A FeFET-based Time-Domain Associative Memory for Multi-bit Similarity Computation

Qingrong Huang[1], Hamza Errahmouni Barkam[2], Zeyu Yang[1], Jianyi Yang[1,3,*], Thomas Kämpfe[4], Kai Ni[5], Grace Li Zhang[6], Bing Li[7], Ulf Schlichtmann[7], Mohsen Imani[2], Cheng Zhuo[1,8,*], Xunzhao Yin[1,8,*]
[1]Zhejiang University, [2]University of California, Irvine, [3]ZJU-Hangzhou Global Scientific and Technological Innovation Center,
[4]Fraunhofer IPMS, [5]University of Notre Dame, [6]Technical University of Darmstadt, [7]Technical University of Munich,
[8]Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province
[*]Corresponding authors, email: {czhuo, yangjy, xzyin1}@zju.edu.cn

Abstract*—The exponential growth of data across various domains of human society necessitates the rapid and efficient data processing. In many contemporary data-intensive applications, similarity computation (SC) is one of the most fundamental and indispensable operations. In recent years, In-memory computing (IMC) architectures have been designed to accelerate SC by reducing data movement costs, however, they encounter challenges with signal domain conversion, variation sensitivity, and limited precision. This paper proposes a ferroelectric FET (FeFET) based time-domain (TD) associative memory (AM) for energy efficient SC. Such TD design can convert its output (i.e., time interval) to digits with relatively simple sensing circuitry thus saves large amount of area and energy compared with conventional IMC designs that process analog voltage/current signals. The variable-capacitance (VC) delay chain structure in our design supports quantitative SC and enhances robustness against variations. Furthermore, by exploiting multi-domain ferroelctric FET (FeFET), our design is capable of performing SC on vectors with multi-bit element, enabling support for higher-precision algorithms. Simulation results show that the proposed TD-AM achieves 13.8×/1.47× energy saving of our design compared to CMOS/NVM based TD-IMC designs. Additionally, our design exhibits good robustness in monte carlo simulation with variation extracted from experimental measurements. Investigation on precision of hyperdimensional computing (HDC) show that higher element precision reduces the size of HDC model when considering to achieve same accuracy, indicating an improved efficiency. Benchmarkings against GPU demonstrate in general 2/3 orders of magnitude speedup/energy efficiency improvement of our design. Our proposed multi-bit TD-AM promises energy-efficient quantitative SC for diverse intensive data processing application, especially in energy-constrained scenarios.

I. Introduction

Similarity computation (SC) has become an essential operation in modern applications such as network routing [1], caches [2], database [3], [4], bioinformatics [5], and, notably, in the field of machine learning (ML) [6]–[9]. This operation typically involves comparing one vector to another, measuring the differences between vectors using various distance metrics, or detecting the vector that is closest to the query among a set of vectors. Many efforts have been dedicated to designing in-memory computing (IMC) circuits and architectures for SC acceleration [10]–[12], as IMC significantly alleviates the costly data movement in conventional Von Neumann architectures. For example, some designs employ crossbars that perform multiplication-and-accumulation (MAC) operations to measure cosine distance [6], [12], while others utilize binary/ternary content addressable memories (B/TCAMs) for query searches [11], [13]–[15]. However, several issues hinder these technical routes from achieving higher efficiency and accuracy: 1) most of IMC designs perform computation in voltage or current domain, necessitating expensive peripherals (e.g. ADCs) to convert the signals between analog and digital domains, and incurring relatively high static power consumption due to DC current in computing phase; 2) analog signals are more vulnerable to variations, leading to trade-off between precision and efficiency; 3) in most AI applications, quantitative SC is required, while many accelerator designs (e.g. CAMs) can only support identifying full match or mismatches within a limited number of cells.

Time-domain (TD) computing is emerging as a promising alternative computing paradigm to address the challenges mentioned above [16]. In TD computing, computation results are represented as cumulative signal propagation delays generated through a series of cascading delay stages. The propagation delay of each delay stage is modulated by the computing output, according to the basic delay formula $t_{delay} = RC$, the computing output can be manifested as either a change in capacitance or a change in resistance. Compared to conventional analog IMC designs, TD-IMCs are mostly digital and are thus more compatible to advanced process technologies and more robust against variations. Moreover, time-digital conversions are more energy efficient than analog-digital conversions. Additionally, TD-IMCs can easily perform quantitative SC, making them an attractive solution for energy-constrained SC application scenarios, including edge AI, energy harvesting device and implantable device.

Many TD-IMC designs have been proposed, which employ SRAM based IMC cell as their logic cells [17]–[20], however, the large SRAM cell size and the need for additional delay circuity (e.g. buffer or inverter) result in a loss of advantages in terms of area cost when compared with other IMC designs. In recent years, breakthroughs in many non-volatile memory (NVM) technologies have paved the way for ultra-compact and efficient IMC cell and array designs. Several previous studies explored the integration of NVMs into TD-IMC designs, for example, [21] proposed a TD-IMC design using spintronics device to accelerate MAC in neural networks, [22] utilized ferroelectric FET (FeFET) for similar purposes, [23] proposed RRAM based TD-CAM, and [24] presented a FeFET based TD-IMC design capable of computing both MAC and hamming distance. However, it is noteworthy that these TD-IMC works are still in the early stages of development, primarily due to the lack of comprehensive robustness analysis and algorithm/system-level benchmarking, furthermore, there is ample room for optimization to fully exploit the capabilities of NVMs in terms of storage density and other potential benefits.

To address these issues, we propose a FeFET-based TD associative memory (AM) that performs quantitative SC based on Hamming distance for energy-constrained AI circumstance. Unlike prior works [21], [22] that use IMC cells as variable resistors, our design has a novel variable-capacitance (VC) delay chain structure, in which IMC cell serves as a control unit for the delay stage, rather than being directly placed in the signal propagation path. This design not only improves the robustness against NVM variations but also maintains a strict linear relationship between delay and the computing result, thus enabling accurate and quantitative SC. We also propose a 2-step computation principle for our TD-AM to maximizes hardware utilization and reduce computation latency. Furthermore, by leveraging multi-domain FeFET

979-8-3503-4859-0/DATE24/© 2024 EDAA

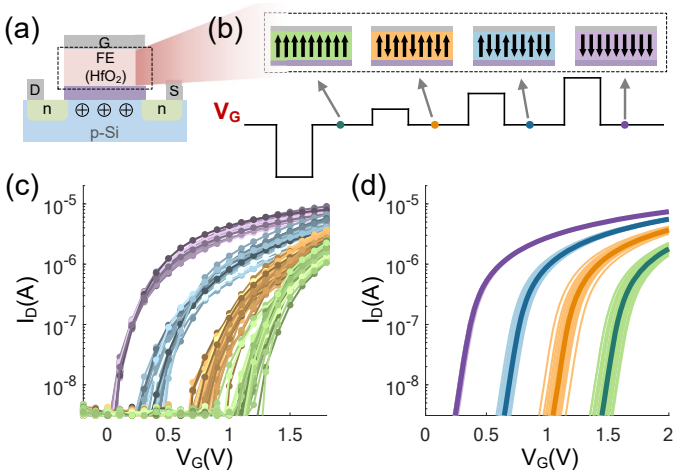

Fig. 1. (a) Physical structure of FeFET. (b) Write pulse for multi-state FeFET and the corresponding polarization in the FE layer. (c) $I_D - V_G$ curves from experimental device-to-device measurement on 60 devices [25]. (d) $I_D - V_G$ curves of FeFET simulation model [26].

device, we propose a 2-FeFET cell design that can support multi-level storage and computing, further enhancing efficiency and enabling higher-precision data processing.

We evaluate the performance of our design through SPICE simulations, the results demonstrate that our multi-bit TD-AM design achieves 13.84× higher energy efficiency compared to other TD-IMC designs. we also conducted Monte Carlo analysis using variation data measured from FeFET prototype chips, demonstrating the robustness of our design against variations. Moreover, we conducted application benchmarking in emerging brain-inspired hyperdimensional computing (HDC) tasks, the data precision investigation on HDC shows reduced dimensional can be achieved by our multi-bit TD-AM design, and benchmarking results against GPU reveal 2 and 3 orders of magnitude improvement on speed and energy efficiency, respectively.

The rest of this paper is organized as follows: Sec. II provides the basics and a brief review of prior works related to this work. Sec. III introduces the proposed FeFET-based multi-bit TD-AM cell and array design. Sec. IV presents the experimental results of the proposed design and benchmarks it for HDC applications. Sec. V concludes the paper.

II. Background

A. FeFET basics

FeFET has gradually emerged as an attractive candidate device in IMC research, thanks to its promising characteristics, including non-volatility, CMOS compatibility, and high ON/OFF ratio. As illustrated in Fig. 1(a), the structure of FeFET resembles that of a MOSFET, with the addition of a ferroelectric (FE) layer on its gate, the coupling between the FE capacitance C_{FE} and MOSFET gate capacitance C_G provides FeFET with tunable hysteresis, resulting in non-volatility. The development of higk-κ FE materials (e.g. HfO_2) has further facilitated the integration of FeFET into advanced technology nodes, such as 28nm bulk [27] and 22nm FDSOI [28]. Information is written into FeFET by applying different write voltages, V_{GS}, resulting in different threshold voltage V_{TH}, as illustrated in Fig. 1(b). Experimental validation has shown that FeFETs can be programmed to exhibit four distinct states [25], the $I_D - V_G$ curves from device-to-device measurement on 60 devices are shown in Fig. 1(c). To capture the non-volatility and multi-state characteristics, an experimentally calibrated multi-domain Preisach FeFET simulation model was proposed by [26], and the simulated $I_D - V_G$ curves are shown in Fig. 1(d).

B. IMC Designs for Similarity Computation

Due to its widespread utilization in numerous applications, considerable research efforts have been directed towards designing IMC circuits and architectures for accelerating similarity computation. A typical representative is CAM [29], which performs parallel associative search between the vectors stored in the array and the input query. Recently, leveraging emerging NVM has significantly improved the density and efficiency of CAMs compared to conventional 16-T TCAM, examples include TCAMs based on 2T-2R cell [30], 1T-1MTJ cell [31], and 2-FeFET cell [32], and such CAM designs have garnered increasing interest due to their demonstrated utility in pattern search during neural network inference phase [15]. FeFET based designs have further evolved to support multi-bit encoding and search [11], [25], [33], [34]. However, these CAMs only identify full match or cases with very few mismatch cells, i.e., do not support quantitative SC, limiting their implementation in more complex application such as deep learning, where full match are rare due to statistical properties.

Some other works designed crossbar arrays for quantitative and parallel SC, for instance, [25] presented a 1-FeFET crossbar based multi-bit CAM design capable of computing the Hamming distances between the query and stored vectors by sensing the mismatch current. However, the current-domain computation leads to high static power, and the cost of sensing unit (i.e., ADC) was not discussed. COSIME [12] employs translinear circuit and winner-take-all circuits with a crossbar, enabling the identification of the vector with the largest cosine similarity to the input query. However like other crossbar-based designs, it inevitably incurs high static power consumption. Furthermore, this design does not output the exact similarity result, which is crucial for parameter update in some machine learning algorithms [35].

C. Motivation of TD-IMC and Related Works

The concept of TD computation was first introduced by [16], to address the challenges encountered by hybrid signal system on chips (SoCs), including: 1) high energy and area cost associated with signal conversion; 2) difficulties in analog circuit design automation and fabrication using advanced technology. TD computation simplifies A/D interface by employing time-digital converter (TDC), and replaces complex voltage/current-domain analog computation cores with TD computing stages that highly compatible to digital circuits. Due to space limitations, for further details, please refer to [16]. As IMC has gained research interests in recent years, some works have attempted to combine these two technical routes for more efficient computing. [17] proposed an SRAM-based TD-IMC design capable of binary MAC for Binary Neural Network (BNN) acceleration, multiplications are executed in TD stages, and summation is achieved by cascading TD stages. Another SRAM-based TD-IMC design, TIMAQ [20] extended support for arbitrary data precision and improved the density of TD stages. However, SRAM still consumes significant area and energy, making it inefficient.

Several recent works have explored the use of NVMs for denser and more efficient TD-IMC. [21] presents a spintronics memory-based design for neural network acceleration, featuring a TD-IMC multi-addend adder. However, the D Flip-Flop based TDC design in this work is much more complex than conventional TDCs due to the adoption of parallel TD units connections rather than cascading, resulting in non-linear time outputs. [22] proposed a FeFET based compact design for MAC operation, the delay stages in this design is similar to [20] but contains only 1 FeFET and 2 MOS by leveraging unique characteristics of FeFET. However, directly connecting FeFETs into the pull-down path and using them as tunable resistors amplify

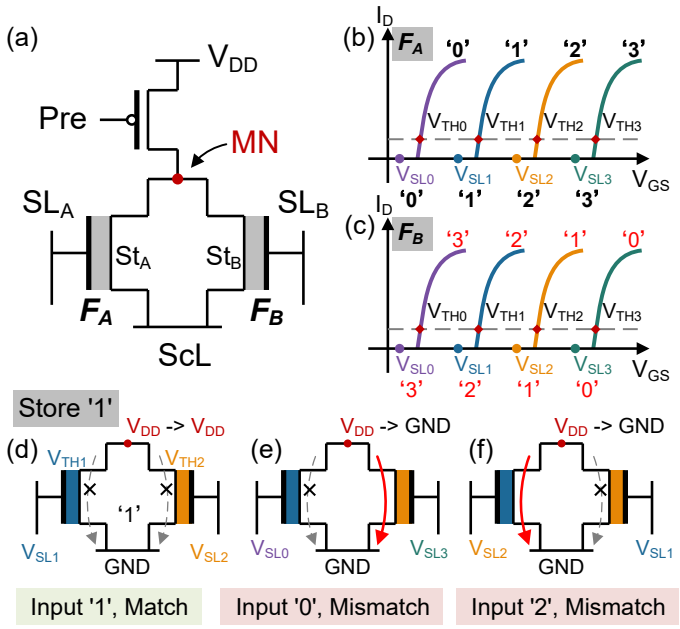

Fig. 2. (a) The proposed multi-bit IMC cell. V_{TH} and V_{SL} configuration for (b) F_A and (c) F_B. (d-f) Voltages and status of a cell storing '1' with different input values.

the impact of FeFET variations, as a slight error on FeFET V_{TH} leads to significant resistance change. Furthermore, due to the large ON/OFF ratio of FeFET, FeFETs in OFF state can fully interrupt signal propagation, resulting in computation failures. [24] partially addressing the above issues by designing a variable capacitance delay chain, and this design can be configured to support MAC and Hamming distance based similarity computation. However, the IMC cell in this work only supports binary operations and does not fully utilize the multi-domain FeFET characteristics. In this paper, we present a FeFET-based TD-AM which can support multi-bit similarity computation based on hamming distance.

III. Multi-bit TD-AM Design

A. Multi-Bit IMC Cell

The structure of our proposed multi-bit IMC cell is illustrated in Fig. 2(a), it consists of 2 FeFETs, referred to as F_A and F_B, connected in parallel, along with a PMOS for precharging. The gates of the FeFETs are connected to the Search Lines (SLs), and the output of the computation is reflected by the drains of FeFETs, referred to as Match Node (MN). The cell operates in two phases: precharge and compute. In the precharge phase, the PMOS is turned on, charging the MN voltage V_{MN} to V_{DD}. In the compute phase, voltages are applied on SL_A and SL_B according to the input operand, the V_{MN} either remains at V_{DD} or drops to GND, depending on the computing result. We illustrate the working principle of this cell using a 2-bit encoding example. For F_A, we define four threshold voltage $V_{TH0} \sim V_{TH3}$ to store the 4 values of a 2-bit number, i.e., '0', '1', '2', '3', as shown in Fig. 2(b), with corresponding SL voltages are $V_{SL0} \sim V_{SL3}$. For F_B, the stored and input values represented by V_{TH} and V_{SL} are reversed compared to F_A, as shown in Fig. 2(c). In this configuration, when the input value and the stored value are equal (indicating a match), both F_A and F_B remains non-conductive and MN remains at V_{DD}, otherwise, F_A/F_B will be turned on, and discharges MN to GND if the input value is larger/smaller than the stored value. Examples of a match (input '1') and mismatches (input '0' and input '2') on a cell storing '1' are provided in Fig. 2(d-f).

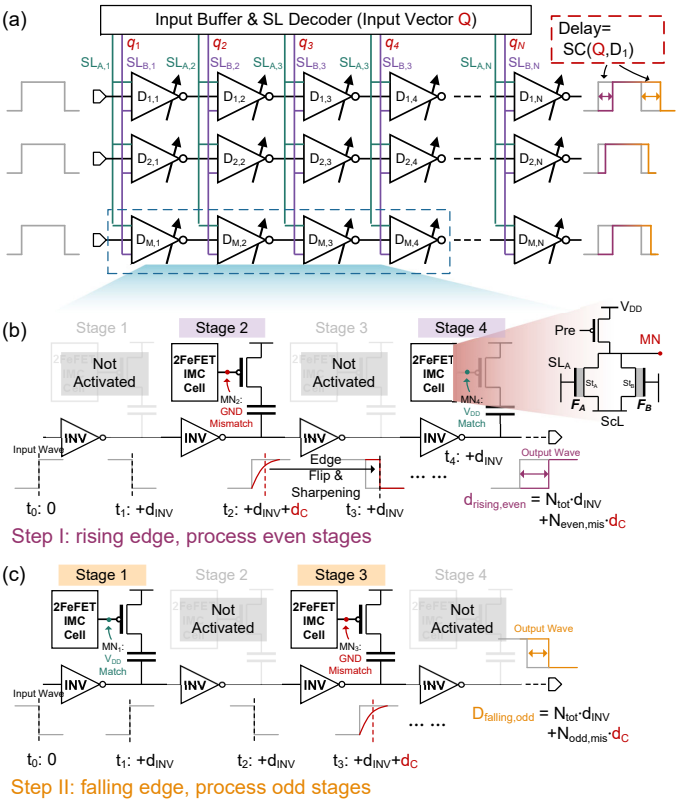

Fig. 3. (a) Structure of the proposed TD-AM. Delay chain status and pulse propagation in (b) step I and (c) step II of the proposed 2-step operation scheme, in which even and odd stages are processed, respectively.

B. TD-AM array

Building upon the 2-FeFET multi-bit IMC cell described above, we have designed the TD-AM array as depicted in Fig. 3(a). The fundamental building block of this array is a delay stage, which compare its stored multi-bit number $D_{i,j}$ with an input multi-bit number q_j, the propagation delay of delay stage is modulated by the comparison result, if it is a match (i.e., $D_{i,j} = q_j$), the delay is short; otherwise it incurs a long delay. N stages are cascaded in rows to form a delay chain. Therefore, the accumulated delay of a delay chain represents the similarity between the input vector Q and a stored vector D_i, both consist of multi-bit elements. The input vector is shared by all delay chains in the TD-AM array through vertical search lines (SLs), enabling the parallel SC between Q and a set of vectors $D_1 \sim D_M$.

The structure of the delay stage can be seen in Fig. 3(b), it comprises an inverter, a load capacitor (C), a PMOS and a 2-FeFET IMC cell as described in Sec. III-A, the MN of the IMC cell connects to the PMOS gate. If the 2FeFET cell outputs a mismatch, MN will be discharged to GND, turning the PMOS on, and the load capacitance will be added to the inverter's output node, introducing an additional signal propagation delay d_C in this stage. In the case of a match, the PMOS will be off, blocking the load capacitor, and the propagation delay of this stage will be the inverter's intrinsic delay d_{INV}.

However, because inverter flip the edge of the input signal, for the inverter-based delay stage, the input signals of two adjacent stages have different edges, resulting in different propagation delays due to the speed mismatch of PMOS and NMOS. Moreover, the output pulse of a mismatched delay stage is not steep enough, and directly using this pulse as the input pulse of the next stage will introduce an error its its delay, this error is exacerbated at multiple consecutive mismatched stages. A straightforward solution is to replace the inverters with

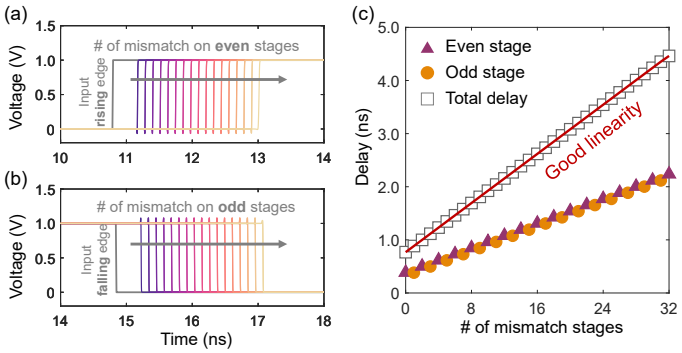

Fig. 4. Transient waveforms of (a) rising/(b) falling edge of output pulse with varying number of mismatch stages. (c) The total delay is linearly dependent on the number of mismatch stages.

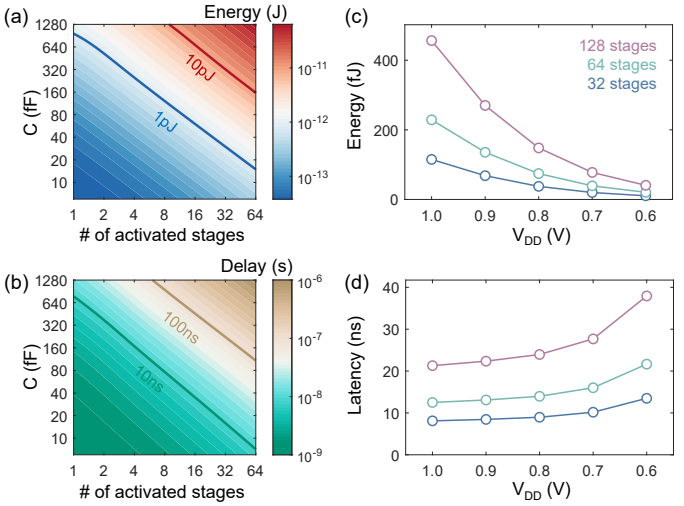

Fig. 5. (a) Energy and (b) delay of the proposed TD-AM array with different array sizes and different load capacitor values. (c) Energy and (c) latency of delay chains with 32/64/128 stages under supply voltage scaling.

buffers, but this would introduce additional area and energy costs. To address these issues, we propose a 2-step operation scheme. In step I, as shown in Fig. 3(a), we process rising edge of the input pulse, all odd stages are deactivated by assigning both SL_A and SL_B of all odd stages to V_{SL0}, consequently, all FeFETs in these stages are non-conductive, and the MN voltage of these stages remains V_{DD}, equivalent to a match that incurs no additional delay. For the even stages, if a stage mismatches (e.g., stage 2 in our example shown in Fig. 3(b)), the smoothly changing output pulse can be sharpened by the next odd stage (i.e., an inverter), therefore avoids delay error. The rising edge delay contributed by even stages is given by

$$d_{rising,even} = N_{tot} \cdot d_{INV} + N_{even,mis} \cdot d_C$$

where N_{tot} is the total number of stages in a delay chain, and $N_{even,mis}$ is the number of mismatched even stages. Similarly, the falling edge is processed in step II, the falling edge delay is contributed by odd stages while all even stages are disabled. The total delay, i,e., the similarity computation result SC(Q, D_i), is obtained by adding the rising edge and falling edge delays together.

$$d_{tot} = d_{rising,even} + d_{falling,odd} = 2 \cdot N_{tot} \cdot d_{INV} + N_{mis} \cdot d_C$$

IV. Evaluation

In this section, we evaluate and validate our proposed design at both circuit level and system level, then we compare the evaluation results with other IMC/TD-IMC works which can perform SC.

A. Circuit-Level Evaluation

We conducted SPICE simulations on the proposed TD-AM using Cadence Spectre Simulator. We employed the compact multi-domain FeFET model proposed by [26], while MOSFET, capacitor and other devices models were obtained from the 40nm UMC processing development kit (PDK). The load capacitor in each stage was set to 6fF unless explicitly mentioned. The FeFET threshold voltages $V_{TH0} \sim V_{TH3}$ shown in Fig. 2(b)(c) are selected as 0.2V, 0.6V, 1.0V and 1.4V. We adopted the write method from [36] to program these V_{TH} values. The search line voltages $V_{SL0} \sim V_{SL3}$ were set to 0V, 0.4V, 0.8V and 1.2V.

To verify the relationship between delay and SC result, we constructed a 32-stage delay chain and fed the input vectors with varying similarity (ranging from 0 to 32) to the stored one. The waveforms of the input pulse and the delayed output pulses are shown in Fig. 4(a)(b), it is evident that the greater the number of mismatched even/odd stages is, the longer the rising/falling edge of the output pulse is delayed. Furthermore, as shown in Fig. 4(c), the total delay is linearly related to the number of mismatched stages, thus demonstrating that our design supports quantitative SC.

We evaluated the performance of our design by measuring the energy and delays for different load capacitor values (ranging from 6fF to 1280fF) and different numbers of stages in a delay chain (ranging from 1 to 64), the measurement results are illustrated in Fig. 5(a)(b), the contour lines that represent a fixed energy consumption or delay are in the diagonal direction, indicating that energy and delay are proportional to the product of the load capacitor value and number of mismatch stages, This relationship also ensures linear relationship between delay and the number of mismatch stages under different load capacitor values, thus demonstrating scalability of our design with respect to array size and load capacitor value. As the load capacitor values grows, both energy and delay increase, suggesting a preference for small load capacitor. However, shorter delay requires sensing unit (e.g., a counter to convert the analog delay to a digital value) to have higher resolution, which typically entails a more complex sensing unit structure and higher energy and area costs. Furthermore, different applications/algorithms may require different computing precision, such as binarized precision for highly quantized models [37]. Therefore, there is a trade-off exists between delay and energy consumption of delay chain, sensing unit complexity and application requirements.

We conducted additional investigations into the performance of our design by varying the supply voltage (V_{DD}). Fig. 5(c)(d) show the average energy and computational latency results for 32/64/128-stage delay chains. The results demonstrate that scaling down V_{DD} leads to a substantial reduction in energy consumption, with a slight increase in delay. Consequently, scaling down V_{DD} is viable method for further improving the energy efficiency. The maximum energy efficiency achieved by our design was recorded as 0.159 fJ/bit. We compared our result with other IMC similarity computation designs, the results are summarized in Table I. Our design achieves 2∼3× higher energy efficiency than CAMs [15], [29] while also offering the capability for quantitative SC. In the realm of TD-IMC counterparts, our proposed design exhibits significantly superior energy efficiency when compared to both CMOS-based designs [20] and even prior FeFET-based design [24]. The enhanced energy efficiency of our design is attributed to its multi-bit capabilities. [22] reports ultra-low energy consumption per bit, it is primarily due to its utilization of advanced 14nm technologies and an optimized measurement configuration, which may not directly

TABLE I
COMPARISON OF THE PROPOSED TD-AM WITH STATE-OF-THE-ART TD-IMC DESIGN.

Designs	Signal domain	Device	Cell/Stage size	SC Type	Energy per bit (fJ)	Techonology (nm)
16T TCAM [29]	Voltage	CMOS	16T	Hamming distance, non-quantitative	0.59 ($\times$3.71)	45
Nat. Electon.'19 [15]	Voltage	FeFET	2FeFET	Hamming distance, non-quantitative	0.40 ($\times$2.52)	45
JSSC'21 [20]	Time	CMOS	20T+4MUX	MAC/Cosine distance, quantitative	2.20 ($\times$13.84)	28
IEDM'21 [22]	Time	FeFET	2T-1FeFET	MAC/Cosine distance, quantitative	0.039 ($\times$0.245)	14
Work [24]	Time	FeFET	3T-2FeFET	MAC/Hamming distance, quantitative	0.234 ($\times$1.47)	40
This work	Time	FeFET	4T-2FeFET	Hamming distance, quantitative	0.159 ($\times$1)	40

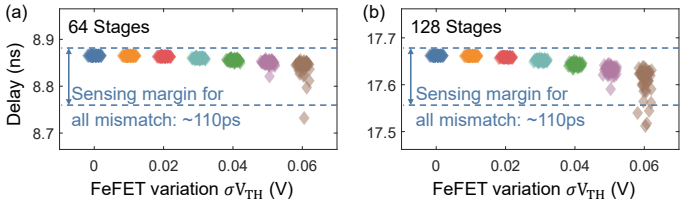

Fig. 6. Distributions of delay of worst case computation in arrays with (a) 64 and (b) 128 stages, with varying levels of FeFET V_{TH} variations.

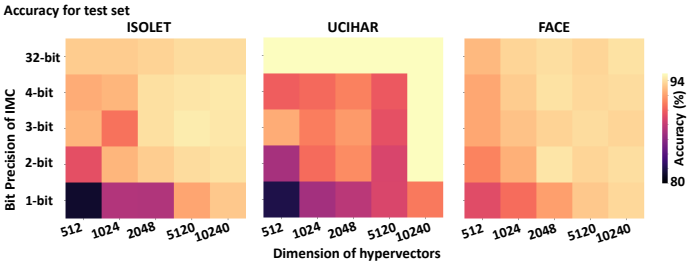

Fig. 7. Accuracy analysis for different bit-precision and dimensionalities for three datasets (ISOLET, UCIHAR, and FACE, respectively).

translate to realistic conditions.

The impact of NVM variations was also analysed through Monte Carlo simulation. We modeled the effect of all FeFET variations as a shift in V_{TH}, which we will refer to as V_{TH} variation for the sake of later discussion. We verified the robustness of our design by introducing different levels of FeFET V_{TH} variations into the delay chain and examined the resulting delay variability in the worst scenario: all stages are mismatched. The results are shown in Fig. 6. As the variation level or the length of delay chain increases, the delays become more widely distributed. Nonetheless, even when considering FeFET V_{TH} variation up to 60mV, the delays of vast majority of Monte Carlo runs remain within the sensing margin. For reference, we derived the V_{TH} variation values from experimentally measured data [25] and fitted them with standard distributions, the standard deviation values for $V_{TH0} \sim V_{TH3}$ are found to be 7.1mV, 35mV, 45mV and 40mV, respectively. Therefore, our design demonstrated sufficient robustness to NVM variations, and also revealed an intriguing potential of our design for supporting higher precision, e.g., 3- or 4-bit storage and computation.

B. Case study: HDC

In the ever-changing realm of ML, we grapple with a key challenge: hardware and computational constraints. Conventional ML models, while delivering top-tier performance, can be resource-intensive, hindering their scalability across diverse devices. To tackle this issue, a new field called brain-inspired learning or hyperdimensional computing (HDC) [7] aims to emulate the brain's remarkable abilities and incorporate them into various aspects of the computational framework. HDC excels in various tasks, spanning graph memorization, reasoning, classification, clustering, and genomic detection [38]–[40]. HDC and

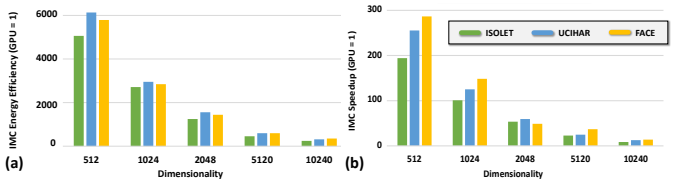

Fig. 8. (a) Energy efficiency and (b) speedup comparison between GPU and our IMC architecture at 0.6V and 128 stages.

IMC devices share a common goal: enhancing the unique qualities that distinguish our brains from conventional computers. While HDC elevates learning and representation capabilities, IMC seamlessly integrates computation and memory, offering functionalities absent in traditional ML software and hardware. Although AI models aspire to excel in IMC devices, they grapple with constraints like limited bit precision (typically 4 or 8 bits) and inherent noise. Nevertheless, HDC's holographic nature and noise resilience position it as an ideal companion for IMC architectures [41].

In our quest to demonstrate the effectiveness of an HDC architecture tailored for IMC devices, we undertook a comprehensive case study. We generated two models—one utilizing the traditional full-precision 32-bit model as a reference point [35]. The quantization process was intricately designed to minimize information loss during the transition from 32-bit to the n-bit design of the IMC circuit. By thoroughly mapping the class hypervector values based on probability distributions into 2^n blocks of equal areas, we achieved a nuanced representation, allocating smaller widths to more significant values.

Our experimentation extended to the testing phase across three diverse datasets—face detection (FACE) [42], voice recognition (ISOLET) [43], and activity recognition (UCIHAR) [44]. Each experiment was executed thrice, utilizing varying bit-precision (ranging from 32 bits for ideal conditions down to 4, 3, 2, and 1-bit configurations) and dimensions (512, 1024, 2048, 5120, and 10240). The results were satisfactory, as shown in Fig. 7, with the reduced-bit models consistently demonstrating performance on par with the 32-bit model, particularly in higher dimensionalities. Our experimental findings reveal an intriguing connection between bit precision and dimensionality in our quantized models. Specifically, we've observed that as we increase the bit precision from 1 to 4 bits, the dimensionality requirements to match the maximum accuracy of the full precision model (at 32 bits) steadily diminish. This trend holds true for most scenarios, although exceptions exist where achieving peak accuracy with just a single bit proves non-successful, such as that of UCIHAR.

Furthermore, the augmentation of bit precision provides us with the opportunity to achieve significant reductions in hypervector dimensionality. This, in turn, translates into substantial memory savings, a critical factor that can help reduce latency and boost energy efficiency. To illustrate, in the case of the ISOLET dataset, both the full precision and 2-bit quantized models converge at an optimal dimensionality of a mere 2048. In stark contrast, the 1-bit quantized model necessitates a dimensionality five times larger (10240) to attain the same maximum

accuracy, highlighting the profound influence of bit precision on dimensionality, memory demands, and overall system performance.

In our experiments, we studied the temporal and energy expenses associated with our GPU (NVIDIA GeForce RTX 4070). To ensure seamless compatibility with PyTorch, we embarked on the development of novel code and software, facilitating seamless integration within our framework. Our empirical findings, presented in Fig. 8(b), show the speedup between the GPU and the IMC system with the configuration of 128 stages, running at 0.6V. This encompassed several dimensionalities and spanned three distinct datasets. Of particular note, for smaller dimensionalities, we observe high speedup enhancements, with gains ranging from 194× in the ISOLET dataset to an impressive 287× in the FACE dataset. However, as we ventured into the realm of higher dimensionalities with 128 stages, a substantial delay emerged, resulting in a gradual attenuation of the speedup effect. Ultimately, this resulted in an average speedup factor of 11.65×, which, though diminished, remains appreciable. Note that even under the circumstances of 3 to 4-bit precision, where we accomplished maximum accuracy across all three datasets with 1024 dimensions, an enduring average speedup of 124.8× was maintained, reaffirming the importance of having an IMC with higher bit-precision.

Regarding GPU energy expenses, we tracked the energy consumption throughout the software's operation. The detailed outcomes and graphical representation of this data can be conveniently located in Fig. 8(a). In tandem with the speedup outcomes, we observed advancements in energy efficiency for smaller dimensionalities, with efficiency enhancements ranging from an astounding 5061× in the ISOLET dataset to an astonishing 5790× for the FACE dataset. Notably, even in the context of the highest dimensionality setting, a resolute average energy efficiency factor of 303× was sustained. Significantly, in the scenario of 3 and 4-bit precision, where maximum accuracy materialized with only 1024 dimensions across all three datasets, enduring average energy efficiency of 2837× persisted, showing substantial efficiency improvements.

V. Conclusion

In this paper, we presented a novel FeFET based TD-AM for efficient and quantitative similarity computation. Leveraging multi-domain FeFET, our 2-FeFET IMC cell demonstrated its ability for multi-bit storage and computation. A variable capacitance delay chain was proposed to support quantitative, reliable, parallel and energy efficient similarity computation between a multi-bit input vector and a multi-bit stored vector. Evaluation results show that our design outperform other IMC based counterparts, with remarkable robustness to NVM variation. Benchmarking for HDC application against GPUs reveals substantial performance improvements and suggests the potential of our design in diverse energy-constrained scenarios.

Acknowledgements

This work was supported in part by NSFC (62104213, 92164203), Zhejiang Provincial Natural Science Foundation (LD21F040003, LQ21F040006).

References

[1] F. Yu, et al., Gigabit rate packet pattern-matching using TCAM, Proc. ICNP, 2004.
[2] V. S. Srinivasavarma, et al., A TCAM-based caching architecture framework for packet classification, ICNP, 20(1): 1-19, 2020.
[3] M. W. Berry, et al., Matrices, vector spaces, and information retrieval, SIAM review, 41(2): 335-362, 1999.
[4] X. Wang, et al., Triangle counting accelerations: From algorithm to in-memory computing architecture, IEEE TC, 71(10): 2462-2472, 2021.
[5] T. Robinson, et al., Hardware acceleration of genomics data analysis: challenges and opportunities, Bioinformatics, 37(13): 1785-1795, 2021.
[6] G. Karunaratne, et al., Robust high-dimensional memory-augmented neural networks, Nature communications, 12(1): 2468, 2021.
[7] P. Kanerva, et al., Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors, Cognitive computation, 1: 139-159, 2009.
[8] Z. Yan, et al., Swim: Selective write-verify for computing-in-memory neural accelerators, Proc. DAC, 2022.
[9] X. Chen, et al., Accelerating graph-connected component computation with emerging processing-in-memory architecture, IEEE TCAD, 41(12): 5333-5342, 2022.
[10] L. Liu, et al., A Reconfigurable FeFET Content Addressable Memory for Multi-State Hamming Distance, IEEE TCAS-I, 2023.
[11] X. Yin, et al., FeCAM: A universal compact digital and analog content addressable memory using ferroelectric, IEEE TED, 67(7): 2785–2792, 2020.
[12] C.-K. Liu, et al., Cosime: Fefet based associative memory for in-memory cosine similarity search, Proc. ICCAD, 2022.
[13] S. Shou, et al., See-mcam: Scalable multi-bit fefet content addressable memories for energy efficient associative search, Proc. ICCAD, 2023.
[14] J. Cai, et al., Energy efficient data search design and optimization based on a compact ferroelectric FET content addressable memory, Proc. DAC, 2021.
[15] K. Ni, et al., Ferroelectric ternary content-addressable memory for one-shot learning, Nature Electronics, 2(11): 521-529, 2019.
[16] D. Miyashita, et al., An LDPC decoder with time-domain analog and digital mixed-signal processing, IEEE JSSC, 49(1): 73-83, 2013.
[17] D. Miyashita, et al., A neuromorphic chip optimized for deep learning and CMOS technology with time-domain analog and digital mixed-signal processing, IEEE JSSC, 52(10): 2679-2689, 2017.
[18] L. R. Everson, et al., An energy-efficient one-shot time-based neural network accelerator employing dynamic threshold error correction in 65 nm, IEEE JSSC, 54(10): 2777-2785, 2019.
[19] J. Song, et al., TD-SRAM: Time-domain-based in-memory computing macro for binary neural networks, IEEE TCAS-I, 68(8): 3377-3387, 2021.
[20] J. Yang, et al., TIMAQ: A time-domain computing-in-memory-based processor using predictable decomposed convolution for arbitrary quantized DNNs, IEEE JSSC, 56(10): 3021-3038, 2021.
[21] Y. Zhang, et al., Time-domain computing in memory using spintronics for energy-efficient convolutional neural network, IEEE TCAS-I, 68(3): 1193-1205, 2021.
[22] J. Luo, et al., Energy-and area-efficient Fe-FinFET-based time-domain mixed-signal computing in memory for edge machine learning, Proc. IEDM, 2021.
[23] Y. Halawani, et al., RRAM-based CAM combined with time-domain circuits for hyperdimensional computing, Scientific reports, 11(1): 19848, 2021.
[24] X. Yin, et al., A homogeneous processing fabric for matrix-vector multiplication and associative search using ferroelectric time-domain compute-in-memory, arXiv preprint arXiv:2209.11971, 2022.
[25] X. Yin, et al., An ultracompact single-ferroelectric field-effect transistor binary and multibit associative search Engine, Adv. Intell. Syst., 5(7):2200428, 2023.
[26] K. Ni, et al., A circuit compatible accurate compact model for ferroelectric-FETs, Proc. VLSI Technology, 2018.
[27] M. Trentzsch, et al., A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs, Proc. IEDM, 2016.
[28] S. Dünkel, et al., A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond, Proc. IEDM, 2017.
[29] K. Pagiamtzis, et al., Content-addressable memory (CAM) circuits and architectures: A tutorial and survey, IEEE JSSC, 41(3): 712-727, 2006.
[30] J. Li, et al., 1Mb 0.41 μm 2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing, Proc. VLSI Technology, 2013.
[31] C. Zhuo, et al., Design of Ultra-Compact Content Addressable Memory Exploiting 1T-1MTJ Cell, IEEE TCAD, 2022.
[32] X. Yin, et al., An ultra-dense 2FeFET TCAM design based on a multi-domain FeFET model, IEEE TCAS-II, 66(9): 1577-1581, 2018.
[33] C. Li, et al., A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing, Proc. IEDM, 2020.
[34] T. Xu, et al., On the challenges and design mitigations of single transistor ferroelectric content addressable memory, IEEE EDL, 45(1): 112-115, 2024.
[35] A. Hernández-Cano, et al., Onlinehd: Robust, efficient, and single-pass online learning using hyperdimensional system, Proc. DATE, 2021.
[36] D. Reis, et al., Design and analysis of an ultra-dense, low-leakage, and fast FeFET-based random access memory array, IEEE JxCDC, 5(2): 103-112, 2019.
[37] I. Hubara, et al., Binarized neural networks, Advances in neural information processing systems, 2016.
[38] G. Karunaratne, et al., In-memory hyperdimensional computing, Nature Electronics, 3(6): 327-337, 2020.
[39] Q. Huang, et al., FeFET Based In-Memory Hyperdimensional Encoding Design, IEEE TCAD, 2023.
[40] L. Ge, et al., Classification Using Hyperdimensional Computing: A Review, IEEE Circuits and Systems Magazine, 20(2): 30-47, 2020.
[41] H. E. Barkam, et al., HDGIM: Hyperdimensional Genome Sequence Matching on Unreliable highly scaled FeFET, Proc. DATE, 2023.
[42] M. Imani, et al., A Framework for Collaborative Learning in Secure High-Dimensional Space, Proc. CLOUD, 2019.
[43] R. Cole, et al., "ISOLET", UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C51G69.
[44] J. Reyes-Ortiz, et al., "Human Activity Recognition Using Smartphones", UCI Machine Learning Repository, 2012, DOI: https://doi.org/10.24432/C54S4K.