ADVANCED INTELLIGENT SYSTEMS
Open Access
www.advintellsyst.com

An Ultracompact Single-Ferroelectric Field-Effect Transistor Binary and Multibit Associative Search Engine

Xunzhao Yin, Franz Müller, Qingrong Huang, Chao Li, Mohsen Imani, Zeyu Yang, Jiahao Cai, Maximilian Lederer, Ricardo Olivo, Nellie Laleni, Shan Deng, Zijian Zhao, Zhiguo Shi, Yiyu Shi, Cheng Zhuo,* Thomas Kämpfe,* and Kai Ni**

Content addressable memory (CAM) is widely used in associative search tasks due to its parallel pattern matching capability. As more complex and data-intensive tasks emerge, it is becoming increasingly important to enhance CAM density for improved performance and better area efficiency. To reduce the area overheads, various nonvolatile memory (NVM) devices, such as ferroelectric field-effect transistors (FeFETs), are used in CAM design. Herein, a novel ultracompact 1FeFET CAM design that enables parallel associative search and in-memory hamming distance calculation is used, as well as a multibit CAM for exact search using the same CAM cell. The proposed CAM design leverages the 1FeFET1R structure, and compact device designs that integrate the series resistor current limiter into the intrinsic FeFET structure are demonstrated to turn the 1FeFET1R structure into an effective 1FeFET cell. A two-step search operation of the proposed binary and multibit 1FeFET CAM array through both experiments and simulations is proposed, showing a sufficient sensing margin despite unoptimized FeFET device variation. In genome pattern matching applications, using the hyperdimensional computing paradigm, the design results in a 89.9× speedup and 66.5× improvement in energy efficiency over the state-of-the-art alignment tools on GPU.

1. Introduction

In the era of artificial intelligence (AI) and Internet of Things (IoT), the rapidly growing amount of data generated by various machine learning (ML) models, edge devices, and data centers is putting a strain on efficient computational hardware and architectures. The traditional Von Neumann architectures, however, are facing significant energy costs and latency issues due to the massive data transfer between storage and processing units, known as "memory wall" issues. To address this issue, emerging computational accelerators, particularly in-memory computing (IMC) circuits and architectures, have been proposed and studied. IMC aims to tackle the memory wall by reducing the data movement and replacing sequential operations with parallel data analytics, thus improving the performance and energy efficiency of the computational cores.[1–3] Moreover, the utilization of emerging nonvolatile memory (NVM) devices, for example, resistive random access memory (ReRAM), spin torque transfer magnetic random access memory (STT-MRAM), and ferroelectric field effect transistor (FeFET), has further enhanced IMC solutions in terms of area, information density, and energy efficiency when compared to traditional complementary metal oxide semiconductor (CMOS)-based cores.[1,3] Crossbar structures, a type of IMC

X. Yin, Q. Huang, C. Li, Z. Yang, J. Cai, Z. Shi
College of Information Science and Electronic Engineering
Zhejiang University
Hangzhou 310027, China
E-mail: xzyin1@zju.edu.cn

F. Müller, M. Lederer, R. Olivo, N. Laleni, T. Kämpfe
IPMS Fraunhofer Center Nanoelectronic Technologies
01109 Dresden, Germany
E-mail: thomas.kaempfe@ipms.fraunhofer.de

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202200428.

DOI: 10.1002/aisy.202200428

M. Imani
Department of Computer Science
University of California Irvine
Irvine 92697, USA

S. Deng, Z. Zhao, K. Ni
Department of Electrical and Microelectronic Engineering
Rochester Institute of Technology
Rochester 14623, USA
E-mail: kxneen@rit.edu

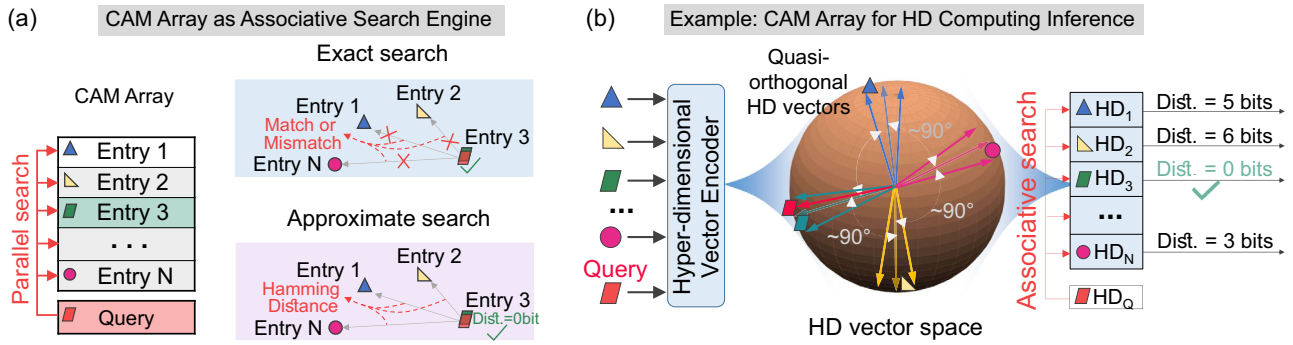
Y. Shi
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame 46556, USA

C. Zhuo
School of Micro-Nano Electronics
Zhejiang University
ZJU-Hangzhou Global Scientific and Technological Innovation Center
Hangzhou 311200, China
E-mail: czhuo@zju.edu.cn

element based on NVM, have been studied for their potential to accelerate the core matrix multiplication operations in data-intensive tasks such as neural networks, signal processing and solving differential equation, etc.[1,3]

Besides matrix multiplications, the search operation over multiple data vectors is also widely used at the core of many applications and improving its efficiency can greatly improve system performance. Content addressable memories (CAMs), as a special type of IMC solutions, can accelerate parallel search operations throughout an entire memory array, making them a promising technology in modern computing platforms.[4–6] Depending on the stored value (i.e., binary, ternary, or multibit), CAMs can be classified as binary CAM (BCAM), ternary CAM (TCAM) with a third "don't care" or wildcard state, or multibit CAM (MCAM).[4,7] When an input query is given, a CAM simultaneously compares each stored memory entry with the input and returns the matching entries, as shown in **Figure 1**a. The search can be performed in either exact mode, where only the entries that match exactly are returned, or approximate mode, where the distances (i.e., Hamming distance for BCAM/TCAM,[5] and a novel distance metric for MCAM[8]) between the stored entries and the input query is calculated, as shown in Figure 1a. In the latter case, the CAMs serve as a distance kernel for various applications.[4,5,8]

One promising application that can greatly benefit from CAM is hyperdimensional computing (HDC), a brain-inspired computing model used for cognition tasks such as image classification and speech recognition.[9–11] In HDC, class vectors are represented as nearly orthogonal hypervectors in a high-dimensional space (e.g., thousands of dimensions and each dimension is independent and identically distributed), as depicted in Figure 1b.[11] The HDC classification process is performed by identifying the closest classes to the input query. CAM can significantly speed up this process by storing class hypervectors (HD_N in Figure 1b and efficiently calculating the distances between the stored class hypervectors and the input search query vector (HD_Q in Figure 1b) through a massively parallel fashion. In addition to cognition tasks, HDC has found broad applications, including genome sequencing. Genome sequencing, which is a common pattern matching problem in bioinformatics, involves searching for entries in the genome library that contain the query sequence. Despite its significance,



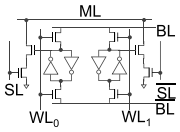
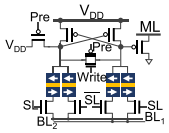
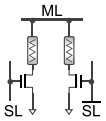
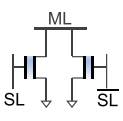
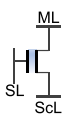
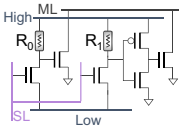
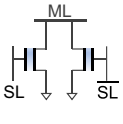
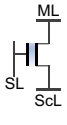
(c)	SRAM	STT-MRAM	ReRAM	FeFET	Proposed
Binary CAM Cell	16 T	10T-4MTJ	2T-2R	2FeFET	1FeFET
Binary Design Schematic					
Multi-Bit CAM Cell	NA	NA	6T-2R	2FeFET	1FeFET
Multi-Bit Design Schematic	NA	NA			

Figure 1. Overview of a CAM-based associative search engine for data-intensive pattern searching applications. a) CAM array features massive parallelism and in-memory computing capabilities and can perform both exact and approximate matching searches for the input query. b) The approximate matching mode, which calculates the distance between the query and stored entries, can find widespread use such as in hyperdimensional computing as an associative memory. c) Overview of existing BCAM and MCAM designs and their respective CAM bit density. The proposed 1FeFET-based CAM design achieves the highest density utilizing only one active memory device.

accelerating the pattern matching process for genome sequencing remains an open challenge. However, HDC has been proposed as a promising solution, as it can transform sequential pattern matching processes into highly parallelizable computation tasks and translate complex pattern distance metrics into hamming distance.[12] These properties make CAM an ideal platform for accelerating genome pattern matching through HDC.

To support complex and data-intensive tasks, such as cognition and genome sequencing, high-density CAM arrays are crucial. However, traditional static random access memory (SRAM)-based CAM designs, which are based on CMOS technology, suffer from high area and power overhead due to the large number of transistors (i.e., 16 transistors) required for a CAM cell and resultant large parasitics. These limitations make traditional SRAM-based CAM designs impractical for widespread use. To address these challenges, two novel approaches have been proposed and studied. The first approach leverages emerging NVM devices to build compact and efficient BCAM/TCAM designs. As shown in Figure 1c, compact CAM designs can be achieved using STT-MRAM (10 T-4MTJ, MTJ: magnetic tunnel junction) or two-terminal resistive memories such as ReRAM or phase-change memory (PCM). The CAM design with the highest density to date uses two FeFETs.[5] These novel approaches show promise for improving system performance, reducing area, and increasing energy efficiency over traditional CMOS-based designs.

The second approach is to explore beyond the conventional binary/ternary CAM designs and exploit the multilevel cell (MLC) capability of NVMs for the design of MCAM. In an MCAM cell, multibit information is encoded, stored, and searched, offering an alternative route to increase information density. Despite being a less explored approach compared to BCAM designs, there have been a few examples demonstrated so far. For instance, the use of resistive memory devices has led to the proposed design of MCAM with 6T-2R,[13] which is different from ReRAM-based binary design. Scalable 2FeFET-based MCAMs have also been demonstrated, featuring a hybrid parallel–serial connection.[14,15] However, their search speed is limited by the serial branch, leading to a slower MCAM. Another example is the parallel 2FeFET CAM cell, which offers a high-speed, universal design that can serve simultaneously as a BCAM/TCAM cell and MCAM cell.[7] The unique advantage of FeFET, along with its superior write energy efficiency and density, makes FeFET-based CAM an attractive option for associative memory.

All the aforementioned CAM designs, whether binary or multibit, require at least two active memory devices. However, a compact design that uses only one memory element to achieve ultimate CAM density is yet to be realized. Such a design, if it existed, would also be preferred to be universal, similar to the 2FeFET CAM cell,[7] which can simultaneously function as a BCAM and MCAM. In this work, we propose an ultimate compact BCAM and MCAM design based on a single FeFET, leveraging its intrinsic three-terminal transistor structure and nonvolatility. Our proposed design is significantly superior to CAM cell designs based on single flash memory transistors,.[16–18] First, the flash memory-based CAM design is a serial design, as the match or mismatch operation is performed bit-by-bit. To search a CAM word with N cells, it would require $2N$ cycles without pipelining or $N+1$ cycles with pipelining, which significantly degrades the search speed of CAMs. Second, our proposed CAM is a universal design that can serve simultaneously as both binary and MCAM. Previous 1 T

CAM designs do not demonstrate this novel concept. Our design is capable of performing hamming distance calculation in-memory, which opens the door for its application in data-centric computation, cognition, and classification tasks. This capability is typically neglected in the prior 1 T CAM literature. Third, the FeFET technology used in this work is more appealing than flash memory-based designs because it is highly energy efficient and CMOS compatible. FeFETs have a much lower write voltage (i.e., ≤ 4 V) than flash and fast write speed (i.e., ≤ 50 ns), and their cell size is smaller due to their better scalability, which is integrated into advanced technology platforms including 28 nm bulk[19] and 22 nm FDSOI.[20] In the following sections, the FeFET device will first be explained, and the 1FeFET universal CAM design will be proposed and validated. Finally, we will leverage the proposed design for genome sequencing applications through HDC.

2. HfO_2-Based FeFET

The discovery of ferroelectricity in doped HfO_2[21] has generated significant interest in its integration into FeFET for high-density, energy-efficient-embedded NVM. FeFET operates by applying a positive/negative pulse to the gate, which sets the ferroelectric polarization toward the channel/gate electrode, and the FeFET threshold voltage V_{TH} to low/high value, respectively. The stored memory state can then be read by sensing the channel current through the application of a read gate bias between the low-V_{TH} (LVT) and high-V_{TH} (HVT) states. With its electric field-driven write mechanism, FeFET exhibits superior write energy efficiency, making it attractive for IMC applications. In this work, FeFET devices were fabricated using a 28 nm-node gate-first high-k metal gate CMOS process on 300 mm silicon wafers. Detailed process information can be found in another study.[19] The gate stack of the fabricated FeFET comprised a polycrystalline Si/TiN (2 nm)/doped HfO_2 (8 nm)/SiO_2 (1 nm)/p-Si structure, as shown in **Figure 2**a. The ferroelectric gate stack process started with the growth of a thin SiO_2-based interfacial layer, followed by the deposition of an 8 nm thick-doped HfO_2 layer. A TiN metal gate electrode was deposited using physical vapor deposition (PVD), followed by the deposition of the poly-Si gate electrode. The source and drain n^+ regions were formed through phosphorous ion implantation and activated through rapid thermal annealing (RTA) at ≈1000 °C, which also resulted in the formation of the ferroelectric orthorhombic phase within the doped HfO_2.

The microstructure of the ferroelectric layer plays a vital role in its application within the FeFET memory. Therefore, we analyzed the microstructure of a planar film using a transmission scanning transmission electron microscope (STEM) equipped with a high-dynamic range detector at each pixel, as demonstrated in Figure 2b. Similar to the transmission Kikuchi diffraction (TKD) technique,[22] which has been utilized to study the granular structure of polycrystalline HfO_2 films, the analysis was performed in transmission, with the detector placed inside the beam. However, this method differs in that it has an

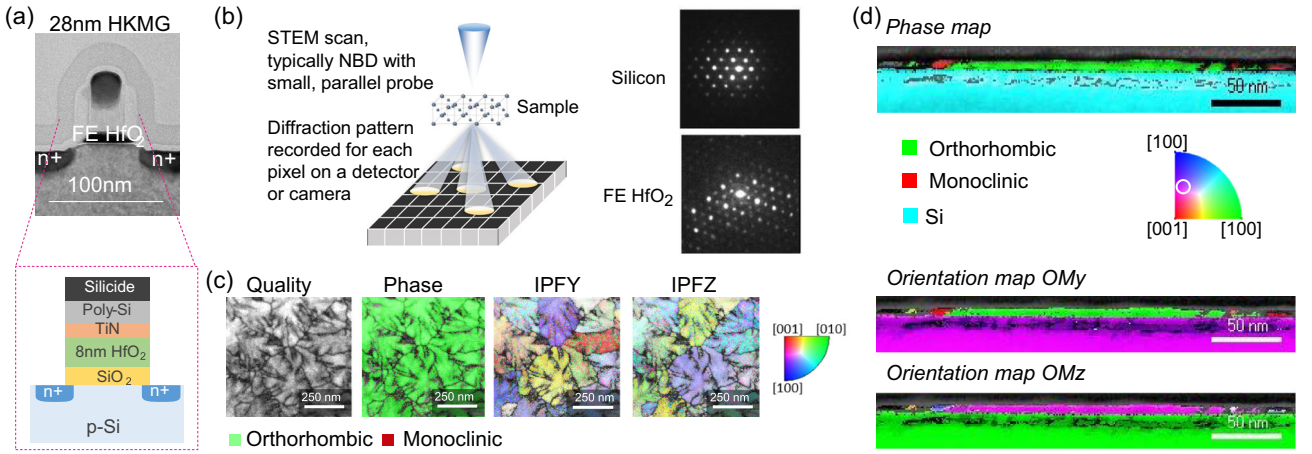


Figure 2. HfO_2-based FeFET. a) Cross-sectional transmission electron microscopy (TEM) and schematic of the FeFET integrated into a 28 nm-high-k metal gate process. b) TEM-based electron diffraction characterization method used to identify the phase and orientation of grains' high spatial resolution. The reference diffraction patterns for silicon and ferroelectric HfO_2 are shown. c,d) Phase map and inverse pole figure/orientation maps extracted from electron diffraction indexation of the hafnium oxide grains in the high-k stack, for an in-plane and cross-sectional view, respectively. The orientation close to [101] of the large polar orthorhombic grain in the cross section can be extracted from the OMz map.

adjustable convergence angle and accelerator voltage, allowing for the tuning of the measured signal between classical electron diffraction and Kikuchi diffraction patterns.[23,24] In this analysis, we utilized the former and indexed the phase and grain orientation based on the extracted diffraction patterns, as shown in Figure 2b. The phase and grain orientation maps are shown in Figure 2c,d for the in-plane and cross-section views, respectively. The results indicate a significant fraction of the highly polar orthorhombic phase in the film, with a small monoclinic phase fraction of less than 5%. The grain orientation is homogeneous within the grains, with a preferred tilted out-of-plane orientation along the [110] axes.

3. Proposal of 1FeFET CAM

3.1. Target FeFET Characteristics for IMC

IMC designs such as the crossbar arrays for matrix-vector multiplications and CAM for bit-wise XNOR operations typically operate in the current domain, as depicted in Figure S1a,b, Supporting Information. The output results of these circuits are determined by the sum of currents from an array of memory devices. In this current mode of operation, it is crucial to have well-controlled ON current variability in FeFET devices to reduce error rates induced by current variations. The ON current variability of a conventional FeFET device with a stored V_{TH} state is directly correlated with the V_{TH} variation, as demonstrated in Figure S1c, Supporting Information. The I_D–V_G characteristics of LVT and HVT states of 60 FeFETs are shown in **Figure 3**a,b. As shown in Figure 3c, a non-negligible device-to-device variation is observed, leading to significant variation in the ON current. Additionally, when the device is read with different gate biases (e.g., V_{SL2} and V_{SL3} in Figure 3b, chosen to turn on the LVT and HVT states, respectively, yielding the same ON current), it is evident that the ON current of the device with LVT is not constant but highly dependent on the gate bias V_G. These findings impose a strict requirement on controlling V_{TH} variation in FeFET devices, which is an ongoing process that involves optimization of integration process, but still faces challenges with FeFET scaling.[20,25] To address these challenges, two approaches are being studied. The first approach is continuous process optimization, such as the control of polycrystalline phase and grain orientation control, among others.

In this work, we explore an alternative method, proposed in another study,[26] to control the ON current variability of FeFET devices. This method involves using a current limiter to ensure that the ON current is independent of both the applied V_G and the stored V_{TH} state. This eliminates the translation of V_{TH} variations into the ON current variations, as demonstrated in Figure S1d, Supporting Information. The implementation of the current limiter is achieved through a series resistor, as shown in Figure 3d. The ON current is determined by the component with the larger resistance value, ensuring that the ON current is dominated by V_D/R_S once FeFET turns ON. Experimental device-to-device measurements on 60 devices with a 1FeFET1R structure were conducted, as shown in Figure 3e. The I_D–V_G characteristic results show that the ON state current is V_G independent with significantly suppressed variability, which can significantly improve the accuracy and robustness of IMC implementations.[26] The extracted ON currents for two read gate voltages, similar to Figure 3c, confirm the effectiveness of the series current limiter in suppressing the ON current variability, as shown in Figure 3f. Moreover, the current limiter is found to be robust against FeFET scaling, as demonstrated in Figure S2a,b, Supporting Information. The results show that even with the gate length L_G scaled down to 80 nm, the current limiter could still suppress the ON current variability and the V_G dependence of ON current.

In order to fully utilize the benefits of the current limiter, the integration of the series resistor with the FeFET is necessary. Previous work[27] proposed using a TiN/SiO_2 tunneling junction-based resistor integrated in the back end of line (BEOL) and connected with FeFET drain, which is effective

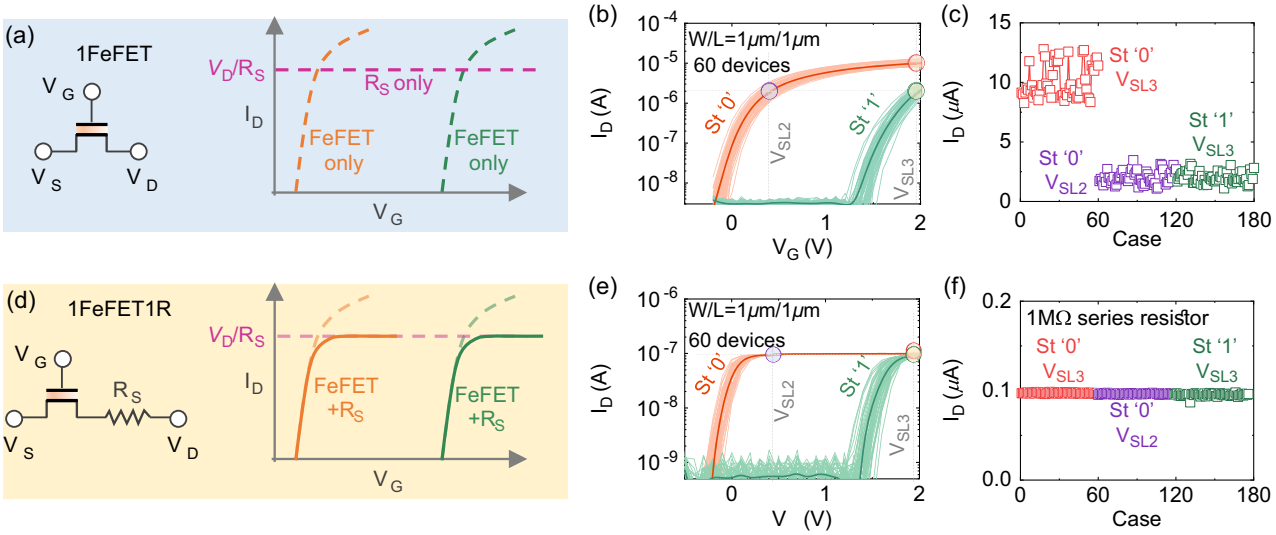


Figure 3. Mitigation of variation and V_G dependency in FeFET I_{ON}. a) A FeFET exhibits the typical transistor behavior, with a constant current flow from the source to drain if the transistor is replaced by a resistor R_S. b) The I_D–V_G curves of 60 FeFETs demonstrate significant V_G dependency and variation in the ON current I_{ON}. V_{SL2} for reading state "0" and V_{SL3} for reading state "1" can yield the same average I_{ON}. c) The corresponding read I_{ON} when applying V_{SL2} and V_{SL3} to both states "0" and "1." d) The implementation of a series current limiter (e.g., a resistor), e) clamps the FeFET I_D–V_G curves, f) making the ON current independent of the FeFET, and solely dependent on the applied drain bias and resistor. $V_D = 0.1$ V for all the measurements.

but difficult to implement. Another design involves using a split-gate structure similar to split-gate embedded FLASH memory,[28] where a conventional transistor is integrated in series with a FeFET, offering increased flexibility through gate-tunable series resistance. However, this design complicates the integration processes. To address these challenges, we propose two designs for integrating the current limiter into the FeFET. The first design involves adopting a Schottky source/drain contact (Figure S2c, Supporting Information), while the second design involves using an underlapped channel region (Figure S2f, Supporting Information). The technology computer-aided design (TCAD) simulation results, shown in Figure S2d, Supporting Information, demonstrated that the Schottky barrier becomes the limiter for carrier transport once the transistor turns ON. The I_D–V_G characteristics at different barrier heights (Figure S2e, Supporting Information) clearly show the effectiveness of Schottky barrier as a current limiter. In the underlapped channel design, the underlapped region limits carrier transport once the gate voltage is high and functions as a current limiter without being controlled by the gate. Figure S2g, Supporting Information, shows the simulated conduction band along the channel, and the I_D–V_G characteristics at different underlap lengths (Figure S2h, Supporting Information) demonstrate the effectiveness of underlapped region in limiting the ON current. These designs are by no means optimal, but they demonstrate that the FeFET-based design for IMC implementation does not have to follow the conventional FeFET practices and that the design space remains largely unexplored.

3.2. Demonstration of Binary CAM

Thanks to its inherent transistor structure, a FeFET with V_G independent ON current can function as both a BCAM and a MCAM simultaneously. This ultracompact CAM cell comprises a single FeFET, which can be fabricated from an 1FeFET1R structure as discussed previously. The proposed design leverages the unique property of the FeFET V_{TH} state, which can be identified through a two-step search scheme. In the first step, a search voltage is applied below V_{TH}, inducing negligible I_{OFF} leakage current. In the second step, a search voltage above V_{TH} is applied, leading to high I_{ON} current. If a low/high I_D flowing through the FeFET is sensed in the first/second search step, respectively, the FeFET V_{TH} state can be uniquely identified. By encoding information into FeFET V_{TH} state and properly choosing the two-step search voltages (i.e., read V_G), either a BCAM or MCAM can be realized. The BCAM leverages the binary state of FeFET and performs hamming distance calculations, while the MCAM leverages the multilevel state of FeFET and performs exact search. In this illustration, we describe the operation principles of a BCAM design using an 1FeFET structure per cell.

Figure 4 presents a schematic of the 1FeFET CAM design and our proposed two-step search scheme for the BCAM search function. Based on the device binary storage level, a state "0" can be programmed into the FeFET in a CAM cell by setting the device to the LVT state, while a state "1" is written by programming it to the HVT state. In the first step, V_{SL1}, which is below V_{TH1} of the LVT state, is applied to the FeFET gate in the CAM cells to search for bit "0". On the other hand, V_{SL2}, which lies between the V_{TH1} of LVT and V_{TH2} of HVT, is applied to search for bit "1", as shown in Figure 4. A match occurs when FeFETs storing "0" are applied with V_{SL1} (denoted as St0Sr0) or when FeFETs storing "1" are applied with V_{SL2} (denoted as St1Sr1). All search voltages (regardless of searching for "0" or "1") in this step are below the respective V_{TH}s, resulting in a negligible matchline (ML) (i.e., I_{ML}) compared to the device on current I_{ON}. A mismatch, causing a high I_{ML}, occurs when one or more FeFETs storing "0"

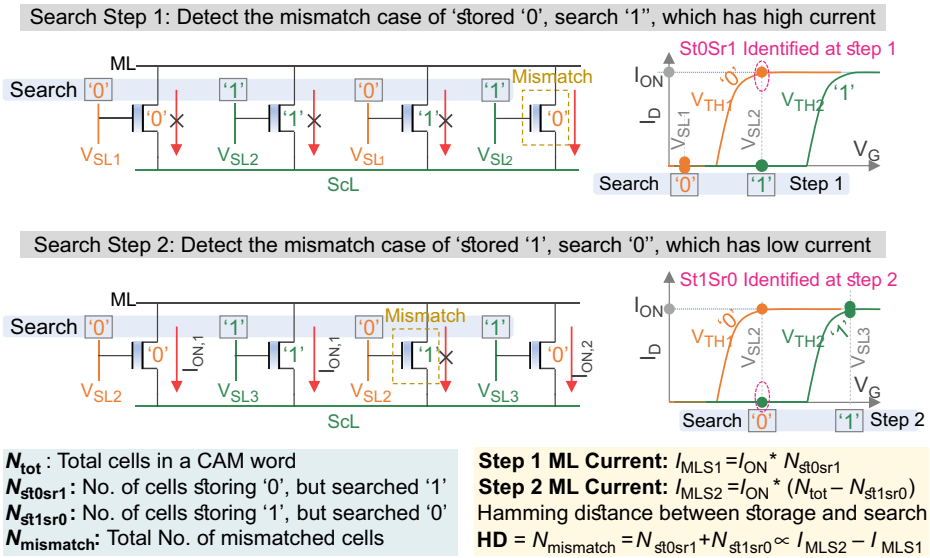


Figure 4. Two-step search in the 1FeFET BCAM with V_G independent ON current. a) The first step identifies the "St0Sr1" mismatch between the stored entries and query by detecting a high ML current. b) The second step identifies the "St1Sr0" mismatch by detecting a low ML current. The difference between the ML currents in the two steps is proportional to the Hamming distance between the stored and query data.

are searched with bit "1" by applying V_{SL2} to the FeFETs in the LVT state (denoted as St0Sr1). The number of St0Sr1 mismatch cases (denoted as N_{st0sr1}) is linearly proportional to I_{ML}. Therefore, the first search step identifies the St0Sr1 mismatch condition. For exact search, the presence of any St0Sr1 instances ($N_{st0sr1} \neq 0$) means a mismatch, while for approximate search where hamming distance is calculated, the ML current I_{ML} in the first step, I_{MLS1}, can be used to determine the number of mismatch bits, N_{st0sr1}.

For the second search step, V_{SL2} and V_{SL3}, above V_{TH2} of the HVT state, are applied to represent searching for bit "0" and "1", respectively, as shown in Figure 4. In this step, three cases conduct the ON current I_{ON}: FeFETs storing bit "0" are applied with V_{SL2} (i.e., St0Sr0), FeFETs storing "1" are applied with V_{SL3} (i.e., St1Sr1), and FeFETs storing "0" are applied with V_{SL3} (i.e., St0Sr1). The only mismatch case with a low I_{ML} is when V_{SL2} is applied to the stored HVT state, that is, when FeFETs storing "1" are searched with bit "0" (denoted as St1Sr0). Sensing the ML current in the second step, I_{MLS2}, detects the number of St1Sr0 mismatch cases (denoted as N_{st1sr0}), as shown in Figure 4. The total mismatch bits of the CAM word, that is, the hamming distance between the stored word and the input query, can be calculated by adding N_{st0sr1} and N_{st1sr0}, which is linearly proportional to the current difference I_{MLS2}–I_{MLS1}.

The proposed two-step search scheme for the BCAM design with Figure 4 is not feasible with traditional FeFET devices, as demonstrated in Figure S3, Supporting Information. While the first step to identify the presence of St0Sr1 is robust due to its sole conducting I_{ON}, the second step to detect St1Sr0 presents a challenge. In this step, the cells that are in the St0Sr1 state conduct a higher current than those in the St0Sr0 and St1Sr1 states. This leads to a significant overlap between the I_{MLS2}'s for different N_{st1sr0} values, making it difficult to differentiate between them, as shown in Figure S3, Supporting Information. This analysis highlights the importance of incorporating a current limiter to mitigate the dependency and variability of ON current on V_G.

The proposed BCAM design has been validated through both experiments and simulations. **Figure 5** provides experimental verification of the proposed BCAM array, which has a word length of 8 bits. Figure 5a illustrates the experiment setup of the CAM array, in which a series resistor is connected with each FeFET in series as a current limiter. All CAM cells are connected in parallel, and the current–voltage characteristic of each cell is presented in Figure 5b. The results clearly demonstrate the presence of a sufficiently large memory window, as well as the independence of ON current I_{ON} from V_G and the suppression of its variation.

Two test cases are considered in this study, with case I representing all cells storing the bit "0" and case II representing all cells storing the bit "1". Figure 5c,d shows the ML currents for case I, with I_{MLS1} and I_{MLS2} representing the first and second search steps, respectively. The I_{MLS1} in the first search step increases linearly with the number of St0Sr1 (i.e., N_{St0Sr1}), demonstrating the consistency with the design principle illustrated in Figure 4. By locating the decision reference between $N_{St0Sr1} = 0$ and $N_{St0Sr1} = 1$, exact match detection can be achieved, and sensing the I_{MLS1} directly allows for the detection of N_{St0Sr1}. In the second search step, as there are no St1Sr0 conditions in case I, all 8 cells conduct I_{ON}, resulting in a full match I_{MLS2} measurement, as shown in Figure 5d. Figure 5e,f shows the measured I_{MLS1} for the first search step and I_{MLS2} for the second search step in case II, respectively. In this case, as no St0Sr1 conditions exist, the I_{MLS1} is low and below the exact match decision boundary, as depicted in Figure 5e. The I_{MLS2} exhibits a linear relationship with the number of St1Sr0 (i.e., N_{St1Sr0}), allowing for the detection of the other mismatch cases in addition to the St0Sr1 condition shown in Figure 5c. These measurements serve as experimental validation of the proposed BCAM design.

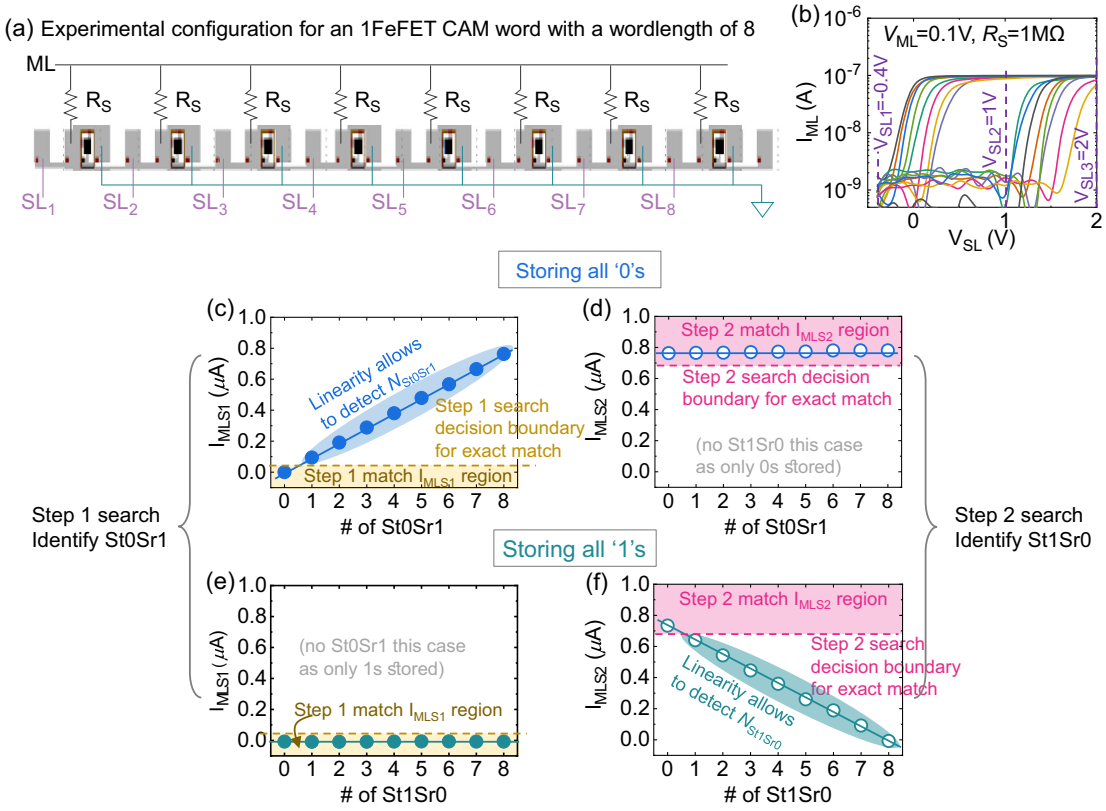


Figure 5. Experimental verification of 1FeFET CAM word with a word length of 8. a) Experimental setup for the measurement; b) The I_D–V_G of the 8 1FeFET1R devices. c,d,e,f) the ML current in step 1/step 2 search when all cells store "0" and "1", respectively. Successful operation of the array is demonstrated. Note that all eight FeFETs have different sizes, for example, W/L = 80 nm/80 nm, 100 nm/200 nm, 80 nm/1 μm, 200 nm/100 nm, 240 nm/240 nm, 500 nm/80 nm, 500 nm/240 nm, and 1 μm/1 μm, and ON currents. By incorporating the series current limiter, the proposed 1FeFET CAM is able to work even under the worst scenario where all the devices have different ON currents.

To further demonstrate the potential of the proposed BCAM design, simulations were conducted using a SPICE Monte Carlo method with a compact model of FeFET that incorporates experimentally measured device variations shown in Figure 3b. The simulations were carried out for a CAM array with a word length of 64 cells per word, which is beyond the limit of the experiment setup. For comparison purposes, a simulation of a conventional 1FeFET CAM array with only two cells per word and without a current limiter was also performed, as shown in Figure S4a, Supporting Information. The results of the simulation demonstrate that while the first step in the search process remains robust, the second step fails due to the V_G dependent ON current and its significant variation. However, with the addition of a current limiter, Figure S4b, Supporting Information, shows that the proposed BCAM design can successfully operate in a CAM array with 64 cells per word, highlighting its great potential.

3.3. Demonstration of Multibit CAM

As previously noted, the proposed CAM cell design can be used as both a BCAM and MCAM cell. The two-step search scheme is used to determine whether the stored V_{TH} state matches the encoded search information. During the first step, the below-V_{TH} search identifies any mismatch cells with a V_{TH} above the search value. During the second step, the above-V_{TH} detects any mismatch cells with a V_{TH} below the search value. If all cells match, both the I_{MLS1} in step 1 and the I_{MLS2} in step 2 will be low and high, respectively. This search principle enables the design of a MCAM with exact search functionality. Furthermore, the proposed design methodology is scalable and can be applied to FeFETs with any number of distinguishable V_{TH} states, given that the device variation is within acceptable limits.

In **Figure 6**, the operation principle is illustrated using 2-bit-per-cell example. Without loss of generality, assuming that all cells store bits "01", corresponding to the second lowest V_{TH} state, the search voltages representing the search bits "00", "01", "10", and "11" are applied to each cell, respectively. During step 1 search, as shown in Figure 6a, a below-V_{TH} search, that is, V_{SL3L} below the V_{TH} corresponding to state "10" and above the V_{TH} corresponding to state "01" is applied to search bit "10". This way, any mismatched cell that is applied by a search voltage above the V_{TH} corresponding to the stored state, that is, the "above-V_{TH}" scenario in this case, represented by search bits "10" and "11", will conduct a high ON current I_{ON}. For step 2 search, an above-V_{TH} search voltage, that is, V_{SL3h} above the V_{TH} corresponding to state "10" and below the V_{TH} corresponding to state "11" is applied to search bit "10." In this case, the only mismatched cells searched with a search voltage below the V_{TH}

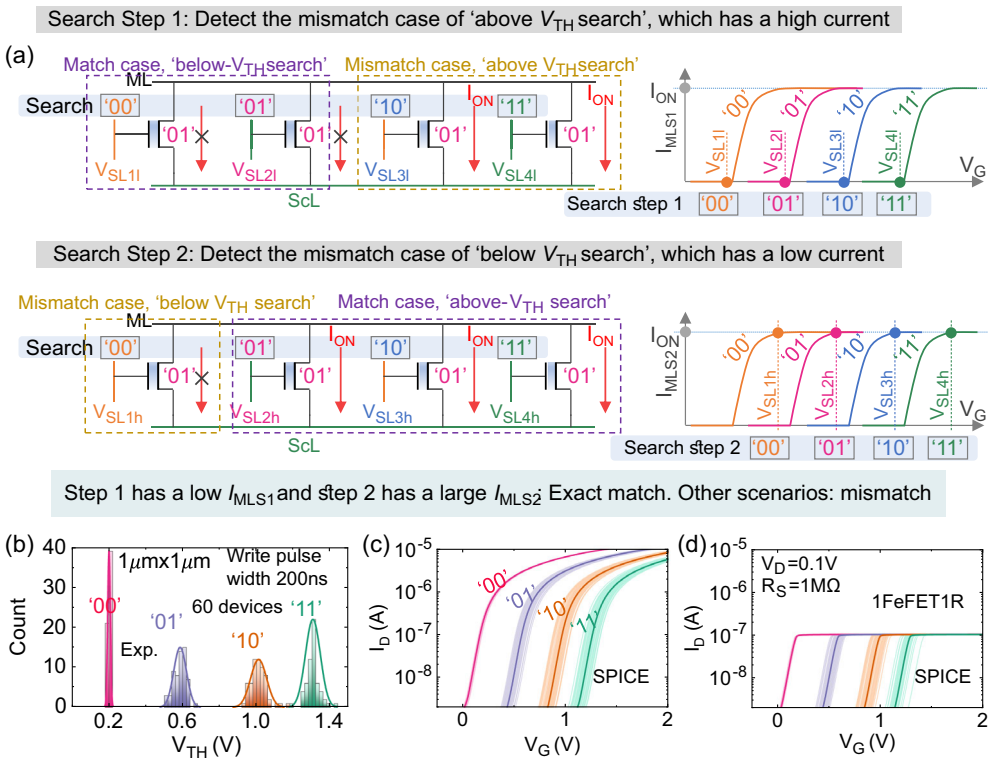


Figure 6. 1FeFET CAM cell for MCAM with two-step exact search. a) Search step 1/step 2 detects the "above-V_{TH}"/"below-V_{TH}" mismatch case, respectively. b) Two bits/cell measured in FeFETs. c,d) SPICE simulation of the 1FeFET/1FeFET1R with measured V_{TH} variation.

corresponding to the stored state, that is, "below-V_{TH}" scenario represented by search bits "00", will yield a low cell current. All other cells conduct a high I_{ON}. The total ML current can then be used to determine whether a "below-V_{TH}" mismatch exists. By combining the results of the two-step ML current, the exact match in the MCAM words can be identified by a low I_{MLS1} in step 1 and a high I_{MLS2} ($\approx I_{ON} \times N$, N is word length) in step 2. Any other ML current combination indicates a mismatch.

We validated the design of the MCAM, which stores two bits per cell, through a combination of experiments and SPICE simulations. The four V_{TH} states representing the stored two bits in FeFETs were demonstrated across 60 devices utilizing the partial polarization switching of a ferroelectric material modulated with pulse amplitudes,[29] as shown in Figure 6b. It is commonly observed that the variation is smaller for extreme states and larger for intermediate states. This is due to the fact that the typical write pulse for extreme states (i.e., lowest and highest V_{TH}) ensures saturated polarization switching for the two states, resulting in low variation. On the other hand, for intermediate states, the write pulse applied cannot finely tune the amount of polarization switched, leading to larger variation.[30] This distribution has also been reported in other literature.[31,32] For this MLC write operation, we used here a gradual erase scheme. We start from programming LVT state ("00" state), which is the extreme state, and then applying different write pulses to switch the device to intermediate states. Since the intermediate states (i.e., "01", "10", "11" states) are in the steeper switching transition and sensitive to write pulses, the corresponding write pulses then result in larger variation. The resulting V_{TH} distributions were incorporated into a FeFET compact model,[33] and Monte Carlo SPICE simulations were performed. The I_D–V_G characteristics of the FeFETs, shown in Figure 6c,d, demonstrate that the ON current variability and V_G dependency for all the V_{TH} states are suppressed after incorporating a series current limiter. This cell design enables the MCAM. The simulation results for a MCAM word with 64 cells per word and 2 bits per cell are shown in Figure S5, Supporting Information. Without loss of generality, all cells store bits "01" and the worst-case search scenarios are considered, where only one cell mismatches with the search query. The results show that if one cell is searched with bits "00", that is, "below-V_{TH}" scenario, this mismatch is not detected in step 1 search, but it shows up in the step 2 search. If one cell is searched with bits "10" or "11", both corresponding to "above-V_{TH}" scenario, these mismatches are detected in step 1 search, which results in a high I_{ON}. Only when all the cells exactly match, as shown in Figure S5b, Supporting Information, step 1 search yields a low I_{MLS1} and step 2 search causes a high I_{MLS2} ($\approx I_{ON} \times 64$), indicating a match scenario. These results validate the proposed MCAM array function even with the current, non-fully optimized FeFET.

4. Benchmarking and Application of 1FeFET CAM Array

Figure 7a depicts the trend of CAM cell footprint scaling based on various emerging devices, in addition to CMOS. Compared to existing CAM designs, which require a minimum of two devices

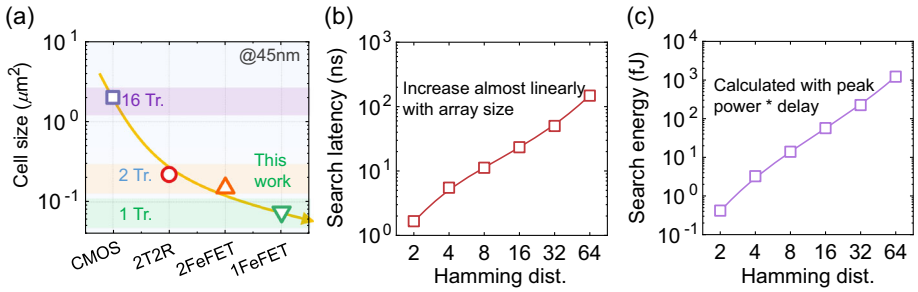


Figure 7. a) Cell-size comparison between 1FeFET CAM, CMOS, 2 T-2RRAM, and 2FeFET CAM designs. The 1FeFET CAM cell is the most compact. Search b) latency and c) energy of 1FeFET CAM array for one-search step with different hamming distance thresholds. 1FeFET-1R CAM cell is considered for the benchmarking, where the series resistance R_s value is set to 1 MΩ. The search latency is measured from the time point when the search voltage is applied to the time point when the sense amplifier detects the voltage output that corresponds to the maximum Hamming distance, that is, word length. The search energy is the product of the peak power consumption of the CAM array and the search latency.

per cell, our 1FeFET CAM design, featuring the proposed two-step search scheme, offers the most compact CAM cell area efficiency and flexible BCAM and MCAM functionalities. Furthermore, with the area-per-bit metric, our 1FeFET CAM design can attain an even smaller size when MCAM is enabled, which is half that of the design presented in another study.[7] To measure the ML currents I_{ML} in the two-step search scheme of our proposed CAM designs, we use a thermometer-code analog-to-digital converter (ADC) as the array sense amplifier. The design converts the ML currents into output voltages, as depicted in Figure S6, Supporting Information. When the 1FeFET CAM is configured as a BCAM, the output voltages of the sense amplifier, guided by the proposed design principles and appropriate reference currents, directly indicate the hamming distance between the input query and the stored words, as shown in Figure 4. When the 1FeFET CAM is configured as a MCAM, the output voltages indicate the match/mismatch condition, as demonstrated in Figure S5, Supporting Information. Note that the sense amplifier adopts a serial design, resulting in a search delay and energy that is proportional to the number of stages (i.e., word length), as shown in Figure 7. If a search delay that is independent of the word length is desired, parallel designs for sensing can be used, although at the cost of increased area and power consumption.[34]

The proposed 1FeFET CAM design is evaluated for its efficiency beyond the array level using it as an associative search engine for HDC in multiple-genome sequencing tasks.[35] In genome sequencing, a query DNA sequence represented by nucleotide bases, that is, A, C, G, T, is searched in a reference DNA string that consists of millions of DNA bases to identify the presence of the query sequence in the reference sequence and accelerate DNA alignment techniques.[36,37] Figure S8a, Supporting Information, illustrates the HDC architecture that parallelizes the genome-sequencing tasks. The sequences from genome databases (*E.coli*,[38] Human CHR14,[38] and COVID-19)[39] are encoded and stored in the associative memory (Figure S8c, Supporting Information), and a new genome query is encoded and searched across multiple CAM banks in parallel. The memory entries with a hamming distance within a threshold with the query are identified, as shown in Figure S8b, Supporting Information. Figure S8d, Supporting Information, shows the benchmarking results for the proposed associative memory architecture for HDC genome-sequencing tasks, revealing that the proposed search engine can achieve on average 89.9× (71.9×) faster and 66.5× (30.7×) higher energy efficiency than state-of-the-art alignment tools NVBIO (GPU-BLAST).[40] Additionally, the use of 1FeFET CAM for HDC application demonstrates the benefits of hardware-algorithm codesign by exploiting the superior robustness of HDC against device variations. Figure S9a,b, Supporting Information, shows the error rates of hamming distance computations for the proposed CAM arrays with word lengths of 32 and 64, respectively, when the variations of both FeFETs and series current limiters are included. Given the experimentally reported 8% resistor variation,[27] the error rate of hamming distance calculation in an array with a word length of 32 can reach 5% when the hamming distance detection threshold is set to be 3-bits. However, even if such error rate is introduced during the hypervector search in HDC, the resulting genome-sequencing error is still low, at only 2% and 0.2% for hypervector dimensions of 512 and 4,000, respectively, as shown in Figure S9c,d, Supporting Information. This highlights the benefits of the proposed 1FeFET CAM in HDC tasks, as it offers high density while taking advantages of HDC's robustness against device variations.

5. Conclusion

In summary, we present a novel 1FeFET CAM design that is both ultracompact and scalable. The design enhances the search function and improves CAM density for low-power pattern matching application through the HDC paradigm. Our design integrates a series resistor into the FeFET structure to minimize the fluctuation of ON state current and features a two-step search operation. Our experiments demonstrate that the 1FeFET CAM array performs both approximate and exact searches, acting as a binary CAM for hamming distance computation and a multibit CAM for improved CAM density. Furthermore, our results show significant improvements in latency and energy efficiency compared to state-of-the-art hardware, emphasizing the potential of the proposed 1FeFET CAM design as a powerful associative search memory engine.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported in part by NSFC No 62104213. This work has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Belgium, Germany, Netherlands, Portugal, Spain, Switzerland.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

X.Y., T.K., and K.N. proposed and supervised the project. F.M., M.L., R.O., N.L., S.D., and Z.Z. performed the experimental verification of the proposed design. Q.H., C.L., Z.Y., J.C., Z.S., Y.S., and C.Z. conducted SPICE simulations and verification. M.I. performed system-level benchmarking. All authors contributed to write-up of the manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

associative search, compute-in-memory, content addressable memory, ferroelectric field effect transistor, hyperdimensional computing

Received: December 9, 2022
Revised: February 28, 2023
Published online:

[1] D. Ielmini, H.-S. P. Wong, *Nat. Electron.* **2018**, *1*, 333.
[2] N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L.-Y. Chen, B. Zhang, P. Deaville, *IEEE Solid-State Circuits Mag.* **2019**, *11*, 43.
[3] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, *Nat. Nanotechnol.* **2020**, *15*, 529.
[4] X. S. Hu, M. Niemier, A. Kazemi, A. F. Laguna, K. Ni, R. Rajaei, M. M. Sharfi, X. Yin, in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2021**, pp. 533–536.
[5] K. Ni, X. Yin, A. F. Laguna, S. Joshi, S. Duenkel, M. Trentzsch, J. Müller, S. Beyer, M. Niemier, X. S. Hu, et al., *Nat. Electron.* **2019**, *2*, 521.
[6] R. Karam, R. Puri, S. Ghosh, S. Bhunia *Proc. IEEE* **2015**, *103*, 1311.
[7] X. Yin, C. Li, Q. Huang, L. Zhang, M. Niemier, X. S. Hu, C. Zhuo, K. Ni, *IEEE Trans. Electron Devices* **2020**, *67*, 2785.
[8] A. Kazemi, M. M. Sharifi, A. F. B. Laguna, F. Muller, X. Yin, T. Kampfe, M. Niemier, X. S. Hu, *IEEE Trans. Comput.* **2021**.
[9] T. F. Wu, H. Li, P.-C. Huang, A. Rahimi, G. Hills, B. Hodson, W. Hwang, J. M. Rabaey, H.-S. P. Wong, M. M. Shulaker, et al., *IEEE J. Solid-State Circuits* **2018**, *53*, 3183.
[10] M. Imani, X. Yin, J. Messerly, S. Gupta, M. Niemier, X. S. Hu, T. Rosing, *IEEE Trans. Comput. Aid. Design Integr. Circuit. Syst.* **2019**, *39*, 2422.
[11] L. Ge, K. K. Parhi, *IEEE Circuit. Syst. Mag.* **2020**, *20*, 30.
[12] Y. Kim, M. Imani, N. Moshiri, T. Rosing, in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, Piscataway, NJ **2020**, pp. 115–120.
[13] C. Li, C. E. Graves, X. Sheng, D. Miller, M. Foltin, G. Pedretti, J. P. Strachan, *Nat. Commun.* **2020**, *11*, 1.
[14] T. Hanyu, H. Kimura, M. Kameyama, in *Proc. 1999 29th IEEE Int. Symp. on Multiple-Valued Logic (Cat. No. 99CB36329)*, IEEE, Piscataway, NJ **1999**, pp. 30–35.
[15] C. Li, F. Müller, T. Ali, R. Olivo, M. Imani, S. Deng, C. Zhuo, T. Kämpfe, X. Yin, K. Ni, in *IEEE Inter. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2020**, pp. 29–33.
[16] T. Hanyu, N. Kanagawa, M. Kameyama, *IEEE J. Solid-State Circuits* **1996**, *31*, 1669.
[17] T. Hanyu, N. Kanagawa, M. Kameyama, in *IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers, ISSCC,* IEEE, Piscataway, NJ **1996**, pp. 264–265.
[18] T. Hanyu, N. Kanagawa, M. Kameyama, *Comput. Electr. Eng.* **1997**, *23*, 407.
[19] M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck et al. in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2016**, pp. 11–15.
[20] S. Dünkel, M. Trentzsch, R. Richter, P. Moll, C. Fuchs, O. Gehring, M. Majer, S. Wittek, B. Müller, T. Melde, et al. in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2017**, pp. 19–27.
[21] T. Böscke, J. Müller, D. Bräuhaus, U. Schröder, U. Böttger, *Appl. Phys. Lett.* **2011**, *99*, 102903.
[22] M. Lederer, T. Kämpfe, R. Olivo, D. Lehninger, C. Mart, S. Kirbach, T. Ali, P. Polakowski, L. Roy, K. Seidel, *Appl. Phys. Lett.* **2019**, *115*, 222902.
[23] C. Ophus, *Microsc. Microanal.* **2019**, *25*, 563.
[24] M. Lederer, A. Reck, K. Mertens, R. Olivo, P. Bagul, A. M. Kia, B. Volkmann, T. Kämpfe, K. Seidel, L. M. Eng, *Appl. Phys. Lett.* **2021**, *118*, 012901.
[25] S. Beyer, S. Dünkel, M. Trentzsch, J. Müller, A. Hellmich, D. Utess, J. Paul, D. Kleimaier, J. Pellerin, S. Müller et al. in *IEEE Int. Memory Workshop (IMW)*, IEEE, Piscataway, NJ **2020**, pp. 1–4.
[26] T. Soliman, F. Müller, T. Kirchner, T. Hoffmann, H. Ganem, E. Karimov, T. Ali, M. Lederer, C. Sudarshan, T. Kämpfe, et al. in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2020**, pp. 29–32.
[27] D. Saito, T. Kobayashi, H. Koga, N. Ronchi, K. Banerjee, Y. Shuto, J. Okuno, K. Konishi, L. Di Piazza, A. Mallik, et al. in *Symp. on VLSI Technology*, IEEE, Piscataway, NJ **2021**, pp. 1–2.
[28] R. Richter, M. Trentzsch, S. Dünkel, J. Müller, P. Moll, B. Bayha, K. Mothes, A. Henke, M. Mazur, J. Paul, et al. in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2018**, pp. 18–25.
[29] M. Jerry, P.-Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, S. Datta, in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2017**, pp. 6–12.
[30] S. Chatterjee, S. Thomann, K. Ni, Y. S. Chauhan, H. Amrouch, *IEEE Trans. Electron Devices* **2022**, *69*, 5316.
[31] T. Ali, P. Polakowski, K. Kühnel, M. Czernohorsky, T. Kämpfe, M. Rudolph, B. Pätzold, D. Lehninger, F. Müller, R. Olivo, et al. in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2019**, pp. 28–37.
[32] C.-Y. Liao, Z.-F. Lou, C.-Y. Lin, A. Senapati, R. Karmakar, K.-Y. Hsiang, Z.-X. Li, W.-C. Ray, J.-Y. Lee, P.-H. Chen, et al. in *Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2022**, pp. 36–46.
[33] K. Ni, M. Jerry, J. A. Smith, S. Datta, in *IEEE Symp. on VLSI Technology*, IEEE, Piscataway, NJ **2018**, pp. 131–132.

[34] H. Jiang, W. Li, S. Huang, S. Cosemans, F. Catthoor, S. Yu, *IEEE Des. Test* **2021**.
[35] M. Imani, S. Gupta, Y. Kim, T. Rosing, in *ACM/IEEE 46th Annual Int. Symp. on Computer Architecture (ISCA)*, IEEE, Piscataway, NJ **2019**, pp. 802–815.
[36] P. Compeau, P. Pevzner, in *Bioinformatics Algorithms: An Active Learning Approach*, Vol. *1*, Active Learning Publishers, La Jolla, CA **2015**.
[37] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, *BMC Bioinf.* **2009**, *10*, 1.
[38] NIH National Library of Medicine, https://www.ncbi.nlm.nih.gov/sars-cov-2/ (accessed: December 2020).
[39] NIH SARS-COV-2 Data, https://www.ncbi.nlm.nih.gov/datasets/docs/command-line-virus/ (accessed: January 2021).
[40] P. D. Vouzis, N. V. Sahinidis, *Bioinformatics* **2011**, *27*, 182.