Hybrid Magneto-electric FET-CMOS Integrated Memory Design for Instant-on Computing

Deniz Najafi[†], Sepehr Tabrizchi[§], Ranyang Zhou[†], Mohammadreza Amel Solouki[*], Andrew Marshall[**], Arman Roohi[§], and Shaahin Angizi[†]

[†] Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA
[§] School of Computing, University of Nebraska–Lincoln, Lincoln, NE, USA
[*] Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy
[**] Department of Electrical and Computer Engineering, The University of Texas at Dallas, TX, USA
dn339@njit.edu,aroohi@unl.edu,shaahin.angizi@njit.edu

ABSTRACT

The surge in the number of normally-off power-constraint Internet of Things (IoT) devices in recent years has amplified the demand for high-performance and energy-efficient in-memory computing architectures built on top of various non-volatile memories. Magneto-Electric Field Effect Transistors (MEFETs) have presented compelling design features suitable for logic and memory integration as an emerging post-CMOS FET. These include high-speed switching, minimal power usage, and non-volatility. This work introduces a new in-memory computing architecture designed for edge applications, leveraging emerging MEFETs. The proposed architecture enables the execution of both Boolean logic operations and Binary Content Addressable Memory (BCAM) operations within a single cycle. Furthermore, the energy consumption during the write operation of the proposed cell is optimized by introducing a new write circuitry. The outcomes of our device-to-architecture evaluation reveal approximately 43.5% and 96.9% reduction in read and write energy consumption, respectively, compared to the counterpart non-volatile memories. At the application level, the proposed architecture is applied to implement Binary Neural Networks (BNNs) based on AlexNet and VGG16. Our results showcase a decrease of approximately 54% in the overall energy consumption when implementing these networks using the proposed design compared to non-volatile in-memory computing designs.

CCS CONCEPTS

• **Hardware → Spintronics and magnetic technologies**.

ACM Reference Format:
Deniz Najafi[†], Sepehr Tabrizchi[§], Ranyang Zhou[†], Mohammadreza Amel Solouki[*],, Andrew Marshall[**], Arman Roohi[§], and Shaahin Angizi[†]. 2024. Hybrid Magneto-electric FET-CMOS Integrated Memory Design for Instant-on Computing. In *Great Lakes Symposium on VLSI 2024 (GLSVLSI '24), June 12–14, 2024, Clearwater, FL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3649476.3660361

GLSVLSI '24, June 12–14, 2024, Clearwater, FL, USA

ACM ISBN 979-8-4007-0605-9/24/06
https://doi.org/10.1145/3649476.3660361

1 INTRODUCTION

In recent years, the surge in the number of sensors and applications within the Internet of Things (IoT) has amplified the necessity to transmit data back and forth between sensors and the cloud. This trend has introduced various issues such as increased energy consumption, heightened latency, and diminished security [2, 11, 17, 30]. The escalating necessity to extend the operational lifespan of battery-constrained IoT edge devices has driven a swift rise in attention toward integrating emerging Non-Volatile Memory (NVM) technologies into edge devices. This interest is primarily propelled by the distinctive characteristics of NVMs, such as non-volatility, durability, extended endurance, high integration density, remarkably low standby power consumption, and compatibility with intermittent computing [7, 9, 22, 23, 25]. Particularly for embedded applications and low-power IoT systems reliant on on-chip cache, integrating robust NVMs has the potential to augment memory capacity and performance [16].

The NVM technologies have evolved significantly, with Resistive RAM (ReRAM) and Phase Change Memory (PCM) emerging as promising alternatives to DRAM/SRAM due to their higher ON/OFF ratio and packing density (~2-4×) [6]. However, they face challenges such as slow write operations, high power consumption, and low endurance (~ 10^5-10^{10}) [23, 31]. Ferroelectric transistor RAMs (FERAMs) offer advantages in endurance and sense margin with a reduced 1-10 ns [28] write time and could be a possible alternative. The downside is their large write voltage ($> 4.0V$) and power consumption [10, 28]. Spin-based NVMs have attracted attention for their sub-nanosecond switching speeds, long retention times (10 years), and low energy consumption. These NVMs utilize Spin-Transfer Torque (STT) or Spin-Orbit Torque (SOT) to manipulate magnetization [15]. However, they exhibit poor ON/OFF ratios and face reliability issues due to high current densities and power dissipation. The Magneto-Electric Field-Effect Transistor (MEFET), leveraging the antiferromagnetic Magneto-Electric (ME) phenomena, has recently undergone experimental investigation [16, 22, 23]. This spintronic device displays promising attributes, characterized by enhanced performance and heightened temperature resilience. What distinguishes the MEFET from conventional spintronic devices are its notably high switching speed (<20 ps), high ON/OFF ratio (e.g., 10^6 for WSe_2), and low energy consumption (<20 aJ) by capitalizing on coherent rotation as the domain switching mechanism, thereby eliminating the necessity for ferromagnetic switching or domain wall movement [8, 22].

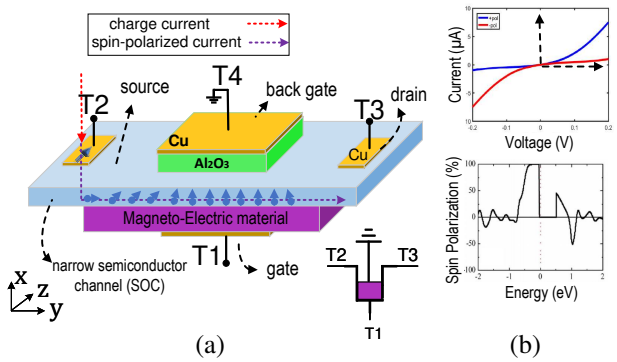

Figure 1: (a) MEFET device and the circuit scheme, (b) Sample source-to-drain current versus voltage at T1 and the induced spin polarization in WSe_2.

Drawing upon the promising attributes of MEFET, this study introduces, for the first time, a novel non-volatile hybrid MEFET-CMOS architecture demonstrating proficiency in executing Boolean logic operations alongside Binary Content Addressable Memory (BCAM) within a single cycle. The study explores its potential for advancing high-performance and energy-efficient IoT applications. The primary contributions of this research are outlined as follows:

(1) We design a hybrid MEFET-CMOS integrated memory cell based on a set of efficient circuit-level and micro-architectural schemes. We develop a new write circuitry to facilitate a one-cycle write operation for the proposed cell, comprising both the data and its complementary counterpart; (2) We enable an in-memory bit-line computing scheme based on the proposed architecture to implement various Boolean logic operations alongside BCAM within a single cycle; and (3) We develop a bottom-up evaluation framework to showcase the performance of the proposed design against the well-known non-volatile memory candidates running the Binary Neural Networks (BNNs) acceleration task.

2 MEFET BASICS

The Magneto-Electric Spin Field Effect Transistor (MEFET) demonstrates structural similarities with the CMOS FET device. Fig. 1(a) illustrates the basic single-source configuration of the MEFET, a four-terminal device comprising gate (T1), source (T2), drain (T3), and back gate (T4) terminals [22, 27], [12]. This device consists of a narrow semiconductor channel situated between two dielectrics: the magneto-electric (ME) material, such as chromia (Cr_2O_3), and the insulator, such as alumina (Al_2O_3). Various materials like PbS, graphene, InP, or WSe_2 can be employed to form the narrow semiconductor channel. Two electrodes are connected to this stacked structure, with T1 located at the bottom gate through the ME layer and T4 at the top via the alumina layer forming the back gate. The channel material, tungsten diselenide (WSe_2), yields a high on-off ratio, enhancing hole mobility [22], [12]. Both conductors and FM polarizers can be utilized for the source and drain materials (at the T2/T3 terminal).

The MEFET functions as a transistor by biasing the semiconductor channel via T1 and T4 terminals (similar to gate biasing in CMOS) and then applying current from T2 to T3 (similar to source-drain biasing in CMOS). Applying a very low voltage of approximately ±100 mV [22], [21] across the gate (T1) and back gate terminals (T4: ground) charges the ME capacitor. The change

Table 1: Compact Verilog-A model parameters.

Parameter	Value	Description of Parameter and Units
ϵ_{ME}	12	Dielectric constant of chromia [5]
$\epsilon_{Al_2O_3}$	10	Dielectric constant of Alumina
t_{ME}	10	thickness of magnetoelectric layer, nm
$W_{ME} \times L_{ME}$	900	area of magnetoelectric layer, nm^2
t_{ox}	2	Oxide barrier thickness, nm
V_{th}	0.05	Threshold of Chromia state inversion, V
V_g	0.1	Voltage applied across ME layer, V
R_{on}	1.05	ON Resistance, $k\Omega$
R_{off}	63.4	OFF Resistance, $M\Omega$

in polarity at T1 generates a vertical electric field across the gate, which switches the paraelectric polarization and anti-ferromagnetic (AFM) order in the ME insulator layer, leading to the reorientation of chromia spin vectors. Through exchange interactions and spin-orbit coupling (SOC), the high boundary polarization of the ME layer can polarize carriers' spins in the semiconductor channel, resulting in preferred conduction along a specific axis, such as lower resistance. The surface magnetization of the MEFET's channel induces directionality in conductance, unlike conventional gate dielectrics, ultimately altering the channel spin vector's orientation. Non-Equilibrium Green's Function (NEGF) transport simulations are employed to investigate the current-voltage relationship (Fig.1(b)), dependent on the direction of ME polarization based on [7], [13]. These simulations are conducted on a 2D ribbon with a width of 20 nm and a band mass of $0.1m_e$, considering a conservative exchange splitting value of 0.1 eV, T3–T2 = 0.1 V, at 300 K. To read out the MEFET, the T2-T3 resistive path can be sensed and compared with a reference. The ON/OFF current ratio for (WSe_2) can reach up to 10^6. Table 1 showcases the experimental parameters utilized for the switching behavior of the Chromia layer and SOC channel in our model.

3 PROPOSED ARCHITECTURE

Fig. 2(a) illustrates the proposed architecture, designed to operate effectively in both memory and computing modes. The memory architecture is constructed using the proposed hybrid CMOS-MEFET memory bit-cell as shown in Fig. (2(b)), Row Decoder (RD), Column

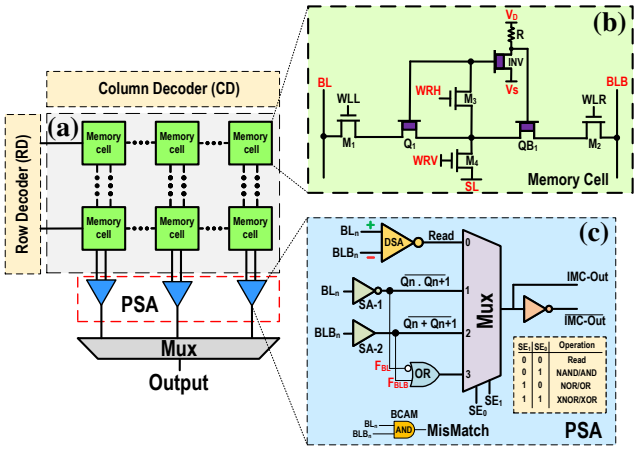

Figure 2: (a) Proposed MEFET-based non-volatile architecture, (b) MEFET memory bit-cell structure, (c) Proposed reconfigurable sense amplifier.

Table 2: Signaling of the Proposed MEFET Cell.

Signals	Hold	Read	Write	Computing (Boolean)	BCAM
WLL	'0'	'1'	'0'	'1'	*Data*
WLR	'0'	'1'	'0'	'1'	$\overline{Data}$
WRH	'1'	'0'	'1'	'0'	'0'
WRV	'0'	'1'	'1'	'1'	'1'
SL	'0'	'0'	V_{WR}	'0'	'0'
BL	'0'	V_{DD}	'0'	V_{DD}	V_{DD}
BLB	'0'	V_{DD}	'0'	V_{DD}	V_{DD}

Decoder (CD), and the Proposed Sense Amplifier (PSA), as defined in the Fig. 2(c). As shown in Fig. 2(b) the suggested hybrid memory cell comprises three MEFETs and four MOSFETs. This configuration allows for the storage of both data and its complementary value within a single cell. This capability facilitates the execution of various logic operations, including X(N)OR, in a single cycle. The memory cell utilizes row-based control signals, namely Word Line Left (WLL), Word Line Right (WLR), Source Line (SL), and Write Line Horizontal (WRH), while the control signals, i.e., Write Line Vertical (WRV), Bit-line (BL) and inverted Bit-line (BLB) operates on a column basis.

As illustrated in Fig. 2(c), the cell value is read out via BL/BLB and transmitted to PSA for sensing and computing. The PSA comprises a 4×1 multiplexer capable of executing read, (N)AND, (N)OR, and X(N)OR logic operations, depending on appropriate control signals. When the control signals are asserted to "00", the BL, and BLB are activated and fed into the Differential Sense Amplifier (DSA) to read the data of the selected cell. Alternatively, if the control signals are set to "01", the BL is connected to the skewed inverter (SA-1), thereby unveiling the (N)AND operation in the output of the PSA. Moreover, a skewed buffer (SA-2) is employed to reveal the NOR output when the PSA control signals are set to "10". Finally, X(N)OR logic is achieved in the PSA by setting the control signals to "11", thereby enabling the OR gate in the PSA. Furthermore, for the implementation of the BCAM operation, the BL and BLB are linked to an AND gate in the PSA to identify data mismatches. Table 2 shows an overview of the necessary signaling for various operations. A comprehensive discussion on the memory and computing modes in the proposed architecture is provided in detail to understand the functionality of the proposed design.

3.1 Memory Mode

In Fig. 3(a), the proposed write circuitry is shown with the simultaneous writing of the intended data and its complement into the cell. For this purpose, the activation of the M4 transistor occurs through WRV, and the SL is connected to the write voltage V_{WR}.

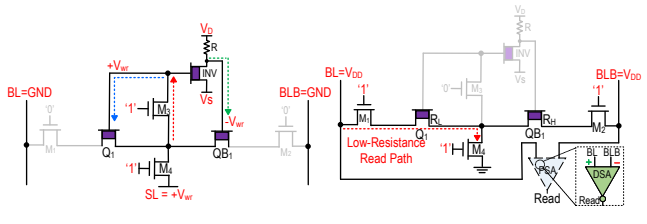

Figure 3: (a) Write and (b) Read operations of the proposed cell.

Simultaneously, the M3 transistor is activated by the WRH. The write voltage and its complement are directed to the gate of Q1 and QB1 MEFETs, respectively. The complement of the write voltage is generated through the MEFET-based inverter, utilizing a fixed resistance in the pull-up network. By appropriately applying voltage to the gate of MEFETs, the resistance of MEFETs is configured. As an illustration, applying positive V_{WR} induces a transition in the resistance state of Q1 to high, concurrently causing the inverted voltage of V_{WR} to impact the gate of the QB1 MEFET (negative voltage), thereby leading to a low-resistance state in this MEFET. In the read process as shown in Fig. 3(b), the SL is connected to the ground, while the BL and BLB are linked to V_{DD} and then left floated. Subsequently, the access transistors (M1 and M2) are activated through the control signals WLL and WLR. Owing to the resistance discrepancy in the MEFETs, which reflects the stored data, either BL or BLB discharges more rapidly than the other. This voltage difference can be detected by the DSA in the PSA, revealing the stored data as shown in Fig. 3(b). For instance, if the data stored in the Q1 MEFET is '1' (high resistance, R_H), and QB1 stores its complementary '0' (low resistance, R_L), BLB discharges much faster than BL, and this differentiation is identified by the PSA, which is set up to function in read mode by control signals SE1 and SE0. These signals are configured to "00" to activate the DSA in the PSA.

3.2 Computing Mode

Bit-line Logic Computation Core: The proposed architecture enables the execution of row-wise X(N)OR, (N)AND, and (N)OR logic operations. To accomplish this, the logic operands need to be initially arranged in corresponding columns. The RD control unit triggers the memory cells containing the operands, as depicted in Fig. 4, across the entire physical memory row. To this end, the BL and BLB are connected to V_{DD} and left floated. Subsequently, two columns within the memory structure which contains the logic operands are activated by applying appropriate control signals to their access transistors and connecting their SL to the ground by activating the column-based WRVn signal. As shown in Fig. 4, depending on the resistance of MEFETs representing the stored data, the BL and BLB either discharge or remain unaffected. If either data stored in MEFETs (Q1 and Qn) is set to '0' (i.e., low resistance), the BL is discharged. By configuring the PSA's congratulation bits to "01", the *BL* is connected to SA-1 that is properly skewed to unveil the output of the NAND operation (F_{BL}).

It is noteworthy that, owing to the symmetry of the proposed bit-cell with inverted inputs, the opposite side connected to BLB simultaneously generates a NOR logic. In this scenario when the appropriate control signal "10" is asserted to the PSA, *BLB* is connected to the skewed buffer (SA-2), thus revealing the NOR logic of the stored data as shown in Fig. 4. This logic is achieved based on Equation 1.

$$F_{BLB} = QB1.QBn = \overline{Q1}.\overline{Qn} = \overline{Q1 + Qn} \quad (1)$$

For instance, when Q1 and Qn are '1' (high resistance, R_H), BL remains untouched. Simultaneously, QB1 and QBn are set to '0' (complementary of Q1 and Qn), leading to the discharge of BLB at the same time. Given that *BL* signifies the NAND operation of stored data after crossing the SA-1 (F_{BL}) and *BLB* represents the AND operation of the inverted stored data of selected rows, the

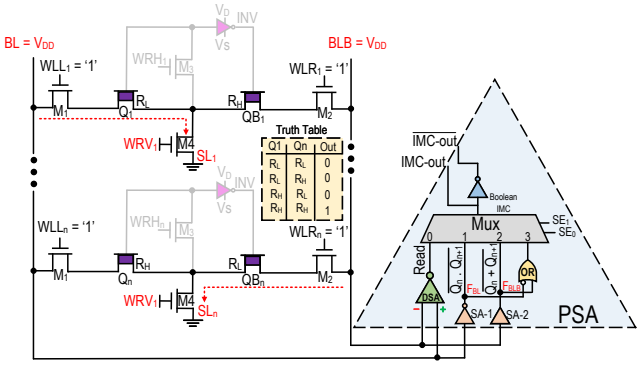

Figure 4: Performing Boolean in-memory computing.

implementation of XNOR logic can be accomplished by connecting the inverted output of SA-1 and the output of SA-2 to an OR gate in the PSA (Fig. 4). Equation 2 represents the output of the XNOR operation.

$$F_{XNOR} = \overline{F_{BL}} + F_{BLB} = \overline{\overline{Q1.Qn}} + \overline{(Q1 + Qn)} = (Q1.Qn) + (\overline{Q1}.\overline{Qn}) \quad (2)$$

BCAM Core: The proposed architecture can be readily reconfigured to realize BCAM to directly compare the stored data with the input provided, enabling swift retrieval of information. To this end, the BL and BLB are connected to V_{DD} and subsequently left floating similar to read and Bit-line logic computing. Following this, the search data is applied to the WLLn and its complementary asserted to WLRn. In cases where the search data matches the stored data, the BL and BLB remain unchanged. However, in instances where a disparity between the stored data and the search data occurs, either the BL or BLB is discharged. As an illustration, let's examine a scenario where the search data is "011" and it is compared against the stored data in a column-based architecture, which includes "010", "011", and "100" as shown in Fig. 5. In the case of a mismatch, BL0 and BL2 are discharged, yielding a '0' output for AND0 and AND2. However, due to the match between the stored data "011" and the search data, no discharge occurs in BL1 and BLB1, resulting in a '1' output for AND1.

4 PERFORMANCE EVALUATION

To assess the efficacy of the proposed design, a bottom-up evaluation framework is developed across device, circuit, and architecture levels. A comparison is made with a cutting-edge non-volatile memory array utilizing RRAM and Magnetic Tunnel Junction (MTJ),

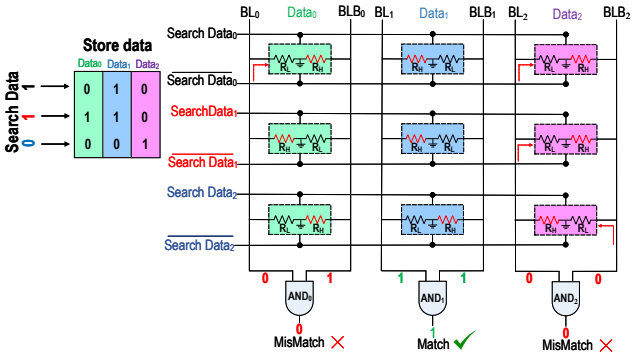

Figure 5: BCAM operation.

Table 3: Evaluation of Delay and Energy consumption in memory mode.

Designs	Technology	Read			Write		
		Delay (ps)	Power (μW)	Energy (fJ)	Delay (ns)	Power (μW)	Energy (fJ)
[32]	RRAM	73.19	35.24	2.57	1.2	97.3	116.76
[33]	STT/SOT MTJ	145.71	18.02	2.62	0.84	139.2	116.928
Proposed	MEFET	71.61	20.36	1.45	0.22	16.2	3.564

specifically tailored for in-memory computing. The assessments are conducted using the HSPICE tool, along with the Verilog-A MEFET model based on [19], using the 45 nm NCSU product design kit at 0.8V supply voltage [1].

4.1 Memory Mode Evaluation

The assessment of the read and write operations for the proposed cell entails a comparison with two designs, as outlined in Table 3. It is evident from the comparison that the proposed design exhibits significantly lower time and power consumption in write operation compared to the other cells. This is because the write operation is executed by applying a voltage below 100 mV to the MEFET gate in the proposed cell [20]. Furthermore, the proposed cell accomplishes the process of writing data and its complementary counterpart within a single cycle. This results in a substantial decrease in energy consumption during the write operation. While the process of writing data into the RRAMs and MTJs is notably characterized by extensive time and power consumption, the proposed design demonstrates a write energy consumption of about 96.9% lower compared to the designs presented in [32] and [33]. The differential nature of the read operation in the proposed design, coupled with the heightened high-to-low resistance ratio in the MEFET, streamlines the sense amplifier for data reading. Consequently, the energy consumption during read operations is significantly reduced in the proposed design when compared to the reference designs. The proposed design demonstrates a noteworthy reduction in read energy, with a 43.5% decrease compared to the [32] design and a 44.6% decrease compared to the [33] design.

4.2 Computing Mode Evaluation

To verify the functionality of the proposed design, Fig. 6 presents the transient waveform showing various logic operations under different input conditions in the proposed design. As illustrated, the evaluation of different logic functions in the proposed design involves considering four distinct inputs. In each input state, upon activation of the WLL and WLR signals, the selected operands and their complements are computed on the BL and BLB, respectively.

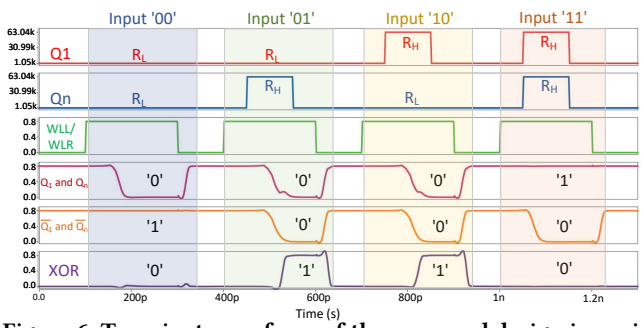

Figure 6: Transient waveform of the proposed design in various logic functions.

Table 4: Delay and power consumption of different Boolean function operations.

Designs	NAND/AND		NOR/OR		XNOR/XOR	
	Delay(ps)	Power(μW)	Delay(ps)	Power(μW)	Delay(ps)	Power(μW)
[32]	74.84	43.7	74.84	43.7	85.61	47.32
[33]	132.19	30.06	132.19	30.06	197.2	44.9
Proposed	60.21	25.99	63.2	26.12	66.8	27.7

Subsequently, these computed results are transmitted to the PSA. The PSA is capable of executing various logic operations depending on the appropriate control signals. An illustrative example of the PSA output representing the result of an XOR operation is shown in Fig. 6.Furthermore, for a more precise evaluation of the proposed design, a comparative analysis is performed against state-of-the-art IMC architectures from recent years. Table 4 presents the delay and power consumption associated with the implementation of Boolean logic gates. As evident from the table, the proposed cell exhibits lower delay and power consumption in all Boolean logic operations compared to the selected designs in [32, 33]. It is worth noting that, the proposed design directly computes various logic functions utilizing the PSA. This diminishes the overhead of the entire memory system since the number of PSAs aligns with the number of columns. In contrast, other designs, except for [32], necessitate additional circuits for executing diverse logic operations. The power consumption of the proposed design exhibits a significant reduction, approximately 40.22% lower than the [32] and 13.1% lower than the [33] when implementing (N)AND, and (N)OR operations. Moreover, the power consumption in implementing the X(N)OR logic in the proposed design is lower than the compared cells. The power consumption for implementing X(N)OR is 41.4% and 38.3% lower than the ones in [32] and [33], respectively. Additionally, as previously mentioned, the proposed design demonstrates proficiency in executing BCAM operations. Fig. 7 illustrates the energy consumption required for implementing a 4-bit BCAM operation. The energy consumption in BCAM operation is notably lower, with a reduction of 43.4% compared to the design in [32] and 44.4% compared to the design in [33].

5 APPLICATION LEVEL ANALYSIS

5.1 Binary Neural Networks

Convolutional Neural Networks (CNNs) operate in two distinct modes: training mode and inference mode. During training, the network learns by processing pre-classified training images to calculate the configuration values of its layers. In inference mode, the network applies this learned configuration to examine new test images. In both modes, the most computationally demanding arithmetic operations are Multiplication and Accumulation (MAC) [24].

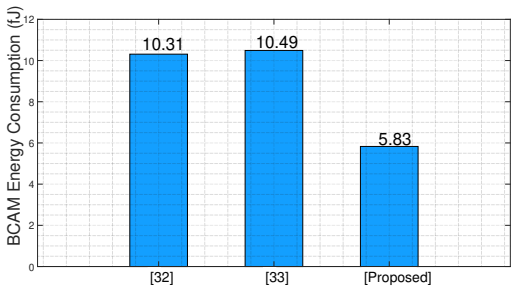

Figure 7: Energy consumption of 4-bit BCAM operation.

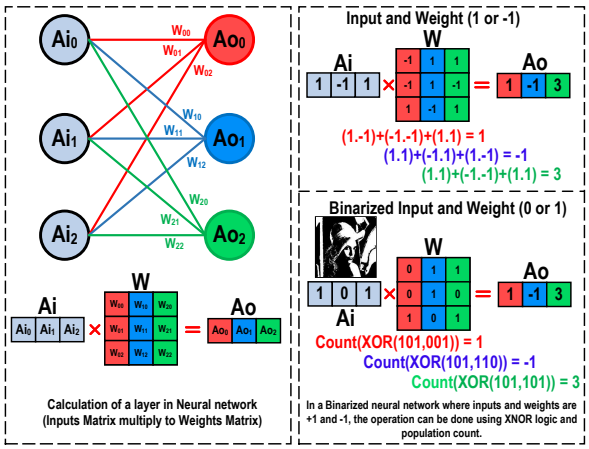

Figure 8: BNN calculation using XNOR and bit counting, where inputs and weights are +1 or -1 [29].

To address the computational and memory challenges associated with these operations, researchers have developed various BNNs, where weights and input activations are constrained to be binary during forward propagation. One such approach, BinaryConnect [14], trains deep neural networks with binary weights (-1, +1) and demonstrates near state-of-the-art performance on popular datasets such as MNIST and CIFAR-10. This methodology holds promise for hardware implementations of CNNs by replacing complex multiplication operations with simpler XNOR operations [18, 26], and significantly reducing the storage requirements for weights. XNOR-NET [18] presents a straightforward and accurate implementation of BNN, achieving comparable results to the full-precision AlexNet on the ImageNet dataset.

The common activation functions in BNNs, i.e., *Sign*, generate binary values of either +1 or -1, rendering neurons (A_i and A_o) in subsequent layers as 1-bit representations. Consequently, computations involve multiplying a binarized input neuron vector A_i by a binarized weight matrix W. This operation can be efficiently conducted utilizing XNOR and a counter. Fig. 8 demonstrates how the matrix-vector operation involving +1 and -1 values can be binarized and executed employing XNOR and counter.

5.2 Experiments

To assess the efficacy of the proposed architecture in real-world applications, we executed Binarized Image classification on neural networks. Specifically, we employed AlexNet as a binarized Convolutional neural network (CNN) and a more intricate architecture, VGG16, to study the performance of the proposed design. AlexNet, characterized by five convolutional layers and three fully connected layers [4], was employed alongside the deeper VGG16 network, comprising 13 convolutional layers and three fully connected layers [3]. The implementation of these networks commenced with a circuit-level simulation in HSPICE, aiming to analyze the delay, power, and energy consumption of the XNOR logic gate embedded within the proposed design. Following this, both network architectures were instantiated using MATLAB software. The data acquired from the circuit-level simulation was subsequently applied to each convolutional layer in both network architectures. As shown in Fig.

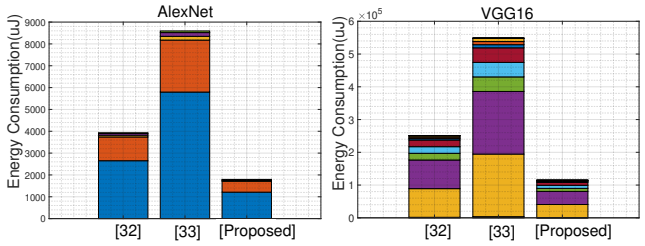

Figure 9: Energy consumption of different convolutional layers in (a) AlexNet (b) VGG16

9, the energy consumption of each convolutional layer in both networks is contrasted with that of other designs. The outcomes reveal a noteworthy reduction in energy consumption for the proposed design compared to the reference cells. The decrease in energy consumption is credited to the diminished energy usage linked with the integration of XNOR logic in the proposed design. This can be attributed to the utilization of an innovative differential circuit for computing XNOR logic within a single cycle. Particularly, the energy consumption of the proposed design during the implementation of Alex-Net is approximately 54.2% less than that of the [32] design. Furthermore, in implementing a more intricate network such as VGG16, the overall energy consumption is reduced by approximately 54.3% and 79.1% compared to the [32] and [33] designs, respectively.

6 CONCLUSION

In this work, we introduce a non-volatile in-memory computing architecture employing MEFETs, specifically designed for cutting-edge IoT devices. Our proposed architecture is adopted to execute Boolean logic operations alongside the BCAM operation within a single cycle. This achievement is realized through the utilization of the innovative sense amplifier and the storage of both data and its complement within a singular cell. Additionally, our designed write circuit enables the proposed cell to write data and its complementary counterpart in just one cycle, leading to a comprehensive reduction in energy consumption. The simulation results demonstrate a notable reduction in energy consumption for the proposed design, with approximately a 43.5% decrease in read operation energy and an impressive 96.9% reduction in write operation energy compared to other designs. Moreover, to assess the practical performance of the proposed architecture, we implemented two BNNs named AlexNet and VGG16 utilizing the suggested design. The simulation results reveal a substantial reduction in energy consumption, approximately 54% when implementing AlexNet and VGG16 using the proposed architecture in comparison to the reference designs.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under Grant No. 2228028, 2216772, 2216773, and 2247156.

REFERENCES

[1] 2011. *NCSU EDA FreePDK45.* http://www.eda.ncsu.edu/wiki/FreePDK45:Contents

[2] Minhaz Abedin et al. 2022. Mr-pipa: An integrated multilevel rram (hfo x)-based processing-in-pixel accelerator. *IEEE JxCDC* (2022), 59–67.

[3] N. Deepa and S.P. Chokkalingam. 2022. Optimization of VGG16 utilizing the Arithmetic Optimization Algorithm for early detection of Alzheimer's disease. *Biomedical Signal Processing and Control* 74 (2022), 103455.

[4] A. Elahi H. et al. 2021. Comparative Analysis of AlexNet, ResNet18 and SqueezeNet with Diverse Modification and Arduous Implementation. *Arabian Journal for Science and Engineering* 47 (2021), 1–10. Issue 2.

[5] A. Iyama et al. 2013. Magnetoelectric hysteresis loops in Cr 2 O 3 at room temperature. *Physical Review B* 87, 18 (2013), 180408. doi: 10.1103/PhysRevB.87.180408.

[6] B. Lee et al. 2009. Architecting phase change memory as a scalable dram alternative. In *Proceedings of the 36th annual international symposium on Computer architecture*. 2–13.

[7] C. Ma et al. 2020. MNFTL: An Efficient Flash Translation Layer for MLC NAND Flash Memory. *ACM TODAES* 25 (2020), 1–19. doi: 10.1145/3398037.

[8] D. Nikonov et al. 2015. Benchmarking of beyond-CMOS exploratory devices for logic integrated circuits. *IEEE JxCDC* 1 (2015), 3–11. doi: 10.1109/JXCDC.2015.2418033.

[9] D. Najafi et al. 2024. Enabling Normally-Off In Situ Computing With a Magneto-Electric FET-Based SRAM Design. *IEEE Transactions on Electron Devices* 71, 4 (2024), 2742–2748.

[10] D. Reis et al. 2018. Computing in memory with FeFETs. In *Proceedings of the International Symposium on Low Power Electronics and Design*. 1–6.

[11] F. Zhou et al. 2020. Near-sensor and in-sensor computing. *Nature Electronics* 3 (2020). Issue 11.

[12] H. Chuang et al. 2016. Low-resistance 2D/2D ohmic contacts: a universal approach to high-performance WSe2, MoS2, and MoSe2 transistors. *Nano letters* 16, 3 (2016), 1896–1902. doi: 10.1021/acs.nanolett.5b05066.

[13] MP. Anantram et al. 2008. Modeling of nanoscale devices. *Proc. IEEE* 96, 9 (2008), 1511–1550. doi: 10.1109/JPROC.2008.927355.

[14] M. Courbariaux et al. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*. 3123–3131.

[15] M. Gargari et al. 2023. An Energy Efficient In-Memory Computing Architecture Using Reconfigurable Magnetic Logic Circuits for Big Data Processing. *IEEE Transactions on Magnetics* 59, 12 (2023), 1–10.

[16] M. Morsali et al. 2023. Design and Evaluation of a Near-Sensor Magneto-Electric FET-Based Event Detector. *IEEE TED* (2023). doi: 10.1109/TED.2023.3296389.

[17] M. Morsali et al. 2023. OISA: Architecting an Optical In-Sensor Accelerator for Efficient Visual Computing. *arXiv:2311.18655v1* (2023).

[18] M. Rastegari et al. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*. Springer, 525–542.

[19] N. Sharma et al. 2017. Verilog-A based compact modeling of the magneto-electric FET device. In *E3S*. IEEE, 1–3. doi: 10.1109/E3S.2017.8246186.

[20] N. Sharma et al. 2018. Compact Modeling and Design of Magneto-Electric Transistor Devices and Circuits. In *IEEE SOCC*. IEEE, 146–151. doi: 10.1109/SOCC.2018.8618494.

[21] N. Sharma et al. 2020. Evolving magneto-electric device technologies. *Semiconductor Science and Technology* (2020). doi: 10.1088/1361-6641/ab8438.

[22] P. Dowben et al. 2018. Towards a strong spin–orbit coupling magnetoelectric transistor. *IEEE JxCDC* 4, 1 (2018), 1–9. doi: 10.1109/JXCDC.2018.2809640.

[23] P. Dowben et al. 2020. Magneto-electric antiferromagnetic spin–orbit logic devices. *Applied Physics Letters* 116, 8 (2020), 080502. doi: 10.1063/1.5141371.

[24] R. Andri et al. 2016. YodaNN: An ultra-low power convolutional neural network accelerator based on binary weights. In *ISVLSI*. IEEE, 236–241.

[25] S. Angizi et al. 2018. Cmp-pim: an energy-efficient comparator-based processing-in-memory neural network accelerator. In *Proceedings of the 55th Annual Design Automation Conference*. 1–6.

[26] S. Angizi et al. 2018. IMCE: Energy-efficient bit-wise in-memory convolution engine for deep neural network. In *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 111–116.

[27] S. Angizi et al. 2021. MeF-RAM: A New Non-Volatile Cache Memory Based on Magneto-Electric FET. *ACM TODAES* 27 (2021), 1–18. doi: 10.1145/3484222.

[28] S. George et al. 2016. Nonvolatile memory design based on ferroelectric FETs. In *Proceedings of the 53rd Annual Design Automation Conference*. 1–6.

[29] S. Jeloka et al. 2016. A 28 nm Configurable Memory (TCAM/BCAM/SRAM) Using Push-Rule 6T Bit Cell Enabling Logic-in-Memory. *IEEE Journal of Solid-State Circuits* 51, 4 (2016), 1009–1021.

[30] T. Wan et al. 2022. In-Sensor Computing: Materials, Devices, and Integration Technologies. *Advanced Materials wiley* 35 (2022). Issue 37.

[31] X. Dong et al. 2008. Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement. In *2008 45th ACM/IEEE Design Automation Conference*. IEEE, 554–559. doi: 10.1145/1391469.1391610.

[32] Y. Chen et al. 2021. A Reconfigurable 4T2R ReRAM Computing In-Memory Macro for Efficient Edge Applications. *IEEE Open Journal of Circuits and Systems* 2 (2021), 210–222.

[33] Z. Yang et al. 2022. A Novel Computing-in-Memory Platform Based on Hybrid Spintronic/CMOS Memory. *IEEE Transactions on Electron Devices* 69, 4 (2022), 1698–1705.