Enabling Normally-Off In Situ Computing With a Magneto-Electric FET-Based SRAM Design

Deniz Najafi, *Student Member, IEEE*, Mehrdad Morsali, *Student Member, IEEE*, Ranyang Zhou, *Student Member, IEEE*, Arman Roohi, *Senior Member, IEEE*, Andrew Marshall, *Life Fellow, IEEE*, Durga Misra, *Life Fellow, IEEE*, and Shaahin Angizi, *Senior Member, IEEE*

***Abstract*— As an emerging post-CMOS Field Effect Transistor, magneto-electric field-effect transistors (MEFETs) offer compelling design characteristics for logic and memory applications, such as high-speed switching, low power consumption, and nonvolatility. In this article, for the first time, a nonvolatile MEFET-based SRAM design named ME-SRAM is proposed for edge applications which can remarkably save the SRAM static power consumption in the idle state through a fast backup-restore process. To enable normally-OFF in situ computing, the ME-SRAM cell is integrated into a novel processing-in-SRAM architecture that exploits a hardware-optimized bitline computing approach for the execution of Boolean logic operations between operands housed in a memory sub-array within a single clock cycle. Our device-to-architecture evaluation results on Binary convolutional neural network acceleration show the robust performance of ME-SRAM while reducing energy consumption on average by a factor of ∼5.3× compared to the best in-SRAM designs.**

***Index Terms*— Magneto-electric field-effect transistor (MEFET), normally-OFF computing, processing-in-SRAM.**

I. Introduction

THE battery-constraint Internet of Things (IoT) edge devices need to operate for extended periods and minimizing power leakage in standby mode leveraging normally-OFF in situ computing is a promising solution for such devices [1], [2]. In the past few years, there has been a notable surge in interest surrounding the integration of emerging nonvolatile memory (NVM) technologies in edge devices primarily driven by the distinctive attributes of NVMs, including nonvolatility, robustness, long endurance, high integration density, exceptionally low standby power consumption, and compatibility with intermittent computing [3], [4], [5]. For embedded applications and low-power IoT systems that rely on an ON-chip cache, the integration of a robust NVM holds the potential to enhance memory capacity and performance.

Recent experiments on spintronics have shown the capability of achieving fast magnetization switching, with switching times in the sub-nanosecond range conducted on magnetic tunnel junction (MTJ) devices, utilizing either the spin-transfer torque (STT) or spin-orbit torque (SOT) switching mechanisms [6], [7]. Such NVM technologies have shown interestingly long retention times (up to 10 years) and low write energy (fJ/bit). However, they suffer from low ON/OFF ratios (less than 10), leading to reliability issues due to the current-driven switching scheme [7], [8]. ReRAM suffers from slower and more power-hungry write operations with lower endurance compared to MTJs [3], though it offers a higher ON/OFF ratio and larger sense margin. The magneto-electric field-effect transistor (MEFET), based on the antiferromagnetic (AFM) magneto-electric (ME) phenomena, has recently been introduced and experimentally studied [2], [3], [5], [9], [10]. This spintronic device shows great promise with superior performance and improved temperature stability. What set the MEFET apart from conventional spintronic devices are its significantly faster switching speed and a notably larger ON/OFF ratio. The MEFET achieves very fast switching times (<20 ps) and low energy consumption (<20 aJ) by utilizing a coherent rotation as the domain switching mechanism, eliminating the need for ferromagnet switching or domain wall movement [5], [8].

In this work, we propose ME-SRAM as a nonvolatile SRAM design, based on MEFET technology for the first time that enables normally-OFF in situ computing in edge applications. The main contributions of this work are listed as follows.

1) We develop an ME-SRAM platform by optimizing the Verilog-A MEFET device model to capture the switching dynamics of the ME layer as well as designing innovative circuit-level and micro-architectural schemes to reduce the static power consumption of ME-SRAM during idle periods through rapid backup and restore process.
2) We design an efficient and parallel processing-in-SRAM scheme that enables bulk bit-wise X(N)OR logic

Manuscript received 21 January 2024; accepted 11 February 2024. Date of publication 19 February 2024; date of current version 26 March 2024. This work was supported in part by the National Science Foundation under Grant 2228028, Grant 2216772, and Grant 2216773; and in part by Semiconductor Research Corporation (SRC). The review of this article was arranged by Editor J. Kang. *(Corresponding author: Shaahin Angizi.)*

Deniz Najafi, Mehrdad Morsali, Ranyang Zhou, Durga Misra, and Shaahin Angizi are with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: dn339@njit.edu; mm2772@njit.edu; rz26@njit.edu; dmisra@njit.edu; shaahin.angizi@njit.edu).

Arman Roohi is with the School of Computing, University of Nebraska–Lincoln, Lincoln, NE 68588 USA (e-mail: aroohi@unl.edu).

Andrew Marshall is with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: andrew.marshall@utdallas.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2024.3366172.

Digital Object Identifier 10.1109/TED.2024.3366172

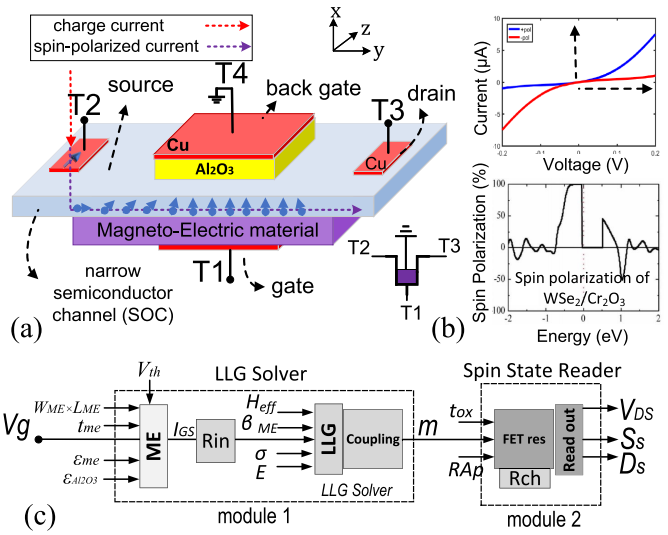

Fig. 1. (a) MEFET device structure and the circuit scheme. (b) Sample source-to-drain current versus voltage at T1 and the induced spin polarization in WSe_2. (c) MEFET Verilog-A modeling.

processing required in various edge applications such as deep learning.

3) We create an extensive bottom-up evaluation framework to analyze the performance of the proposed ME-SRAM architecture compared with state-of-the-art designs.

II. MEFET Device and Modeling

The MEFET shares structural similarities with the CMOS FET device. Fig. 1(a) shows the basic single-source version of MEFET, which is a four-terminal device with gate (T1), source (T2), drain (T3), and back gate (T4) terminals [2], [5], [11] along with the simplified three-terminal design schematic used in this work. This device comprises a narrow semiconductor channel positioned between two dielectrics: the ME material, such as Chromia (Cr_2O_3), and the insulator, for example, Alumina (Al_2O_3). Various materials, including PbS, graphene, InP, $WSe2_2$ can be used to construct the narrow semiconductor channel named spin-orbit coupling (SOC) in the MEFET. One electrode is attached to the gate (T1) through the ME layer, while the other electrode is connected to the back gate (T4) via the alumina layer. In this work as shown in Fig. 1(b), the MEFET utilizes tungsten diselenide (WSe_2) as the channel material, providing a high ON-OFF ratio and high hole mobility [5], [11]. For the source and drain, both conductors and ferromagnetic (FM) polarizers can be used.

The MEFET functions as a transistor by initially biasing the SOC channel through the T1 and T4 terminals, similar to the gate biasing process in CMOS. Subsequently, the current is applied from the T2 to T3 terminals, resembling the source-drain biasing in CMOS. It has been shown that by applying a very low voltage of approximately ±100 mV [5] across the gate (T1) and back gate terminal (T4: ground), the ME capacitor is charged. In fact, by applying a voltage, a vertical electric field is created across the gate, depending on whether T1 is positively or negatively charged. This electric field induces a change in the paraelectric polarization and AFM order within the ME insulator layer. Hence, the reorientation of spin vectors occurs in Chromia as a consequence. This reorientation is facilitated by exchange interactions and SOC.

TABLE I
Compact Verilog-A Model Parameters

Parameter	Value	Description of Parameter and Units
ϵ_{ME}	12	Dielectric constant of chromia [13]
$\epsilon_{Al_2O_3}$	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, $M\Omega$

Subsequently, the high boundary polarization of the ME layer polarizes the spins of carriers within the semiconductor channel. This polarization induces a favored conduction path along a specific axis, resulting in a notable change in resistance in that particular direction.

We use nonequilibrium green's function (NEGF) transport simulations to explore the current–voltage relationship [Fig. 1(b)] dependent on the direction of ME polarization based on [5] and [12]. These simulations are conducted on a 2-D ribbon with a width of 20 nm and a band mass of $0.1m_e$. To account for the effects, we considered a conservative exchange splitting value of 0.1 eV and a voltage difference of T3-T2 = 0.1 V at a temperature of 300 K. Therefore, the MEFET's surface magnetization on the channel induces a directionality in the conductance as shown in Fig. 1(b). Moreover, an exceptionally high level of spin polarization is brought about by the ME layer to the WSe_2 channel. To readout the MEFET, the T2-T3 resistive path can be sensed and compared with a reference. The ON/OFF current ratio for WSe_2 can be extended up to 10^6.

The new enhanced MEFET Verilog-A model is depicted in Fig. 1(c) and comprises two modules to enable: 1) write process by controlling and inducing polarization through ME dynamics in the semiconductor channel, and enabling source-to-drain spin injection, as described in [9]; and 2) read process by reading out FET resistance. Table I showcases the experimental parameters utilized for the switching behavior of the Chromia layer and SOC channel in our model.

Module 1: Landau-Lifshitz-Gilbert (LLG) Solver is designed to capture the electrical charging of the ME capacitor and the dynamics involved in switching at the interface between the ME layer and the SOC channel. The relevant capacitance of the ME layer is represented by a resistor–capacitor circuit network as $C_{\mathrm{ME}} = ((\varepsilon_{\mathrm{ME}} A))/(t_{\mathrm{ME}})$. Here, $\varepsilon_{\mathrm{ME}}$ denotes the dielectric constant of the ME layer, which has a thickness of t_{ME}, and A represents the cross-sectional area. Additionally, R_{in} denotes the load resistance at the input driving level. When a voltage difference is applied to the gate electrodes, the capacitor undergoes a charging process. The model compares the gate-source voltage (V_{g}) as the input and the threshold voltage for Chromia state inversion (V_{th} = 0.050 V [14]) to initialize the memory and determine the resulting voltage across the drain and source terminals. The spin dynamics (m) is modeled by the widely used LLG equation and takes into account thermal fluctuation, electron/spin transport, and the voltage-controlled ME effect [5], [15]

$$\frac{dm}{dt} = -|\gamma| m \times H_{\mathrm{eff}} + \alpha\left(m \times \frac{dm}{dt}\right) + \sigma\beta_{\mathrm{ME}}(m \times E) \quad (1)$$

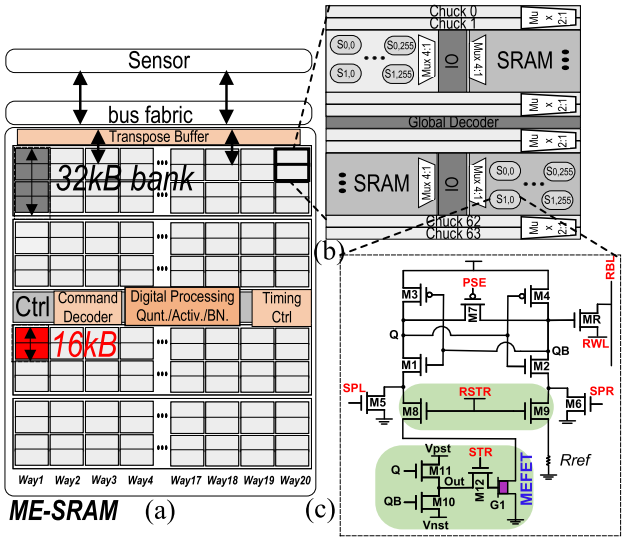

Fig. 2. (a) ME-SRAM cache slice architecture with 2.5 MB capacity. (b) 16 KB memory matrix design. (c) Proposed ME-SRAM cell.

here, γ represents the gyromagnetic ratio, and H_{eff} denotes the effective magnetic field. The ME susceptibility, β_{ME}, depends on temperature, as discussed in [3] and [5]. To align the calculated ME-induced momentum magnitude with our experimental data, we utilize the scaling factor σ. We further modeled the temperature variation on the dynamics of MEFET as a random magnetic field with each spatial component (x,y,z) drawn from a Gaussian distribution of zero mean and standard deviation $(2\alpha K_B T/\gamma M_s V \Delta t)^{1/2}$ where α is the damping factor, M_s is the saturation magnetization, K_B is Boltzmann's constant, T is absolute temperature, V is the volume, and Δt is the simulation time step.

Module 2: Spin State Reader is responsible for determining the appropriate channel resistance (R_{ch}) and calculating related electrical parameters, including the output voltage at the drain terminal. The channel resistance (R_{ch}) is computed in series with the input resistance (R_{in}) to establish the switching boundary conditions. Furthermore, the spin states at the source and drain terminals, denoted as "Ss" and "Ds," are verified using two spin-state terminals. In Fig. 1(c), the "Ss" terminal is configured as "+1 V" for the "up" spin and "−1 V" for the "down" spin. Our model incorporates a fixed delay of 200 ps to account for the processional delay across the FM layer, which is estimated based on reliable coupling delay data [8].

III. ME-SRAM Architecture

We propose ME-SRAM as a near-sensor cache-based architecture that enables normally-OFF in situ computing to accelerate two-input bulk bit-wise X(N)OR operations in various X(N)OR-intensive applications such as data encryption and deep neural networks (DNNs). The proposed geometry of a single 2.5 MB ME-SRAM aligned with the cache architecture for application-level analysis connected to a vision sensor is shown in Fig. 2(a). ME-SRAM features cache slices with 80 memory banks, each comprising 32 KB of storage organized into 20 ways. Each bank includes two 16 KB memory matrices [highlighted in Fig. 2(b)] with 8 KB computational sub-arrays. A shared digital control unit centrally times and controls data transfer with extra processing units such as quantization and activation function for neural network processing. Fig. 2(c) shows the structure of the proposed nonvolatile ME-SRAM cell operating in two modes: *memory mode* (supporting read/write and check-pointing) and *computing mode* (enabling bit-wise in-memory operations).

TABLE II
SIGNALING OF THE ME-SRAM CELL FOR MEMORY MODE

Signals	Hold	Read	Write	Store	Restore
RBL	V_{DD}	Pre-Charge	V_{DD}	V_{DD}	V_{DD}
RWL	V_{DD}	'0'	V_{DD}	V_{DD}	V_{DD}
PSE	V_{DD}	V_{DD}	⎍	V_{DD}	⎍
SPL	V_{DD}	V_{DD}	$\overline{Data}$	V_{DD}	'0'
SPR	V_{DD}	V_{DD}	$\overline{Data}$	V_{DD}	'0'
STR	'0'	'0'	'0'	V_{DD}	'0'
RSTR	'0'	'0'	'0'	'0'	V_{DD}

A. Memory Mode

The ME-SRAM bit-cell as depicted in Fig. 2(c), comprises two primary components: the volatile component, which includes the 8T SRAM cell (M1–M7 and MR), and transistors contributing to nonvolatile component (highlighted in green) responsible for the storage and retrieval of data to/from the ME-SRAM cell. This component comprises one MEFET and five transistors (M8–M12). Table II shows the signaling of the ME-SRAM in various memory mode operations.

1) Normal Operation: During normal operation, the ME-SRAM cell acts as a typical memory performing hold, read, and write operations, where the nonvolatile component is deactivated via M8, M9, and M12 transistors in Fig. 2(c).

a) Read: The read operation is performed by pre-charging the read bitline (RBL) to the supply voltage and connecting the read word line (RWL) to the ground as depicted in Fig. 3(a). Now based on the data stored in memory nodes (Q/QB), the RBL is discharged or remains unchanged. If the data stored in QB is "1," the RBL is discharged through the MR transistor. In contrast, if the data in QB is "0," the MR transistor is off and the RBL remains untouched. To maximize the read reliability and handle the impact of sneak current as will be analyzed in Section IV, the read operation is performed when the RBL is discharged at 10% of its initial value.

b) Write: For a write operation, SRAM pull-down network left (SPL) and SRAM pull-down network right (SPR) signals are grounded which results in turning the M5 and M6 transistors off. Meanwhile, the M7 transistor is turned on connecting the Q and QB to $(V_{\text{DD}})/(2)$, as shown in Fig. 3(b). When the voltage of the Q and QB nodes becomes the same, the data and its complementary are tied to SPL and SPR, respectively. Based on the data being stored, the M5 or M6 transistor will turn on and the related data node (Q or QB) will discharge. When the desired node is discharged, the other transistor is turned on resulting in activating the positive feedback of the cross-coupled inverters. This positive feedback will drive the node with a lower voltage level to "0" while pushing the node with a higher voltage to "1." As tabulated in Table II, to write "0" in the Q node, the SPL, SPR, and pre-charge sense amplifier enable (PSE) signals are tied to the ground and setting the Q and QB to $(V_{\text{DD}})/(2)$. Then, PSE and SPR are deactivated and the SPL is tied to "1" resulting in discharging the Q node. Then, the SPR signal is tied to "1" which activates the ME-SRAM cross-coupled inverters.

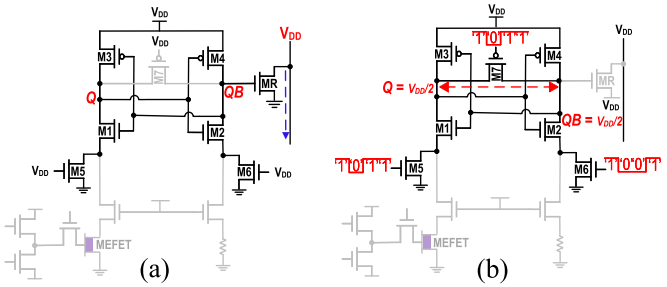

Fig. 3. ME-SRAM in memory mode. (a) Read. (b) Write.

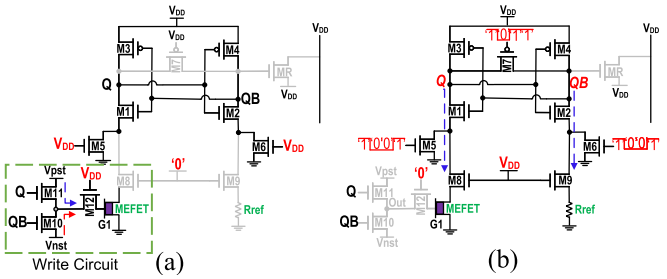

Fig. 4. ME-SRAM in memory mode. (a) Store. (b) Restore.

2) Check-Pointing Operation: ME-SRAM can be readily reconfigured to perform a fast and efficient check-pointing operation based on the signaling listed in Table II.

a) Store: To back up the data in the ME-SRAM cell and store it in the MEFET, the proposed write circuitry in Fig. 4(a) assigns proper MEFET write voltage (V_{pst} or V_{nst}) based on Q and QB values to the MEFET's gate. So, the resistance of MEFET changes to low/high states, and data is stored in the nonvolatile element. For example, if Q = "1" the activation of the M12 transistor causes the transmission of V_{pst} to the gate of MEFET. This action changes the device to the high resistance state (R_{off}).

b) Restore: To restore the data from the MEFET to the SRAM cell, the SPL and SPR are tied to the ground, and the PSE signal is activated as shown in Fig. 4(b). This causes Q and QB to be floating. Meanwhile, the restore signal (RSTR) is activated resulting in turning on the M8 and M9 transistors. Here, ME-SRAM operates to compare the MEFET resistance (R_{ME}) with a reference resistance (R_{Ref}) on the right branch set to $(R_{\text{on}} + R_{\text{off}})/(2)$. Based on the difference between MEFET resistance and reference resistance either the Q or QB discharges faster than the other. This results in storing the desired data in the SRAM cell. For instance, when the MEFET holds a"1," its resistance sets to R_{on} (see Table I). Consequently, the left path's resistance becomes lower than that of the right path ($R_{\text{ME}} < R_{\text{Ref}}$). Therefore, the Q node undergoes a faster discharge compared to QB.

B. Computing Mode

The ME-SRAM is capable of performing massively parallel X(N)OR logic as shown in Fig. 5(a). To this end, two ME-SRAM cells holding operands (here A_n and B_n) in the same memory sub-array column are activated. To realize efficient bitline computing, we propose to tie RWL1 to V_{DD}, where the RWL2 is connected to the ground. Meanwhile, the RBL is pre-charged to $(V_{\text{DD}})/(2)$ as shown in Fig. 5(b). This will form a voltage divider as shown in Fig. 5(c). Now if $QB1 = QB2$ the RBL remains at its initial value.

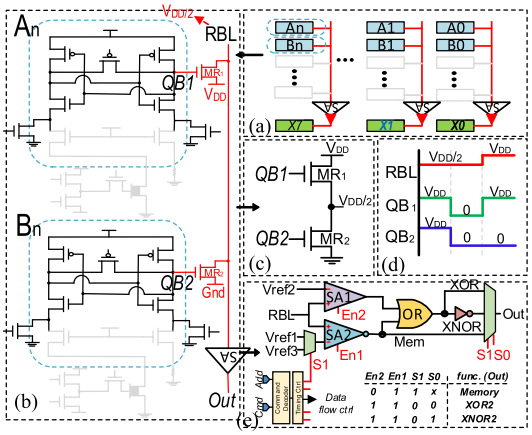

Fig. 5. Realizing parallel in-memory X(N)OR operation. (a) Block diagram. (b) Circuit schematic. (c) Equivalent circuit while data is read from cells. (d) Timing diagram. (e) Proposed SA.

If $QB1 = QB2$ = "1," MR1 and MR2 are activated which results in voltage division connecting RBL to its initial value. Moreover, if $QB1 = QB2$ = "0" both MR1 and MR2 are not activated which results in remaining the RBL in $(V_{\text{DD}})/(2)$. In contrast, if the data in $QB1$ is not equal to $QB2$, the RBL is either tied to V_{DD} or the ground [Fig. 5(d)]. Under these circumstances, when $QB1$ and $QB2$ are identical, the outcome of XOR/XNOR operations will be "0"/"1." Conversely, if the data are dissimilar, the result for XOR/XNOR will be "1"/"0," respectively. We propose a sense amplifier (SA) to distinguish the achieved voltage states leveraging two comparators to sense the disparity between RBL and referenced voltages, along with an OR gate connected to their output, which allows the extraction of X(N)OR logic output [Fig. 5(e)]. The proposed SA can be readily configured through the four control bits issued by the controller ($En2$,$En1$,$S1$,$S0$). By selecting different reference voltages, the SA can perform basic memory and X(N)OR functions according to the configuration bits shown in Fig. 5(e). For X(N)OR operation, V_{ref1} and V_{ref2} are set to satisfy $V_{\text{ref1}} < V_{\text{ref3}} < V_{\text{ref2}}$. For instance, when the XOR input data are the same, the voltage on the RBL stabilizes at $(V_{\text{DD}})/(2)$. As a result, both SA1 and SA2 outputs settle at "0" according to the reference voltages resulting in the "0" output. In contrast, if the input data are dissimilar, the RBL connects to the supply voltage or ground. This action causes one of SA's outputs to shift to "1." Consequently, the OR gate is activated, driving its output to "1" and producing the intended XOR output.

IV. Experimental Results

To evaluate the performance of the ME-SRAM architecture, a comprehensive bottom-up evaluation framework is developed as depicted in Fig. 6. At the device level, we develop a Verilog-A compact model for the MEFET-RAM based on Section II to co-simulate with other peripheral CMOS circuits displayed in Fig. 2 in SPICE. At the circuit level, we use 45 nm NCSU product development kit (PDK) library to fully design and verify the ME-SRAM arrays in HSPICE and to extract performance parameters such as delay and energy consumption. We use the Synopsys Design Compiler to design the ME-SRAM controller using standard industry-level technology. At the architecture level, we extensively

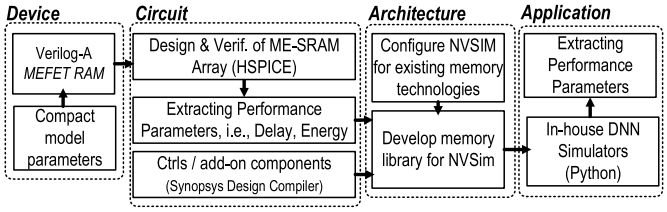

Fig. 6. Proposed evaluation framework.

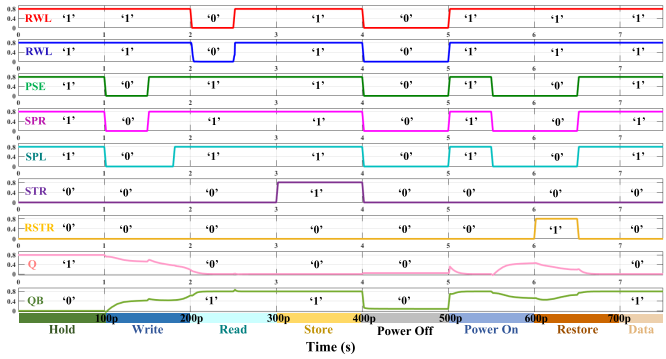

Fig. 7. Transient waveform of the ME-SRAM bit-cell.

modify NVSIM [16] as a memory performance evaluation tool to take memory configuration and circuit data for the MEFET library and report the array-level read/write energy and latency. At the application level, we develop HW/SW python simulators for ME-SRAM taking the architecture-level data for the ME-SRAM to estimate the system performance while running DNN workloads.

A. Device-to-Circuit-Level Analysis

1) Memory Mode: Fig. 7 illustrates the transient waveform of the ME-SRAM cell during various operational states: hold, write, read, store, and restore. Initially, the cell is in the hold state storing Q = "1." Following this, the ME-SRAM write operation commences, and data "0" is written into the Q node followed by a read operation to validate the correctness of the write operation. Subsequently, before the initiation of the power gating process, the data is safeguarded and stored in the MEFET. During this phase, the SRAM cell remains in its hold state. Upon re-activating the cell, the data within the ME-SRAM cell is erased and requires restoration from the MEFET. This restoration process is accomplished by triggering the RSTR signal and initiating a race condition between the path connected to the MEFET and the path linked to the reference. It is important to note that the device characteristics of the chromia layer render this material intriguing for incorporation into the back end of the line (BEoL). The experimental demonstration of the feasibility of integrating chromia with silicon has been conducted in [17] and [18].

a) Normal operation: Various parameters of the ME-SRAM cell, including static noise margin (SNM), delay, and power consumption, are compared with the conventional 6T SRAM cell in Table III due to its predominant use in designing nonvolatile SRAM cells [19], [20], [21]. We observe that: 1) in ME-SRAM bit-cell, the isolation of the data node from the RBL results in a read SNM (RSNM) of 288 mV which is nearly equivalent to the hold SNM (HSNM); 2) contrarily, conflicts between read and write operations in the conventional 6T SRAM cell result in a notable reduction in RSNM by 126.5 mV; 3) ME-SRAM exhibits a significantly higher combined word-line margin (CWLM) of 374.8 mV compared to the 6T SRAM cell (261.7 mV), achieved by floating the data nodes. It is noteworthy that achieving optimal writability, coupled with a reduction in half-select issues, is attainable in the SRAM cell when the CWLM is maintained at approximately $(V_{DD})/(2)$. At a reduced supply voltage of 800 mV, the CWLM is measured at 374.8 mV, underscoring the excellent writability of the cell while addressing the lower half-selected issues; and 4) by addressing the inherent conflict between read and write operations and employing transistors with minimal dimensions in ME-SRAM, the power-delay product (PDP) for both read and write operations is significantly reduced compared to the baseline cell. ME-SRAM's overall PDP is lower than 6T SRAM, despite the latter's possible shorter write delay due to differential write techniques.

b) Check-pointing operation: To assess the performance of store and restore operations, the delay, power consumption, and PDP of the ME-SRAM are compared with those of six cutting-edge nonvolatile SRAM cells relying on MRAM [19], [20], [21], [22], [23], [24]. As listed in Table IV, we observe that the collective latency of the ME-SRAM during store and restore operations demonstrates a noteworthy decrease when contrasted with the designs scrutinized in the context of this specific configuration.

1) The delay of the proposed design in the store mode is ∼94% and 91.7% less than the designs presented in [21] and [24], respectively. This superiority arises from the fact that the write operation in MRAMs demands a substantial current for altering the orientation of the MTJ, while the MEFET requires ±100 mV to change its resistance.
2) Furthermore, the restore delay of the proposed design is 13.7% lower than that of the [24] design, which represents the minimum delay among the cells considered for comparison. In the restore operation, the substantial resistance ratio of the MEFET leads to a notable disparity between the reference resistance and data path resistance, resulting in a degradation of restore time. It is noteworthy that the PDP of the proposed design in both store and restore modes is lower than that of all the designs used for comparison.
3) The PDP of the ME-SRAM during the store operation is approximately 80%, 89.5%, and 78% lower than that of [19], [21], and [24], respectively.
4) During the restore mode, the PDP of the proposed design is lower than that of the compared cells. Specifically, the PDP of the proposed design is 30% lower than the [24] design, which holds the second position in the table. It is noteworthy that the utilization of a single MEFET contributes to an overall decrease in the PDP of ME-SRAM in comparison to alternative designs.

2) Computing Mode: Fig. 8 depicts in situ X(N)OR computing power consumption and performance of ME-SRAM compared with selected in-SRAM computing platforms supporting X(N)OR operation including, XSRAM [25], Compute Cache [26], Neural Cache [27], and NS-LBP [28]. The findings shown in Fig. 8(a) and (b) highlight the delay and power consumption characteristics of our design. We observe; 1) ME-SRAM stands as the second-fastest design after

TABLE III
ME-SRAM VERSUS 6T-SRAM

Parameters	6T SRAM	ME-SRAM
HSNM (mV)	288	288
RSNM (mV)	126.5	288
CWLM (mV)	261.7	374.8
Read Delay (ps)	24	14.8
Write Delay (ps)	7	22
Read Power (μW)	10.34	11.9
Write Power (μW)	4	1.2
Read PDP (aJ)	284.16	176.12
Write PDP (aJ)	28	26.6

TABLE IV
STORE/ RESTORE PERFORMANCE COMPARISON

Design	Technology	Store			Restore		
		Delay(ns)	Power(μW)	PDP(fJ)	Delay(ns)	Power(μW)	PDP(fJ)
[19]	SOT-MTJ	1.86	2.39	4.44	0.373	0.64	0.23
[21]	STT/SOT-MTJ	1.81	4.69	8.48	0.085	9.32	0.79
[20]	SOT-MTJ	2.44	6.33	15.44	0.095	10.43	0.99
[22]	STT-MTJ	5.58	19.12	106.68	0.373	0.86	0.32
[23]	STT-MTJ	4.09	3.48	14.23	0.93	0.79	0.73
[24]	STT/SOT-MTJ	1.34	3.02	4.04	0.058	4	0.23
ME-SRAM	MEFET	0.11	8.1	0.89	0.05	3.25	0.16

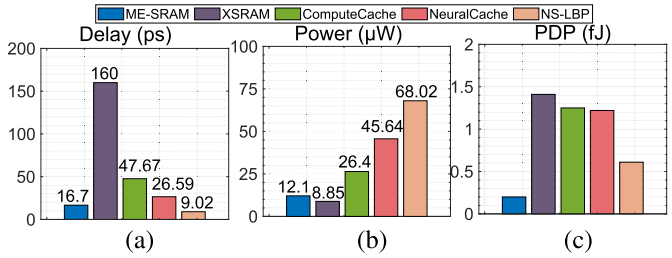

Fig. 8. Performance comparison of computing mode in various platforms. (a) Delay. (b) Power consumption. (c) PDP.

NS-LBP [28] with 16.7 ps, where it can easily outperform inverter SA-based XSRAM [25] and multicycle Compute Cache [26] designs; 2) ME-SRAM as the only SRAM cell supporting check-pointing mode outperforms Neural Cache [27] and NS-LBP [28] with 73.4% and 82.2% lower power consumption. This mainly comes from the relatively larger CMOS circuitry used in BL SAs to enable X(N)OR logic. When compared with XSRAM [25], an increase in SA complexity and current flow in our design, attributable to the reduced resistance path, leads to higher overall power consumption for ME-SRAM compared with the inverter-based design in [25]; and 3) despite ME-SRAM elevated power consumption, the overall PDP for executing X(N)OR operations is 85.8%, 84%, 83.4%, and 67.2% smaller than that of XSRAM [25], Compute Cache [26], Neural Cache [27], and NS-LBP [28] [Fig. 8 (c)].

3) Variation Analysis: A comprehensive Monte-Carlo statistical analysis with 1000 iterations is conducted in HSPICE on critical transistor parameters, i.e., width, length, and threshold voltage of bit-cell and SA incorporating Gaussian-distributed variations (3σ = 0%–70%). We test all 256 bit-lines within each ME-SRAM's sub-array, covering all possible bit-value combinations in memory. Considering that, in the design of SRAM cells, the RSNM exhibits greater sensitivity to process variation compared to the HSNM. The investigation of the RSNM for ME-SRAM in the presence of process variation is depicted in Fig. 9(a). Fig. 9(b) thoroughly examines the CWLM of ME-SRAM in the presence of process variations. We observe that the voltage levels at the Q and QB nodes are closely matched before introducing the data and its complement through SPL and SPR signals. Subsequently,

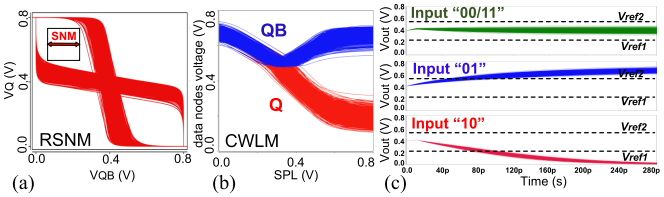

Fig. 9. Monte-Carlo simulations in SPICE for (a) RSNM, (b) CWLM of the proposed cell, and (c) XOR logic outputs.

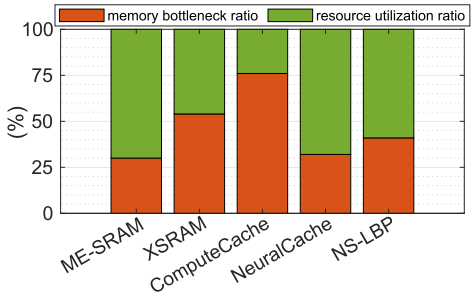

Fig. 10. Memory bottleneck ratio for under-test platforms.

depending on the activation of SPL or SPR, the desired data is written into the SRAM cell. Notably, there were no instances of failure observed during the read-and-write operations. To evaluate the computing mode's effectiveness for implementing X(N)OR logic amidst variations, a separate Monte-Carlo simulation [Fig. 9(c)] is performed. The results show no failures in XOR operation across diverse input combinations.

4) Memory Bottleneck: We conducted an analysis of the memory bottleneck ratio, which represents the time fraction during which computation is delayed due to ON- and OFF-chip data transfer hindering performance (referred to as the memory wall). Moreover, the resource utilization ratio can be readily calculated on top of it. This assessment was based on peak performance and derived data for our platform, XSRAM [25], ComputeCache [26], Neural Cache [27], and NS-LBP [28], considering the number of memory accesses. The results reported in Fig. 10 highlight the effectiveness of the ME-SRAM in addressing the memory wall issue. Specifically, we observe that the ME-SRAM solution requires less than approximately 30% of the time for memory access and data transfer offering the highest resource utilization ratio, while XSRAM [25] and ComputeCache [26] accelerators spend over 54% of their time waiting for data loading. This discrepancy arises from two factors: 1) an increased number of computational cycles and 2) an imbalance between computation and data movement in previous in-memory accelerators. Besides, Neural Cache demonstrates a ~32% memory bottleneck ratio as the second-best in-memory accelerator.

B. Architecture-to-Application-Level Analysis

We assess the efficiency of ME-SRAM by executing Binarized AlexNet, a DNN architecture featuring five convolutional layers, where every multiply-accumulate (MAC) operation is performed equivalently using XNOR and addition operations [28]. The execution time and energy consumption of ME-SRAM are compared with state-of-the-art in-SRAM processing accelerators in Fig. 11. We observe

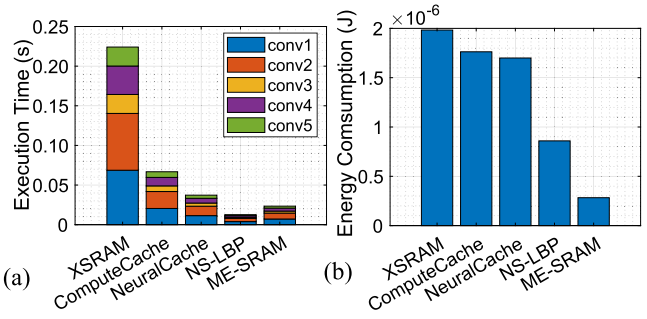

Fig. 11. (a) Execution time. (b) Energy consumption of binarized AlexNet.

that: 1) ME-SRAM shows a remarkable speedup compared to all under-test counterparts except NS-LBP [28], e.g., ME-SRAM outperforms Neural Cache [27] and XSRAM [25], respectively, by ∼39% and 89.7× reduction in execution time [Fig. 11(a)]; and 2) As shown in Fig. 11(b), ME-SRAM imposes ∼0.28 μJ to process the five convolutional layers of AlexNet and reduces the energy consumption by a factor of 7×, 6.1×, 3× compared with XSRAM [25], Compute Cache [26], and NS-LBP [28].

V. Conclusion

This article presents ME-SRAM, a nonvolatile SRAM design that minimizes static power consumption during idle states through rapid backup-restore. Integrated into a novel processing-in-SRAM architecture, ME-SRAM enables normally-OFF computing with optimized bitline computing, resulting in robust performance and significant energy and time savings compared to counterparts. On DNN acceleration, ME-SRAM achieves on average ∼5.3× higher energy efficiency compared to the best designs.

References

[1] M. K. Q. Jooq, M. H. Moaiyeri, and K. Tamersit, "A new design paradigm for auto-nonvolatile ternary SRAMs using ferroelectric CNTFETs: From device to array architecture," *IEEE Trans. Electron Devices*, vol. 69, no. 11, pp. 6113–6120, Nov. 2022, doi: 10.1109/TED.2022.3207703.

[2] M. Morsali, S. Tabrizchi, A. Marshall, A. Roohi, D. Misra, and S. Angizi, "Design and evaluation of a near-sensor magneto-electric FET-based event detector," *IEEE Trans. Electron Devices*, vol. 70, no. 9, pp. 4822–4828, Sep. 2023, doi: 10.1109/TED.2023.3296389.

[3] P. A. Dowben, D. E. Nikonov, A. Marshall, and C. Binek, "Magneto-electric antiferromagnetic spin–orbit logic devices," *Appl. Phys. Lett.*, vol. 116, no. 8, Feb. 2020, Art. no. 080502, doi: 10.1063/1.5141371.

[4] C. Ma, Y. Wang, Z. Shen, R. Chen, Z. Wang, and Z. Shao, "MNFTL: An efficient flash translation layer for MLC NAND flash memory," *ACM Trans. Design Autom. Electron. Syst.*, vol. 25, no. 6, pp. 1–19, Nov. 2020, doi: 10.1145/3398037.

[5] 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, doi: 10.1109/JXCDC.2018.2809640.

[6] Y. Pan et al., "A multilevel cell STT-MRAM-based computing in-memory accelerator for binary convolutional neural network," *IEEE Trans. Magn.*, vol. 54, no. 11, pp. 1–5, Nov. 2018, doi: 10.1109/TMAG.2018.2848625.

[7] X. Fong et al., "Spin-transfer torque devices for logic and memory: Prospects and perspectives," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 35, no. 1, pp. 1–22, Jan. 2016, doi: 10.1109/TCAD.2015.2481793.

[8] D. E. Nikonov and I. A. Young, "Benchmarking of beyond-CMOS exploratory devices for logic integrated circuits," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 1, pp. 3–11, 2015, doi: 10.1109/JXCDC.2015.2418033.

[9] N. Sharma, C. Binek, A. Marshall, J. Bird, P. Dowben, and D. Nikonov, "Compact modeling and design of magneto-electric transistor devices and circuits," in *Proc. IEEE SOCC*, Sep. 2018, pp. 146–151, doi: 10.1109/SOCC.2018.8618494.

[10] A. Mahmood et al., "Voltage controlled Néel vector rotation in zero magnetic field," *Nature Commun.*, vol. 12, no. 1, pp. 1–8, Mar. 2021, doi: 10.1038/s41467-021-21872-3.

[11] H.-J. Chuang et al., "Low-resistance 2D/2D ohmic contacts: A universal approach to high-performance WSe_2, MoS_2, and $MoSe_2$ transistors," *Nano Lett.*, vol. 16, no. 3, pp. 1896–1902, Mar. 2016, doi: 10.1021/acs.nanolett.5b05066.

[12] M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov, "Modeling of nanoscale devices," *Proc. IEEE*, vol. 96, no. 9, pp. 1511–1550, Sep. 2008, doi: 10.1109/JPROC.2008.927355.

[13] A. Iyama and T. Kimura, "Magnetoelectric hysteresis loops in Cr_2O_3 at room temperature," *Phys. Rev. B, Condens. Matter*, vol. 87, no. 18, May 2013, Art. no. 180408, doi: 10.1103/physrevb.87.180408.

[14] N. Sharma, A. Marshall, J. Bird, and P. Dowben, "Verilog-A based compact modeling of the magneto-electric FET device," in *Proc. E3S*, Oct. 2017, pp. 1–3, doi: 10.1109/E3S.2017.8246186.

[15] X. Fong, S. K. Gupta, N. N. Mojumder, S. H. Choday, C. Augustine, and K. Roy, "KNACK: A hybrid spin-charge mixed-mode simulator for evaluating different genres of spin-transfer torque MRAM bit-cells," in *Proc. IEEE SISPAD*, Sep. 2011, pp. 51–54, doi: 10.1109/SISPAD.2011.6035047.

[16] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012, doi: 10.1109/TCAD.2012.2185930.

[17] A. K. Panda et al., "Crystallographic texture study of pulsed laser deposited Cr_2O_3 thin films," *Thin Solid Films*, vol. 660, pp. 328–334, Aug. 2018, doi: 10.1016/j.tsf.2018.06.030.

[18] S. Punugupati, J. Narayan, and F. Hunte, "Strain induced ferromagnetism in epitaxial Cr_2O_3 thin films integrated on Si(001)," *Appl. Phys. Lett.*, vol. 105, no. 13, Sep. 2014, Art. no. 132401, doi: 10.1063/1.4896975.

[19] K. Ali, F. Li, S. Y. Lua, and C.-H. Heng, "Energy efficient reduced area overhead spin-orbit torque non-volatile SRAMs," in *Proc. 46th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2020, pp. 2275–2280, doi: 10.1109/IECON43393.2020.9254623.

[20] C. Wang et al., "Magnetic nonvolatile SRAM based on voltage-gated spin-orbit-torque magnetic tunnel junctions," *IEEE Trans. Electron Devices*, vol. 67, no. 5, pp. 1965–1971, May 2020, doi: 10.1109/TED.2020.2982683.

[21] S. Tripathi, S. Choudhary, and P. K. Misra, "A novel STT-SOT MTJ-based nonvolatile SRAM for power gating applications," *IEEE Trans. Electron Devices*, vol. 69, no. 3, pp. 1058–1064, Mar. 2022, doi: 10.1109/TED.2022.3140407.

[22] A. Raha, A. Jaiswal, S. S. Sarwar, H. Jayakumar, V. Raghunathan, and K. Roy, "Designing energy-efficient intermittently powered systems using spin-Hall-effect-based nonvolatile SRAM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 2, pp. 294–307, Feb. 2018, doi: 10.1109/TVLSI.2017.2767033.

[23] W. Kang, W. Lv, Y. Zhang, and W. Zhao, "Low store power high-speed high-density nonvolatile SRAM design with spin Hall effect-driven magnetic tunnel junctions," *IEEE Trans. Nanotechnol.*, vol. 16, no. 1, pp. 148–154, Jan. 2017, doi: 10.1109/TNANO.2016.2640338.

[24] S. Tripathi, S. Choudhary, and P. K. Misra, "Highly reliable, stable, and store energy efficient 8T/9T-2D-2MTJ NVSRAMs," *IEEE Trans. Nanotechnol.*, vol. 23, pp. 89–94, 2024, doi: 10.1109/TNANO.2023.3345304.

[25] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: Enabling in-memory Boolean computations in CMOS static random access memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 12, pp. 4219–4232, Dec. 2018, doi: 10.1109/TCSI.2018.2848999.

[26] S. Aga et al., "Compute caches," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 481–492, doi: 10.1109/HPCA.2017.21.

[27] J. Wang et al., "A 28-nm compute SRAM with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 76–86, Jan. 2020, doi: 10.1109/JSSC.2019.2939682.

[28] S. Angizi, M. Morsali, S. Tabrizchi, and A. Roohi, "A near-sensor processing accelerator for approximate local binary pattern networks," *IEEE Trans. Emerg. Topics Comput.*, early access, Jun. 16, 2023, doi: 10.1109/TETC.2023.3285493.