Reconfigurable Compute-In-Memory on Field-Programmable Ferroelectric Diodes

Xiwen Liu
University of Pennsylvania
John Ting
University of Pennsylvania
Yunfei He
University of Pennsylvania
Merrilyn Mercy Fiagbenu
University of Pennsylvania
Jeffrey Zheng
University of Pennsylvania
Dixiong Wang
University of Pennsylvania
Jonathan Frost
University of Pennsylvania
Paria Musavigharavi
University of Pennsylvania
Giovanni Esteves
Sandia National Laboratories
Kim Kisslinger
Brookhaven National laboratory
Surendra Anantharaman
University of Pennsylvania
Eric Stach
University of Pennsylvania https://orcid.org/0000-0002-3366-2153
Roy Olsson
University of Pennsylvania
Deep Jariwala (dmj@seas.upenn.edu)
University of Pennsylvania https://orcid.org/0000-0002-3570-8768

Article

Keywords: Compute in memory, ferroelectric diode, ternary content-addressable memory, neural network, nonvolatile, reconfigurable architecture, parallel search

Posted Date: February 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1328791/v1

Reconfigurable Compute-In-Memory on Field-Programmable Ferroelectric Diodes

Xiwen Liu,[a] John Ting,[a] Yunfei He,[a] Merrilyn Mercy Adzo Fiagbenu,[a] Jeffrey Zheng,[b] Dixiong Wang,[a] Jonathan Frost,[a] Pariasadat Musavigharavi,[a,b] Giovanni Esteves,[d] Kim Kisslinger,[e] Surendra B. Anantharaman,[a] Eric A. Stach,[b,c] Roy H. Olsson III,[a]* Deep Jariwala[a]*

[a] Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA

[b] Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA, USA

[c] Laboratory for Research on the Structure of Matter, University of Pennsylvania, Philadelphia, PA, USA

[d] Microsystems Engineering, Science and Applications (MESA), Sandia National Laboratories, Albuquerque, New Mexico, USA

[e] Brookhaven National Laboratory, Center for Functional Nanomaterials, Upton, NY, USA

*Corresponding author: rolsson@seas.upenn.edu, dmj@seas.upenn.edu

Abstract

The deluge of sensors and data generating devices has driven a paradigm shift in modern computing from arithmetic-logic centric to data centric processing. At a hardware level, this presents an urgent need to integrate dense, high-performance and low-power memory units with Si logic-processor units. However, data-heavy problems such as search and pattern matching also require paradigm changing innovations at the circuit and architecture level to enable compute in memory (CIM) operations. CIM architectures that combine data storage yet concurrently offer low-delay and small footprint are highly sought after but have not been realized. Here, we present Aluminum Scandium Nitride (AlScN) ferroelectric diode (FeD) memristor devices that allow for storage, search and neural network-based pattern recognition in a transistor-free architecture. Our devices can be directly integrated on top of Si processors in a scalable, back-end-of-line process. We leverage the field-programmability, non-volatility and non-linearity of FeDs to demonstrated circuit blocks that can support search operations in-situ memory with search delay times < 0.1 ns and a cell footprint < 0.12 μm^2. In addition, we demonstrate matrix multiplication operations with 4-bit operation of the FeDs. Our results highlight FeDs as promising candidates for fast, efficient, and multifunctional CIM platforms.

KEYWORDS: Compute in memory, ferroelectric diode, ternary content-addressable memory, neural network, nonvolatile, reconfigurable architecture, parallel search.

Introduction

The convergence of big-data with artificial intelligence (AI) has led to multiple emerging technologies across a range of computing applications [1-3]. The increasingly ubiquitous presence of sensors and edge/IoT devices has created a flood of data, which has exposed a wide efficiency gap in computing hardware, ranging from mobile and edge devices to data centers and cloud computing hardware [4-5]. In addition, the slowdown in the miniaturization of silicon-based complementary metal–oxide–semiconductor (CMOS) devices further accentuates the gap between resource requirements based on conventional von Neumann computing hardware architectures, specifically the central processing unit (CPU), graphics processing unit (GPU), and field-programmable gate arrays (FPGA) [6]. Furthermore, it is well-known that for many data-centric tasks in such von Neumann architectures, most of the energy and time is consumed in memory access and data movement, rather than in actual computation [7-9]. Several solutions have been proposed to mitigate and overcome this bottleneck, with a prominent one being placing memory and logic units in close physical proximity. While significant progress has been made along those lines at both materials and device levels, a transformative approach would be to perform computing functions using in-situ memory. This is popularly known as compute-in-memory (CIM). The overarching goal of CIM is to radically transform the computing architecture by completing computations in-situ, exactly where the data are stored, instead of re-engineering conventional von Neumann architecture by individual optimizations in memory bandwidth, novel non-volatile memory (NVM) technology, and data parallelism [8-9]. While several demonstrations of CIM architectures using NVMs have been made, the bulk of the effort has been constrained to a single type of computing task, an example being matrix multiplication accelerators, typically achieved using memristive crossbar arrays [10-12]. However, AI computational tasks that exploit 'big-data' generally require more than one data-intensive computational operation on the same chip, preferably using the same architecture to process information in the pipeline. Three of the most important functions or operations are: 1) on-chip storage, 2) parallel search, and 3) matrix multiplication. A key challenge in the construction of CIM architectures is the conflicting trade-off between the performance and the flexibility required to achieve these three functions. Consequently, while CIM accelerators have been demonstrated that can achieve high performance on matrix multiplication acceleration, they are are fundamentally ill-suited for other big data operations such as parallel search [13-14]. Hence, it is important to conceptualize and develop reconfigurable and operationally flexible hardware for CIM to simultaneously support essential data operations such as on-chip memory, parallel search and matrix multiplication acceleration.

In this work, we leverage the unique characteristics of aluminum scandium nitride (AlScN) ferroelectric diodes (FeD) devices – specifically their field-programmability, non-volatility and non-linearity – and demonstrate circuit blocks based on FeD devices which support multiple,

essential primitive data operations in-situ memory in a transistor free design (Fig. 1). Specifically, first, we demonstrate FeD devices which are non-volatile and show self-rectifying behavior with a non-linearity > 10^6, a high ON/OFF ratio over 10^2, endurance over 10^4 cycles and field-programming speed faster than 500 ns, and which are compatible with CMOS back-end-of-line (BEOL) processing. Then, we exploit these unique properties and demonstrate a non-volatile ternary content addressable memory (TCAM) using 0-transistor/2-FeD cells. These serve as a key building block in hardware implementation of in-memory computing for the parallel search process in big data applications. This transistor-free approach is a key merit of our device and memory cell design. Consequently, the 2-FeD TCAMs have the most compact design known (0.12 μm^2/cell for 45 nm node), along with a significantly reduced search delay (< 0.1 ns for 45 nm node) compared to 2-transistor/2-resistors (2T-2R) based TCAM cells, as evaluated using integrated circuit emphasis (SPICE) simulations. Finally, we also show that FeD devices can be programmed into 4-bit, distinct, conductive states with superior linearity and symmetry via electrical pulsing. Using this programmable, multi-bit attribute of the ferrodiodes, we demonstrate a hardware implementation of neural network computation in the form of analog voltage-amplitude matrix multiplication, which is a crucial kernel in neural network computation. We demonstrate accuracies approaching ideal, software-based, neural networks. The matrix multiplication operation is benchmarked by mapping neural network weights to experimental FeD conductance states in a convolutional neural network architecture for both inference and the in situ learning task, and shows that our accuracies approach ideal software-level simulation on the MNIST dataset. Our results indicate that the AlScN-based, field-programmable, non-volatile FeDs offer unique opportunities to build reconfigurable CIM architectures with superior balance between performance and flexibility.

Field-programmable AlScN FeDs for memory

Our FeD devices consist of a 45-nm thick layer of a sputter-deposited, ferroelectric AlScN layer sandwiched between top and bottom aluminum electrodes. This forms a metal-insulator-metal (MIM) structure, as shown in the left panel of Fig. 2a. AlScN is a recently discovered ferroelectric material with nearly ideal ferroelectric hysteresis loops, record values of remnant polarization, and a composition-tunable coercive field [15-17]. Furthermore, it can be integrated directly in a CMOS BEOL-compatible process technology over 8-inch wafers. It has also been shown to be one of the most promising candidates for high-performance ferroelectric memory devices, scalable down to < 10 nm in thickness [18]. The AlScN films were characterized electrically and exhibit large coercive fields, E_C, of 2-4.5 MV/cm [15-19]. This is important for scaling to thinner ferroelectric layers, all while maintaining a large memory window, high ON/OFF ratio and good retention [16, 19]. When combined with high measured remnant polarizations – P_r of 80-150 $\mu C/cm^2$ [15-19] – this leads to a significant tunneling electro-resistance effect, based on strong tunnel barrier modulation, thus giving a high ON/OFF ratio (Supplementary Note 1).

Additional details of film deposition, characterization and device fabrication are provided in methods and supplementary information. Figure 2(a) presents a representative cross-section transmission electron microscopy (TEM) image of the MIM FeD device comprised of an AlScN film with an Al top electrode, deposited on Al/AlScN/Si substrate. An atomic-resolution TEM image of the AlScN film is shown in Figure 2(b1). Figure 2(b2) shows ~2 nm thick interfacial layer at AlScN/bottom Al interface.

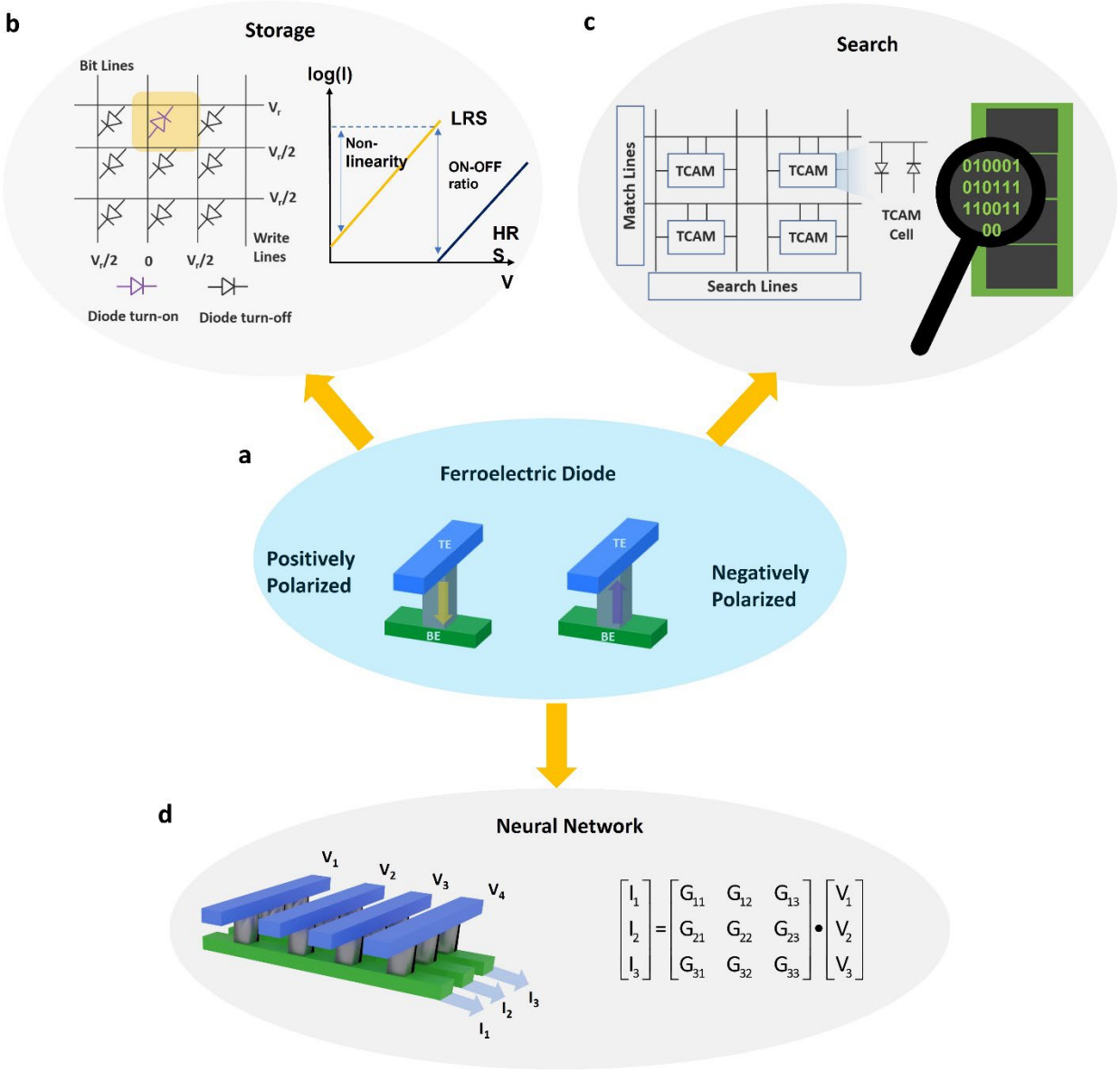

Fig. 1. Reconfigurable CIM on Field-Programmable Ferroelectric Diodes. a**,** Schematic diagram of FeD devices in a cross-bar structure with up and down polarization of the ferroelectric AlScN. The field programmability, non-volatility and non-linearity of these devices can be leveraged for

multiple, primitive data operations such as storage, search, and neural networks without the need for additional tansistors, as shown in b-d. b. The two-terminal FeD devices show a diode-like self-rectifying behavior with non-linearity > 10^6 concurrently with a ON/OFF ratio over 10^2 and endurance over 10^4 cycles, making FeD devices well placed in the memory hierarchy for storage. In addition, the high non-linearity can suppress sneak currents without the need for additional access transistors or selectors. C. For search operations, a non-volatile TCAM can be built upon 0-transistor/2-FeD cells, which serves as a building block in hardware implementation of in-memory computing for parallel search in big data applications. D. For neural networks, FeD devices can provide programmability to distinct multiple conductive states with a high degree of linearity with respect to number of electrical pulses. This allows mapping the matrix multiplication operation, a key kernel in neural-network computation, into reading the accumulated currents at each bitline of a FeD device by encoding an input vector into analog voltage amplitudes and the matrix elements into conductances of an array of FeD devices.

The ferroelectric response of the 45 nm AlScN thin film was characterized by a positive-up, negative-down (PUND) measurement on a circular metal/ferroelectric/metal capacitor with radius of 25 µm, using a square wave with 2 µs delay and pulse widths of 400 ns (Supplementary Fig. S11).PUND testing is preferred over a polarization-electric field hysteresis loop (P-E loop) measurment because the P-E loop of 45 nm AlScN shows a polarization-dependent leakage which hinders the observation of polarization saturation for positive applied fields that switch the material into the metal-polar state [16-17]. The PUND result indicates a remanent polarization ~150 µC/cm^2, as shown in Fig. 2c, and in agreement with prior observations [15-19]. To further verify the ferroelectric switching, a dynamic current response was carried out, in which peaks corresponding to ferroelectric switching were observed (Supplementary Fig. S12). To further charactierze the memory effect and reliability, we performed endurance tests between the positive and negative polarization states, as shown in Fig. 2d. Fig. 2d presents remanent positive and negative polarization extracted from 20,000 PUND cycles. Cyclic set/reset operations of the same AlScN FeD device indicate that both positive and negative polarization states are stable and rewritable for a significant number of cycles. As shown in Fig. 2e, we repeatedly set/reset the FeD device between the low resistance state (LRS) and high resistance state (HRS) by applying negative/positive voltages on the top electrode while grounding the bottom electrode, for 100 cycles, using quasi d.c. voltage sweeps. The FeD device shows ultralow operating current and self-rectifying behavior with non-linearity > 10^6 between 9 V and 0 V, which helps suppress sneak currents without the need for additional access transistors or selectors. The distributions of LRS and HRS resistances are summarized in Fig. 2f showing a tight distribution on the cycle-to-cycle variation of the ratio between the LRS and HRS.

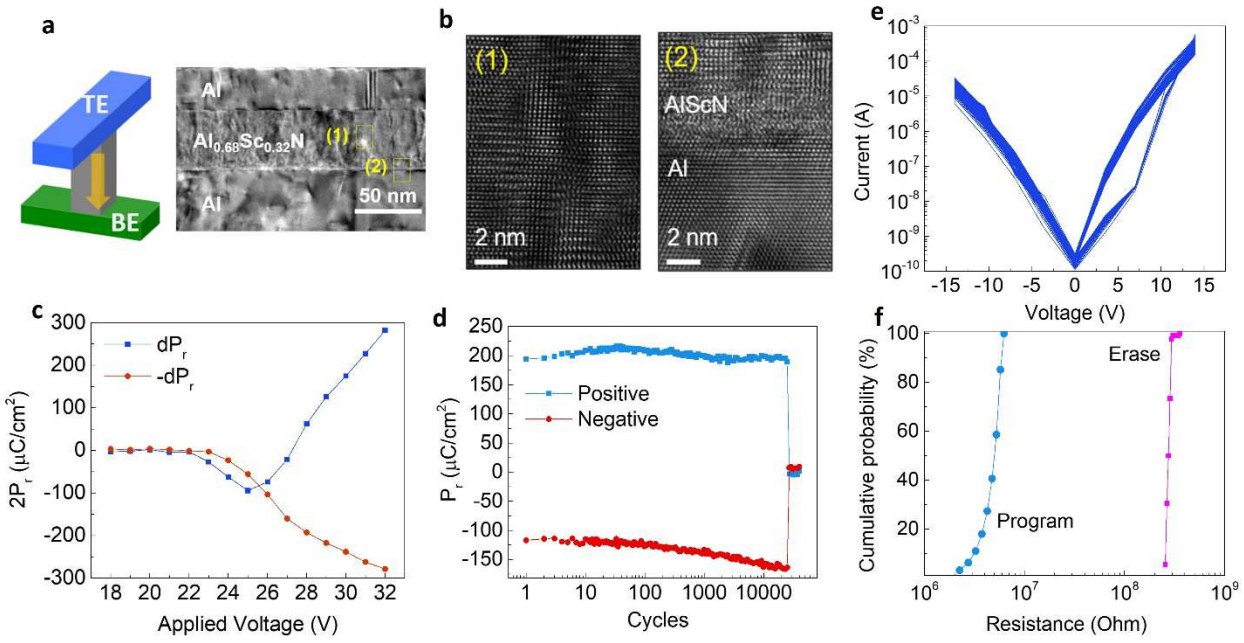

Fig. 2. Room-temperature electrical characterization of AlScN/MoS_2 FE-FETs. a, 3D schematic illustration of the AlScN FeD device and cross-sectional TEM image of the AlScN FeD, showing 45 nm AlScN as the ferroelectric switching layer. b, The high-resolution phase-contrast TEM image obtained from the denoted regions (1) and (2) in (a) where the atomic structure of the ferroelectric and interface are visible. c, PUND results of a 45 nm AlScN thin film with a pulse width of 400 ns and 2 µs delay between pulses. The PUND test reveals a saturated remanent polarization of 150 µC/cm^2. d, The extracted remanent polarizations from PUND measurements during the endurance test of AlScN films using 1.5 µs pulse width and 26 V amplitude. e, 100 cycles of program and erase measurements over the 45 nm AlScN-based FeDs. f, Distribution of HRS and LRS resistances during program and erase measurements in e.

2-FeD TCAM cell for search

Next we focus on CIM circuit architectures and computing applications comprising the above-described FeDs serving as non-volatile memories. We first demonstrate a TCAM implementation using our FeDs. TCAM is a key building block in hardware implementation of CIM for fast and energy-efficient parallel searches in big data applications [20]. TCAM performs the search function by comparing the input data with the stored data in the memory array in parallel, and returning the data address when a match is detected. Such parallel search allows TCAMs to perform a look-up table function in a single clock cycle. Unlike a binary content-addressable memory cell, which stores bit values of either '0' or '1', a TCAM cell can store an additional 'X' ('don't care') bit, which results in a match state regardless of the input search data, and makes TCAM much more powerful in searching applications. However, in conventional Si CMOS

architecture multiple transistors (~16) are required to construct a single TCAM cell with static random-access memories (SRAMs) (Fig. 3a). This configuration results in large footprints and high-power consumption due to charging and discharging of the transistor and due to interconnect parasitic capacitances. This limits the use of this configuration in high speed, massive scale and power-constrained systems [21]. Non-volatile memories (NVMs) are promising alternatives for implementing TCAMs as they are more area-and energy-efficient. This is because they form a single TCAM cell in a more compact architecture and because they retain stored information even if power is removed [22]. TCAMs based on resistive random-access memories (RRAMs) [22-23], magnetic tunnel junction (MTJ) RAMs [24], floating gate transistor memory (FLASH) [25], and phase change memories (PCMs) [26] have already been demonstrated. However, all of those architectures are still constructed on top of front-end-of-line transistors and none of them are fully BEOL compatible.

In this work, the cell structure of TCAM can be significantly simplified by using just two FeDs, which does not require the incorporation of a transistor due to the large non-linearity of the FeDs (Fig. 3a). The operation of the single FeD TCAM cell is demonstrated in Fig. 3b. The cell structure makes it natural to utilize the FeD crossbar memory array, in which the signal lines connecting to the anode and to the cathode are in parallel in a bit-search for the TCAM demonstration, as shown in Supplementary Fig. S3. First, we discuss how the FeDs based TCAM stores and searches a '0' or '1' bit (Fig. 3b). During the cell operation, complementary states are first written into the two FeDs, and if the search data biased on search lines (SL and $\overline{SL}$) matches with the stored information, the match line (ML) remains high; otherwise, the ML is pulled down. As we have shown *supra vide,* the FeD devices are highly self-rectifying and sustain high ON/OFF ratios. Thus, a discharge on the ML can only happen if the FeD is programmed to a low resistance state and the read voltage is higher than the turn-on voltage of the FeD.

As shown in Fig. 3b, we write the logic '1' state into the FeD TCAM cell by setting the left/right FeD to a low-resistance/high-resistance state, respectively. During the search operation, the match lines are biased by a read voltage V_S which is higher than the turn-on voltage of the FeD. Next, we search logic '1' by applying a high/low voltage to the left/right FeD, respectively, and search logic '0' by applying a low/high to the left/ right FeD, respectively. In this context 'high voltage' refers to the read voltage V_S, which is higher than the turn-on voltage of the FeD, but is lower than the write voltage. Conversely, a 'low voltage' refers to a read voltage near zero, much lower than the turn-on volatge of the FeD. Since the left FeD is in parallel with the right FeD, a match state is observed only if both of the FeDs in a cell are cut-off (Fig. 3b, left panel). Based on these write and search schemes, when the stored data and search data match (as shown in the left panel in Fig. 3b, the stored bit is logic '1' and the search bit is logic '1'), the FeD with the low-resistance state is turned-off, as the voltage drop beween its anode and cathode is near zero and lower than its turn-on voltage. Further, the FeD with the high-resistance state is also cut-off,

because the current is naturaly low when passing through the FeD in the high-resistance state. Therefore, the discharge currents at the two channels are both minimal and the ML stays high. However, when search data do not match the stored data, even though the right FeD with the high-resistance state is still cut off, the left FeD is not. The left FeD with the low-resistance state is turned-on as the voltage drop beween its anode and cathode is V_S and is higher than its turn-on voltage. Therefore, the discharge current is signifcant and the ML volatage is low (Figure 3b, middle panel). We also demonstrate a ternary 'don't care' state in the two FeD-based TCAM. As shown in the right panel of Fig. 3b, we write the logic 'don't care' state into the FeD TCAM cell by setting the left/right FeD both to a high-resistance state. With the above write schemes and the same search schemes as logic '1' and '0', whatever singals arrive at the two FeDs, both the FeDs are always cut off as they are in the high-resistance state. Fig. 3c shows repeated quasi-DC reading of the resistance of the two FeD-based TCAM cell for both match and mismatch states between the search data and the stored data bit '1', using moderate search voltages of 7 V on FeDs. Fig. 3d shows repeated quasi-DC reading of the resistance of the two FeD TCAM cell for the stored data bit 'Don't care', using both query bit '1' and '0'. This shows that for both queries, the ML resistance of the two FeDs-based TCAM remains high and thus no discharging occurs through any of the two FeDs. Hence, the TCAM cell with two FeDs is fully functional with all three states. The full look up table of the two FeD-based TCAM cell is summarized in Supplementary Table. 1.

Conventional two-terminal NVMs (memristors) are always paired together with a front-end transistor to construct TCAM cells. This is because transistors are required to cut-off the channel, as they are in series with the two-terminal NVMs. The FeD-based design benefits from a high self-rectifying ratio which cuts off the channel without the need for any transistors. In other words, the FeD abstracts the functionality of the transistors into its self-rectifying behavior. The absence of a transistor leads to smaller cell footprints and area efficiency, and increases the search speed of the FeD-based TCAM. Using a SPICE simulation, we verify that the search delay in our FeD-based TCAM is reduced in comparison to prior TCAM architectures based on 2-transistors-2-resistors (2T-2R). A benchmark comparison chart of lateral footprint of various TCAM cells vs. search delay is shown in Figure 3e. The superior performance of our two FeD-based TCAM of over CMOS SRAM and other transistor + NVM devices based architectures is evident.

The sensing margin of our FeDs-based TCAM is a function of both the self-rectifying ratio and the ON/OFF conductance (or current) ratios [26]. Per our detailed compact model (See supplementary note 1) the ON/OFF ratio of a FeD can be further imporved by integrating a non-ferroelectric insulator on top of the ferroelectric layer and engineering both the thickness ratio between these ferroelectric and non-ferroelectric insulator layers as well as the coercive field of

the ferroelectric layer. Future studies will focus on further improving the sense margins by engineering these variables.

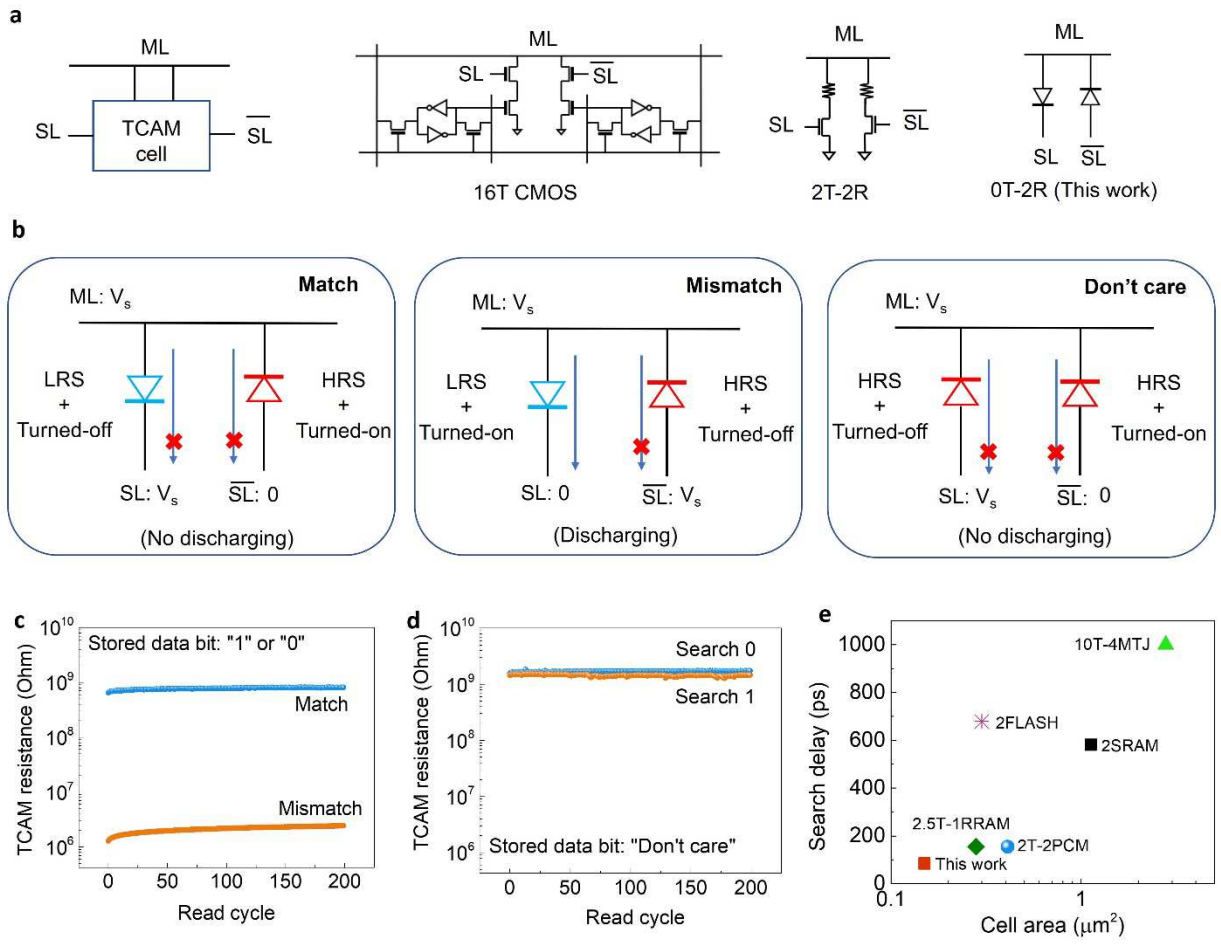

Fig. 3. 2-FeD TCAM cell for search operation a, A box schematic representation of a TCAM cell with match line (ML), search line (SL) and search line (SL bar) electrodes (left). Circuit diagrams of a single 16 transistor (16T) TCAM cell based on CMOS volitle static-random-access-memory (SRAM) technology, and 2-transistor-2-resistor (2T2R) TCAMs based on resistive storage elements such as PCM and RRAM. (center). The two ferrodiode-based TCAM cell proposed in this work (right) significantly simplifies the TCAM design by using two FeDs connected in parallel but oppositely polarized. b, Operation of a single TCAM cell comprising 2 FeDs for “match”, “mismatch” and “don’t care” states. c, Repeated quasi-DC reading of the resistance of the two FeDs TCAM cell for both match and mismatch states between the search data and the stored data bit ‘1’, showing a >100 X difference over ML resistances. d, Repeated quasi-DC reading of the resistance of the two ferro-diodes TCAM cell for the stored data bit ‘Don’t care’, using both query bit ‘1’ and ‘0’, which turns out that for both two queries the ML resistance two FeDs TCAM

is high and thus no discharging through any of the two FeDs. e, A benchmark comparison chart of lateral footprint of TCAM cells in various memory technlogies vs search delay [21, 24-27]. A single FeD area of 0.0081 μm^2 is assumed for this estimate.

Neural Network

Next we focus on the application of our FeD device arrays for deep neural network (DNN) inference, which involves repeated matrix multiply/accumulate (MMAC) operations. MMAC operations and DNNs are typically implemented at the software level. However, their software implementation makes it particularly challenging to deploy them in power and resource contrained devices or environments. Once again, this is in large part due to the traditional von Neumann computing hardware approaches, which are intensive in terms of memory access and are difficult to parallelize. Conducting MMAC operations in the analog domain offers a promising alternative: memristors with analog condutances have been shown to be a superior hardware medium in which to perform MMAC operations [8, 11-12]. By leveraging Kirchhoff's current law's (KCL's) high parallelism, the MMAC operation can be significantly reduced to reading the accumulated currents in a single time clock at each bit-line of a memristor. This is accomplished by encoding an input vector into analog voltage amplitudes and matrix elements into conductances of the memristor array.

The ideal MMAC-suitable memristive devices should perform linearly arranged conductance values over electrical programming, linear dependence of current on the drive voltage, and high resistance to suppress the amount of current. Prior research in this domain has primarily focused on memristive devices which exhibit excellent ohmic behavior and a large number of conductance states, such as RRAM [11] and PCM [29]. In the context of DNN inference accuracy, a linear relationship between current and voltage is necessary to minimize the distortion of the input datum and a large number of conductances will minimize the precision loss on the weight matrix, which are essential to perform a highly accurate inference task. However, from the power and areal efficiency point of view, an excellent ohmic behavior and a large number of conductance states will hurt the architecture metrics for power efficiency and low latency per computation. There are a few reasons for this. First, memristive devices with excellent ohmic behavior come at the cost of high device conductance, meaning that high operating currents impose limitations on array scaling [11, 13, 29]. Second, a large number of conductances will correspondingly require high precision analog-to-digial converters (ADCs) [30-32]. It is already known from prior work that the energy and area costs are dominated by the ADCs on a circuit level in memristor array systems [30-32]. Therefore more conductance states means more power overhead at the architecture level in a DNN inference engine. Thus, there is an obvious trade-off between the DNN inference accuracy with the power and area efficiency.

Here we show that FeD memristors can be used to perform an optimal trade-off between these metrics. First, to realize the trade-off on the device condutance, it is important to decrease the operating conductance of memristive devices while maintaining linear behaviour. The former condition is readily met for highly self-rectifying devices, an inherent property of the FeDs; the latter condition can be met by applying a encoder on the input voltage amplitudes to linearize the current-voltage relationship (See Supplementry Note 2). Second, to relax the trade-off on the number of conductance states, a few but sparsely and linearly arranged conductance states are necessary. This approach can achieve equivalent inference accuracy in comparison to approaches implementing a large number of conductance states [33-34].

Fig. 4a shows the gradual switching in a FeD by stepwise voltage pulse modulation. The FeD cells are gradually programmed into 16 distinct conductance states using stepwise voltage pulses. These conductance states show a high degree of linearity with number of programming pulses as discussed below. The figure (left) shows a sequence of programming operations in which the stepwise voltage pulses (ranging from 16 V to 19 V) are applied on the top electrodes on the FeDs followed each time by an erase operation. The callout window (right) shows the conductance versus pulse number for a representative cycle. Fig.4b shows that the FeD device is capable of voltage pulse-induced analog bipolar switching (ranging from 16 V to 19 V, left). The callout window (right) shows one cycle of gradual programming and gradual erasing. The FeD device showed superior linearity (R^2 score of 0.9997 for a linear fit) for 16 distinct conductance states over bidirectional modulation. Conductance retentions for 16 distinct conductance states are shown in Fig. 4c, and show no obvious degradation. Fig.4d shows conductance states distribution of five separate FeD devices subjected to the identical sequences of 16 program pulses (2 μs pulse width) with interleaved reads (8 V). The results show negligible device-to-device variations between those FeD devices. We note that the range of conductance in FeD devices used to program these states (~25-250 nS) is much smaller as compared to the range of conductance used for TCAM operations (~2-250 nS). This is primarily because this linearity in operation is better achieved in a smaller range of conductance. Further the DNN inference application does not necessarily require a high range of conductance modulation [10-11, 34]. We simulate the performance of arrays comprising such FeD devices in a practical application where a trained convolutional neural network (CNN) is used for computor vision. A CNN (including two convolutional layer and one fully connected layer) was trained on the MNIST dataset (MNIST, Modified National Institute of Standards and Technology database) [35], which is followed by transferring the pretrained weights to the FeD conductance range. An illustration of the network is shown in Fig. 4e. We analyse the accuracy degradation due to weight transfer to low-precision conductance values with an added varying factor A, which is an indicator of non-linearity. The relationship between the A factor and the non-linearity has been discussed in Supplementry Note 3 in detail. The weights of the full-precision trained network are therefore quantized to a number of conductance states (varying from 1 to 9-bit). Then, the network's accuracy on the MNIST

testing dataset is re-evaluated. Convolutional neural networks are generally robust to low-precision weight transfer with low non-linearity (A > 0.5), as can be seen in Fig. 4f, where for low-weight transfer variation, the full-precision testing accuracy of 97.5% on single-precision floating-point format (FP32) is recovered with just 3 bits of weight precision. For high non-linearity (A < 0.35), there requires one or two bits weight precision to recover full-precision testing accuracy on FP32, which shows that sparsely but linearly arranged conductance states with superior linearity can replace a large amount of conductance states to perform equivalent inference accuracy. Furthermore, we simulate the in-memory implementations of in-situ training on FeD arrays where the same convolutional neural network is being trained and weight updates are directly mapped to the realistic conductance states of the FeDs after each backpropagation. As shown in Fig. 4g, for the demonstrated 16 separate conductance states in Fig. 4a in the FeD devices, the in-situ learning accuracy suffers a ~2% degradation compared to the accuracy trained on FP32. However, with more advanced low-precision training techniques and model compression techniques on software [33], we believe this number can be substantially reduced, allowing for little to no accuracy degradation when performing low-precision weight transfer to FeD devices in the training phase.

Conclusion

In conclusion, we demonstrate AlScN based ferrodiode devices as a novel, BEOL compatible platform for multi-functional CIM in a transistor free architecture. Our experimental demonstration of search function is realized via a TCAM circuit with lateral cell footprint and search delays bettering all exisiting and experimental NVM technologies. Finally, we demonstrate a stable, pulse-programmable 4-bit memory from ferrodiode devices combined with hardware implementation of a convolutional neural network with inference accuracy comparable to software. Our work therefore opens new possibilities in CIM platforms by enabling architectures with novel ferroelectric materials and diode devices made using them.

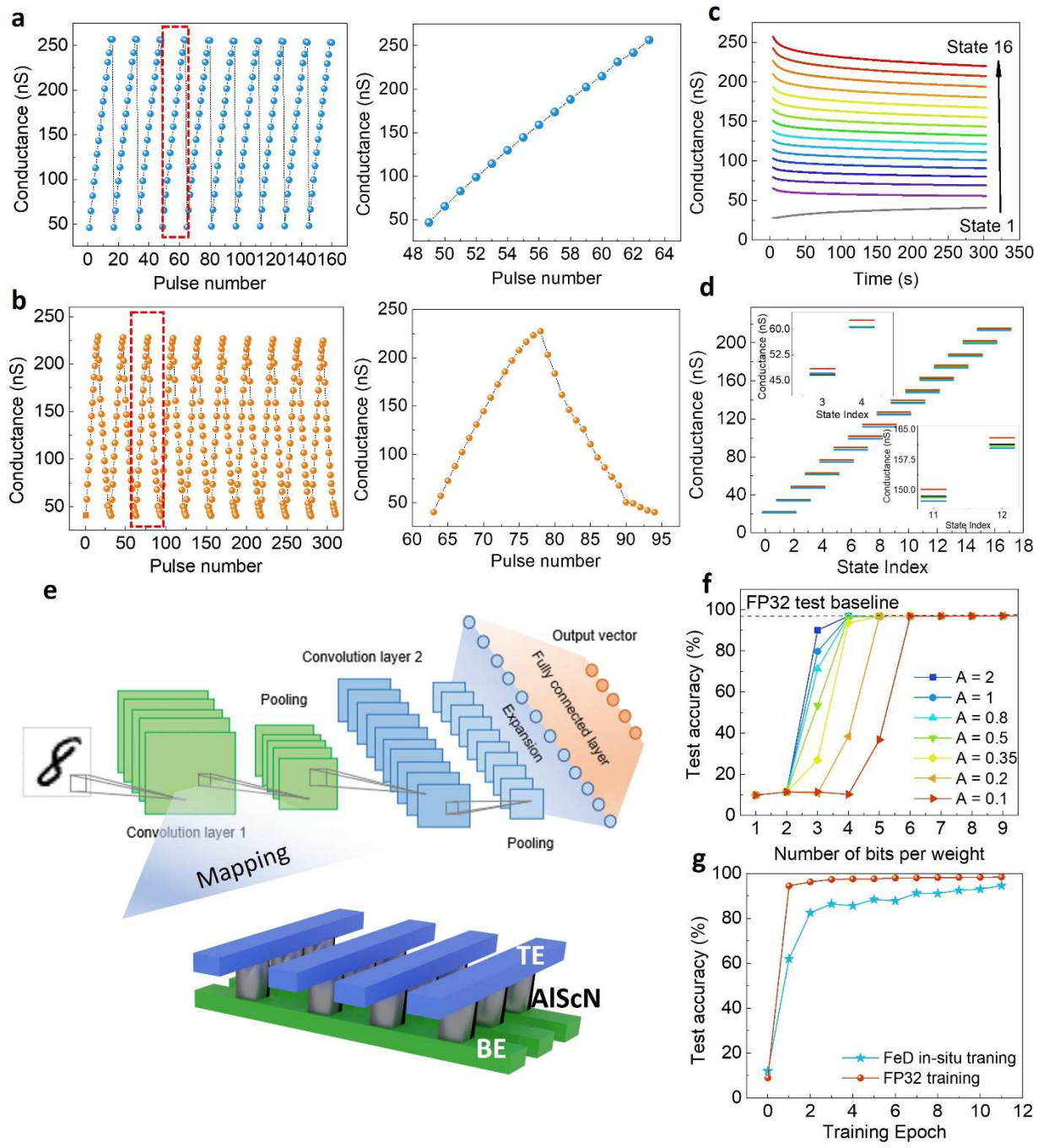

Fig. 4. FeD-based neural network. a, Gradual switching in a ferrodiode (FeD) by stepwise voltage modulation pulses. The FeD cells are gradually programmed into various conductance states using stepwise voltage pulses. The left panel shows a sequence of programming operations in which the stepwise voltage pulses are biased on the top electrodes on the FeDs followed each time by an erase operation. Callout window (right panel) shows conductance versus pulse number for a representative cycle. b, The FeD is demonstrated to be capable of voltage pulse-induced analog bipolar switching (left). Callout window (right) shows one cycle of gradual

programming and gradual erasing. The FeD device showed superior linearity over 16 distinct states. c, Resistance retentions for 16 distinct resistance states. d, Resistance states distribution of five separate FeDs subjected to sequences of 16 program pulses (2 μs pulse width) with interleaved reads (8 V). e, Illustration of a CNN trained for the MNIST dataset. A hardware implemention of a neural network using ferro-diode arrays for the matrix multiplications can operate in a fully analog mode without the peripheral analog-to-digital converters. f, Simulated performance of inference efficacy of the network in **e.** The simulation comprises of FeD devices implementing analog weight layers, with inaccurate weight mapping of the network trained on MNIST with FP32 compute. The simulations in (f) demonstrate that degradation of the network's inference accuracy is less than 1% for low weight precision of just 3 bits for $A < 0.5$. g, Simulations of in-situ training of the network in (e) directly with the FeD devices implementing analog weight layers. Leveraged by the superior linearity in the gradual programming of the FeDs, the analog weight layers with 16 resistance states have been simulated to perform at an inference accuracy comparable to the FP32 compute baseline.

METHODS

Device fabrication

FeD consisted of a film stack of Al (80 nm)/$Al_{0.68}Sc_{0.32}N$ (45 nm)/Al (30 nm) on top of Si/$Al_{0.8}Sc_{0.2}N$ (85 nm) substrate. To prepare this stack, we start by sputter depositing a layer of 85-nm thick $Al_{0.80}Sc_{0.20}N$ template on the top of a 6" Si <100> wafer. The $Al_{0.8}Sc_{0.2}N$ was deposited using pulsed-DC reactive sputter deposition of a single-alloyed $Al_{0.8}Sc_{0.2}$ target using 5 kW of target power, a pressure of $7.47x10^{-3}$ mbar, and a deposition temperature of 375°C in a N_2 atmosphere. The first layer of 85 nm $Al_{0.8}Sc_{0.2}N$ serves to orient the subsequent 80-nm thick Al layer into a {111}-orientation. This Al (80 nm thick) layer serves as the bottom electrode for the second layer of $Al_{0.68}Sc_{0.32}N$ (45 nm thick), which is the ferroelectric layer used in this device. The 45-nm thick ferroelectric $Al_{0.68}Sc_{0.32}N$ film was co-sputtered from separate 4-inch Al and Sc targets in an Evatec CLUSTERLINE® 200 II pulsed DC Physical Vapor Deposition System. The Al and Sc targets were operated at 1250 W and 695 W, respectvely, at a chuck temperature of 350 °C with 10 sccm of Ar gas flow and 25 sccm of N_2 gas flow. The chamber pressure was maintained ~$1.45x10^{-3}$ mbar. This sputter condition resulted in a deposition rate of 0.3 nm/sec. The highly oriented {111} Al layer promotes the growth of AlScN with its [0001] axis direction being perpendicular to the substrate, thus, yielding a highly textured FE film. Then, without breaking vacuum, a layer 30 nm Al layer was sputtered as the top electrode and as a capping layer to prevent the oxidation of ferroelectric $Al_{0.68}Sc_{0.32}N$.

Device characterization

Current-voltage measurements were performed in air at ambient temeprature using a Keithley 4200A semiconductor characterization system. P-E hysteresis loops and PUND measurements of ferroelectric AlScN were conducted using Keithley 4200A semiconductor characterization system and a Radiant Precision Premier II testing platform. TEM cross-sectional sample was prepared in a FEI Helios Nanolab 600 focused ion beam (FIB) system using the in-situ lift-out technique. The sample was coated with thin carbonaceous protection layer by writing a line on the surface with a Sharpie® marker. Subsequent electron beam and ion beam deposition of Pt protection layers were used to prevent charging and heating effects during FIB milling. At the final cleaning stage, a low-energy Ga+ ion beam (5 keV) was used to reduce FIB-induced damage. TEM characterization and image acquisition were carried out on a JEOL F200 operated at 200 kV accelerating voltage. The sample was orientated to the [001] zone axis for imaging. All of the captured TEM images were collected using Digital Micrograph software.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the journal website at DOI:

AUTHOR INFORMATION

Corresponding Author

*E-mail: rolsson@seas.upenn.edu, dmj@seas.upenn.edu

Author Contributions

X.L., D.J., R.O. and E.S. conceived the idea. X.L. designed FeD devices, TCAM circuits and the reconfigurable CIM architecture. X.L. performed all neural network simulations and SPICE simulations. X.L., J.T., Y.H., M.M.A.F., S.B.A performed all the electrical meaurements. X.L., J.F. and Y.H. perfomed all compact modeling. D.W., J.Z., and G.E. performed microfabrication, ferroelectric material and metal thin film growth. P.M. and K.K. performed transmission electron microscopy imaging (P.M.) and sample preparation (K.K). D.J., R.O., E.S. and X.L. analyzed and interpreted the data. All others contributed to writing of the manuscript.

Notes

Competing interests

The authors declare the following competing interests: D.J., X.L., R.O. and E.A.S. have a provisional patent filed based on this work. The authors declare no other competing interests.

Data Availability

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

ACKNOWLEDGEMENTS

This work was supported by the Defense Advanced Research Projects Agency (DARPA), Tunable Ferroelectric Nitrides (TUFEN) Program under Grant HR 00112090047. The work was carried out in part at the Singh Center for Nanotechnology at the University of Pennsylvania which is supported by the National Science Foundation (NSF) National Nanotechnology Coordinated

Infrastructure Program (NSF grant NNCI-1542153). The authors gratefully acknowledge use of facilities and instrumentation supported by NSF through the University of Pennsylvania Materials Research Science and Engineering Center (MRSEC) (DMR-1720530). D.J. acknowledges support from the Intel RSA program. This research used resources of the Center for Functional Nanomaterials, which is a US Department of Energy Office of Science User Facility, at Brookhaven National Laboratory under contract no. DE-SC0012704. The data that support the conclusions of this study are available from the corresponding authors upon request. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436-444 (2015).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Comm. ACM* **60**, 84–90 (2012).
3. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778 (2016).
4. V. Sze, Y.-H. Chen, T.-J. Yang, J. S. Emer, Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE* **105**, 2295-2329 (2017).
5. Y.-H. Chen, T. Krishna, J. S. Emer, V. Sze, Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE journal of solid-state circuits* **52**, 127-138 (2016).
6. T. N. Theis, H.-S. P. Wong, The end of moore's law: A new beginning for information technology. *Computing in Science & Engineering* **19**, 41-50 (2017).
7. Wong, H.-S. P. & Salahuddin, S. Memory leads the way to better computing. *Nat. Nanotechnol*. **10**, 191 (2015).
8. D. Ielmini, H.-S. P. Wong, In-memory computing with resistive switching devices. *Nature Electronics* **1**, 333-343 (2018).
9. A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, Memory devices and applications for in-memory computing. *Nature Nanotechnology* **15**, 529-544 (2020).
10. P. Yao *et al.*, Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641-646 (2020).
11. C. Li *et al.*, Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. *Nature Communications* **9**, 1-8 (2018).
12. S. Ambrogio *et al.*, Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60-67 (2018).

13. R. Berdan *et al.*, Low-power linear computation using nonlinear ferroelectric tunnel junction memristors. *Nature Electronics* **3**, 259-266 (2020).
14. Y. Zha, E. Nowak, J. Li, Liquid silicon: A nonvolatile fully programmable processing-in-memory processor with monolithically integrated ReRAM. *IEEE Journal of Solid-State Circuits* **55**, 908-919 (2020).
15. Fichtner, S.; Wolff, N.; Lofink, F.; Kienle, L.; Wagner, B., AlScN: A III-V semiconductor based ferroelectric. *Journal of Applied Physics* **2019,** *125* (11), 114103.
16. X. Liu *et al.*, Post-CMOS Compatible Aluminum Scandium Nitride/2D Channel Ferroelectric Field-Effect-Transistor Memory. *Nano Letters* **21**, 3753-3761 (2021).
17. D. Wang *et al.*, Sub-Microsecond Polarization Switching in (Al, Sc) N Ferroelectric Capacitors Grown on Complementary Metal–Oxide–Semiconductor-Compatible Aluminum Electrodes. *Physica Status Solidi (RRL)–Rapid Research Letters* **15**, 2000575 (2021).
18. S. Yasuoka *et al.*, Effects of deposition conditions on the ferroelectric properties of (Al1−x Sc x) N thin films. *Journal of Applied Physics* **128**, 114103 (2020).
19. X. Liu *et al.*, Aluminum scandium nitride-based metal–ferroelectric–metal diode memory devices with high on/off ratios. *Applied Physics Letters* **118**, 202901 (2021).
20. K. Pagiamtzis, A. Sheikholeslami, Content-addressable memory (CAM) circuits and architectures: A tutorial and survey. *IEEE journal of solid-state circuits* **41**, 712-727 (2006).
21. K. Ni *et al.*, in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 240-241 (2014).
22. R. Yang *et al.*, Ternary content-addressable memory with mos2 transistors for massively parallel data search. *Nature Electronics* **2**, 108-114 (2019).
23. M.-F. Chang *et al.*, A 3T1R nonvolatile TCAM using MLC ReRAM for frequent-off instant-on filters in IoT and big-data processing. *IEEE Journal of Solid-State Circuits* **52**, 1664-1679 (2017).
24. B. Song, T. Na, J. P. Kim, S. H. Kang, S.-O. Jung, A 10T-4MTJ nonvolatile ternary CAM cell for reliable search operation and a compact area. *IEEE Transactions on Circuits and Systems II: Express Briefs* **64**, 700-704 (2016).
25. Fedorov, V. V., Abusultan, M. & Khatri, S. P. An area-efficient ternary CAM design using floating gate transistors. In *2014 IEEE 32nd Int. Conference on Computer Design (ICCD)* 55–60 (IEEE, 2014).
26. J. Li, R. K. Montoye, M. Ishii, L. Chang, 1 Mb 0.41 μm² 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing. *IEEE Journal of Solid-State Circuits* **49**, 896-907 (2013).
27. K. Ni *et al.*, Ferroelectric ternary content-addressable memory for one-shot learning. *Nature Electronics* **2**, 521-529 (2019).

28. Ni, K.; Li, X.; Smith, J. A.; Jerry, M.; Datta, S., Write disturb in ferroelectric FETs and its implication for 1T-FeFET AND memory arrays. IEEE Electron Device Letters **2018,** 39 (11), 1656-1659.
29. V. Joshi *et al.*, Accurate deep neural network inference using computational phase-change memory. *Nature Communications* **11**, 1-13 (2020).
30. A. Shafiee *et al.*, ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Computer Architecture News* **44**, 14-26 (2016).
31. A. Ankit *et al.*, in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. (2019), 715-731.
32. P. Houshmand *et al.*, in *2020 IEEE International Electron Devices Meeting (IEDM)*. (IEEE, 2020), 29.21. 21-29.21. 24.
33. S. Han, H. Mao, W. J. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, (2015).
34. P.-Y. Chen, X. Peng, S. Yu, 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* **37**, 3067-3080 (2018).
35. Y. LeCun, The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FDTCAMSIv7.pdf