Received 19 March 2024; revised 16 May 2024; accepted 2 June 2024.
Date of publication 6 June 2024; date of current version 21 June 2024.

Digital Object Identifier 10.1109/JXCDC.2024.3410681

MEFET-Based CAM/TCAM for Memory-Augmented Neural Networks

SAI SANJEET, JONATHAN BIRD, and BIBHU DATTA SAHOO

Department of Electrical Engineering, University at Buffalo, Buffalo, NY 14260 USA

CORRESPONDING AUTHOR: B. D. Sahoo (bibhu@buffalo.edu)

This article has supplementary downloadable material available at https://doi.org/10.1109/JXCDC.2024.3410681, provided by the authors.

ABSTRACT Memory-augmented neural networks (MANNs) require large external memories to enable long-term memory storage and retrieval. Content-addressable memory (CAM) is a type of memory used for high-speed searching applications and is well-suited for MANNs. Recent advances in exploratory nonvolatile devices have spurred the development of nonvolatile CAMs. However, these devices suffer from poor ON–OFF ratio, large write voltages, and long write times. This work proposes a nonvolatile ternary CAM (TCAM) using magnetoelectric field effect transistors (MEFETs). The energy and delay of various operations are simulated using the ASAP 7-nm predictive technology for the transistors and a Verilog-A model of the MEFET. The proposed structure achieves orders of magnitude improvement in search energy and $>45\times$ improvement in search energy-delay product compared with prior works. The write energy and delay are also improved by $8\times$ and $12\times$, respectively, compared with CAMs designed with other nonvolatile devices. A variability analysis is performed to study the effect of process variations on the CAM. The proposed CAM is then used to build a one-shot learning MANN and is benchmarked with the Modified National Institute of Standards and Technology (MNIST), extended MNIST (EMNIST), and labeled faces in the wild (LFW) datasets with binary embeddings, giving $>99\%$ accuracy on MNIST, a top-3 accuracy of 97.11% on the EMNIST dataset, and $>97\%$ accuracy on the LFW dataset, with embedding sizes of 16, 64, and 512, respectively. The proposed CAM is shown to be fast, energy-efficient, and scalable, making it suitable for MANNs.

INDEX TERMS Content-addressable memory (CAM), ferroelectric field effect transistor (FeFET), magnetoelectric field effect transistors (MEFETs), magnetoelectric magnetic tunnel junction field effect transistor (ME-MTJ-FET), memory-augmented neural network (MANN), resistive random access memory (ReRAM), ternary CAM (TCAM).

I. INTRODUCTION

MEMORY-AUGMENTED neural networks (MANNs) [1], [2], [3], [4] are a class of neural networks that use external memory to store and retrieve information. MANNs are well-suited for tasks that require long-term memory, such as language translation, question-answering, and image recognition. External memory facilitates MANNs to be used for one-shot learning [5], [6], [7], where a model is trained to learn a class from one or a few examples.

One-shot learning is a type of learning where a model is trained to learn the similarities and differences between the input data points, by training the model to generate embeddings of the input data such that the embeddings of the same class are close to each other and the embeddings of different classes are far from each other. Once a model is trained to generate "*good*" embeddings; the model can generate the representative embeddings of a class from one or a few examples, hence the name one-shot learning. The model can then classify the test data based on the distance from the representative embeddings. This paradigm is well-suited for tasks that require learning from a small dataset, such as face recognition, signature verification, and fingerprint recognition.

The representative embeddings used in MANNs must be stored in memory, and the distance between the embeddings must be computed to classify the test data. As the number of classes to be classified increases, the memory size and the number of distance computations increase, making it the bottleneck compared with the computation of embedding from the input data. Therefore, the memory used to store the embeddings must be fast, energy-efficient, and scalable. One such memory is the content-addressable memory (CAM),

which is used for high-speed searching applications.

Traditional CAM uses modified static random access memory (SRAM) cells [8]. However, such conventional CAM has several disadvantages, such as high power consumption, large area, and vulnerability to side-channel attacks. CAM must be nonvolatile for a one-shot learning MANN, as the representative embeddings must be stored for a long time and ideally be updated infrequently.

The use of nonvolatile devices for memory applications has gained a lot of attention in the last decade. Previous works have discussed the design of nonvolatile memories for machine learning applications [9], [10], [11], [12], [13]. Many works also proposed nonvolatile CAMs using exploratory devices, such as: 1) spin–torque transfer (STT) [14], [15], [16], [17], which have a poor ON–OFF ratio (∼2); 2) phase-change memory (PCM) [18], [19], [20], which has an ON–OFF ratio of $\sim 10^2$; 3) resistive random access memory (ReRAM) [21], [22], [23]; 4) ferroelectric field effect transistors (FeFETs) [24], [25], [26], [27], which require high write voltages and have long write time; and 5) magnetoelectric magnetic tunnel junction (ME-MTJ) [28], [29], which also have poor ON–OFF ratio.

The magnetoelectric field effect transistor (MEFET) [30], [31], [32] is an emerging device that is predicted to require very small voltages to switch its state (∼100 mV) and to have a large ON–OFF ratio ($\sim 10^5$). The write time of the MEFET is also projected to be significantly smaller (∼3 ps) than the FeFET and the ME-MTJ and is estimated to have a very high write endurance [45]. MEFET proposed by Nikonov, Dowben, and colleagues is a four-terminal device that uses the magnetoelectric effect to control the resistance of a 2-D transistor channel. Fig. 1(a) shows a cross-sectional view of the MEFET [31]. When a potential difference is applied between the gate (G) and the back gate (BG), the resulting electric field changes the alignment of the chromia spin vectors in either the ‘‘up’’ or ‘‘down’’ direction. This programs the spin polarization of the narrow channel, which results in a low or high resistance state. The circuit symbol of the MEFET used in this article is shown in Fig. 1(b). The BG terminal is omitted from the circuit symbol as it is typically connected to the ground. Fig. 1(c) shows the calculated current through the MEFET as a function of gate voltage [32]. The BG terminal is assumed to be grounded. However, for this work, a simpler piecewise linear model [31] is used to simulate the MEFET.

Previous works have proposed nonvolatile random access memory (RAM) architectures using MEFETs [33], [34], as shown in Fig. 2(a) and (b). The 2T-1M MEFET-based RAM cell shown in Fig. 2(a) has separate transistors to perform the read the write operations. The 1T-1M MEFET-based RAM cell shown in Fig. 2(b) uses a single MEFET to perform the read and write operations. The write operation is performed by applying a programming voltage to the bitline (BL) when the word line (WL) is high and the source line (SL) is grounded. The read operation is performed by applying a sense voltage to the BL when WL is high and SL is grounded. The sense current on SL is measured

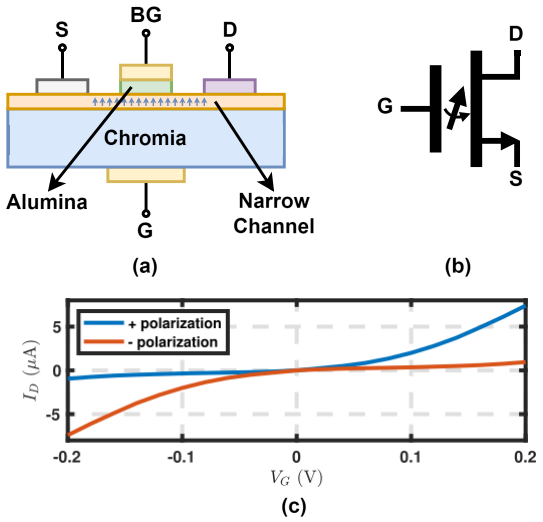


FIGURE 1. (a) Cross-sectional view of an MEFET, (b) circuit symbol of the MEFET, and (c) drain current as a function of the gate voltage, figure adapted with permission from [32].

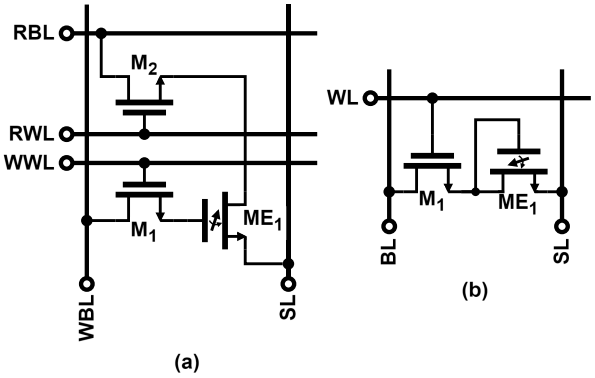
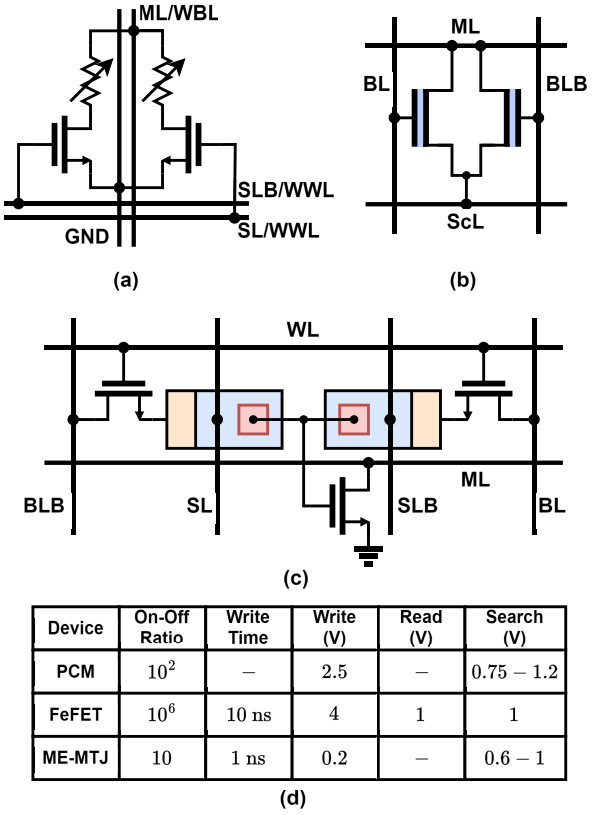


FIGURE 2. MEFET-based (a) 2T-1M RAM cell [33] and (b) 1T-1M RAM cell [34].

to determine the resistance state of the MEFET. However, during the write operation, the on-resistance of M_1 is in series with the resistance of the MEFET, forming a voltage divider. When the MEFET is in the low resistance state, the required voltage at the BL increases, making it difficult to write the data to the MEFET. This issue can be addressed either by adding a large resistance to SL or by driving SL to the same voltage as the BL during the write operation.

This work presents a nonvolatile ternary CAM (TCAM) using the MEFET and benchmarks its performance for one-shot learning MANNs. This article is organized as follows. Section II gives an overview of existing CAMs based on non-volatile devices, Section III presents the proposed MEFET-based CAM cell, Section IV shows the design of a CAM array with the proposed CAM cell, Section V discusses the effect of variability on the performance of the proposed CAM array, and Section VI covers the one-shot learning network structures along with system-level results obtained using proposed TCAM. The word ‘‘transistor’’ in this article refers to a CMOS transistor unless otherwise specified.



Device	On-Off Ratio	Write Time	Write (V)	Read (V)	Search (V)
PCM	10^2	–	2.5	–	0.75 – 1.2
FeFET	10^6	10 ns	4	1	1
ME-MTJ	10	1 ns	0.2	–	0.6 – 1

(d)

FIGURE 3. Structure of CAM cell using (a) PCM [19], (b) FeFET [25], (c) ME-MTJ [29], and (d) ON–OFF ratio, the time required to change the state of the device, and the voltages required for read, write, and search operations.

II. OVERVIEW OF NONVOLATILE CAM TOPOLOGIES

The design of CAM using exploratory nonvolatile devices has gained a lot of attention in the last decade. Fig. 3 shows the structure of a CAM cell using (a) PCM, (b) FeFET, and (c) ME-MTJ devices along with the ON–OFF ratio of the devices, write delay, and their write/read/search voltages.

Early works in the literature have proposed CAMs using STT devices [14], [15], [16], [17]. However, these devices have a poor ON–OFF ratio (∼2), leading to a strict requirement on sense amplifiers. PCM has a larger ON–OFF ratio (∼10^2) and can be used to design a CAM [19], as shown in Fig. 3(a). The write operation is performed by applying a large voltage (∼2.5 V) to the write bitline (WBL) while maintaining the corresponding write word line (WWL) high. It takes two write cycles to store the data and its complement. The search operation is performed by precharging the match line (ML) to a small voltage (∼0.75–1.2 V) and applying the search bit to the search line (SL) and the complement of the search bit to the complement search line (SLB). The ML is discharged if the search bit does not match the stored bit. The ML is connected to a sense amplifier to detect the match condition.

The FeFET is a three-terminal device that uses the change in polarization of a ferroelectric material to control the current through its channel. The FeFET has a large ON–OFF ratio (∼10^6) and can be used to design a CAM [25] as shown in Fig. 3(b). The bit "1" is stored in an FeFET if its gate–source voltage is high (∼4 V) and the bit "0" is stored if the gate—source voltage is low (∼−4 V). The write operation is performed by storing the data and its complement in two FeFETs in two write cycles by applying the corresponding voltages to the BL, the complementary bitline (BLB), and the source line (ScL). The search operation is performed by precharging the ML and applying the search bit to BLB and its complement to BL. The ML is discharged if the search bit does not match the stored bit.

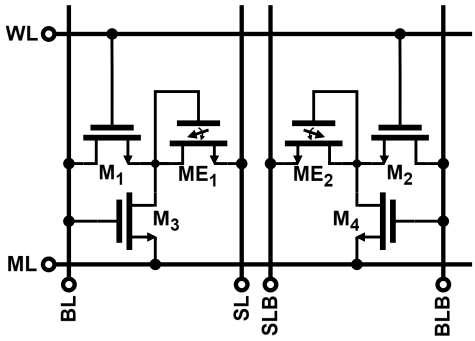


FIGURE 4. Proposed 4T-2M MEFET-based CAM cell.

The ME-MTJ is a three-terminal device that uses the magnetoelectric effect to control the tunnel resistance of a magnetic tunnel junction. The ME-MTJ requires a small voltage (∼0.2 V) to change its resistance state and can be used to design a CAM [29], as shown in Fig. 3(c). The write operation is performed by applying a small voltage (∼0.2 V) to the BL and SL when the WL is high. The write operation is done in a single cycle. The search operation is performed by precharging the ML and applying the search bit to SL and the complement of the search bit to SLB. With careful sizing of the transistors, the ML is discharged if the search bit does not match the stored bit.

III. PROPOSED MEFET-BASED CAM

The 1T-1M MEFET-based RAM cell shown in Fig. 2(b) can be extended to build a 4T-2M CAM cell. As the CAM computes an XOR operation, the data and its complement must be stored. Therefore, two MEFETs are used to store these data. The structure of the proposed CAM cell is shown in Fig. 4. The cell consists of two MEFETs (ME_1 and ME_2), two access transistors (M_1 and M_2), and two XOR transistors (M_3 and M_4).

A. WRITE OPERATION

The write operation of the proposed 4T-2M MEFET-based CAM cell is performed by applying a programming voltage ($\pm V_w$) to the BL when the WL is high. The required voltage at the BL increases when the MEFET is in the low resistance state to perform a successful write. Therefore, to minimize the driving voltage of BL, the SL is driven to the same voltage as the BL during the write operation. The complementary data are written to the complementary MEFET using the BLB and SLB. Table 1 shows the voltages of various lines during

TABLE 1. 4T-2M MEFET-based CAM cell write operation.

Data	WL	BL	SL	BLB	SLB	R_{ME_1}	R_{ME_2}
0	V_{DD}	$-V_w$	$-V_w$	V_w	V_w	R_H	R_L
1	V_{DD}	V_w	V_w	$-V_w$	$-V_w$	R_L	R_H
X	V_{DD}	$-V_w$	$-V_w$	$-V_w$	$-V_w$	R_H	R_H

$V_{DD} = 700$ mV, $V_w = 200$ mV.

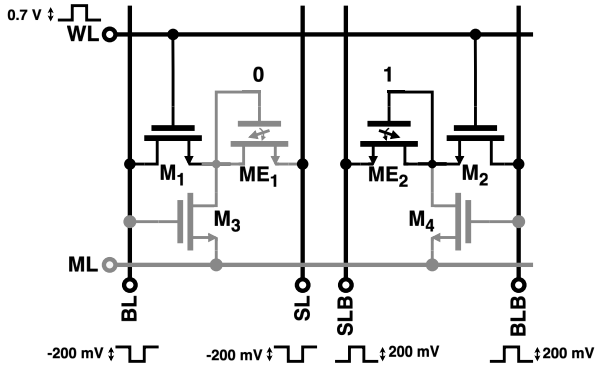


FIGURE 5. Write operation of a single CAM cell. The devices in lighter colors are OFF.

TABLE 2. 4T-2M MEFET-based CAM cell read operation.

Stored Data	WL	SL	SLB	Condition
0	V_{DD}	V_s	V_s	$V_{BL} < V_{BLB}$
1	V_{DD}	V_s	V_s	$V_{BL} > V_{BLB}$
X	V_{DD}	V_s	V_s	$V_{BL} = V_{BLB}$

$V_{DD} = 700$ mV, $V_s = 50$ mV.

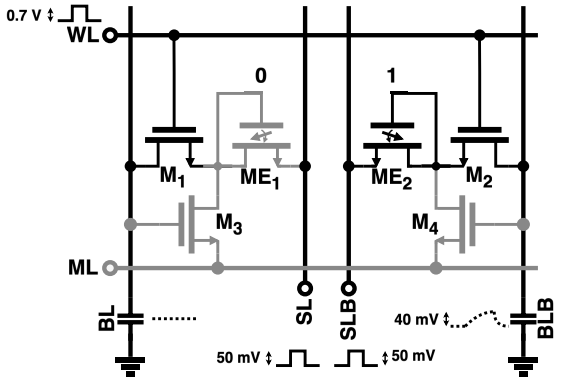


FIGURE 6. Read operation of a single CAM cell. The devices in lighter colors are OFF.

the write operation, along with the resistance states of the MEFETs. R_H is the high resistance state when the resistance of the MEFET is ≈100 MΩ. R_L is the low resistance state, when the resistance of the MEFET is ≈1 kΩ. The do not care state, "X," is represented by storing the bit "0" in both MEFETs. It is to be noted that the drain and source of the access transistors reach negative voltages during the write operation. However, V_w is chosen such that the body diode of the access transistors is not forward biased.

Fig. 5 shows the write operation of a single CAM cell if the bit "0" is being written. BL and SL are driven to −200 mV, BLB and SLB are driven to 200 mV, while the word is driven to V_{DD}. Since BL and SL are at −200 mV, the gate of ME_1 would be at −200 mV. As the BG terminals of the MEFETs are connected to the ground, ME_1 switches to a high resistance state. Similarly, the gate voltage of ME_2 is 200 mV, and ME_2 switches to a low resistance state.

B. READ OPERATION

The read operation is performed by applying a sense voltage (V_s) to SL and SLB when WL is high. The BL and BLB capacitances are discharged before the read operation. A small current flows through the MEFETs, charging the BL and BLB capacitances. A small MEFET resistance results in a larger voltage on the bit lines, while a large MEFET resistance results in a smaller voltage. These voltage levels are measured by a sense amplifier to determine the resistance state of the MEFETs. Table 2 shows the voltages of various lines during the read operation.

Fig. 6 shows the read operation of a single CAM cell if the bit stored is "0." Both SL and SLB are driven to 50 mV, and the WL is driven to V_{DD} to read the data. As shown in Fig. 6, the transistors M_3 and M_4 would be OFF, and the MEFET ME_1 would be in the high resistance state. Since the BL and BLB are not driven in the read phase, the BL and BLB capacitances would be charged through the MEFETs. Since the resistance of ME_1 is high, BL would be charged to a much lower voltage than BLB.

C. SEARCH OPERATION

The XOR transistors M_3 and M_4 in Fig. 4 are used to perform the search operation, which is performed by applying the search bit to the BLB and the complement of the search bit to the BL while the WL is low. The ML capacitance is discharged, and a sense voltage is applied to SL and SLB. The ML capacitance will be charged if the search bit does not match the stored bit. Table 3 shows the voltages of various lines during the search operation. The ML voltage is 0 when the search bit matches the stored bit and starts increasing when the search bit does not match the stored bit. Therefore, the voltage at ML would not increase if there is a perfect match.

Fig. 7 shows the search operation in a single CAM cell. The stored bit is "0" and the search bit is "1." The BL is driven to 0 V, and the BLB is driven to V_{DD} while the WL is low. The SL and SLB are driven to 50 mV. The transistor M_3 would be OFF, and the MEFET ME_2 would be in the low resistance state. Since the transistor M_4 is ON, there exists a path from SLB to ML for the current to flow, charging the ML capacitance. In the worst case of only one mismatch, the ML voltage would be the same as the read operation. With more mismatches, the ML would charge faster.

Fig. 8 shows the simulation results of the write, read, and search operations of one cell in a CAM array. The simulation was done in Cadence Spectre using the Verilog-A model of the MEFET [31] and ASAP 7-nm predictive PDK [35] for the access transistors and peripheral circuitry. The write voltage,

TABLE 3. 4T-2M MEFET-based CAM cell search operation.

Search Bit	Stored Data	WL	BL	BLB	SL	SLB	V_{ML}
0	0	0	V_{DD}	0	V_s	V_s	0
	1	0	V_{DD}	0	V_s	V_s	↑
	X	0	V_{DD}	0	V_s	V_s	0
1	0	0	0	V_{DD}	V_s	V_s	↑
	1	0	0	V_{DD}	V_s	V_s	0
	X	0	0	V_{DD}	V_s	V_s	0
X	0	0	0	0	V_s	V_s	0
	1	0	0	0	V_s	V_s	0
	X	0	0	0	V_s	V_s	0

$V_{DD} = 700$ mV, $V_s = 50$ mV.

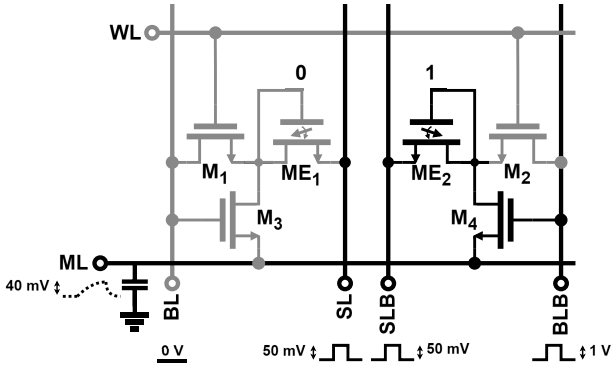


FIGURE 7. Search operation in a single CAM cell. The devices in lighter colors are OFF.

V_w, is taken to be 200 mV, sense voltage, V_s, is taken to be 50 mV, and the supply voltage, V_{DD}, is taken to be 700 mV.

IV. CAM ARRAY

The proposed 4T-2M MEFET-based CAM cell is used to build a CAM array, as shown in Fig. 9. The major blocks in the peripheral circuitry are the word-line decoder and the sense amplifier. The decoder is used to select the row of the CAM array for the read and write operations. The sense amplifier is used to detect the high and low sense voltages during the read and search operations.

A. WORD-LINE DECODER

The word-line decoder is used to generate the word lines based on the address to be selected for reading and writing data. These lines must be synchronous and monotonically increasing. Fig. 10 shows the dynamic NOR gate [36], [37] used in the word line decoder. This gate has a monotonically rising output with a race-based structure.

B. SENSE AMPLIFIER

The read and search operations require the measurement of small voltages (<50 mV) on the bit lines and ML, respectively. A sense amplifier is used to detect the voltage levels on BL and BLB lines to determine the stored bit. The reference voltage for the sense amplifier is generated using a stack of 5 "ON"-state CAM cells charging a dummy BL

TABLE 4. Comparison of the proposed MEFET-based CAM with the previous nonvolatile CAM structures.

Device	MEFET	ME-MTJ [29]	FEFET [25]	SRAM [29]
Technology	7 nm	14 nm	14 nm	14 nm
Word Line (V)	0.7	0.7	0.7	0.7
Array Size	64 × 64	64 × 64	64 × 64	64 × 64
Cell Area (μm^2)	0.049	0.131	0.161	0.21
Write Voltage (V)	0.2	0.2	4	0.7
Write Energy (fJ)	**12.74**	105	183	1510
Write Delay (ns)	**0.08**	1	10	1
Search Voltage (V)	**0.05**	0.8	1	0.7
Search Energy	**4.12 fJ**	537 fJ	353 fJ	723 fJ
Search Delay	594 ps	433 ps	328 ps	420 ps
Search EDP* (fJ-ns)	**2.45**	232.52	115.78	303.66
Leakage Power (nW)	**2.38**	0.71	6.8	695

*EDP: Energy-delay product.

capacitance. The sense amplifier is designed using offset cancellation techniques [38] and a cascade of a preamplifier and a strongARM latch-based comparator structure [39], as mentioned in [33].

Table 4 shows the simulation results of a 64 × 64 CAM array using the typical process corner at 27°C and compares it with other CAM structures. The write energy reported is the energy required to write the data and its complement to one CAM cell. The search energy reported is the energy required per ML during the search operation, when there is a mismatch in 50% of the cells in the row. The leakage power report is the per-cell leakage power.

The SRAM-based structure in Table 4 is a 16T CMOS structure taken from [29]. As shown in Table 4, the write energy of the proposed MEFET-based CAM is 12.74 fJ, which is significantly lower than the write energy of prior works. The write delay improvement of the proposed CAM is >12× that of prior works, due to the quick switching of the MEFET. The search energy of the proposed CAM is two orders of magnitude lower than the search energy of the ME-MTJ-based CAM and the FEFET-based CAM, due to the low voltage requirement of the search operation (50 mV). The energy-delay product of the proposed CAM is >45× lower than that of the prior works.

V. VARIABILITY ANALYSIS

Fig. 11 shows the variation of the energy and delay of the write, read, and search operations with temperature, process

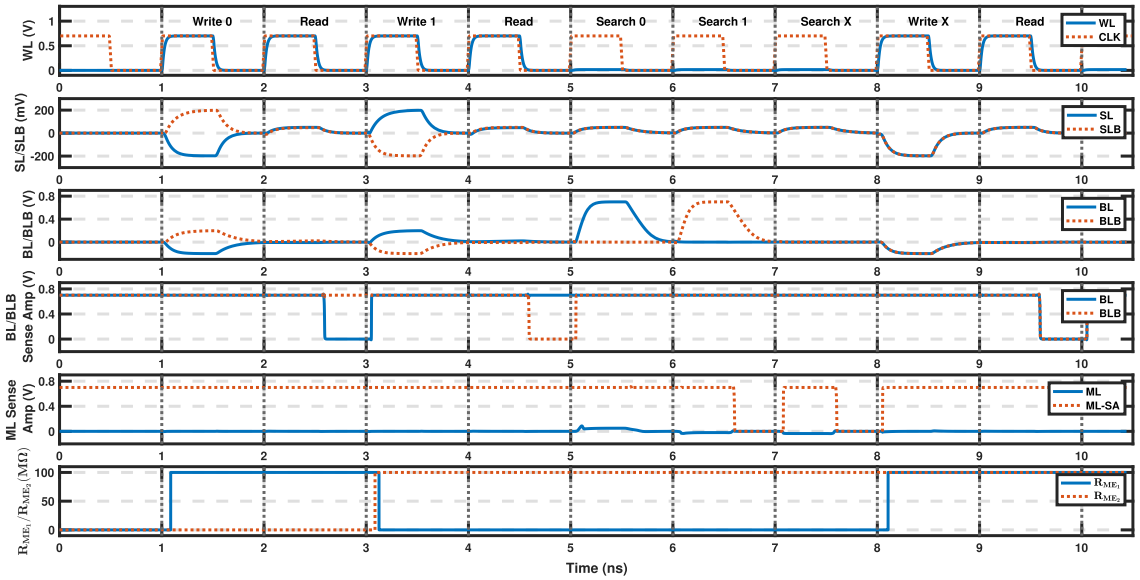


FIGURE 8. Proposed CAM cell operations for write, read, and search operations showing voltage levels of word-line, bitline, source line, and ML outputs of the sense amplifier on the source line and ML, and resistance of ME_1 and ME_2. All the lines are discharged in the clock-low phase.

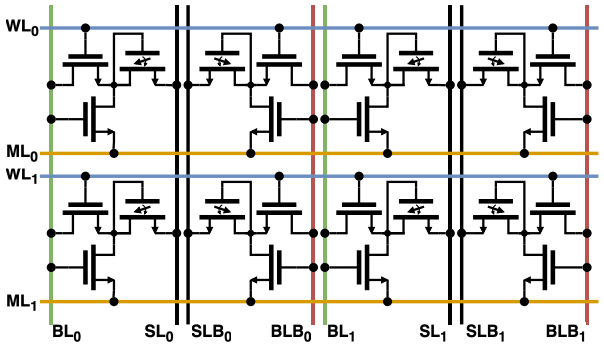


FIGURE 9. 4T-2M MEFET-based 2 x 2 CAM array.

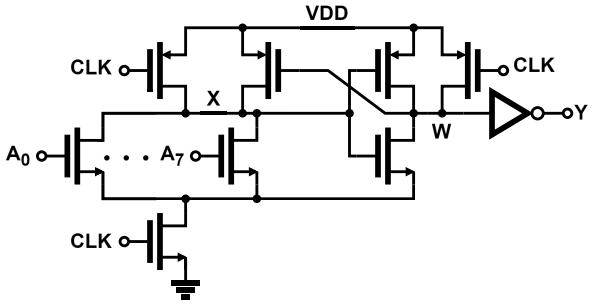


FIGURE 10. Eight-input dynamic NOR gate.

corners, and capacitance variation of the BL, SL, and ML. The temperature is varied from −40 °C to 125 °C. The process corners are typical, slow, and fast. The capacitances of the BL, SL, and ML are varied by ±10% from the nominal capacitance of 100 fF and are indicated by the error bars in Fig. 11. It is noted that the energy and delay of the write, read, and search operations are not affected by the variability of the MEFET resistance.

The read and search delays do not vary (<0.2% variation) with the capacitance variation of the BL, SL, and ML as the sense amplifier senses the voltage levels in the clock low phase. The final low and high voltages on the bit line and ML during the read and search operations are not significantly affected by the capacitance variation of BL and ML. The read and search energy are also not affected by the capacitance variation of BL, SL, and ML as the sense voltage is low (50 mV). The write energy and delay, however, are affected by the capacitance variation of BL as the required voltages at the BL and BLB during the write are higher (200 mV).

VI. ONE-SHOT LEARNING

The one-shot learning paradigm is a type of learning where a neural network model is used to learn from one or a few examples. The model is trained to generate embeddings of the input data such that the embeddings of the same class are close to each other and the embeddings of different classes are far from each other. The model is then used to classify the input data based on the distance between the embeddings. If the learned embeddings are binary, the classification can be performed using a CAM to search for the closest match. The network structure used in this work is a Siamese network [5], which consists of two identical subnetworks that share the same weights as shown in Fig. 12. The input data are passed through the subnetworks to generate the embedding vectors. The distance between the embeddings will be small if the input data is from the same class and large if the input data is from different classes.

The one-shot learning paradigm can be broken down into three stages: training the network to learn the embeddings, generating the embeddings of a class with one or few examples, and classifying the test data based on the distance between the embeddings. Fig. 13 shows the overall one-shot learning framework. In this work, two networks described in Table 5 are used, targeting the Modified National Institute of Standards and Technology (MNIST) [40] and extended MNIST (EMNIST) datasets [41]. N_{emb} is the size of the

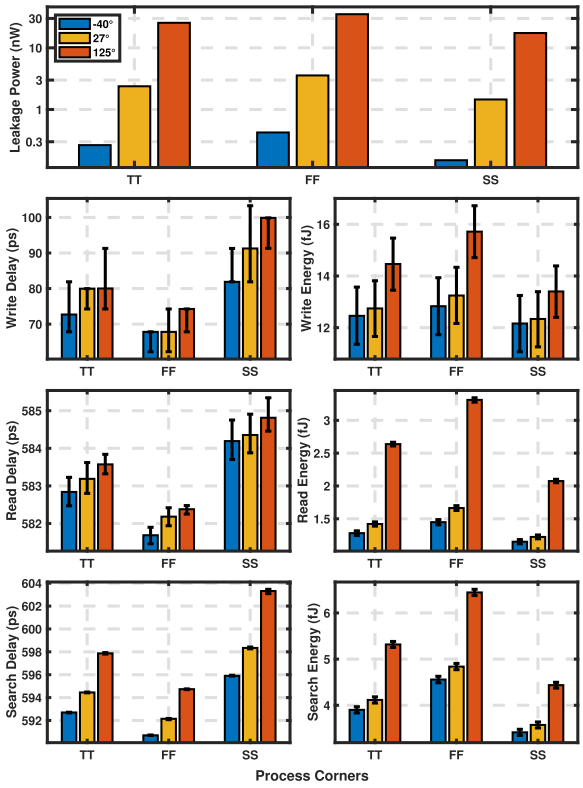


FIGURE 11. Variation of the energy and delay of the write, read, and search operations, and the leakage power with temperature, process corners, and capacitance variation of BL, SL, and ML. The temperature is varied from −40 °C to 125 °C. The process corners are typical, slow, and fast. The capacitance of the BL, SL, and ML is varied by ±10%. The error bars indicate the variation of the parameters with the capacitance variation.

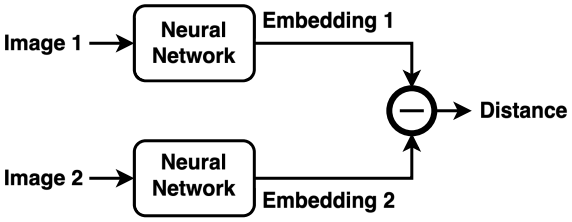


FIGURE 12. Siamese network structure.

embedding vector.

To facilitate binarization of the embeddings, sigmoid activation is used in the output layer of the networks. The distance metric used is the $L1$ distance, given by s the following equation:

$$d\,(x, y) = \sum_{i=1}^{N_{\text{emb}}} |x_i - y_i|\,. \tag{1}$$

During evaluation, the embeddings are binarized using a threshold of 0.5, which is equivalent to the sign function in place of the sigmoid function. The $L1$ distance between the binarized embeddings is the Hamming distance.

The training data provided in the dataset is split into training and validation sets for both networks. The training set is used to train the network, and the validation set is used to tune the hyperparameters in the step shown in Fig. 13(a). The trained network is then used to generate the embeddings of the input data in the step shown in Fig. 13(b). The embeddings are generated from one randomly selected example of each class from the validation set. The embeddings are stored in memory and then used to classify the test data in the step shown in Fig. 13(c). The test data are passed through the network to generate the embeddings, and the distance from the stored embeddings is computed to classify the test data.

Table 6 shows the classification accuracies of the one-shot learning paradigm for the MNIST and EMNIST datasets for different embedding sizes. A classification accuracy of >99% is achieved for the MNIST dataset with an embedding size of 16. For the EMNIST dataset, a classification accuracy of 79.03% is achieved with an embedding size of 128. The EMNIST dataset is more challenging than the MNIST dataset as it contains more classes and variations. The top-3 accuracy of the EMNIST dataset is 97.11% with an embedding size of 64. The top-3 accuracy is the percentage of test data that has the correct class in the top-3 predictions. The top-3 accuracy is targeted for the EMNIST dataset in this work.

TABLE 5. Details of the networks used for the two datasets. N_{emb} is the size of the embedding vector.

Dataset	Layer	Output Shape	# Params
MNIST	Conv 1	$(28 \times 28 \times 8)$	80
	Maxpool	$(14 \times 14 \times 8)$	-
	Conv 2	$(14 \times 14 \times 16)$	1168
	Maxpool	$(7 \times 7 \times 16)$	-
	FC 1	128	100,608
	Output	N_{emb}	$129 \times N_{emb}$
EMNIST	Conv 1	$(28 \times 28 \times 8)$	80
	Maxpool	$(14 \times 14 \times 8)$	-
	Conv 2	$(14 \times 14 \times 16)$	1168
	Maxpool	$(7 \times 7 \times 16)$	-
	Conv 3	$(7 \times 7 \times 32)$	4640
	Maxpool	$(3 \times 3 \times 32)$	-
	FC 1	256	73,984
	Output	N_{emb}	$257 \times N_{emb}$

TABLE 6. Classification accuracy of the two datasets.

Embedding Size	Accuracy (%)		
	MNIST	EMNIST	EMNIST (Top-3)
4	98.20	30.46	82.04
8	98.82	64.03	94.35
16	**99.06**	72.19	96.57
32	99.08	77.09	96.74
64	99.05	75.35	**97.11**
128	99.08	**79.03**	97.32
256	99.09	79.17	97.02

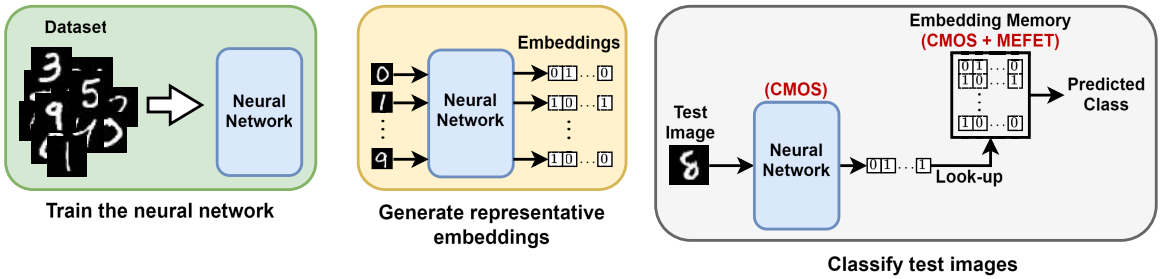


FIGURE 13. One-shot learning framework with (a) training the network on a large dataset, (b) generation of representative embeddings with one or few examples, and (c) classification of the test data based on the distance between the embeddings.

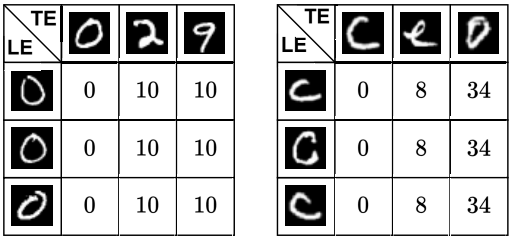


FIGURE 14. Hamming distance of the embeddings for example images from (a) MNIST dataset and (b) EMNIST dataset. LE stands for learning examples, and TE stands for testing examples. The maximum possible Hamming distance for MNIST and EMNIST is 16 and 64, respectively.

Fig. 14 shows the Hamming distance of the embeddings, for example, images from the MNIST and EMNIST datasets. The embedding size for MNIST is 16, and the embedding size for EMNIST is 64. As shown in Fig. 14(a), the Hamming distance between the testing example (TE) "0" and the three learning examples (LEs) "0" is 0, and the distance between the TEs "2" and "9" and the LEs "0" is 10 (the maximum possible distance in this case is 16). As shown in Fig. 14(b), the Hamming distance between the TE "*C*" and the three LEs "*C*" is 0. However, the distance between the TE "*e*" and LEs "*C*" is also small as they are visually similar. The distance between the TE "0" and the LEs "*C*" is 34. (The maximum possible distance in this case is 64.)

The paradigm is also tested on a more complex dataset, labeled faces in the wild (LMW) [42], using a pretrained model that generates embeddings from images of faces [6]. The pretrained model generates embeddings of size 512. However, these embeddings are 32-bit floating-point numbers. The embeddings are binarized by simple sign-based thresholding to fit into the proposed methodology. The classification accuracy postbinarization is 97.1%, which is less than 1% worse than the floating-point embeddings.

The representative embeddings generated in Fig. 13(b) are stored in the proposed MEFET-based CAM. The EMINST dataset is chosen for simulation as it has 47 classes, and each having 64-bit wide embeddings. One image from each class is randomly picked, and the embeddings are stored in a 64 × 64 array. The same images are given as search patterns to estimate the search energy. The average mismatch with these patterns is 38.58%, resulting in a search energy of 3.18 fJ. This result shows that the search energy is a linear function of the average mismatch as a mismatch of 50% in Table 4 had a search energy of 4.12 fJ.

The number of rows in the CAM increases with the number of classes to be classified. The neural network used to generate the embeddings in Fig. 13(c) can be implemented in CMOS using hardware accelerator architectures [43], [44]. For a large number of classes, the area, delay, and power of the memory become the bottleneck instead of the neural network. The proposed MEFET-based CAM can be used to store the embeddings and classify the test data with low energy and delay and with a smaller area than a CMOS CAM.

VII. CONCLUSION

This work proposed a novel MEFET-based TCAM cell that can perform read, write, and search operations with very low programming voltages of ≤200 mV. The proposed MEFET-based TCAM cell was used to build a 64 × 64 CAM array, and the energy and delay of the read, write, and search operations were evaluated via simulation. The improvement in write energy and write delay of the proposed CAM array were shown to be 8× and 12×, respectively, compared with other nonvolatile device-based CAM structures. The search energy of the proposed CAM array was shown to be orders of magnitude lower than that of prior works, and the search energy–delay product was shown to be 45× lower than that of prior works. The energy and delay of the array were evaluated with temperature, process corners, and capacitance variation. The proposed MEFET-based CAM array was used to store the representative embeddings generated by a neural network in the one-shot learning paradigm. The classification accuracies of the one-shot learning paradigm for the MNIST, EMNIST, and LMW datasets were evaluated for different embedding sizes. The accuracy was found to be >99% for the MNIST dataset with an embedding size of 16, the top-3 accuracy was 97.11% for the EMNIST dataset with an embedding size of 64, and the accuracy was 97.1% on the LMW dataset with an embedding size of 512. The proposed MEFET-based CAM array can be used to store the embeddings and classify the test data with low energy and delay and with a smaller area than a CMOS CAM. Future work would include improving the search delay and developing better techniques to improve the search margin.

REFERENCES

[1] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing machines," 2014, *arXiv:1410.5401*.

[2] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *Proc. Intl. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.

[3] A. Graves et al., "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, Oct. 2016.

[4] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, Jun. 2016, pp. 1842–1850.

[5] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, 2015.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[7] O. Vinyals et al., "Matching networks for one shot learning," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Barcelona, Spain, 2016, pp. 1–9.

[8] C. Afghahi and B. Sahoo, "Content addressable memory cell techniques," U.S. Patent 6 529 395 B1, Mar. 4, 2003.

[9] Y. Chen, L. Lu, B. Kim, and T. T. Kim, "Reconfigurable 2T2R ReRAM architecture for versatile data storage and computing in-memory," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 12, pp. 2636–2649, Dec. 2020.

[10] K. Mishty and M. Sadi, "Designing efficient and high-performance AI accelerators with customized STT-MRAM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 10, pp. 1730–1742, Oct. 2021.

[11] T. Li, Y. Ma, K. Yoshikawa, and T. Endoh, "Hybrid signed convolution module with unsigned divide-and-conquer multiplier for energy-efficient STT-MRAM-based AI accelerator," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 31, no. 7, pp. 1078–1082, Jul. 2023.

[12] Z. Lu et al., "An RRAM-based computing-in-memory architecture and its application in accelerating transformer inference," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 32, no. 3, pp. 485–496, Mar. 2024.

[13] C. Matsui, K. Toprasertpong, S. Takagi, and K. Takeuchi, "FeFET local multiply and global accumulate voltage-sensing computation-in-memory circuit design for neuromorphic computing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 32, no. 3, pp. 468–479, Mar. 2024.

[14] W. Xu, T. Zhang, and Y. Chen, "Design of spin-torque transfer magnetoresistive RAM and CAM/TCAM with high sensing and search speed," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 1, pp. 66–74, Jan. 2010.

[15] B. Yan, Z. Li, Y. Chen, and H. Li, "RAM and TCAM designs by using STT-MRAM," in *Proc. 16th Non-Volatile Memory Technol. Symp. (NVMTS)*, Pittsburgh, PA, USA, Oct. 2016, pp. 1–5.

[16] L. Xue, Y. Cheng, J. Yang, P. Wang, and Y. Xie, "ODESY: A novel 3T-3MTJ cell design with optimized area density, scalability and latency," in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD)*, Austin, TX, USA, Nov. 2016, pp. 1–8.

[17] J. Min, C. Kim, S.-Y. Kim, and K.-W. Kwon, "A study of read margin enhancement for 3T2R nonvolatile TCAM using adaptive bias training," in *Proc. IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, Aug. 2019, vol. 27, no. 8, pp. 1840–1850.

[18] B. Rajendran et al., "Demonstration of CAM and TCAM using phase change devices," in *Proc. 3rd IEEE Int. Memory Workshop (IMW)*, Monterey, CA, USA, May 2011, pp. 1–4.

[19] J. Li, R. K. Montoye, M. Ishii, and L. Chang, "1 Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 896–907, Apr. 2014.

[20] P. Junsangsri, J. Han, and F. Lombardi, "Design and comparative evaluation of a PCM-based CAM cell," *IEEE Trans. Nanotechnol.*, vol. 16, no. 2, pp. 359–363, Mar. 2017.

[21] L.-Y. Huang et al., "ReRAM-based 4T2R nonvolatile TCAM with 7x NVM-stress reduction, and 4x improvement in speed-wordlength-capacity for normally-off instant-on filter-based search engines used in big-data processing," in *Proc. Symp. VLSI Circuits Dig. Tech. Papers*, Honolulu, HI, USA, Jun. 2014, pp. 1–2.

[22] C.-C. Lin et al., "A 256b-wordlength ReRAM-based TCAM with 1ns search-time and 14x improvement in wordlength-energyefficiency-density product using 2.5T1R cell," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Feb. 2016, pp. 136–137.

[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 J. Solid-State Circuits*, vol. 52, no. 6, pp. 1664–1679, 2017.

[24] X. Yin, M. Niemier, and X. S. Hu, "Design and benchmarking of ferroelectric FET based TCAM," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2017, pp. 1444–1449.

[25] X. Yin, K. Ni, D. Reis, S. Datta, M. Niemier, and X. S. Hu, "An ultra-dense 2FeFET TCAM design based on a multi-domain FeFET model," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 9, pp. 1577–1581, Sep. 2019.

[26] A. J. Tan et al., "Experimental demonstration of a ferroelectric HfO_2-based content addressable memory cell," *IEEE Electron Device Lett.*, vol. 41, no. 2, pp. 240–243, Feb. 2020.

[27] X. Yin et al., "FeCAM: A universal compact digital and analog content addressable memory using ferroelectric," *IEEE Trans. Electron Devices*, vol. 67, no. 7, pp. 2785–2792, Jul. 2020.

[28] S. Matsunaga et al., "MTJ-based nonvolatile logic-in-memory circuit, future prospects and issues," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, Nice, France, Apr. 2009, pp. 433–435.

[29] S. Narla et al., "Modeling and design for magnetoelectric ternary content addressable memory (TCAM)," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 8, no. 1, pp. 44–52, Jun. 2022.

[30] J. A. Kelber, C. Binek, P. A. Dowben, and K. Belashchenko, "Magneto-electric voltage controlled spin transistors," U.S. Patent 0 23 188 8A1, Aug. 21, 2014.

[31] N. Sharma, "Circuit level modeling of spintronic devices," Ph.D. dissertation, Dept. Elect. Eng., Univ. Texas Dallas, Richardson, TX, USA, 2017.

[32] P. A. Dowben et al., "Towards a strong spin–orbit coupling magneto-electric transistor," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 4, no. 1, pp. 1–9, Jun. 2018.

[33] S. Angizi, N. Khoshavi, A. Marshall, P. Dowben, and D. Fan, "MeF-RAM: A new non-volatile cache memory based on magneto-electric FET," *ACM Trans. Design Autom. Electron. Syst.*, vol. 27, no. 2, pp. 1–18, Mar. 2022.

[34] S. Angizi, D. Fan, A. Marshall, and P. Dowben, "Nonvolatile memory based architectures using magnetoelectric FETs," in *Advances in Semiconductor Technologies: Selected Topics Beyond Conventional CMOS*. Hoboken, NJ, USA: Wiley, 2023, pp. 79–92.

[35] L. T. Clark et al., "ASAP7: A 7-nm finFET predictive process design kit," *Microelectron. J.*, vol. 53, pp. 105–115, Jul. 2016.

[36] H. Nambu et al., "A 1.8-ns access, 550-MHz, 4.5-Mb CMOS SRAM," *IEEE J. Solid-State Circuits*, vol. 33, no. 11, pp. 1650–1658, 1998.

[37] B. S. Amrutur and M. A. Horowitz, "Fast low-power decoders for RAMs," *IEEE J. Solid-State Circuits*, vol. 36, no. 10, pp. 1506–1515, 2001.

[38] S. Konwar, H. Roy, S. W. Chiang, and B. D. Sahoo, "Deterministic dithering-based 12-b 8-MS/s SAR ADC in 0.18-μm CMOS," *IEEE Solid-State Circuits Lett.*, vol. 5, pp. 243–246, 2022.

[39] B. Razavi, "The StrongARM latch [a circuit for all seasons]," *IEEE Solid State Circuits Mag.*, vol. 7, no. 2, pp. 12–17, Jun. 2015.

[40] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[41] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: An extension of MNIST to handwritten letters," 2017, *arXiv:1702.05373*.

[42] G. B. Huang, M. Ramesh, T. Berg, and E. L. Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Manning College Inf. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[43] X. Zhang et al., "DNNbuilder: An automated tool for building high-performance DNN hardware accelerators for FPGAs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Diego, CA, USA, 2018, pp. 1–8.

[44] S. Sanjeet, B. D. Sahoo, and M. Fujita, "Energy-efficient FPGA implementation of power-of-2 weights-based convolutional neural networks with low bit-precision input images," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 2, pp. 741–745, Feb. 2023.

[45] Y. Zhang et al., "Switching endurance of the molecular spin crossover complex [Fe(HB(Tz)3)2]: From single crystals to thin films and electronic devices," *Mater. Adv.*, vol. 3, no. 22, pp. 8193–8200, 2022.