IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 6 September 2024; revised 29 October 2024; accepted 13 November 2024. Date of publication 18 November 2024; date of current version 6 December 2024.

Digital Object Identifier 10.1109/JXCDC.2024.3502053

Ferroelectric Transistor-Based Synaptic Crossbar Arrays: The Impact of Ferroelectric Thickness and Device-Circuit Interactions

CHUNGUANG WANG and SUMEET KUMAR GUPTA (Senior Member, IEEE)

Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA

CORRESPONDING AUTHOR: C. WANG (wang4015@purdue.edu)

This work was supported in part by the Center for Brain-Inspired Computing (C-BRIC); and in part by the Center for the Co-Design of Cognitive Systems (COCOSYS), funded by Semiconductor Research Corporation (SRC) and Defense Advanced Research Projects Agency (DARPA) under Grant AWD-004311-S4.

This article has supplementary downloadable material available at https://doi.org/10.1109/10.1109/JXCDC.2024.3502053, provided by the authors.

ABSTRACT Ferroelectric transistors (FeFETs)-based crossbar arrays have shown immense promise for computing-in-memory (CiM) architectures targeted for neural accelerator designs. Offering CMOS compatibility, nonvolatility, compact bit cell, and CiM-amenable features, such as multilevel storage and voltage-driven conductance tuning, FeFETs are among the foremost candidates for synaptic devices. However, device and circuit nonideal attributes in FeFETs-based crossbar arrays cause the output currents to deviate from the expected value, which can induce error in CiM of matrix-vector multiplications (MVMs). In this article, we analyze the impact of ferroelectric thickness (T_{FE}) and cross-layer interactions in FeFETs-based synaptic crossbar arrays accounting for device-circuit nonidealities. First, based on a physics-based model of multidomain FeFETs calibrated to experiments, we analyze the impact of T_{FE} on the characteristics of FeFETs as synaptic devices, highlighting the connections between the multidomain physics and the synaptic attributes. Based on this analysis, we investigate the impact of T_{FE} in conjunction with other design parameters, such as number of bits stored per device (bit slice), wordline (WL) activation schemes, and FeFETs width on the error probability, area, energy, and latency of CiM at the array level. Our results show that FeFETs with T_{FE} around 7 nm achieve the highest CiM robustness, while FeFETs with T_{FE} around 10 nm offer the lowest CiM energy and latency. While the CiM robustness for bit slice 2 is less than bit slice 1, its robustness can be brought to a target level via additional design techniques, such as partial wordline activation and optimization of FeFETs width.

INDEX TERMS Computing-in-memory (CiM), crossbar array, error probability, ferroelectric thickness, ferroelectric transistors.

I. INTRODUCTION

COMPUTING-IN-MEMORY (CiM) is a promising technique to eliminate the overheads of memory-processor transactions. In the context of deep neural networks (DNNs), CiM has been mainly utilized for the computation of the most dominant kernel, i.e., matrix-vector multiplications (MVMs) or dot products between the synaptic weights and the neuron activations. The most common approach relies on simultaneous assertion of multiple rows in a crossbar memory array (storing the synaptic weight matrix) by applying the input voltage vector on the word-lines (WLs), leading to seamless and massively parallel computation of dot products on the sense-lines (SLs) [1].

While promising in enhancing the energy efficiency of DNN accelerators, CiM based on synaptic crossbar arrays poses its own challenges. As an example, standard 6T static random access memory (SRAM)-based CiM leads to severe stability concerns and aggravated design conflicts. This requires the design of more stable SRAMs, such as 8T SRAMs [2] (albeit with reduced area efficiency) or utilizing nonvolatile memory (NVM) technologies for crossbar array design. Various NVM device candidates have been explored in this context, which include spin-based memories [3], resistive RAMs (ReRAMs) [4], phase change memory [5], ferroelectric transistors (FeFETs) [6], and so on. While each technology has its own pros and cons, hafnium zirconium oxide (HZO)-based FeFETs have demonstrated a particularly great potential for synaptic crossbar design by virtue of their CMOS compatibility, electric-field driven programming, multilevel storage, and compact bit cell [7] (details in Supplementary S1).

The analyses in the works [6], [8], [9], [10], [11] have shown a large promise of FeFETs in realizing energy-efficient synaptic arrays. However, several design aspects have not been systematically analyzed and need to be studied to understand the benefits and limitations of FeFETs-based synaptic arrays. One such aspect is the analysis of device-circuit interactions. In other words, there is a need to comprehensively study how the unique device-level attributes of FeFETs interact with the circuit-level properties of the crossbar arrays, which, in turn, dictate the CiM energy, latency, area, and computational robustness.

Such a cross-layer analysis has been performed in this context of other memory technologies, such as SRAMs [12], spin-orbit torque magnetic RAMs (SOT-MRAMs) [3], and ReRAMs [4]. These works have highlighted the importance of accounting for device-circuit nonidealities (such as wire resistance, driver/sink loads, and device nonlinearities) to study their impact on computational errors and system accuracy. More details on non-idealities in crossbar arrays for CiM can be found in Supplementary S2. However, such studies for FeFETs are limited and often ignore important device or circuit properties.

In this article, we fill this gap by performing the device-circuit analysis for FeFETs-based synaptic crossbar arrays. Based on a cross-layer simulation flow (calibrated to device experiments), we analyze the implications of different device-circuit design knobs, including ferroelectric thickness (T_{FE}), number of bits stored per device (bit slice), WL activation schemes and FeFETs width (W) on array-level robustness, area, energy, and latency of CiM. The key contributions of this work are as follows.

1) We analyze the impact of T_{FE} on the characteristics of FeFETs based on a physical model calibrated with experiments and phase-field simulations.
2) We establish a cross-layer simulation flow to evaluate CiM robustness, area, energy, and latency for FeFETs crossbar arrays, which capture the interactions between the unique device attributes of FeFETs and nonidealities in crossbar array circuits.
3) We perform the design space exploration and co-optimization of key design knobs of FeFETs-based crossbar arrays considering the tradeoffs among CiM robustness, area, energy, and latency, and establishing the connection between the device-level attributes to the array characteristics.

II. CROSS-LAYER SIMULATION FLOW FOR FEFET-BASED CROSSBAR ARRAYS

We establish a cross-layer simulation flow based on which we perform a comprehensive analysis on the impact of various design knobs associated with FeFETs-based crossbar array on CiM robustness, area, energy, and latency considering device-circuit nonidealities. First, based on the experimental results of metal–ferroelectric–insulator–metal (MFIM) structures, we calibrate our in-house phase-field model of MFIM. Then, utilizing the trends obtained from MFIM phase-field simulations, we develop our in-house compact model of FeFETs calibrated with experiments. Next, we integrate the FeFETs compact model into the FeFETs-based crossbar array in HSPICE and perform extensive circuit simulations considering parasitic resistances and capacitances in the array. Utilizing this flow, we evaluate CiM robustness, area, energy, and latency of the FeFETs-based crossbar array and analyze the impact of various device-circuit design knobs. Next, we will elaborate each step in details.

A. PHASE-FIELD MODELING OF MFIM AND FEFET

To extensively capture the physics that govern the multidomain interactions in HZO, we utilize our in-house phase field model [13], which solves 2-D time-dependent Ginzburg–Landau (TDGL) equations self-consistently with the Poisson and semiconductor charge equations. Our framework captures T_{FE}-dependent multidomain properties including domain interactions and their role in determining domain patterns and polarization (P) switching via domain nucleation and domain growth. These properties, in turn, affect the macroscopic properties of HZO, such as memory window (MW), minor polarization–voltage (P–V) loop formations, permittivity, and so on.

Fig. 1(a) shows the calibration of the phase-field model with the experimental data [13] capturing the trends of Q–V characteristics for an MFIM structure (10-/7-/5-nm HZO + 5-nm Al_2O_3). Here, we would like to point out two important properties, which play an important role in the operation of FeFETs-based synaptic crossbar arrays. First, as observed in the experimental characteristics and phase-field simulations in Fig. 1(a), the effective dielectric permittivity of HZO (ε_D) in the MFIM stack increases with the decrease in T_{FE}. These trends are plotted in Fig. 1(b). (Note that ε_D is associated with the slope of charge-density versus electric field in the region where no polarization switching takes place.) This phenomenon is related to the multidomain interactions in HZO [13] (details in Supplementary S4).

The second important effect is related to the coercive voltage (V_C) of the FeFETs as a function of T_{FE}. As pointed out in several earlier works [13], V_C decreases with T_{FE} scaling. However, this dependence is not linear and needs to be properly captured. Our phase field model predicts the nonlinear dependency of the hysteretic window with respect to T_{FE}, as shown in Fig. 1(c). This nonlinear trend accounts for the T_{FE}-dependent domain patterns and polarization switching mechanisms in HZO. Specifically, in addition to the linear effect of geometry scaling on V_C, the nonlinear relationship is dictated by an increase in domain density with T_{FE} scaling [13] and the consequent change on how the polarization switches (i.e., combined effect of domain nucleation and domain wall motion for large T_{FE} versus domain wall motion-dominated switching for scaled T_{FE}). The nonlinear dependence of V_C on T_{FE} affects the set voltage (V_{SET}), i.e., how much programming voltage needs to be applied to achieve a target polarization switching. If a simple linear assumption is made between V_C and T_{FE} and V_{SET} is scaled

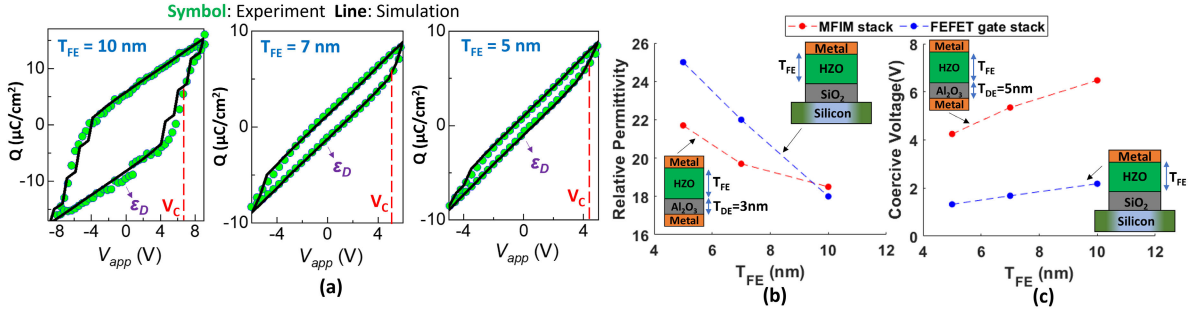

FIGURE 1. **(a) Q–V characteristics of MFIM stack (phase-field simulation matches with experiment). The method of identification of ε_D and V_C can be found in Supplementary S3. Dependence of (b) ε_D and (c) V_C on T_{FE} for MFIM stack and gate-stack of FeFET. ε_D decreases and V_C increases as T_{FE} increases.**

using that assumption, there will be some deviation from target values, potentially leading to incorrect computations. In this work, we obtain V_{SET} accounting for the nonlinear dependence of V_C on T_{FE} and based on some targets for the device currents, the details of which are presented later.

Now, based on the trends obtained from MFIM simulations (calibrated to the experiments), we model the T_{FE} dependency of ε_D and V_C for the gate-stack of FeFETs. For this, we utilize the experimentally calibrated parameters for HZO (obtained from the MFIM analysis) and couple HZO with 0.5-nm SiO_2 (interfacial oxide in FeFETs) in our simulation framework. The trends for ε_D and V_C as a function of T_{FE} for the gate-stack of FeFETs are shown in Fig. 1(b) and (c).

B. COMPACT MODELING OF FEFET

The values for ε_D and V_C with respect to T_{FE} scaling obtained from the phase-field model (Fig. 1) are utilized in a circuit-compatible model of FeFETs to carry out the simulations of synaptic crossbar arrays. The overall compact model of FeFETs is shown in Fig. 2(a). Our compact model [14] of multidomain FeFETs is based on modified Miller's equations [15], which capture the multidomain effects including minor P–V loop formation in ferroelectrics. FeFETs are modeled by integrating Miller's model with MOSFET following the approach in [14]. Predictive technology models (45-nm PTM HP) are used in modeling the underlying MOSFET in FeFET. The total capacitance of the gate-stack of FeFETs consists of two components: dielectric capacitance (C_{DE}) and capacitance induced from polarization switching (C_{FE}). We also consider a resistor (R_{FE}) as in [14] in series with C_{FE} to capture the delay associated with polarization switching.

The major P–V loop of FE is modeled by Miller's equation [15], which yields the maximum charge density (Q_M) at a particular applied electric field across FE (E_{FE})

$$Q_M^{\pm} = P_S \left[\tanh\left(\frac{E_{FE} \mp E_C(T_{FE})}{2\delta} \right) \right] + \varepsilon_D(T_{FE})\, \varepsilon_0 E_{FE} \quad (1)$$

$$\delta = \alpha \times \left[\ln\left[\frac{P_S + P_R}{P_S - P_R} \right] \right]^{-1} \quad (2)$$

where P_R is the remnant polarization, P_S is the saturated polarization, E_C is the coercive electric field, ε_0 is the vacuum

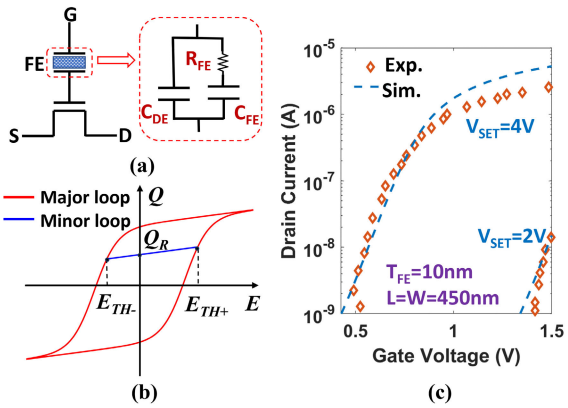

FIGURE 2. **(a) Compact model of FeFET integrating modified Miller's model with MOSFET. (b) Major and minor Q–E loop in ferroelectrics. (c) Simulated FeFET transfer characteristics show a reasonably good match with experiment [11].**

permittivity, and α controls the slope of Q_M versus E_{FE}, where polarization switching occurs.

The minor loops are modeled based on trends from the experiments and phase-field simulations. One limitation of the original Miller's model for minor loops is that it predicts a polarization increase for multiple voltage pulses with the same voltage amplitude. This is observed neither in the experiments nor in the phase-field models of FeFETs. Therefore, to alleviate this limitation, we model the charge density (Q) evolution in a minor loop in response to the electric field (E) by first obtaining $E_{TH+(-)}$ for a minor loop, as illustrated in Fig. 2(b). This is defined as the positive (negative) threshold field below (above) which no polarization switching takes place, i.e., HZO shows a dielectric response. For this, we start with the remnant charge density (Q_R) of that minor loop and follow its dielectric response in the positive (negative) directions till it intersects with the major loop [given by (1) and (2)]. This point of intersection gives $E_{TH+(-)}$. Once we have these threshold values, we model $Q = Q_R + \varepsilon_{DE}$ for $E_{TH-} < E < E_{TH+}$. For $|E| > |E_{TH+/-}|$, then Q follows the major Q–E loop. The parameters in the model have been calibrated with experiments [11] and validated using self-consistent phase-field simulations [13]. As shown in Fig. 2(c), a reasonably good fit of the transfer characteristics of FeFETs in the region of interest for this analysis (V_{GS}

TABLE 1. Parameters in FeFET compact model.

FE Thickness (nm)	10	7	5
Dielectric Permittivity of HZO	18	22	25
Coercive Voltage (V)	2.18	1.68	1.325
Saturated Polarization ($\mu C/cm^2$)	30		
Remnant Polarization ($\mu C/cm^2$)	27		

between 0 and 1 V) is observed between simulations and experiments [11]. Note that some overestimation in simulations is observed at higher gate voltages and could be due to assumptions in our phase-field model of FeFETs, such as homogeneous strain in FE and absence of charge traps as well as the abstraction of the parameters from the phase-field models to Miller's model (to make it circuit-compatible).

Based on the insights from phase-field simulations, we capture the effect of T_{FE}-dependent dielectric permittivity and MW by obtaining $E_C\ (T_{FE}) = V_C\ (T_{FE})/T_{FE}$ and $\varepsilon_D\ (T_{FE})$ from Fig. 1(b) and (c) (for FeFETs gate-stack) and using them in (1) and (2). The parameters used in the FeFETs compact model are summarized in Table 1. It is noteworthy that apart from V_C and ε_D, other parameters in (1) and (2) may also be a function of T_{FE}; however, here, we focus on the T_{FE}-dependency of ε_D and V_C to highlight the corresponding effects on the synaptic device behavior and crossbar array characteristics.

C. FEFET-BASED CROSSBAR ARRAY MODELING

To perform scalar multiplication of 1-bit input and 2-bit weight utilizing FeFETs, input bit stream is applied at the gate of FeFETs, with bits 0 and 1 corresponding to 0 and 1 V, respectively. The synaptic weight is encoded as the conductance of FeFETs, which can be programmed to different states utilizing the standard write technique: 1) reset voltage ($V_{RESET} = -5$ V) is applied to the gate of FeFETs to write weight $= 0$ and 2) then, V_{SET} is applied to the gate of FeFETs to program weight $= 1$, 2, or 3. A more detailed elaboration on the write operation of FeFETs can be found in Supplementary S5.

The output of scalar product of input and weight is obtained by sensing the FeFETs current (I_{DS}) with $V_{DS} = 0.25$ V. We define $I_{LRS,k}$ as I_{DS} when input $= 1$ and weight $= k$ ($k = 1$–3), I_{HRS} as I_{DS} when input $= 1$ and weight $= 0$, and I_{OFF} as I_{DS} when input $= 0$. I_{HRS} and I_{OFF} should be close to 0 (ideal values of corresponding scalar products being 0). However, in reality, the nonzero values of I_{HRS} and I_{OFF} in FeFETs lead to nonidealities and can degrade CiM robustness, especially when they are not negligible compared to $I_{LRS,1}$ (more details later). We design the 64×64 FeFETs-based crossbar array (Fig. 3) at 45-nm technology node. Our design employs input stream of 1 bit, weight slices of 1 or 2 bits, and the current-based sensing scheme. The compact FeFET bit cell is designed with a single FeFET by virtue of its self-selecting functionality. The gate terminals of FeFETs in a row are connected to WL running horizontally, while the drain (source) terminals in a column are connected to BL (SL) running vertically. The vertical height of the layout of

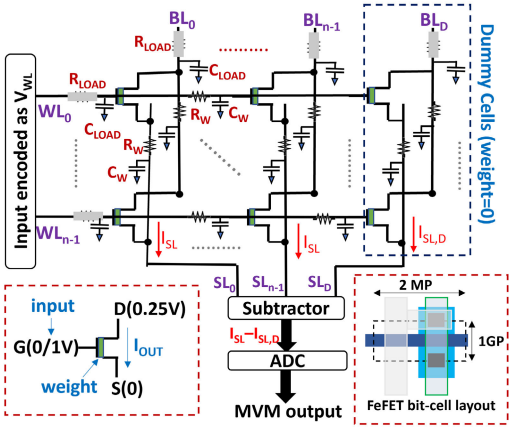

FIGURE 3. FeFETs-based crossbar array with dummy column. Layout of FeFET bit cell with height of one GP and width of two MP. The subtractor and ADC are implemented by utilizing behavioral models.

TABLE 2. Parameters in FeFET-based crossbar arrays.

Technology	45nm	Array Size	64×64
Metal Pitch	160nm [16]	Gate Pitch	160nm [16]
R_{LOAD}	500 Ω	C_{LOAD}	0.65fF
Bits/input	1b	Bits/device	1b/2b

the FeFET bit cell is one gate pitch (GP). Based on SCMOS layout rules, we determine that the horizontal dimension of the layout of FeFET bit cell is determined by the metal pitch (MP) when W is no more than $3 * W_{MIN}$ ($W_{MIN} = 67.5$ nm). This is important for W optimization, as we will discuss later.

The parameters used in the HSPICE simulations of FeFETs-based crossbar arrays are shown in Table 2. The parasitic resistance of peripheral circuits, such as driver (R_{LOAD}), their parasitic capacitance (C_{LOAD}), wire resistance (R_W), and wire capacitance (C_W), are considered in the circuit simulations of the crossbar array. We estimate the wire lengths for the WLs, BLs, and SLs based on the width and height of the layout of FeFET bit cell. This is used to obtain R_W and C_W for each cell, which are used in a distributed fashion, as illustrated in Fig. 3. At the 45-nm technology node, R_W is 3.3 Ω/μm [16] and C_W is 0.2 fF/μm [16]. We also consider the loading effect of the peripheral circuits, such as driver by using series-connected R_{LOAD}. These resistances (in addition to R_W) lead to computational errors.

Note that I_{HRS} is not negligible compared to $I_{LRS,1}$ especially for FeFET with $T_{FE} = 5$ nm, which can lead to error in CiM. Specifically, multiple HRS cells can produce total current, which is more than $I_{LRS,1}$, erroneously giving an output of 1 (when the output should be 0). To mitigate this effect, we follow the design in [1] to use a dummy column, in which all cells store weight 0. By subtracting the current through dummy column ($I_{SL,D}$) from the current through SL (I_{SL}), the multiply–accumulate (MAC) output can be obtained from $I_{SL} - I_{SL,D}$. Therefore, the ideal scalar product when weight is 0 corresponds to I_{HRS} (from real column) $-$ I_{HRS} (from dummy column) ~ 0, thus mitigating the effect of high I_{HRS}. (In reality, the wire resistances lead to a nonzero output for

this case, which we consider in our analysis.) When input is 0, the scalar product corresponds to the difference between the off-currents of the real and dummy columns, which is quite small. The scalar product when weight is $k = 1$, 2, or 3 and input is 1 is represented by $(I_{\mathrm{LRS,k}} - I_{\mathrm{HRS}})$. By tuning V_{SET}, the ratio of $(I_{\mathrm{LRS},3} - I_{\mathrm{HRS}})$:$(I_{\mathrm{LRS},2} - I_{\mathrm{HRS}})$:$(I_{\mathrm{LRS},1} - I_{\mathrm{HRS}})$ is designed to be 3:2:1 to achieve good device linearity. For a fair comparison, $I_{\mathrm{LRS},1} - I_{\mathrm{HRS}}$ of FeFET for $T_{\mathrm{FE}} = 5/7/10$ nm is designed to be the same ($\sim$3.3 μA), which defines the ideal current quantum corresponding between neighboring MAC output states. To get the same $I_{\mathrm{LRS},1} - I_{\mathrm{HRS}}$ for different T_{FE} values, we optimize V_{SET} with I_{HRS}, $I_{\mathrm{LRS},1}$, $I_{\mathrm{LRS},2}$, and $I_{\mathrm{LRS},3}$ evaluated at fixed $V_{\mathrm{GS}} = 1$ V and $V_{\mathrm{DS}} = 0.25$ V. The values of V_{SET} for different weights and T_{FE} are shown in Table S1 in the Supplemental Material. Linearly separated reference levels are utilized in analog-to-digital converters (ADCs) to convert I_{SL} into digital values and the reference current ($I_{\mathrm{REF},n}$) to distinguish MAC output $= n - 1$ and MAC output $= n$ is set as

$$I_{\mathrm{REF},n} = (I_{\mathrm{LRS},1} - I_{\mathrm{HRS}}) \times (n - 0.5). \tag{3}$$

The ideal sense margin is $(I_{\mathrm{LRS},1} - I_{\mathrm{HRS}})/2$, but due to crossbar nonidealities, the sense margin and the CiM robustness are degraded, which are discussed next.

D. EVALUATION OF CIM ROBUSTNESS

In order to evaluate the CIM robustness, we use error probability (P_{E}) as a statistical metric to quantify the computational robustness of crossbar arrays. The steps to obtain P_{E} are as follows.

1) First, we obtain the distribution of MAC output by profiling the output of the MVM operations (>10 000) of 1-bit 1 × 64 input vectors (subset of CIFAR-10 dataset) and 64 × 64 1-bit/2-bit weight matrices (subset of weight submatrices corresponding to ResNet20 network). Higher MAC output has a lower probability of occurrence due to the sparsity in the input vectors and weight matrices. We use exponential distribution to fit the histogram of MAC output [Fig. 4(b)], which yields the occurrence probability of MAC output (P_{O}).
2) From HSPICE circuit simulations of FeFETs-based crossbar arrays, we obtain I_{SL} for the MAC output. Due to various nonidealities in the FeFETs-based crossbar array, such as *IR* drop on wire resistances and nonzero I_{OFF} and I_{HRS}, different permutations of the input and weight vectors lead to different I_{SL}, despite corresponding to the same ideal output (detailed discussions can be found in [1]). To model the distribution of currents for each state of the MAC output, we use the following process. For MAC output > 0, we find that the Gaussian distribution offers a reasonable fit [Fig. 4(a) and (c)] and can be attributed to the central limit theorem [17]. For MAC output = 0, a combination of exponential model and Gaussian model is utilized to fit the histogram of I_{SL} [Fig. 4(d)]. Note that due to high input and weight sparsity, a large percentage of the currents are closer to 0. Furthermore, the dummy column used in our design reduces the effect of nonzero I_{HRS}, leading to a further increase in this percentage. This necessitates the use of exponential function in conjunction with the Gaussian function for output = 0. We define f_n as the probability density function of I_{SL} corresponding to MAC output $= n$.
3) In addition to this range of currents due to different permutations of the input and weight bits, we consider process variations in our analysis. For that, we use the Gaussian function (f_{PV}) to model the probability distribution function of I_{SL} with variations (I_{PV}). The mean of f_{PV} is $I_{\mathrm{SL},n}$ which is I_{SL} for MAC output $= n$ (without variations) and is obtained by sampling f_n. The standard deviation (σ_n) of f_{PV} models the device-to-device variation, where n refers to the MAC output. Let us define $I_{\mathrm{LRS},1} - I_{\mathrm{HRS}}$ for FeFETs with W_{MIN} as I_1 (which, recall, is matched for all T_{FE}). We also define I_0 as the maximum of I_{HRS} and I_{OFF} for FeFETs with W_{MIN}. For our analysis in Section III, we assume $\sigma_n = s * I_1 * \sqrt{n} * \sqrt{(W/W_{\mathrm{MIN}})}$ for $n > 0$. Here, s is a factor representing the relative deviation in the current. The erroneous MAC output value is sensed in ADC when $I_{\mathrm{SL},n}$ is not between $I_{\mathrm{REF},n}$ and $I_{\mathrm{REF},n+1}$. The error probability for a given $I_{\mathrm{SL},n}$ ($n > 0$) is calculated as

$$P\left(I_{\mathrm{SL},n}, \sigma_n\right) = 1 - \int_{I_{\mathrm{REF},n}}^{I_{\mathrm{REF},n+1}} f_{\mathrm{PV}}\left(I_{\mathrm{PV}}, I_{\mathrm{SL},n}, \sigma_n\right) dI_{\mathrm{PV}}. \tag{4}$$

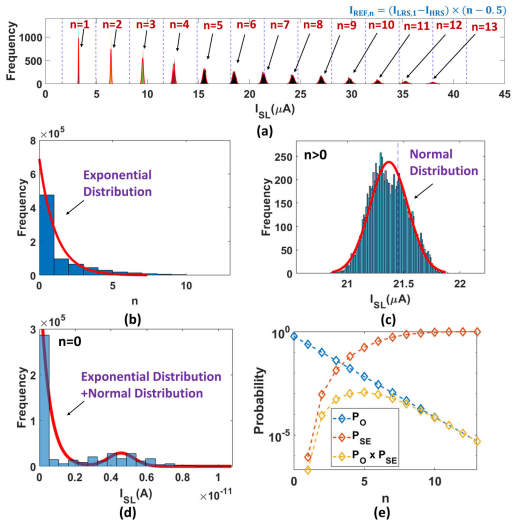

FIGURE 4. Taking FeFETs-based crossbar arrays (T_{FE} = 5 nm, 64 WLs activated, bit slice 1, $W = W_{\mathrm{MIN}}$, and $s = 0.1$) as an example. (a) Histogram of I_{SL} corresponding to nonzero MAC output and their respective normal distribution fit. (b) Histogram of MAC output (n) with fit of exponential distribution. (c) Typical histogram of I_{SL} corresponding to nonzero output with fit of Gaussian model. (d) Typical histogram of I_{SL} corresponding to output = 0 with fit of combination of exponential model and Gaussian model. (e) P_{O}, P_{SE}, and $P_{\mathrm{O}} \times P_{\mathrm{SE}}$ versus MAC output.

For $n = 0$, we assume $\sigma_n = s * I_0 * \sqrt{(W/W_{\mathrm{MIN}})}$, and erroneous MAC output value is sensed in ADC when $I_{\mathrm{SL},0}$ is more than $I_{\mathrm{REF},1}$. The error probability for $I_{\mathrm{SL},0}$ is calculated as

$$P\left(I_{\mathrm{SL},0}, \sigma_n\right) = 1 - \int_0^{I_{\mathrm{REF},1}} f_{\mathrm{PV}}\left(I_{\mathrm{PV}}, I_{\mathrm{SL},0}, \sigma_n\right) dI_{\mathrm{PV}}. \tag{5}$$

4) Last, we combine all the probabilities to calculate the overall error probability. The sensing error probability for MAC output $= n$ is

$$P_{\mathrm{SE}}\left(n, \sigma_n\right) = \int_{-\infty}^{\infty} P\left(I_{\mathrm{SL},n}, \sigma_n\right) \times f_n\left(I_{\mathrm{SL},n}\right) dI_{\mathrm{SL},n}. \tag{6}$$

Total error probability (P_{E}) can be calculated as

$$P_{\mathrm{E}}(\sigma_n) = \sum_{n=0}^{\max(n)} P_{\mathrm{SE}}(n, \sigma_n) \times P_{\mathrm{O}}(\mathrm{n}) \tag{7}$$

where $P_{\mathrm{SE}}(n, \sigma_n)$ and $P_{\mathrm{O}}(n)$ are the sensing error probability for MAC output $= n$, respectively, and $\max(n)$ is the maximum MAC output obtained from crossbar array simulations. Note that $\max(n)$ is less than the theoretical maximum MAC output due to the sparsity in input vectors and weight matrices. Fig. 4(e) shows an example of the values of P_{O}, P_{SE}, and their product ($P_{\mathrm{O}} \times P_{\mathrm{SE}}$) for each n. On the one hand, P_{O} decreases as n increases. On the other hand, as n increases, P_{SE} increases due to the increase of nonidealities from *IR* drop. The two opposing effects lead to the nonmonotonic trend between the product ($P_{\mathrm{O}} \times P_{\mathrm{SE}}$) and n.

Due to the inherent error tolerance of DNNs, DNNs exhibit negligible impact on system accuracy as long as error probability in partial sums is reasonably low [18]. Following the method in [18], we perform system-level simulations (CIFAR-10 dataset and ResNet20 network) by introducing errors in the partial sums based on the error probability obtained from our P_{E} analysis. Our study shows that $P_{\mathrm{E}} <$ 0.03 has little impact on the DNNs accuracy (<0.5% accuracy degradation). Thus, 0.03 is set as the threshold P_{E} and CiM with $P_{\mathrm{E}} < 0.03$ is considered as a robust design in our comparison. It is worth mentioning that the magnitude of threshold P_{E} is dependent on the dataset and network.

III. DESIGN SPACE EXPLORATION OF FEFET-BASED CROSSBAR ARRAY

In this section, we comprehensively analyze the impact of T_{FE} on CiM robustness (P_{E}), area, energy, and latency in conjunction with other design parameters, such as bit slice, number of activated WLs, and W. The analysis in this section will highlight the cross-layer interactions and the design tradeoffs associated with different design knobs.

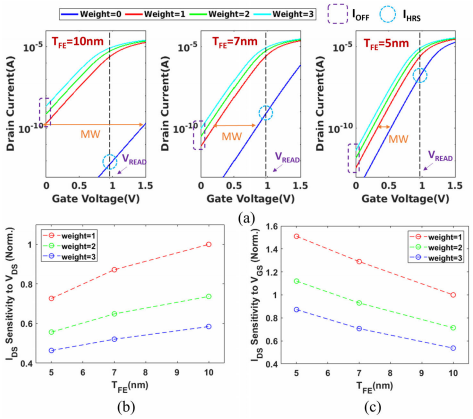

FIGURE 5. (a) FeFET transfer characteristics for different weight (0, 1, 2, 3) and T_{FE} (10, 7, and 5 nm) at $V_{\mathrm{DS}} = 0.25$ V ($L = 45$ nm and $W = 67.5$ nm). (b) Sensitivity of I_{DS} to V_{DS}. (c) Sensitivity of I_{DS} to V_{GS} for different weights (1, 2, 3) and T_{FE} (10, 7, and 5 nm) at $V_{\mathrm{DS}} = 0.25$ V and $V_{\mathrm{GS}} = 1$ V. S_{DS} and S_{GS} for $T_{\mathrm{FE}} = 10$ nm, weight = 1 is normalized as 1.

A. IMPACT OF FERROELECTRIC THICKNESS

We first analyze the impact of T_{FE} on I–V characteristics of FeFETs. The transfer characteristics of FeFETs, which store four states (weight = 0/1/2/3) for $T_{\mathrm{FE}} = 10$, 7, and 5 nm, are compared in Fig. 5(a). We observe two important trends with respect to T_{FE} scaling.

1) As T_{FE} decreases (i.e., C_{FE} increases), the MW reduces as the effect of the series dielectric/channel capacitance increases, which increases the depolarization fields and reduces the polarization switching. Thus, the threshold voltage shift due to polarization switching decreases, reducing the MW. This, in turn, means that for the same I_{LRS}, I_{HRS} increases as T_{FE} is scaled (due to lower threshold voltage shift).
2) With T_{FE} scaling, I_{OFF} decreases since the gate capacitance of FeFETs increases with T_{FE} scaling leading to the reduction in short channel effects and leakage. Note that this effect captures the dependence of ε_{D} on T_{FE} [Fig. 1(b)]. Specifically, the reduction in I_{OFF} with T_{FE} scaling obtained from our analysis is more than the scenario, in which ε_{D} is assumed to be a constant. This is because as ε_{D} increases with T_{FE}, it leads to a larger improvement in short-channel effects compared to simple geometric scaling of T_{FE}.

Another important factor affecting the nonidealities is the sensitivity of I_{DS} to the *IR* drops. *IR* drop on parasitic resistance in the crossbar array yields the deviation of the drain voltage (V_{D}) and the source voltage (V_{S}) of FeFETs from their ideal value ($V_{\mathrm{D}} = 0.25$ V and $V_{\mathrm{S}} = 0$), which in turn results in the variations of I_{DS}. To understand this, we evaluate the sensitivity of I_{DS} to V_{DS} and V_{GS}. Higher sensitivity leads to more impact of *IR* drop on the output current. We define S_{DS} as the sensitivity of I_{DS} to V_{DS} and S_{DS} is calculated as $(dI_{\mathrm{DS}}/dV_{\mathrm{DS}})/(I_{\mathrm{LRS}}/V_{\mathrm{DS}})$. Here, $dI_{\mathrm{DS}}/dV_{\mathrm{DS}}$ is the slope of FeFETs I_{DS}–V_{DS} curve at $V_{\mathrm{DS}} = 0.25$ V. Similarly, we define S_{GS} as the sensitivity of I_{DS} to V_{GS} and S_{GS} is calculated as

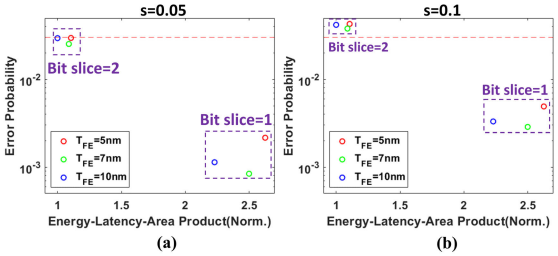

FIGURE 6. P_E versus energy–latency–area product comparison for T_{FE} = 5, 7, and 10 nm, bit slice 1, 2, 64 WLs activated, $W = W_{MIN}$. (a) $s = 0.05$ and (b) $s = 0.1$.

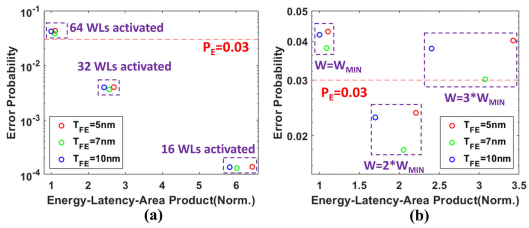

FIGURE 7. P_E versus energy–latency–area product comparison for T_{FE} = 5/7/10 nm, bit slice 2, $s = 0.1$, (a) 64, 32, and 16 WLs activated, $W = W_{MIN}$ and (b) $W = W_{MIN}$, $2 * W_{MIN}$, $3 * W_{MIN}$, 64 WLs activated.

$(dI_{DS}/dV_{GS})/(I_{LRS}/V_{GS})$. Here, dI_{DS}/dV_{GS} is the slope of FeFETs I_{DS}–V_{GS} curve at $V_{GS} = 1$ V. As shown in Fig. 5(b) and (c), S_{DS} decreases and S_{GS} increases with T_{FE} scaling because the gate capacitance of FeFETs increases with T_{FE} scaling. In other words, FeFETs with lower T_{FE} show better gate control and are more sensitive to the deviation in V_{GS}. In contrast, FeFETs with larger T_{FE} are more sensitive to the deviation in V_{DS} due to short-channel effects. Note that the sensitivities are also impacted by the dependence of ε_D on T_{FE}, which we account for in our analysis.

It is important to note that I_{HRS} and I_{OFF} show opposing trends with T_{FE} scaling, i.e., the former increases while the latter decreases (when ideally, both should be zero). In addition, S_{DS} and S_{GS} show opposing trends with T_{FE} scaling, i.e., the former decreases while the latter increases. The question, therefore, is: what the overall impact of T_{FE} scaling is on CiM robustness. We explore this question next.

As shown in Fig. 6, among different T_{FE}, P_E for T_{FE} = 7 nm is the minimum. We can understand this from the trend of I_{OFF}, I_{HRS}, S_{DS}, and S_{GS} with T_{FE} scaling. On the one hand, I_{OFF} and S_{DS} decrease as T_{FE} scales. On the other hand, I_{HRS} and S_{GS} increase as T_{FE} scales. Note that the subtraction of current through dummy column [1] partially compensates the impact of high I_{HRS}, but this compensation is not complete due to different *IR* drop in the real and dummy columns. It turns out that T_{FE} = 7 nm is the sweet spot with minimum P_E considering the two-opposing trend of I_{OFF}/I_{HRS} and S_{DS}/S_{GS} with T_{FE} scaling. We also observe that the energy–latency–area product increases with T_{FE} scaling. The reason is that the gate capacitance of FeFETs increases as T_{FE} decreases leading to higher energy for charging WLs. The procedure to evaluate the CiM energy can be found in Supplementary S6.

B. IMPACT OF BIT SLICE

Bit slice is an important design knob (recall, bit slice *m* implies m bits stored per bit cell). Increasing bit slice can increase the memory density and parallelism in CiM. However, previous work shows that ReRAMs with bit slice beyond 2 degrade accuracy significantly [4]. Hence, we consider bit slice 1 and 2 in FeFETs in our analysis. To maintain sufficient sense margin, the conductance for weight = 0 and weight = 1 is kept the same while designing 1-bit and 2-bit FeFETs. Compared to $I_{LRS,1}$ in bit slice 1, high $I_{LRS,2}$ and $I_{LRS,3}$ in bit slice 2 leads to an increase in I_{SL} and thus worsens the *IR* drop in the crossbar array. From Fig. 6, we observe that the P_E for bit slice 2 is higher than bit slice 1 due to the worsening of *IR* drop. As bit slice increases from 1 to 2, energy–latency–area product decreases because a lower number of crossbar arrays are needed to perform MVM. The area of crossbar array for bit slice 1 is twice that of bit slice 2.

For bit slice 1, P_E for T_{FE} = 5 nm is more than P_E for T_{FE} = 10 nm, while for bit slice 2, P_E for T_{FE} = 5 and 10 nm is comparable. The reason is that the weight sparsity in bit slice 1 is higher than bit slice 2. In other words, compared with bit slice 2, an array with bit slice 1 encounters more cases for I_{HRS} (input = 1 and weight = 0) and less cases for I_{OFF} (input = 0 and weight = 1). Hence, the impact of I_{HRS} on nonidealities is more dominant than I_{OFF} for bit slice 1. Since I_{HRS} is higher for T_{FE} = 5 nm than T_{FE} = 10 nm, P_E for T_{FE} = 5 nm is also higher for bit slice 1. For bit slice 2, the impact of I_{HRS} is reduced, leading to comparable P_E for T_{FE} = 5 and 10 nm. To account for process variations, we consider relative deviation in the current $s = 0.05$ (5%) and 0.1 (10%) in our analysis. As shown in Fig. 6, P_E is less than 0.03 (the threshold value with <0.5% inference accuracy drop) for bit slice 1 at both values of *s* across all T_{FE}. However, for bit slice 2, $P_E > 0.03$ at $s = 0.1$ (but <0.03 for $s = 0.05$). Therefore, if the variations are large, bit slice 2 may not meet accuracy targets. To further reduce P_E for bit slice 2, we analyze other design knobs viz. number of activated WLs and FeFETs width next.

C. IMPACT OF NUMBER OF ACTIVATED WLs

With full WL activation (FWA) in which all WLs in the crossbar array are asserted simultaneously during CiM, minimum energy and latency can be achieved, due to maximum parallelism. However, P_E can be high for FWA, especially for bit slice 2. To reduce P_E in CiM, we can utilize the partial WL activation (PWA) [19] technique, in which a subset of WLs are activated in one cycle and multiple cycles are employed to perform MVM operations. For example, if *k* WLs are asserted in one cycle, then m/k cycles are needed for CiM for $m \times m$ crossbar arrays.

Fig. 7(a) shows the comparison of P_E and normalized energy–latency–area product for different numbers of activated WLs. As the number of activated WLs decreases, P_E decreases due to the reduction in I_{SL} and *IR* drop. On the other hand, energy–latency–area product increases due to the loss in parallelism. By reducing the number of activated WLs

from 64 to 32, P_E in CiM can be reduced to below 0.03, albeit with the penalty of increasing energy and latency. Note that P_E for $T_{FE} = 5/7/10$ nm becomes comparable when the number of activated WLs decreases. Recall for FWA, $T_{FE} =$ 7 nm yields minimum P_E; but when the number of activated WLs is reduced to 16, the P_E values are almost equal for the three T_{FE}. This is because the reduction of *IR* drop in PWA results in less variations of V_{DS} and V_{GS} of FeFETs in the crossbar array. Therefore, the impact of the difference of S_{DS} and S_{GS} for different T_{FE} is reduced. In other words, while opposing S_{DS} and S_{GS} trends yield minimum P_E for $T_{FE} =$ 7 nm for FWA, reduced *IR* drop in PWA diminishes the role of S_{DS} and S_{GS}, reducing the difference in P_E for the three T_{FE} values.

D. IMPACT OF FEFET WIDTH

The on-resistance of a bit cell is an important design variable from the perspective of CiM robustness since it has a direct impact on I_{SL} [1], [3]. FeFETs width is an important design knob to control the on-resistance. However, it may also impact the bit-cell area. As discussed earlier, the horizontal dimension of FeFET bit cell is determined by the MP and not by W when $W < 3{*}W_{MIN}$. In other words, W can be increased up to $3 * W_{MIN}$ without any impact on the bit-cell area. Therefore, for this analysis, we sweep W from W_{MIN} (=67.5 nm) to $3 * W_{MIN}$ to understand the effect of W at isoarea. From Fig. 7(b), we observe that P_E with respect to W shows a nonmonotonic trend. $2 * W_{MIN}$ is the optimal W with the minimum P_E. As W increases, I_{SL} increases, which has the following opposing effects. On the one hand, it exacerbates the nonidealities due to *IR* drop on parasitic resistances. On the other hand, the current quantum between two neighboring output states is enhanced, then augmenting the sense margin for the ADC. In addition, the random device process variation (σ/μ) decreases as W increases, which reduces the sensing errors. Considering the tradeoff between these opposing effects, $W = 2 * W_{MIN}$ is the sweet spot with minimum P_E (<0.03). Among different T_{FE}, $T_{FE} =$ 7 nm yields the lowest P_E due to the reasons discussed in Section III-A.

As shown in Fig. 7(b), the CiM energy–latency–area product increases as W increases. This is due to (a) an increase in the gate and drain capacitance, which leads to higher energy for charging BLs and WLs and (b) larger I_{BL}.

E. ISO-ROBUSTNESS ENERGY–LATENCY–AREA COMPARISON

By utilizing PWA (32 WLs activated) or optimizing W ($W =$ $2 * W_{MIN}$), P_E for bit slice 2 can be reduced below the threshold (0.03) with the penalty of increasing energy and latency. We perform isorobustness ($P_E < 0.03$) comparison of energy–latency–area product for three cases: 1) bit slice 1, FWA, $W = W_{MIN}$; 2) bit slice 2, PWA (32 WLs activated), $W = W_{MIN}$; and 3) bit slice 2, FWA, $W = 2 * W_{MIN}$. As shown in Fig. 8, case 3) achieves the lowest energy–latency–area product. Among $T_{FE} = 5/7/10$ nm, $T_{FE} = 10$ nm is optimal in the context of energy–latency–area product due to the lowest capacitance.

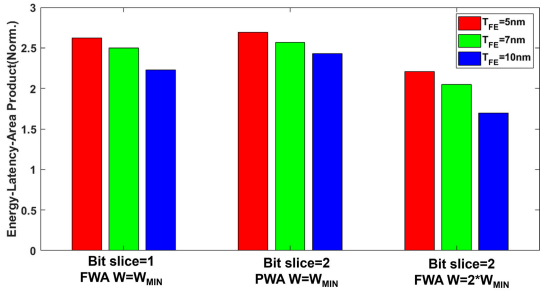

FIGURE 8. Isorobustness array-level energy–latency–area product comparison. Energy–latency–area product is normalized as 1 for the condition $T_{FE} = 10$ nm, bit slice 2, full-WL activation, $W = W_{MIN}$, and $s = 0.1$.

It is worth mentioning that here, we focus only on the area, energy, and latency of the crossbar array (excluding the peripheral circuits), with an objective to illustrate the effect of various device-array design knobs on CiM robustness and array-level area/energy/latency. However, it is worthwhile to mention the tradeoffs between bit slices 1 and 2 considering the ADC and other overheads. While bit slice 1 requires lower precision ADCs (the output being from 0 to m for an $m \times m$ array) compared to bit slice 2 (with output ranging from 0 to $3m$), the number of ADCs needed for bit slice 2 is half that of bit slice 1. Overall, the ADC costs for bit slice 2 are higher. However, bit slice 2 allows for larger number of DNN model parameters to be stored. This can reduce the off-chip dynamic random access memory (DRAM) access, leading to significant energy and latency reduction. Thus, the overall choice of the bit slice needs to consider such macro/system-level aspects along with the device-array co-design strategies presented in this work.

IV. CONCLUSION

In this article, we perform a design space exploration of FeFET-based crossbar arrays for CiM considering device-circuit nonidealities and cross-layer interactions. Based on the physics-based model of FeFETs, we analyze the T_{FE} dependence of ε_D in FeFETs, which also accounts for the trends of I_{OFF}, I_{HRS}, S_{GS}, and S_{DS} with T_{FE} scaling. We compare the CiM robustness, area, energy, and latency for FeFETs-based crossbar arrays with different design knobs. We show that T_{FE} around 7 nm is optimal in the context of CiM robustness, while T_{FE} around 10 nm offers the lowest CiM energy and latency. Compared with bit slice 2, bit slice 1 achieves higher CiM robustness at the cost of higher CiM energy–latency–area product. By utilizing PWA or optimization of FeFETs width, CiM robustness for bit slice 2 can be improved at the cost of increasing energy and latency.

ACKNOWLEDGMENT

The authors would like to thank Jeffry L. Victor and Atanu K. Saha at Purdue University, West Lafayette, IN, USA, for their insights on system-level accuracy and FeFET modeling, respectively.

REFERENCES

[1] C. Wang, J. Victor, and S. K. Gupta, "Comparative evaluation of memory technologies for synaptic crossbar arrays—Part I: Robustness-driven device-circuit co-design and system implications," 2023, *arXiv:2307.04261*.

[2] A. Jaiswal, I. Chakraborty, A. Agrawal, and K. Roy, "8T SRAM cell as a multibit dot-product engine for beyond von Neumann computing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 11, pp. 2556–2567, Nov. 2019, doi: 10.1109/TVLSI.2019.2929245.

[3] T. Sharma, C. Wang, A. Agrawal, and K. Roy, "Enabling robust SOT-MTJ crossbars for machine learning using sparsity-aware device-circuit co-design," presented at the *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2021, doi: 10.1109/ISLPED52811.2021.9502492.

[4] I. Chakraborty, M. Fayez Ali, D. Eun Kim, A. Ankit, and K. Roy, "GENIEx: A generalized approach to emulating non-ideality in memristive xbars using neural networks," presented at the *Proc. 57th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2020, doi: 10.1109/DAC18072.2020.9218688.

[5] A. Sebastian, M. Le Gallo, and E. Eleftheriou, "Computational phase-change memory: Beyond von Neumann computing," *J. Phys. D, Appl. Phys.*, vol. 52, no. 44, Oct. 2019, Art. no. 443002, doi: 10.1088/1361-6463/ab37b6.

[6] M. Jerry, P. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, and S. Datta, "Ferroelectric FET analog synapse for acceleration of deep neural network training," presented at the *IEDM Tech. Dig.*, 2018, doi: 10.1109/IEDM.2017.8268338.

[7] C. Wang et al., "FeFET-based synaptic cross-bar arrays for deep neural networks: Impact of ferroelectric thickness on device-circuit non-idealities and system accuracy," presented at the *Device Res. Conf. (DRC)*, Jun. 2023, pp. 1–2, doi: 10.1109/DRC58590.2023.10187042.

[8] A. Keshavarzi, K. Ni, W. Van Den Hoek, S. Datta, and A. Raychowdhury, "FerroElectronics for edge intelligence," *IEEE Micro*, vol. 40, no. 6, pp. 33–48, Nov. 2020, doi: 10.1109/MM.2020.3026667.

[9] T. Soliman et al., "Ultra-low power flexible precision FeFET based analog in-memory computing," presented at the *IEDM Tech. Dig.*, Dec. 2020.2.1–29.2.4, doi: 10.1109/IEDM13553.2020.9372124.

[10] F. Müller et al., "Multi-level operation of ferroelectric FET memory arrays for compute-in-memory applications," in *Proc. Inst. Elect. Electron. Eng. (IEEE)*, Jun. 2023, pp. 1–4, doi: 10.1109/imw56887.2023.10145940.

[11] K. Ni et al., "In-memory computing primitive for sensor data fusion in 28 nm HKMG FeFET technology," presented at the *IEDM Tech. Dig.*, Dec. 2018.1.1–16.1.4, doi: 10.1109/IEDM.2018.8614527.

[12] A. Kneip and D. Bol, "Impact of analog non-idealities on the design space of 6T-SRAM current-domain dot-product operators for in-memory computing," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 5, pp. 1931–1944, May 2021, doi: 10.1109/TCSI.2021.3058510.

[13] A. K. Saha, M. Si, K. Ni, S. Datta, P. D. Ye, and S. K. Gupta, "Ferroelectric thickness dependent domain interactions in FEFETs for memory and logic: A phase-field model based analysis," presented at the *IEDM Tech. Dig.*, Dec. 2020.3.1–4.3.4, doi: 10.1109/IEDM13553.2020.9372099.

[14] A. K. Saha and S. K. Gupta, "Modeling and comparative analysis of hysteretic ferroelectric and anti-ferroelectric FETs," presented at the *Proc. 76th Device Res. Conf. (DRC)*, Jun. 2018, doi: 10.1109/DRC.2018.8442136.

[15] S. L. Miller, J. R. Schwank, R. D. Nasby, and M. S. Rodgers, "Modeling ferroelectric capacitor switching with asymmetric nonperiodic input signals and arbitrary initial conditions," *J. Appl. Phys.*, vol. 70, no. 5, pp. 2849–2860, Sep. 1991, doi: 10.1063/1.349348.

[16] P. Moon et al., "Process and electrical results for the on-die interconnect stack for Intel's 45 nm process generation," *Intel Technol. J.*, vol. 12, pp. 87–92, Jan. 2008.

[17] S. G. Kwak and J. H. Kim, "Central limit theorem: The cornerstone of modern statistics," *Korean J. Anesthesiol.*, vol. 70, no. 2, pp. 144–156, Apr. 2017.

[18] S. Jain, S. K. Gupta, and A. Raghunathan, "TiM-DNN: Ternary in-memory accelerator for deep neural networks," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 7, pp. 1567–1577, Jul. 2020, doi: 10.1109/TVLSI.2020.2993045.

[19] J. Victor, C. Wang, and S. K. Gupta, "Comparative evaluation of memory technologies for synaptic crossbar arrays—Part 2: Design knobs and DNN accuracy trends," 2024, *arXiv:2408.05857*.