prior FeFET-based CAM designs [22], [23], [18], [17], our proposed 2FeFET-1T SEE-MCAM ad-

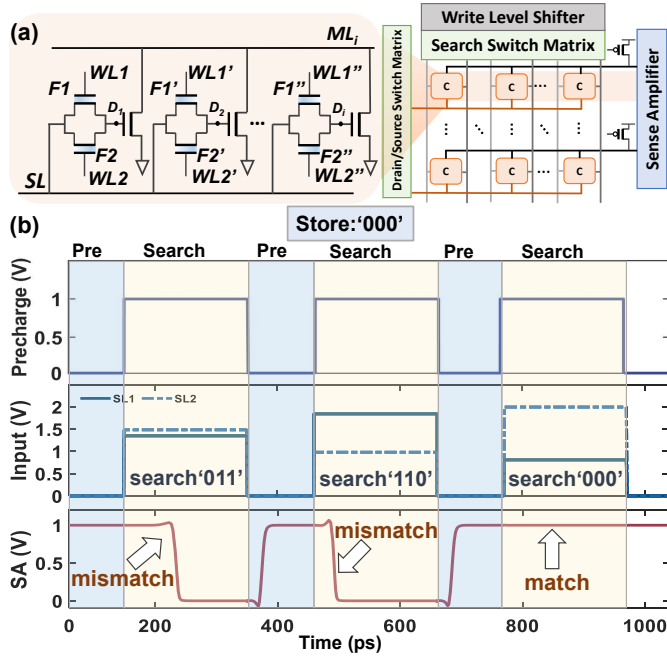


Fig. 5. The proposed NOR type SEE-MCAM design. (a) Schematic of the 2FeFET-1T SEE-MCAM array. (b) Transient waveform of a 3-bit 2FeFET-1T SEE-MCAM.

addresses their respective drawbacks, thus achieving the multi-bit search operation with significant energy efficiency. The BCAM/TCAM design from [22], [23] is limited to single-bit CAM functionality, and requires two complementary SL s associated with the CAM cell to perform the search. Our proposed SEE-MCAM, on the other hand, increases the data density without any additional transistors and peripherals, and only requires one supply rail SL to perform the search operation. The MCAM designs from [17] consumes significant precharge energy as it associates two FeFETs with ML, resulting in a large ML capacitance to be precharged:

$$C_{ML} \approx C_{d,P} + N \times (2C_{FeFET} + C_{parasitic}) \quad (1)$$

where $C_{d,P}$, $C_{parasitic}$ and C_{FeFET} are the drain capacitance of the precharge PMOS, the parasitic capacitance associated with ML, and the drain capacitance of the FeFET, respectively. N is the number of cells within a word. The 2FeFET-1T MCAM from [18] consumes the same device count as the proposed design, but the two ML branches of the MCAM design incur both high precharge energy and latency. Our proposed 2FeFET-1T SEE-MCAM array reduces the transistor associated with ML to only 1, therefore consumes less precharge energy while maintaining the latency:

$$C_{ML} \approx C_{d,P} + N \times (C_{NMOS} + C_{parasitic}) \quad (2)$$

C. Precharge-Free 2FeFET-2T SEE-MCAM

To further improve the energy efficiency of SEE-MCAM, we propose a NAND type 2FeFET-2T SEE-MCAM array that embeds the 2FeFET MIBO structure and leverages a precharge-free scheme [23] to eliminate the energy consuming

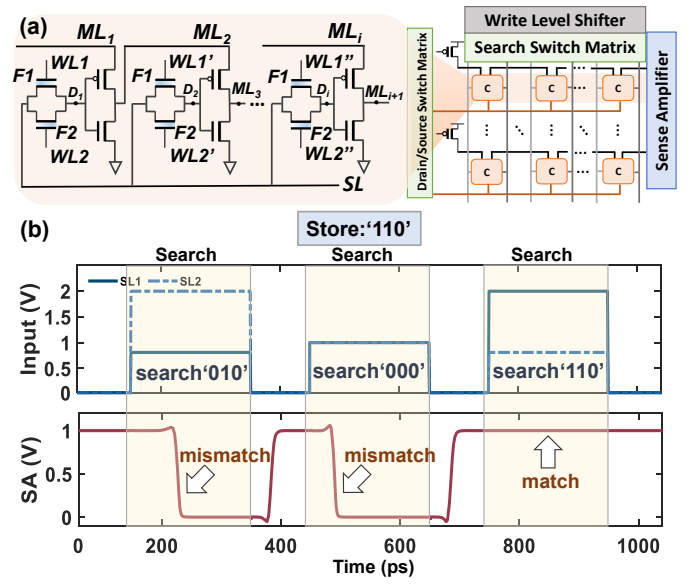


Fig. 6. Precharge-free multibit CAM. (a) Schematic of 2FeFET-2T precharge-free MCAM. (b) Transient waveforms of the 3-bit 2FeFET-2T precharge-free MCAM.

precharge phase. Figure 6(a) shows the schematic of the proposed 2FeFET-2T SEE-MCAM cell and array, respectively. The cell consists of the 2FeFET MIBO structure and an inverter. The array ML adopts the NAND type connection, where the ML of the previous cell is the supply rail of the inverter in the current cell. The wordlines $WL1$ and $WL2$ are shared by each column, and the sourceline SL is shared by a row facilitate the operation of the 2FeFET MIBO.

The write scheme and the inhibition scheme of the 2FeFET-2T SEE-MCAM design are similar to the 2FeFET-1T SEE-MCAM. During the search operation, the ML state of the current cell is formulated as below:

$$ML_i = ML_{i-1} \times \bar{D} \quad (3)$$

Figure 6(b) shows the transient waveforms of the proposed 2FeFET-2T SEE-MCAM with 3bit density, validating the MCAM functionality.

The 2FeFET-2T SEE-MCAM design eliminates the need for precharge in most cases since the ML state ML_i is determined by both the output of 2FeFET MIBO structure D and the previous cell's ML voltage ML_{i-1} . In consecutive searches, the ML state of the previous cell ML_{i-1} only changes when a mismatch case in the last search discharges ML_{i-1} , and then a match in the current search operation charges ML_{i-1} again. If consecutive searches yield the same match/mismatch state of the previous cell, ML_{i-1} remains unchanged. Precharging of the current cell C_i occurs only when two conditions are met: (i) the supply rail of C_i , ML_{i-1} , transitions from a mismatch state (Low) to a match state (High) during a search, activating the PMOS of C_i to charge ML_i , and (ii) the MLs of all previous $i - 1$ cells are at a match state, enabling a charging path from the voltage supply to ML_i . These strict conditions significantly reduce the chances of charging, resulting in much

lower energy consumption. However, it is important to note that the NAND type ML connection introduces higher latency as the sense amplifier needs to wait for the ML state transition to propagate through the entire word.

IV. EVALUATION & BENCHMARKING

In this section, we evaluate and compare the search energy per bit and delay of the proposed SEE-MCAM arrays with existing BCAM/TCAM and MCAM designs to validate the benefits of exploiting MLC FeFET devices and energy efficient design schemes. We then benchmark the proposed design in the context of a quantized HDC model at application level.

A. SEE-MCAM Evaluations

Our proposed SEE-MCAM designs utilize the 45nm Preisach FeFET model [9]. Different threshold voltage states corresponding to multi-bit values are written to FeFETs via different write pulses [17], [18]. The 40nm UMC processing development kit (PDK) is adopted for all CMOS transistors. All the circuits have been simulated with Cadence. Wiring parasitics associated with MLs are extracted from DESTINY [36].

Figure 7 and Figure 8 show the search energy and latency of the proposed SEE-MCAM arrays with varying number of rows and number of cells per row. The latency is reported under the worst case, i.e. one mismatch. It can be seen from Figure 7(a) and Figure 8(a) that the respective search energy of both 2FeFET-1T and 2FeFET-1T SEE-MCAM arrays increase linearly as the number of rows increases. As the rows of both arrays are independent of each other, the search latency only slightly changes with the number of rows. On the other hand, it can be seen from Figure 7(b) and Figure 8(b) that, as the number of cells increases, the associated ML capacitance increases, slowing down the precharge/discharge and signal propagation speed, and resulting in an increase of search latency. The increasing number of cells also leads to larger precharge energy associated with ML capacitance, as well as larger load energy associated with SLs.

Table II summarizes and compares different NVM based CAM designs with our proposed SEE-MCAM designs in terms of search energy per bit, latency and area overhead per bit. The cell sizes are estimated based on a 2X2 SEE-MCAM array layout. The area per bit of our proposed 2FeFET-1T SEE-MCAM is 8% of the conventional 16T CMOS CAM. Moreover, the proposed 2FeFET-1T SEE-MCAM achieves $9.8\times$ more energy efficiency and $1.6\times$ less search latency than CMOS CAM, respectively. Regarding the NVM-based counterparts, on one hand, multi-bit CAM functionality leads to much higher energy efficiency compared to the BCAM and TCAM designs. For example, our proposed 2FeFET-1T SEE-MCAM design is $6.7\times$ more energy efficient than the typical 2FeFET TCAM design [10], respectively. On the other hand, the ML capacitance reduction and precharge-free design schemes applied to the SEE-MCAM designs also brings significant energy saving and speedup when compared to other MCAM designs. Our approach can achieve $8.7\times$ and $4.9\times$

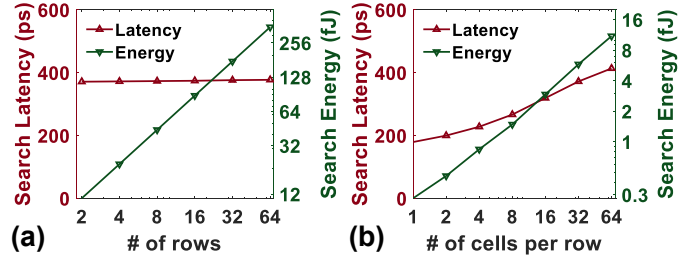


Fig. 7. Search latency and energy of 2FeFET-1T SEE-MCAM (a) with varying number of rows; (b) with varying number of cells per row.

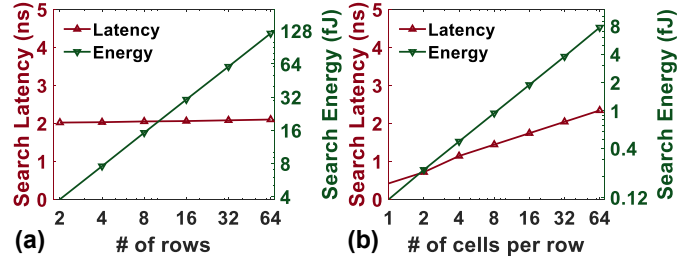


Fig. 8. Search latency and energy of 2FeFET-2T SEE-MCAM (a) with varying number of rows; (b) with varying number of cells per row.

more energy efficiency than ReRAM based [15] and FeFET-based [18] MCAM designs, respectively. Though ReRAM-based MCAM consumes less search latency, this is due to the high sensing current of 6T-2R cell structure. Overall, these evaluation results validate the efficiency of our SEE-MCAM approaches for associative search applications.

We also validate the robustness of our proposed SEE-MCAM design and the ML scheme. The variations of all the CMOS transistors, including the sense amplifier, are modeled using the PDK with TT process corner at a temperature of 27°C. The variations of FeFET devices are obtained from experimentally measured devices, with a standard deviation $\sigma = 54mV$ for the low/high V_{TH} state [37]. Additionally, smaller FeFET variation can be achieved through the use of a write-and-read verify scheme [29]. Figure 9 depicts the transient waveforms of the SA output including process variations during the search operations in the proposed SEE-MCAM array. The transient results of 100 Monte Carlo simulations shown in Figure 9 demonstrate sufficient sense margin of our proposed design even in the worst search case, confirming the

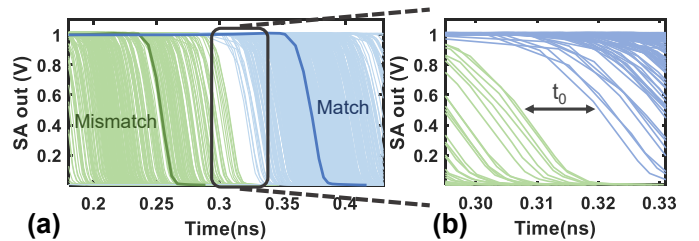


Fig. 9. Transient waveforms of the proposed SEE-MCAM array with device variability under the worst case.

TABLE II
COMPARISONS OF CAM DESIGNS

| Designs | Device | Cell Structure | Type | Search energy per bit (fJ) | Latency (ps) | Area per bit (μm^2) | NVM/MOS node (nm) |
|--------------------|--------|------------------------|-----------|---------------------------------------|------------------------|--|-----------------------|
| 16T CMOS [8] | CMOS | 16T | BCAM | 0.59 ($\times 9.8$) | 582.4 ($\times 1.6$) | 1.12 ($\times 9.3$) | -/45 |
| DAC'22 [32] | FeFET | 2T-1FeFET | BCAM | 0.116 ($\times 1.9$) | 401.4 ($\times 1.1$) | 0.36 ($\times 3$) | 45/45 |
| Nat Ele'19 [10] | FeFET | 2FeFET | TCAM | 0.40 ($\times 6.7$) | 360 ($\times 1$) | 0.15 ($\times 1.2$) | 45/- |
| DATE'21 (P#) [22] | FeFET | 2FeFET-1T | TCAM | 0.195 ($\times 3.3$) | 252.8 ($\times 0.7$) | 0.36 ($\times 3$) | 45/45 |
| DATE'21 (PF#) [22] | FeFET | 2FeFET-2T | TCAM | 0.073 ($\times 1.2$) | 1430 ($\times 3.8$) | 0.44 ($\times 3.6$) | 45/45 |
| JSSC'13 [13] | PCM | 2T-2R | TCAM | 0.55 ($\times 9.2$) | 350.6 ($\times 0.9$) | 0.41 ($\times 3.4$) | 90/90 |
| NC'20 [15] | ReRAM | 6T-2R | ACAM | 0.52 ($\times 8.7$) | 110 ($\times 0.3$) | 0.51 [¶] ($\times 4.2$) | 50/180 |
| TED'20 [17] | FeFET | 2FeFET | MCAM/ACAM | 0.182/0.069 ($\times 3/\times 1.2$) | - | 0.05 ($\times 0.4$) | 45/45 |
| IEDM'20 [18] | FeFET | 2FeFET-1T [‡] | MCAM | 0.292 ($\times 4.9$) | 422 ($\times 1.1$) | 0.03 ^{&} ($\times 0.2$) | 28/- |
| This work (P) | FeFET | 2FeFET-1T | MCAM | 0.06 [‡] ($\times 1$) | 371.8 ($\times 1$) | 0.12 ($\times 1$) | 45/40 |
| This work (PF) | FeFET | 2FeFET-2T | MCAM | 0.039 [‡] | 2040 | 0.146 | 45/40 |

†: Results are evaluated under 32 cells per word. ‡: Two ML branches for one stored vector. #: Precharge and precharge-free. ¶: Reported based on 16nm design rules. &: Smaller cell size due to 28nm technology node.

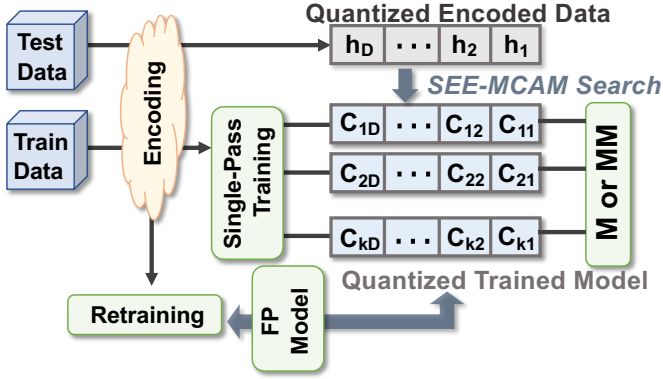


Fig. 10. Overview of the quantized HDC framework leveraging the proposed SEE-MCAM design.

robustness of our proposed design against device variations.

B. Benchmarking: Quantized Hyperdimensional Computing

We further benchmark our proposed SEE-MCAM as an associative memory in the context of a novel hyperdimensional computing (HDC) framework. Inspired by the fact that the brain computes based on patterns that are not readily related to numbers, HDC is built upon a set of transparent operations and is extremely robust to hardware noise due to the *holographic* and *redundant* nature. It has been proposed as an alternative computing paradigm for resource-constrained edge scenarios.

Previous IMC designs [26] are benchmarked with a full precision HDC model [38]. However, due to the limited precision of IMC designs, assuming full precision for benchmarking may lead to less accurate results. In this work, we employ our approach in a calibrated quantized HDC model to conduct a more practical benchmarking. The framework is implemented in Python with Pytorch packages and supports both full precision and quantized HDC inference. Figure 10 shows the overall quantized HDC framework. Non-linear quantization is performed on the encoded query and the class hypervectors that are stored in the design-based associative memory. The multi-bit exact match scheme of SEE-MCAM is adopted. Here, we first discuss the basics of the quantized HDC.

TABLE III
DATASETS (n : FEATURE SIZE, K : NUMBER OF CLASSES)

| Dataset | n | K | Train Size | Test Size | Description |
|---------|-----|-----|------------|-----------|-----------------------------------|
| ISOLET | 617 | 26 | 6,238 | 1,559 | Voice Recognition [39] |
| UCI HAR | 561 | 12 | 6,213 | 1,554 | Physical Activity Monitoring [40] |
| PAMAP | 75 | 5 | 611,142 | 101,582 | Human Activity Recognition [41] |

Encoding: HDC encoding refers to mapping the feature from low dimensional space $\mathcal{F} \subset \mathbb{R}^d$ into high dimensional space $\mathcal{H} \subset \mathbb{R}^D$ where dimensionality $D \gg d$. For instance, a vector $\vec{F} = [f_1, \dots, f_n]$ with n features is multiplied with an $n \times D$ matrix \vec{B} , where every element in \vec{B} is sampled from i.i.d Gaussian distribution with $\mu = 0$ and $\sigma = 1$.

Training and Retraining: The lightweight training of HDC often refers to single-pass training, which is amenable to edge devices. Hypervector \vec{H}_l associated to the label l is generated after the encoding phase. For single-pass training in Figure 10, all the \vec{H}_l are aggregated (k \vec{H}_l s is presented here): $\vec{C}_l = \sum_k \vec{H}_l$. For iterative training, HDC trains the data as follows:

$$\begin{aligned} \vec{C}_l &\leftarrow \vec{C}_l + \eta(1 - \delta)\vec{Q} \\ \vec{C}_{l'} &\leftarrow \vec{C}_{l'} - \eta(1 - \delta)\vec{Q}. \end{aligned} \quad (4)$$

Where η is the learning rate set to 0.03 in this work, l' is the mispredicted label, and l is the correct label.

Inference: After the hypervectors are generated via the encoding phase and trained based on Equation 4, they are then stored in the SEE-MCAM array for inference. When a new query \vec{Q} comes in for classification, it will first be encoded in the encoding phase, and then searched in the associative memory by applying multi-bit voltages corresponding to the element values to the SEE-MCAM array.

In this work, we adopted the quantized HDC model where there exists a full-precision model for training and a quantized model stored in the SEE-MCAM for inference. We quantize each element of a hypervector to the desired bit precision based on its Z-score ($Z = \frac{x - \mu}{\sigma}$) over the Gaussian distribution. Take the proposed 3-bit SEE-MCAM as an example, element values that drop beneath 12.5% of the cumulative distribution function (CDF) will be assigned to '000'.

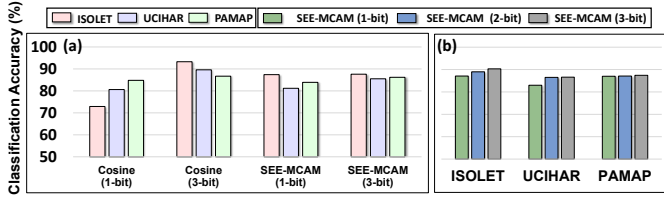


Fig. 11. Quantized HDC benchmarking accuracy with (a) binary cosine similarity, 3-bit cosine similarity, binary SEE-MCAM, and 3-bit SEE-MCAM under $D = 1024$, respectively; (b) SEE-MCAM implementation supporting varying dimensionality ($D = 1024$, $D = 2048$, and $D = 4096$, respectively).

1) *Classification Accuracy*: Figure 11(a) illustrates the quantized HDC accuracy on three different datasets, whose descriptions are summarized in Table III. As HDC typically exploits cosine distance between hypervectors as the optimal similarity function during the inference, we hereby compare the our SEE-MCAM based quantized HDC implementation with the quantized HDC framework based on cosine distance. Both implementations quantize the elements of hypervectors after training to 3 bits, respectively, using the aforementioned non-linear quantization scheme. Inference accuracy results shown in Figure 11 indicate that the proposed 3-bit SEE-MCAM based implementation has on average 3.43% accuracy degradation compared to the 3-bit cosine similarity-based implementation in GPU. To make a fair comparison, We also implement COSIME, a binary cosine similarity-based associative memory [26], in our quantized HDC framework, and compare it with our proposed SEE-MCAM based implementation. The result shown in Figure 11(a) indicates that the binary SEE-MCAM based implementation achieves on average 2.26% accuracy improvement over COSIME-based implementation.

Moreover, as the proposed SEE-MCAM increases the data density compared to BCAM/TCAM, the SEE-MCAM based HDC implementation can actually implement higher dimensionality without extra hardware cost. With the same amount of CAM cells employed (e.g. 1024) in the HDC framework, SEE-MCAM can represent and store much more elements per hypervector (e.g. $D = 2048$ for 2-bit SEE-MCAM and $D = 4096$ for 3-bit SEE-MCAM). For HDC, increasing the dimensionality of the hypervector leads to higher algorithmic accuracy. As a result, the 3-bit SEE-MCAM achieves on average 2.41% accuracy improvement over the binary SEE-MCAM as shown in Figure 11(b), showcasing the superiority of the proposed SEE-MCAM over existing BCAM/TCAM.

2) *Hardware Acceleration*: We further investigate the speedup and energy efficiency improvement of the SEE-MCAM based quantized HDC implementation over the Nvidia Pascal microarchitecture GTX 1080ti GPU. Nvidia System Management Interface is used for accurate power consumption measurement, and Pytorch profiler is used for algorithmic delay breakdown. The profiler delay for exact matching is extracted from Pytorch Aten's API.

We compare the proposed SEE-MCAM with various CAM

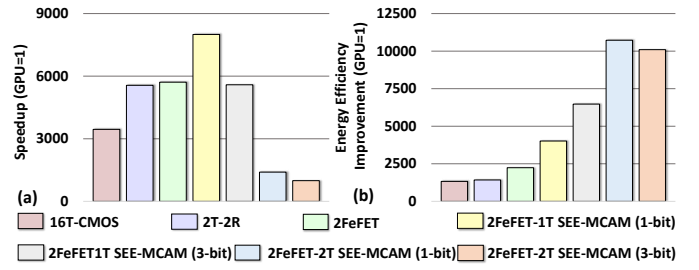


Fig. 12. (a) Computational speedup and (b) energy efficiency improvement of SEE-MCAM compared to a GPU implementation.

designs, including 16T CMOS [8], 2T-2R [13], 2FeFET [10], in the context of the quantized HDC framework. All results are respective to the same tasks running on the GPU. Both binary and 2-bit SEE-MCAM designs have been incorporated into the HDC framework and evaluated. It can be seen from Figure 12 that our proposed approach offers up to 3 orders of magnitude speedup and energy efficiency improvement than GPU. As shown in Figure 12(b), the proposed SEE-MCAM implementations significantly improve the energy efficiency than other CAM-based implementations. Both Figure 12 (a) and (b) illustrate the benefit trend of BCAM, TCAM and MCAM designs over a GPU implementation.

V. CONCLUSION

In this work, we propose SEE-MCAM, scalable energy-efficient MCAM designs that exploit FeFETs as proxy to improve the CAM density and energy efficiency over existing BCAM/TCAM and MCAM designs. We leverage the MLC property of FeFETs to construct a 2FeFET MIBO structure, which is the key part of SEE-MCAM. We then propose NOR type 2FeFET-1T and NAND type 2FeFET-2T SEE-MCAM designs to implement multi-bit CAM functionality and achieve significant energy efficiency at the same time. The functionality and robustness of the proposed approaches have been validated. Evaluation results at array level and quantized HDC application benchmarking suggest that our proposed SEE-MCAM designs achieves better data density, energy efficiency and performance when compared with other state-of-the-art CAM designs.

ACKNOWLEDGEMENTS

This work was supported in part by Zhejiang Provincial Natural Science Foundation (LD21F040003, LQ21F040006), NSFC (62104213, 92164203), National Key Research and Development Program of China (2022YFB4400300). Liu and Wan were supported by COCOSYS, one of seven centers in JUMP2.0, a SRC program sponsored by DARPA. This work was supported in part by National Science Foundation #2127780 and #2312517, Semiconductor Research Corporation (SRC), and generous gifts from Xilinx and Cisco.

REFERENCES

- [1] M. Imani, A. Rahimi, D. Kong, T. Rosing, and J. M. Rabaey, "Exploring hyperdimensional associative memory," in *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2017, pp. 445–456.
- [2] Z. Zou, H. Chen, P. Poduval, Y. Kim, M. Imani *et al.*, "Biohd: an efficient genome sequence search platform using hyperdimensional memorization," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, 2022, pp. 656–669.
- [3] M. Hersche, G. Karunaratne, G. Cherubini, L. Benini, A. Sebastian *et al.*, "Constrained few-shot class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9057–9067.
- [4] M. Li, A. Kazemi, A. F. Laguna, and X. S. Hu, "Associative memory based experience replay for deep reinforcement learning," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–9.
- [5] H. E. Barkam, S. Yun, P. Genssler, Z. Zou, C.-K. Liu *et al.*, "Hdgm: Hyperdimensional genome sequence matching on unreliable highly scaled fefet," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023.
- [6] Z. Wan, B. Yu, T. Y. Li, J. Tang, Y. Zhu *et al.*, "A survey of fpga-based robotic computing," *IEEE Circuits and Systems Magazine*, vol. 21, pp. 48–74, 2021.
- [7] Z. Shi, X. Chang, C. Yang, Z. Wu, and J. Wu, "An acoustic-based surveillance system for amateur drones detection and localization," *IEEE transactions on vehicular technology*, vol. 69, pp. 2731–2739, 2020.
- [8] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (cam) circuits and architectures: A tutorial and survey," *IEEE journal of solid-state circuits*, vol. 41, pp. 712–727, 2006.
- [9] K. Ni, M. Jerry, J. A. Smith, and S. Datta, "A circuit compatible accurate compact model for ferroelectric-fets," in *2018 IEEE symposium on VLSI technology*. IEEE, 2018, pp. 131–132.
- [10] K. Ni, X. Yin, A. F. Laguna, S. Joshi, S. Dünkel *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, pp. 521–529, 2019.
- [11] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu *et al.*, "Metal–oxide rram," *Proceedings of the IEEE*, vol. 100, pp. 1951–1970, 2012.
- [12] M.-F. Chang, C.-C. Lin, A. Lee, Y.-N. Chiang, C.-C. Kuo *et al.*, "A 3t1r nonvolatile tcam using mlc rram for frequent-off instant-on filters in iot and big-data processing," *IEEE Journal of Solid-State Circuits*, vol. 52, pp. 1664–1679, 2017.
- [13] J. Li, R. K. Montoye, M. Ishii, and L. Chang, "1mb 0.41 μm^2 2t2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing," *IEEE JSSC*, vol. 49, pp. 896–907, 2013.
- [14] E. Garzón, M. Lanuzza, A. Teman, and L. Yavits, "Am 4: Mram crossbar based cam/tcam/acam/ap for in-memory computing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 13, pp. 408–421, 2023.
- [15] C. Li, C. E. Graves, X. Sheng, D. Miller, M. Foltin *et al.*, "Analog content-addressable memories with memristors," *Nature communications*, vol. 11, p. 1638, 2020.
- [16] G. Pedretti, C. E. Graves, S. Serebryakov, R. Mao, X. Sheng *et al.*, "Tree-based machine learning performed in-memory with memristive analog cam," *Nature communications*, vol. 12, p. 5806, 2021.
- [17] X. Yin, C. Li, Q. Huang, A. F. Zhang, M. Niemier *et al.*, "Fecam: A universal compact digital and analog content addressable memory using ferroelectric," *IEEE Transactions on Electron Devices*, vol. 67, pp. 2785–2792, 2020.
- [18] C. Li, F. Müller, T. Ali, R. Olivo, M. Imani *et al.*, "A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 29–3.
- [19] A. Kazemi, M. M. Sharifi, A. F. Laguna, F. Müller, X. Yin *et al.*, "Fefet multi-bit content-addressable memories for in-memory nearest neighbor search," *IEEE Transactions on Computers*, vol. 71, pp. 2565–2576, 2021.
- [20] K. Ni, X. Li, J. A. Smith, M. Jerry, and S. Datta, "Write disturb in ferroelectric fets and its implication for 1t-fefet and memory arrays," *IEEE Electron Device Letters*, vol. 39, pp. 1656–1659, 2018.
- [21] Y. Xiao, Y. Xu, Z. Jiang, S. Deng, Z. Zhao *et al.*, "On the write schemes and efficiency of fefet 1t nor array for embedded nonvolatile memory and beyond," in *2022 International Electron Devices Meeting (IEDM)*. IEEE, 2022, pp. 13–6.
- [22] Y. Qian, Z. Fan, H. Wang, C. Li, M. Imani *et al.*, "Energy-aware designs of ferroelectric ternary content addressable memory," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021, pp. 1090–1095.
- [23] X. Yin, Y. Qian, M. Imani, K. Ni, C. Li *et al.*, "Ferroelectric ternary content addressable memories for energy efficient associative search," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2022.
- [24] A. I. Khan, A. Keshavarzi, and S. Datta, "The future of ferroelectric field-effect transistor technology," *Nature Electronics*, vol. 3, pp. 588–597, 2020.
- [25] H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, "Reconfigurable frequency multiplication with a ferroelectric transistor," *Nature Electronics*, vol. 3, pp. 391–397, 2020.
- [26] C.-K. Liu, H. Chen, M. Imani, K. Ni, A. Kazemi *et al.*, "Cosime: Fefet based associative memory for in-memory cosine similarity search," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–9.
- [27] X. Chen, K. Ni, M. T. Niemier, Y. Han, S. Datta *et al.*, "Power and area efficient fpga building blocks based on ferroelectric fets," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, pp. 1780–1793, 2018.
- [28] Y. Fang, J. Gomez, Z. Wang, S. Datta, A. I. Khan *et al.*, "Neuro-mimetic dynamics of a ferroelectric fet-based spiking neuron," *IEEE Electron Device Letters*, vol. 40, pp. 1213–1216, 2019.
- [29] X. S. Hu, M. Niemier, A. Kazemi, A. F. Laguna, K. Ni *et al.*, "In-memory computing with associative memories: a cross-layer perspective," in *2021 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2021, pp. 25–2.
- [30] H. H. Li, Y. Chen, C. Liu, J. P. Strachan, and N. Davila, "Looking ahead for resistive memory technology: A broad perspective on rram technology for future storage and computing," *IEEE Consumer Electronics Magazine*, vol. 6, pp. 94–103, 2016.
- [31] A. J. Tan, K. Chatterjee, J. Zhou, D. Kwon, Y.-H. Liao *et al.*, "Experimental demonstration of a ferroelectric hfo 2-based content addressable memory cell," *IEEE Electron Device Letters*, vol. 41, pp. 240–243, 2019.
- [32] J. Cai, M. Imani, K. Ni, G. L. Zhang, B. Li *et al.*, "Energy efficient data search design and optimization based on a compact ferroelectric fet content addressable memory," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 751–756.
- [33] L. Liu, A. F. Laguna, R. Rajaci, M. M. Sharifi, A. Kazemi *et al.*, "A reconfigurable fefet content addressable memory for multi-state hamming distance," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2023.
- [34] X. Yin, K. Ni, D. Reis, S. Datta, M. Niemier *et al.*, "An ultra-dense 2fefet tcam design based on a multi-domain fefet model," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, pp. 1577–1581, 2018.
- [35] A. Kazemi, F. Müller, M. M. Sharifi, H. Errahmouni, G. Gerlach *et al.*, "Achieving software-equivalent accuracy for hyperdimensional computing with ferroelectric-based in-memory computing," *Scientific reports*, vol. 12, p. 19201, 2022.
- [36] M. Poremba, S. Mittal, D. Li, J. S. Vetter, and Y. Xie, "Destiny: A tool for modeling emerging 3d nvm and edram caches," in *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2015, pp. 1543–1546.
- [37] T. Soliman, F. Müller, T. Kirchner, T. Hoffmann, H. Ganem *et al.*, "Ultra-low power flexible precision fefet based analog in-memory computing," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 29–2.
- [38] A. Hernandez-Cane, N. Matsumoto, E. Ping, and M. Imani, "Onlinehd: Robust, efficient, and single-pass online learning using hyperdimensional system," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021, pp. 56–61.
- [39] "Uci machine learning repository," <http://archive.ics.uci.edu/ml/datasets/ISOLET>.
- [40] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*. Springer, 2012.
- [41] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th international symposium on wearable computers*. IEEE, 2012, pp. 108–109.