FeReX: A Reconfigurable Design of Multi-bit Ferroelectric Compute-in-Memory for Nearest Neighbor Search

***Abstract*—Rapid advancements in artificial intelligence have given rise to transformative models, profoundly impacting our lives. These models demand massive volumes of data to operate effectively, exacerbating the data-transfer bottle-neck inherent in the conventional von-Neumann architecture. Compute-in-memory (CIM), a novel computing paradigm, tackles these issues by seamlessly embedding in-memory search functions, thereby obviating the need for data transfers. However, existing non-volatile memory (NVM)-based accelerators are application specific. During the similarity search operation, they support a single, specific distance metric, such as Hamming, Manhattan, or Euclidean distance in measuring the query against the stored data, calling for the development of reconfigurable in-memory solutions, adaptable to various applications. To overcome such a limitation, in this paper, we present FeReX, a reconfigurable associative memory (AM) that accommodates Hamming, Manhattan, and Euclidean distances. Leveraging multi-bit ferroelectric field-effect transistors (FeFETs) as the proxy and a hardware-software co-design approach, we introduce a constrained satisfaction problem (CSP)-based method to automate AM search voltage and stored voltage settings for different distance functions. Device-circuit co-simulations first validate the effectiveness of the proposed FeReX methodology for reconfigurable search distance functions. Then, we benchmark FeReX against k-nearest neighbor (KNN) and hyperdimensional computing (HDC), which highlights the robustness of FeReX and demonstrates up to 250$\times$ speedup and 10^4 energy savings compared with GPU.**

I. Introduction

The field of artificial intelligence has witnessed remarkable advancements, giving rise to models that yield a profound influence over various aspects of our lives. These models, however, frequently require vast amounts of data for their operation, thus exacerbating the data-transfer bottleneck inherent in the traditional von Neumann architecture. As our computing needs continue to evolve and generate ever-increasing volumes of data, the limitations of the von Neumann machine become increasingly evident. Consequently, there is a growing demand for a departure from the conventional computing paradigm, one that seamlessly integrates the critical functionalities of emerging machine learning models with the memory itself. This shift is not only desirable but also essential to keep pace with the demands of modern computing.

Compute-in-memory (CIM) has emerged as an alternative computing paradigm that integrates the separated computing unit and memory that exists in Von Neuman machine altogether. Several CIM primitives have demonstrated their potential for accelerating inferences in novel machine learning algorithms [1], even accelerating the training on-chip [2]. While it is attractive to push the designs of the domain-specific accelerator toward different applications, a collaborative design that is universal for different existing workloads remains unexplored.

Hamming distance-based CIM design has been originally proposed [3] for memory-augmented neural networks (MANN), but it suffers from non-negligible classification accuracy degradation. Recently, CIM design that implements Manhattan distance for MANN classification has been experimentally verified [4], and CIM design realizes Euclidean distance for hyperdimensional computing (HDC) has been demonstrated at the device level [5]. These designs aim to address the non-negligible algorithmic accuracy degradation with complex distance functions used in a certain application.

However, existing NVM-based AMs are limited to a specific classification problem, as one AM design can only support a single distance metric, including Hamming [3], Manhattan [4], Euclidean [5], and sigmoid [6]. A CIM search engine that is able to achieve a reconfigurable distance function is highly desirable. Based on the nature of applications, different distance functions may be used during the search, and, within a certain application, several distance functions may be exploited for various datasets.

In this paper, we proposed FeReX, a reconfigurable CIM-based associative memory (AM) for Hamming, Manhattan, and Euclidean distance search is presented, and the multi-bit ferroelectric field-effect transistor (FeFET) device is adopted as the proxy. A hardware-software co-design scheme is adopted, with the critical similarity search operation between the query and stored vectors being projected to the search voltages and the stored threshold voltage V_{th} of the FeFET device. A constrained satisfaction problem (CSP) based on device and circuit constraints that leverages backtracking and AC-3 algorithms is built to automate the input voltages of the AM, i.e. the search voltages and stored V_{th} of the FeFET, given a desired distance function of a specific application.

Our contributions in this paper are three-fold:

- We propose an AM cell encoding scheme that dynamically determines the cell structure within the proposed FeReX, which is capable of implementing various search distance functions by exploiting FeFET characteristics.
- We propose a new algorithm to detect whether the distance metric can be realized in the crossbar. This algorithm is not limited to FeFET, but generalizable to other non-volatile memory (NVM) devices with multi-bit properties.
- We benchmark the proposed FeReX with two different applications, KNN and HDC, showcasing the robustness and effectiveness of the proposed methodology and circuit designs. To the best of our knowledge, this is the first work to present a reconfigurable distance search function with NVM-based AM.

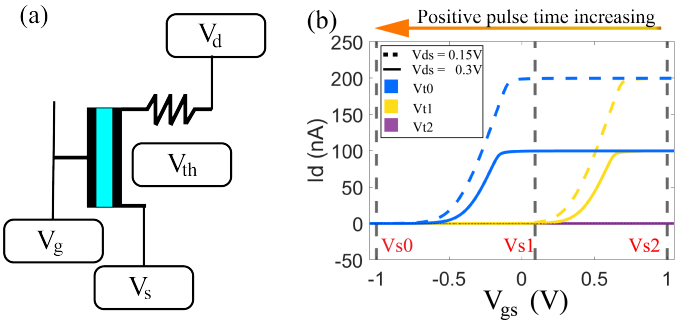

Fig. 1. **(a)** 1FeFET1R structure. **(b)** 1 FeFET 1R multi-bit I-V curve, where V_{t0}, V_{t1}, V_{t2} represent different V_{th} stored in the FeFET, and V_{s0}, V_{s1}, V_{s2} represent different search voltage (i.e, V_{gs}) applied to the FeFET during the search operation.

II. Background

In this section, we review the FeFET basics and its characteristics that are exploited within the FeReX. Then, recent AM designs for NN search are briefly summarized.

A. FeFET Characteristics

Excellent CMOS compatibility, outstanding scalability, and superior energy efficiency [7] of HfO_2 ferroelectric materials elucidate the competitiveness of Ferroelectric FET (FeFET) among other NVMs. Based on the conventional CMOS transistor, a FeFET is made with ferroelectric materials added to the CMOS gate stack. The stored value is captured by the threshold voltage (V_{th}) of a FeFET, and can be altered by applying a positive or negative voltage pulse at the gate terminal, which in turn changes the polarization of the Fe layer. Specifically, the value of V_{th} is determined by the duration and magnitude of the applied voltage pulse [3]. For instance, if the duration of a given positive voltage pulse increases, the V_{th} will decrease accordingly.

Recently, Soliman et al. propose a cell that integrates a resistor with a single FeFET [8], as shown in Fig 1(a). It is demonstrated in [8], [9] that by connecting a large resistor at the source (or effectively, drain) terminal of the FeFET, the ON state current I_{ds} is significantly reduced and thus is independent of the V_{th} variation [8]. Saito et al. further demonstrate a back-end-of-line (BEOL) 1FeFET1R structure, incurring no additional area penalty with an $M\Omega$ resistor integrated with a FeFET [10]. Given a V_{ds} and resistance R, The conducting current of a FeFET can be approximated as Min$\{I_{sat}$,$V_{ds}/R\}$ due to the fact that it is possible when $I = V_{ds}/R$ under a given V_{gs}, the FeFET operates in the linear region. Fig. 1(b) illustrates an MLC 1FeFET1R SPICE simulation. When $V_{gs} > V_{th}$, various V_{th} and V_{gs} values can be explored, where I is approximately equivalent to V_{ds}/R, while I approaches 0 if $V_{gs} < V_{th}$.

B. Existing AM Designs

Associative memory (AM) has been deployed in a variety of scenarios such as HDC [11], [13], MANN [4], few-shot learning [6], and so on. Table I summarizes existing AM based on single-level cell/multi-level cell (SLC/MLC) NVMs with different distance functions. Matching-based MLC 2FeFET-1T AM has been fabricated in [12]. To further achieve algorithmic level accuracy, AM with intricate distance functions in an MLC cell have been proposed including sigmoid and Euclidean functions [5], [6], etc. However, these efforts are designed for a fixed distance function. In this work, FeReX is able to support multiple distance functions as shown in Table I. Below we elaborate on the designed AM and its peripherals first. Then, the proposed algorithm for programming the AM is elucidated in Section IV.

TABLE I
Existing Associative Memories with Different Distance Functions

Design	NVM	Cell structure	MLC	Distance function
Nat. Ele. [11]	PCM	1PCM	No	Hamming
IEDM'20 [12]	FeFET	2FeFET-1T	Yes	Best-match
TED'21 [4]	RRAM	2RRAM	Yes	Manhattan
TC'21 [6]	FeFET	2FeFET	Yes	Sigmoid
SR'22 [5]	FeFET	2FeFET	Yes	Euclidean
FeReX (This work)	FeFET	1FeFET-1R	Yes	HD/L_1/L_2

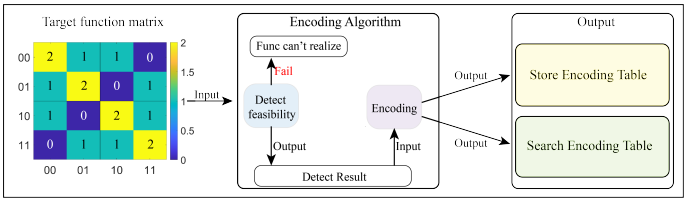

Fig. 2. Workflow of the proposed FeReX.

III. FeReX: Reconfigurable In-Memory Search Engine

In this section, we first elaborate on the required peripherals for FeFET-based AM design. Then, the FeReX circuit for reconfigurable distance search is exhibited. Finally, the proposed programming algorithm will be illustrated in Section IV.

The required peripherals for the NVM array are often driven by the fact that NVM write voltage is incompatible with conventional IO voltage [14]. Level shifters are required to enhance the voltage level from the power ring. In addition,

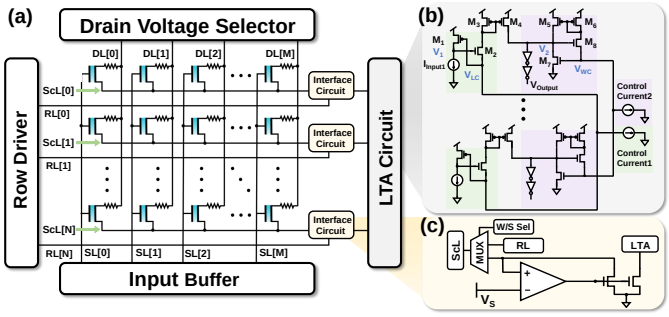

Fig. 3. (a) FeReX associative memory overview. (b) Loser-take-all (LTA) circuit. (c) Interface circuit.

several isolation mechanisms are required to prevent damaging the circuitries that are made of core voltage [15]. Column switch matrix for selecting columns and input decoder (or digital-to-analog converter) are also required [16].

Figure 3 shows the detailed circuit implementation of the proposed FeReX. FeReX consists of a crossbar array with the drain voltage selector and the inference circuit blocks for each row. NN search operation is then realized through the loser-take-all (LTA) circuitry. The search lines (SLs) and drain lines (DLs) are shared by the FeFETs within the same column, while the source lines (ScLs) link the FeFETs in the same row, as shown in Figure 3(a).

During the write phase, the MUX chooses RL. In this configuration, the RL voltage of the selected row is maintained at 0V, For rows that are intended to be left unselected, the RL voltage is raised to half of the V_{write}. Such a writing scheme is known as the write disturbance [17], where the V_{gs} of the unselected cell is raised to avoid disturbing. Whereas during the search phase, based on the results generated by the proposed FeReX encoding algorithm, different levels of search voltages are applied to the gate terminal of the FeFET through SLs. As discussed in Section IV, the ON state current conducts from DL to the ScL only when the applied V_{search} exceeds the threshold voltage V_{th}. Otherwise, the FeFET remains in the cut-off state. The magnitude of current I_{ON} through the FeFET is determined by the voltage difference V_{ds} between the Drain Line DL and the Source Line SL, as discussed in Section II where $I_{ON} \approx V_{ds}/R$. Since all Source Lines share the same voltage during searching, FeFETs in the same column exhibit the same I_{ON} when they are conducting. The currents flowing through FeFETs in the same row are aggregated at ScL and then enter the interface circuit. We propose to add an inference circuitry in between the 1FeFET1R crossbar and the LTA circuitry to inhibit ScL voltage fluctuation. ScL voltage fluctuation is detrimental to the proposed FeReX, as the change in V_{ds} will change the I_{ON} accordingly, resulting in inaccurate LTA sensing. The counterpart of the LTA circuitry is the WTA (winner-take-all), which has been utilized for NN detection as well. Readers interested in its detailed transistor-level operations can refer to [18] due to limited space.

IV. FeReX Encoding Algorithm

A. Encoding Algorithm Design

Unlike conventional AM designs that come with a fixed number of NVM per cell, FeReX allows for the flexibility to configure the number of FeFETs present in each AM cell. Furthermore, the distance measurement method can be customized through the use of the Distance Matrix, which is shown in Fig. 4(a). Within this matrix, columns stand for distinct stored values, and rows correspond to various search values, with each element in the matrix denoting the distance between a stored value and a search value. Fig. 4(a) shows the Distance Matrix corresponding to a 2-bit Hamming distance. When we input the search vector '00' with a store vector '11', we will receive distance 2 as the result.

INPUT: The $M \times N$ Distance Matrix DM to be implemented by each cell, which includes K FeFETs, with a current range $CR = [C_1, C_2, \cdots C_n]$ allowed to flow through each FeFET
OUTPUT: *Feasible Region* or $False$

```
for i from 0 to M-1 do
    for j from 0 to N-1 do
        DMCurs[i, j] ← DecomposeDM(K, DM[i, j], CR)
    end
    Searchlines[i] ← Backtracking(DMCurs[i])
end
FeasibleRegion ← AC3(Searchlines)

if Feasible Region not exist then
    return False
end
return Feasible Region
```

Algorithm 1: FeReX Feasibility Detection Algorithm

Fig. 4(b) illustrates the relationship between the encoding result and the FeReX circuit. The stored encoding is represented by V_{th} value in each FeFET device, while the search encoding consists of the FeFET's V_{ds} and V_{gs} values. The V_{ds} value directly determines the current flow through the FeFET when the FeFET is in ON-state. By investigating the interactions between the I_{ds}, V_{ds}, V_{gs}, V_{th}, and the give Distance Matrix, we argue that there exist several constraints that are bounded by the device and circuit. Specifically, encoding the Distance Matrix involves solving a constrained satisfaction problem (CSP) with three specific constraints. To facilitate the explanation, we assume that the number of FeFETs in one cell is k, the element value of the row r and column c in the Distance Matrix is denoted as $I_{r,c}$, and the current flowing through the FeFET i in this condition is $I_{r,c,i}$.

In the FeReX encoding algorithm, several constraints arise from the FeFET device are considered. Fig. 4(c) illustrates the process of decomposing an element (see the '2' distance in Fig. 4(c)) in the Distance Matrix into a set of FeFETs (three FeFETs are used in this example). The decomposition must satisfy $I_{r,c} = \sum_{i=1}^{k} I_{r,c,i}$, which is known as the Kirchhoff current law. Notice that '0' indicates a FeFET is in the OFF-state and the value of the decomposed result (in this case [I_0, I_1, I_2]) indicates the ratio of the drain-source current I_{ds} between each FeFET. Due to the physical constraint in the FeFET device, the current flows through the FeFET can not be infinitely large, which results in the limited range of available $I_{r,c,i}$ values, and thus the feasible options for the decomposition are constrained. The set of all feasible options for $I_{r,c}$ forms DMCurs[r, c]. We refer to this constraint as the first constraint.

Secondly, considering that the search encoding includes V_{ds}, it implies that when searching for the same value, the applied V_{ds} on FeFETs at the same position must remain the same.

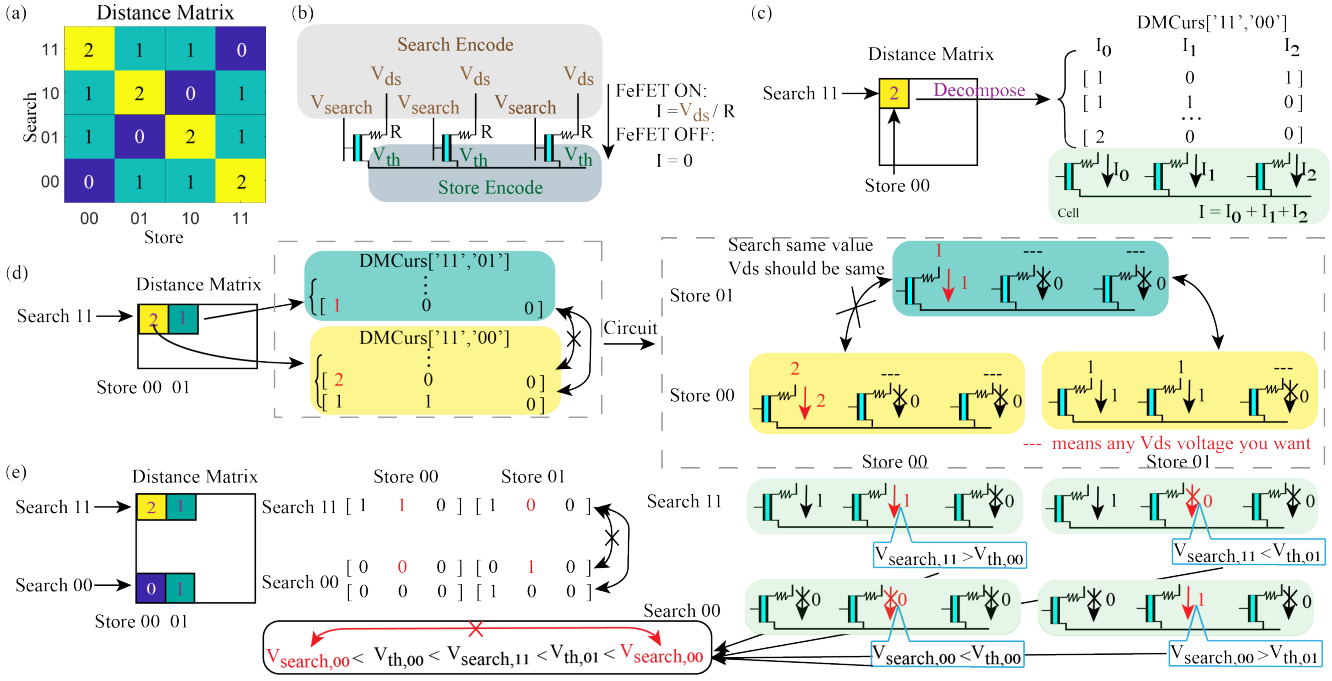

Fig. 4. (a) Distance Matrix of 2-bit Hamming Distance. (b) Encoding with FeReX circuit. The stored encoding involves the V_{th} values, while the search encoding consists of the FeFET's V_{ds} and V_{gs} values. (c) Decomposition process of elements in the Distance Matrix based on the number of FeFETs contained in the AM cell. (d) and (e) describe the two constraints during encoding, where (d) the constraint stipulates that for the same search value, the current through the FeFETs at the same position must either be the same or 0 (i.e., 0 represents OFF-state), and (e) if $\text{FeFET}_{Search11,Store00,2}$ is ON while $\text{FeFET}_{Search11,Store01,2}$ is OFF, it is impossible for $\text{FeFET}_{Search00,Store00,2}$ to be OFF while $\text{FeFET}_{Search00,Store01,2}$ is ON.

Since V_{ds} directly determines the current during conduction, it means that $I_{r_a,c_a,i}$ must be equal to $I_{r_a,c_b,i}$, i.e. same row different column, or one of them must be 0, indicating the i^{th} FeFET is in the OFF-state, as illustrated in Fig. 4(d). We refer to this as the second constraint.

Fig. 4(e) refers to the final constraint, which arises from the multi-level nature of FeFETs and can be expressed as follows: if there exists a $V_{search,11}$ that is greater than $V_{th,00}$ and less than $V_{th,01}$, it is impossible to have a $V_{search,00}$ that is less than $V_{th,00}$ and greater than $V_{th,01}$. In the context of encoding the FeReX circuit, this constraint implies that if $I_{r_a,c_a,i} > 0$ (i.e., conducting) while $I_{r_a,c_b,i} = 0$ (i.e, not conducting), then it is impossible for $I_{r_b,c_a,i} = 0$ while the $I_{r_b,c_b,i} > 0$, as shown in Fig. 4(e). We refer to this as the third constraint.

The CSP has many classical solution methods. Here, we have chosen Backtracking [19] and AC3 [20], [21] for solving the CSP that is bounded by the device and circuit, as shown in Alg. 1. Backtracking is utilized to obtain all solutions that fulfill the second constraint, while AC3 is utilized to swiftly identify solutions that satisfy the third constraint. If the objective is to obtain all possible solutions, AC3 can be substituted with backtracking. The output of the algorithm is the *Feasible Region*, which consists of mutually satisfying items selected from the $DMCurs$ corresponding to each element in the Distance Matrix.

Fig. 5 demonstrates the process of post processing the output *Feasible Region* to obtain the encoding result. Notice that this figure elucidates all the possible search/store combinations for a single FeFET device. During the V_{th} encoding process, the number of ON-state FeFET is tallied in each column. Columns with a higher frequency of ON-state FeFET will correspond to lower V_{th} values. During the V_{search} encoding process, similarly, the occurrence of the OFF-state FeFETs in each row is calculated. Rows with a greater number of OFF-state FeFETs will correspond to lower V_{search} values.

Tab. II displays the encoding results for 2-bit Hamming Distance with the proposed FeReX methodology. FeReX iteratively explores the number of FeFETs within a cell, starting from fewer and gradually increasing, to determine that a 3FeFET-3R cell structure is the optimal solution for the input Distance Matrix. The FeFET is in ON-state only if the V_{t_i} and V_{s_j} (where $i, j \in \{0, 1, 2\}$) satisfy inequality $i < j$. This method has also been extended to compute other distances such as multi-bit Manhattan and multi-bit Euclidean. We leverage results of multi-bit Manhattan and multi-bit Euclidean in the Sec. V.

V. Evaluation & Benchmarking

In this section, we evaluated the FeReX using Cadence Virtuoso to assess the circuit's accuracy, robustness, power consumption, and layout area. The Preisach FeFET model [22] was employed for FeFETs, while the 45nm PTM model [23] was used for all MOSFETs. Wiring parasitics for the 45nm technology node were extracted from DESTINY [24]. The operational amplifier (op-amp) utilized in the simulations was based on the design proposed in the literature [25], scaled down to the 45nm process. The controller of the proposed FeReX is implemented with system verilog and synthesized with

TABLE II
3FeFET-3R 2bit Hamming Distance encoding table

	Store Encoding			Search Encoding					
	$V_{th,FET1}$	$V_{th,FET2}$	$V_{th,FET3}$	$V_{g,FET1}$	$V_{g,FET2}$	$V_{g,FET3}$	$V_{ds,FET1}$	$V_{ds,FET2}$	$V_{ds,FET3}$
"00"	Vt_2	Vt_2	Vt_0	Vs_2	Vs_2	Vs_0	V	V	V
"01"	Vt_2	Vt_0	Vt_2	Vs_1	Vs_0	Vs_2	$2V$	V	V
"10"	Vt_0	Vt_2	Vt_2	Vs_0	Vs_1	Vs_2	V	$2V$	V
"11"	Vt_1	Vt_1	Vt_1	Vs_1	Vs_1	Vs_1	V	V	$2V$

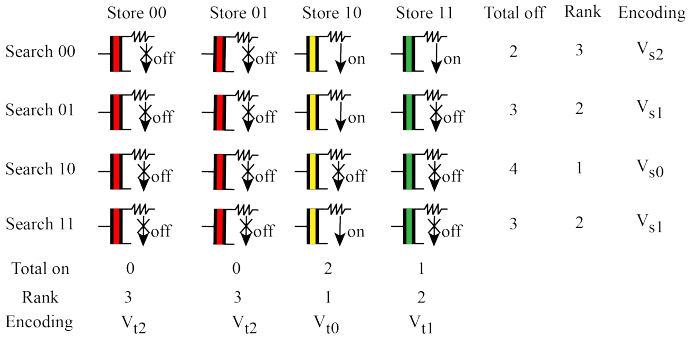

Fig. 5. Encoding the Alg. 1's *Feasible Region* to get the store/search encoding table for a single FeFET device.

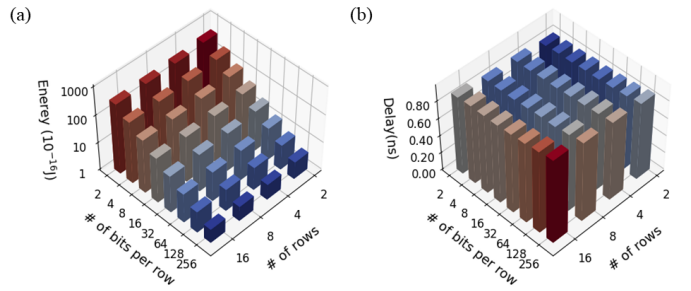

Fig. 6. Search energy and delay for FeReX: (a) Energy consumption per bit per row. (b) Delay trend with varying number of rows and dimensions.

Synopsys Design Compiler with TSMC low-power process development kit (PDK). The total power of the FeReX array controller, including the internal, switching, and leakage power, takes $0.018mW$. The area takes $61\mu m^2$ and the positive slack times are met.

Fig. 6(a) demonstrates that increasing the number of rows in the FeReX can reduce the average energy consumption per bit per row. LTA power consumption grows insignificantly as the number of rows increases, leading to reduced average energy consumption per row as more rows are involved. The ReFeX search delay has two main components. About 60% of the total delay comes from the op-amp's ScL voltage stabilization, which is constrained by the op-amp's slew rate. The remaining delay lies in the LTA circuitry. As shown in Fig. 6(b), the total delay increases gradually as the circuit expands.

To further elucidate the effectiveness of the proposed FeReX, we benchmark it in a bottom-up and top-down manner, where we first conduct Monte Carlo (MC) simulation bottom-up in the context of K-nearest neighbor (KNN), by taking device-to-device variation into account. Then, to further demonstrate the usefulness of the proposed FeReX, we benchmark it with the vector-symbolic architecture (VSA) framework [26], also known as hyperdimensional computing (HDC).

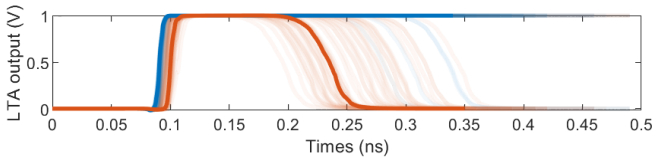

Fig. 7. Monte Carlo simulations considering device-to-device variations: FeReX achieves 90% accuracy when confronted with the worst situation of KNN's search cases in MNIST.

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 [27]
UCIHAR	561	12	6,213	1,554	Physical Activity Monitoring [28]
MNIST	784	10	60,000	10,000	Handwritten Recongition [29]

Fig. 7 illustrates the MC simulations of FeReX with 100 iterations. In the simulations, the device-to-device variation for FeFET threshold voltage was set to 0.054V [8], and the resistance variation for the 1FeFET1R structure was extracted from fabricated data [10] and set to 8%. The results of FeReX at the array level demonstrate 90% search accuracy when comparing distances 5 and 6, representing the most challenging scenarios among KNN's search cases in MNIST. This performance results in only a 0.6% accuracy degradation compared to the software-based implementation.

For top-down benchmarking, we first briefly discuss its advantages and algorithmic flow. In HDC, low dimensional features are first projected to high dimensional representations randomly, enabling *holographicness* across the high dimensional feature vectors. HDC is pre-defined through a set of transparent operations, and due to its holographicness, it has been reported to be robust against hardware noise [30].

The algorithmic flow of HDC can generally be categorized into three steps: first, the data is projected to a high dimension as mentioned above. Second, single-pass training is performed, where the encoded high dimensional vectors that belong to a certain class are aggregated. For higher algorithmic accuracy, iterative training can be performed before storing the trained vectors in the proposed FeReX. Finally, the inference phase that performs classification output the predicted class vector that is

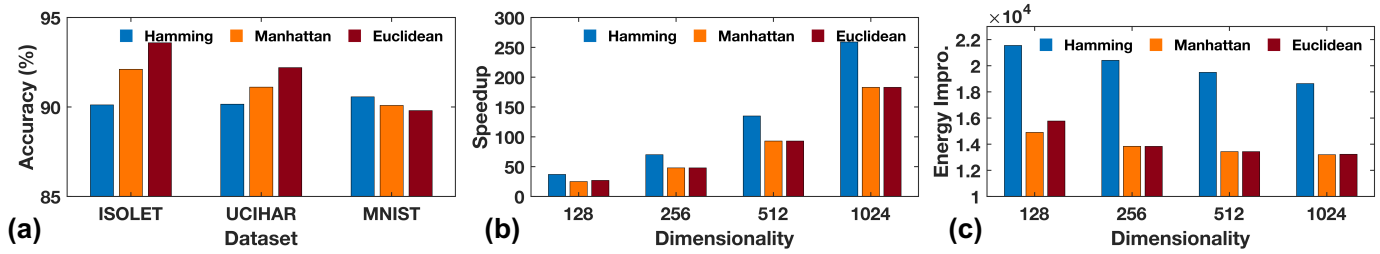

Fig. 8. (a) Classification accuracy with different FeReX distance function. (b) Computation speedup and (c) energy-efficiency improvement over the GPU implementation.

nearest to the query vector with the FeReX distance function.

Here, we benchmark the proposed FeReX in the context of HDC with Nvidia 3090 GPU [31] over three large-scale datasets given in Table III. By extracting the latency of the relevant operations through *Pytorch Profiler* package, the energy is obtained with the Nvidia System Management Interface. Figure 8(a) demonstrates the usefulness of the reconfigurable search engine. Conventional HDC CIM-based accelerator implements Hamming distance, yet different distance metrics may be useful across different datasets, as shown in Figure 8(a). Figure 8(b) and (c) further illustrate the effectiveness of CIM-based AM design, showcasing up to 250x speedup and 10^4 energy improvement over the GPU implementation.

VI. Conclusion

In this paper, we propose FeReX, a FeFET-based associative memory for reconfigurable distance NN search. Based on the observed device and circuit constraints, FeReX judiciously selects write and read voltages for the targeted search distance function in the form of constraint satisfaction problem. Evaluations from bottom-up at the device-to-circuit level with the proposed read/write methodology corroborate the effectiveness of the FeReX, and from top-down, benchmark results illustrate the usefulness of the FeReX. To the best of our knowledge, this is the first that presents a reconfigurable search distance function with the NVM-based AM. We envision that FeReX will pave the way for reconfigurability of the NVM-based circuits and systems.

Acknowledgements

Liu was supported by CoCoSys, one of seven centers in JUMP2.0, an SRC program sponsored by DARPA.

References

[1] A. S. Lele *et al.*, "A heterogeneous rram in-memory and sram near-memory soc for fused frame and event-based target identification and tracking," *IEEE Journal of Solid-State Circuits*, 2023.

[2] X. Peng *et al.*, "Dnn+ neurosim v2. 0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 11, pp. 2306–2319, 2020.

[3] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, no. 11, pp. 521–529, 2019.

[4] H. Li *et al.*, "Sapiens: A 64-kb rram-based non-volatile associative memory for one-shot learning and inference at the edge," *IEEE Transactions on Electron Devices*, vol. 68, no. 12, pp. 6637–6643, 2021.

[5] A. Kazemi *et al.*, "Achieving software-equivalent accuracy for hyperdimensional computing with ferroelectric-based in-memory computing," *Scientific reports*, vol. 12, no. 1, p. 19201, 2022.

[6] A. Kazemi *et al.*, "Fefet multi-bit content-addressable memories for in-memory nearest neighbor search," *IEEE Transactions on Computers*, vol. 71, no. 10, pp. 2565–2576, 2021.

[7] T. Böscke *et al.*, "Ferroelectricity in hafnium oxide: Cmos compatible ferroelectric field effect transistors," in *2011 International electron devices meeting*. IEEE, 2011, pp. 24–5.

[8] T. Soliman *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.

[9] X. Yin *et al.*, "An ultracompact single-ferroelectric field-effect transistor binary and multibit associative search engine," *Advanced Intelligent Systems*, p. 2200428, 2023.

[10] D. Saito *et al.*, "Analog in-memory computing in fefet-based 1t1r array for edge ai applications," in *2021 Symposium on VLSI Technology*. IEEE, 2021, pp. 1–2.

[11] G. Karunaratne *et al.*, "In-memory hyperdimensional computing," *Nature Electronics*, vol. 3, no. 6, pp. 327–337, 2020.

[12] C. Li *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.

[13] S. Shou *et al.*, "See-mcam: Scalable multi-bit fefet content addressable memories for energy efficient associative search," *arXiv preprint arXiv:2310.04940*, 2023.

[14] H. E. Barkam *et al.*, "Hdgim: Hyperdimensional genome sequence matching on unreliable highly scaled fefet," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–6.

[15] S. D. Spetalnick *et al.*, "A 40nm 64kb 26.56 tops/w 2.37 mb/mm 2 rram binary/compute-in-memory macro with 4.23 x improvement in density and¿ 75% use of sensing dynamic range," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65. IEEE, 2022, pp. 1–3.

[16] P.-Y. Chen *et al.*, "Neurosim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 12, pp. 3067–3080, 2018.

[17] K. Ni *et al.*, "Write disturb in ferroelectric fets and its implication for 1t-fefet and memory arrays," *IEEE Electron Device Letters*, vol. 39, no. 11, pp. 1656–1659, 2018.

[18] C.-K. Liu *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.

[19] J. R. Bitner *et al.*, "Backtrack programming techniques," *Communications of the ACM*, vol. 18, no. 11, pp. 651–656, 1975.

[20] A. K. Mackworth, "Consistency in networks of relations," *Artificial intelligence*, vol. 8, no. 1, pp. 99–118, 1977.

[21] R. Soto *et al.*, "A hybrid ac3-tabu search algorithm for solving sudoku puzzles," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5817–5821, 2013.

[22] S. Salahuddin *et al.*, "The era of hyper-scaling in electronics," *Nature Electronics*, vol. 1, no. 8, pp. 442–450, 2018.

[23] R. Vattikonda *et al.*, "Modeling and minimization of pmos nbti effect for robust nanometer design," in *Proceedings of the 43rd annual Design Automation Conference*, 2006, pp. 1047–1052.

[24] M. Poremba *et al.*, "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.

[25] B. H. Kassiri *et al.*, "Slew-rate enhancement for a single-ended low-power two-stage amplifier," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2013, pp. 1829–1832.

[26] D. Kleyko *et al.*, "Vector symbolic architectures as a computing framework for emerging hardware," *Proceedings of the IEEE*, vol. 110, no. 10, pp. 1538–1571, 2022.

[27] "Uci machine learning repository," http://archive.ics.uci.edu/ml/datasets/ISOLET.

[28] D. Anguita *et al.*, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*. Springer, 2012.

[29] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[30] Z. Zou *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.

[31] A. Hernández-Cano *et al.*, "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.