A Study on the Impact of Temperature-Dependent Ferroelectric Switching Behavior in 3D Memory Architecture

Varun Darshana Parekh*, Yi Xiao*, Yixin Xu*, Zijian Zhao†, Zhouhang Jiang†, Rudra Biswas*,
Sumitha George‡, Kai Ni†, Vijaykrishnan Narayanan*
*The Pennsylvania State University, USA †University of Notre Dame, USA ‡North Dakota State University, USA

***Abstract*—The flourishing development of neural networks that require exponentially growing amounts of data has presented an elevated demand for memory footprint. To address this, researchers have been exploring hardware accelerators with innovative memory architectures like 3D memory. These 3D memory architectures offer enhanced storage capacity and processing capabilities, at a cost of rising on-chip temperature during operation. Hafnium Zirconium Oxide (HZO) based Ferroelectric Random Access Memory (FeRAM) is a promising nonvolatile memory candidate in neural network hardware accelerators for its outstanding write performance and reliability. However, its implementation in the architecture regarding the temperature-dependent ferroelectric switching behavior has not been well studied. In this work, we study the thermal impacts on polarization switching through experimental devices and simulation results. We conduct the circuit and architecture-level simulations to showcase that one can exploit this temperature rise to reduce FeRAM's write voltage and write energy due to its unique temperature-activated polarization switching mechanisms. As the on-chip temperature increases to 351K (ambient temperature at 300K) due to neural network workloads, the access energy per bit can be reduced by 27.6% when a dynamic write voltage is applied.**

***Index Terms*—FeRAM, 3D memory, DNN hardware accelerator, temperature dependence, thermal-aware floor plan**

I. Introduction

Neural networks have been widely adopted in various fields due to their advancements. These deeper models with billions of parameters have a cost and a very high memory footprint [1]. By offloading intensive computations from central processing units (CPUs) to specialized hardware, such as graphics processing units (GPUs), field-programmable gate arrays (FPGAs), or application-specific integrated circuits (ASICs), these accelerators can significantly speed up the inference process while reducing power consumption. This is especially critical in applications requiring real-time processing and low latency, such as autonomous driving and real-time language translation [2], [3]. ASICs, such as Google's Tensor Processing Units (TPUs) [4], are specially fabricated to execute neural network operations with optimal efficiency, demonstrating significant performance gains and energy savings over conventional hardware [5]. Researchers have also explored various approaches to manage power and accuracy trade-offs, including approximate computing during inference phases using quantization and fixed-point multipliers [6].

Alongside the advancements in hardware accelerators, there is a pressing need to enhance the underlying memory architectures, including 3D memories, to support these advanced neural network models' extensive memory footprint. And as we push the limits of memory architectures to meet the demands of advanced neural networks, the resulting higher power densities introduce significant thermal challenges. While there is extensive research on using 3D memory for hardware accelerators [7]–[9], these studies often overlook the thermal characteristics of these memory systems, necessitating sophisticated thermal management strategies to maintain system reliability and performance [10]–[12]. Studies by Liu et al. [13] have demonstrated the severe implications of not addressing thermal issues, with inference loss reaching up to 90% under high temperatures. Another study by A. Abdurrob et al. [14] implies that DRAM, when used in a 3D hardware accelerator, can reach an extremely high peak temperature of $170°C$. The refresh overhead, in this case, can be increased 4 times.

Given the significant impact of thermal issues on system performance and reliability, there is a crucial need to integrate innovative thermal management strategies to not only mitigate the risks associated with high temperatures but also pave the way for exploring new design paradigms that balance thermal effects with energy efficiency and computational accuracy. Layer-wise approximation and thermal-aware floor plan design have been developed to enhance energy efficiency [7], [15]. A particular focus on heat management by Zervakis et al. [16] illustrates a methodology to prioritize energy efficiency while managing thermal impacts effectively. However, these methods can lead to some loss in computational accuracy and often do not effectively capitalize on the potential benefits of increased temperatures within the system.

Nonvolatile memories, such as resistive random access memories (ReRAM) and ferroelectric random access memories (FeRAM), consume less energy than their DRAM equivalents. HfO_2 based FeRAM sharing a similar cell structure to DRAM but being nonvolatile, is preferable as memories in neural network hardware accelerators for its CMOS compatibility and excellent write performance, such as low operating voltage and high reliability [17], [18]. Recently, Ramaswamy et al. have presented a two-tier stacked FeRAM array [19] to enable the high-density, high-performance requirements for a near-DRAM memory solution. Despite the excellent read/write performance, managing thermal effects in such stacked FeRAM is crucial, as it directly impacts device reliability and energy efficiency. Chen et al. study the ferroelectricity and polarization-switching behavior in $Hf_{0.5}Zr_{0.5}O_2$ films from 25°C to 150°C [20]. It is revealed that a strong thermal activation of oxygen vacancies causes the temperature dependent leakage current in $Hf_{0.5}Zr_{0.5}O_2$ films. Hur et al. investigate the polarization switching of ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ (HZO) thin film in wide-ranging temperatures from 4K to 400K regarding the reliability effects such as endurance, retention, and small-signal response [21]. A strong temperature dependence is demonstrated with these characteristics. Ali et al. report the temperature-dependent operation for fluorite-structure-based ferroelectric FET (FeFET) [22], where the study of memory window (MW) indicates a strong dependence on temperature intrinsic to the ferroelectric polarization. While these works primarily focus on ferroelectrics' electrical characteristics and suggest techniques to mitigate the impact of temperature differences, our work provides a perspective to leverage the temperature-dependent polarization switching behavior and their integration in advanced memory systems.

The major contributions of this paper are as follows:

- We fabricate a 10-nm-thick HZO thin film capacitor and measure its characteristics at different temperatures to show the temperature dependence of polarization switching.
- We propose a revised multi-domain Monte Carlo model to capture the temperature impact on ferroelectric switching behavior.
- We exploit the temperature rise as a resource that can reduce the FeRAM's write voltage and write energy and validate in circuit simulations.
- We perform a case study of the FeRAM-based hardware accelerator to demonstrate the rise in temperature on-chip for DNN workloads.

The rest of the paper is organized as follows. Section II details device fabrication and experimentation setup. Section III provides in-detail explanation of our proposed design followed by discussion on experimental results in section IV. Finally, we conclude our work in section V.

II. Device Fabrication and Experimental Details

The Metal-Ferroelectric-Metal (MFM) capacitor under measurements is fabricated on low-resistivity silicon substrate. The top and bottom metal electrodes are two 100-nm-thick tungsten (W) layers, and are sputtered by DC sputter under 300 W. The 10-nm-thick $Hf_{0.5}Zr_{0.5}O_2$ layer is deposited through atomic layer deposition (ALD) at temperatures of 250°C. Post-metallization annealing (PMA) is carried out in N_2 atmosphere at 500°C for 1 minute to facilitate the crystallization of the ferroelectric material. The cross-sectional transmission electron microscopy (TEM) image of the fabricated $Hf_{0.5}Zr_{0.5}O_2$ capacitor is shown in Fig. 1(a).

To measure the characteristics of the fabricated MFM, its top and bottom electrode are connected to two pulse measure units (PMU), which apply voltage pulses shown in Fig. 1(b) on the MFM and measure the current. The applied electrical sequence is composed of multiple symmetrical triangular pulses whose amplitude gradually increases from 1V to 3V with a step of 0.4V. The pulse width of one triangular pulse is 20 μs. Then, the polarization versus voltage (P-V) curves can be derived by integrating the current data collected by PMUs over time. By varying the wafer chuck temperature, the temperature-dependent ferroelectric switching behavior is studied. Fig. 1(c) demonstrates the P-V loops of all 3 temperatures (300K, 330K and 360K) implemented in this article. Fig. 2 compares the P-V curves under different temperature settings with 1.4V/3V applied voltage. When 1.4V is applied on the MFM, the remnant polarization (P_r) at 300K, 330K and 360K are 2.5, 3.7 and 5.8 μC/cm^2 respectively. However, when 3V is applied, P_r at 300K, 330K and 360K are all 21 μC/cm^2.

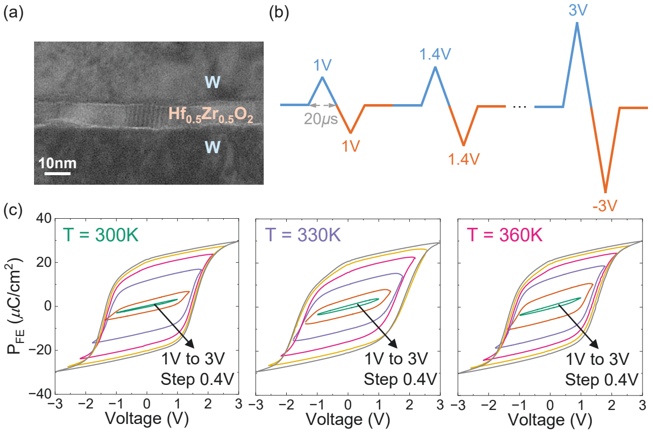

Fig. 1. (a) Cross-sectional TEM image of the fabricated $Hf_{0.5}Zr_{0.5}O_2$ capacitor. (b) The electrical sequence applied on the MFM. The test sequence are repeated 3 times under the ambient temperature at 300K, 330K and 360K. (c) The measured P-V curves under different temperature settings.

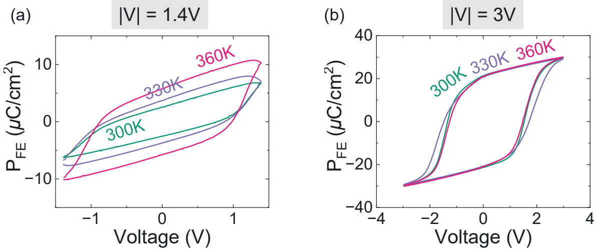

Fig. 2. The experimental P-V curves at 300K, 330K and 360K when the applied voltage amplitude is (a) 1.4V and (b) 3V.

III. Proposed Design

To capture the temperature impact on ferroelectric switching behavior, a revised model based on a reported multi-domain Monte Carlo framework [23], [24] is proposed. In that framework, the field-dependent nucleation-limited switching (NLS) model [25] is generalized for arbitrary input waveforms by calculating each domain's switching probability at each time step and simuating the switching event. Given a domain has not switched until time t, the probability of its switching time

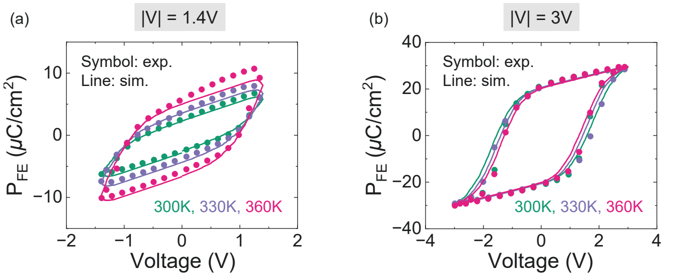

Fig. 3. The P-V curves obtained by the proposed model can be well calibrated with experimental data when applying (a) 1.4V and (b) 3V.

t_S in the time interval $[t, t+\Delta t]$ is:

$$P(t_S < t+\Delta t | t_S > t) = 1 - \exp\left[(h(t))^{\beta} - (h(t+\Delta t))^{\beta}\right] \quad (1)$$

where β is the shape parameter of the Weibull process, and $h(t)$ is an auxiliary history parameter which is defined as:

$$h(t) = \int_{t_0}^{t} \frac{dt'}{\tau} \quad (2)$$

where t_0 is the time when the switching voltage is applied, and τ is the switching time constant of one domain which describes the switching rate of that domain and is defined as:

$$\tau(E_a, E, T) = \tau_{\infty} \exp\left[\left(\frac{RT}{T}\right)^{c}\left(\frac{E_a}{E}\right)^{\alpha}\right] \quad (3)$$

where E_a is the activation field, E is the local field, T is the temperature, τ_{∞} is the time constant obtained for an infinite applied field, and RT is the room temperature (300K). c and α are empirical parameters. The Arrhenius equation is included in the calculation of τ to capture the thermal activation of polarization switching. Additionally, an empirical factor d is used to reflect the exponential decay of saturated polarization (P_s) as T increases [22] as following:

$$P_s'(T) = P_s \exp\left[-d(T - RT)\right] \quad (4)$$

IV. Results and Discussion

A. Parametric Study of the Proposed Model

By exploiting equations (3) and (4), the temperature dependence of both τ and P_s is reflected, two dominant factors in the ferroelectric switching behavior. The increase of P_r with temperature with 1.4 V applied voltage shown in Fig. 2(a) is due to the reduction of τ with temperature. When the applied electric field is small, τ decreases as the operating temperature increases, so more domain switching occurs. However, for large applied electric field, almost all reversible domains are switched (i.e., characteristics in the saturated P-V loop), so the impact of P_s is more significant. Given the P-V curves in Fig. 2(b) are identical, d in equation (4) is supposed to be a small value. The parametric extraction shows that when $c = 4.2, d = 0.001K^{-1}$, our model can be well calibrated with experimental data (Fig. 3). The parametric study of c in equation (3) (Fig. 4(a)(b)) and d in equation (4) (Fig. 4(c)(d)) are then conducted. It is implied by equation (3) that as c increases, the drop of τ is more relevant to temperature rise. The coercive field (E_c) of ferroelectrics is the applied field which enables the reversible dipole switching. Since it is observed that E_c decreases with c in both cases, τ can be used to reflect E_c. However, only for unsaturated P-V loops (Fig. 4(a)), smaller E_c results in higher P_r, which means that the increase of c will have a stronger effect on P_r under a relatively smaller applied electric field. Equation (4) exhibits the inverse correlation between d and P_s, as is verified in Fig. 4(c)(d), and smaller P_s causes the drop of P_r for both small and large applied voltage.

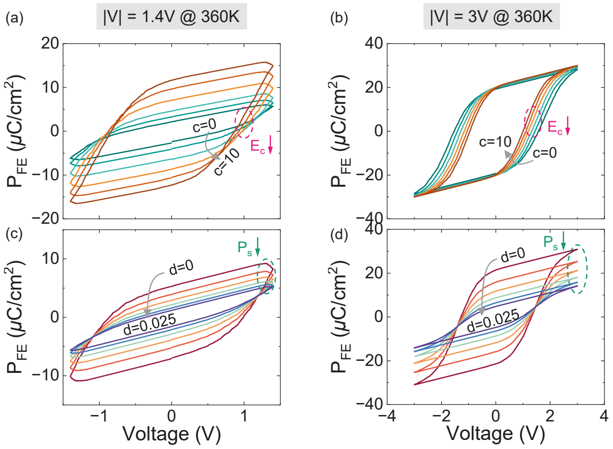

Fig. 4. The parametric study of (a)(b) c in equation (3) and (c)(d) d in equation (4) at 360K. c changes from 0 to 10 with the step size of 2, and d changes from 0 to 0.025 K^{-1} with the step size of 0.005 K^{-1}. The P-V curves when applying (a)(c) ±1.4V and (b)(d)±3V.

B. Circuit Simulation

The circuit-level simulation is implemented by Cadence Spectre Simulator. The MFM capacitor is connected to an access transistor and is programmed to positive P_r (i.e., data '1') by applying a positive electric field (Fig. 5(a)). On the contrary, when the applied electric field is negative (Fig. 5(b)), the MFM is written to negative P_r (i.e., data '0'). Fig. 5(c) shows the simulation waveform applied on word line (WL), bit line (BL) and plate line (PL). It is indicated by the P_{FE} waveform in Fig. 5(d) that the charge memory window (MW) at 300K/330K/360K is 4.7/7.2/10.6 μC/cm^2 when applying ±1.4V. The observed MW rise with temperature means that for the conventional FeRAM write operation, the write voltage (V$_{\text{write}}$) can be reduced and the MW is still above a certain threshold for correct sensing. It is implied by Fig. 6(a) that if the desired MW is about 4.7 μC/cm^2, V$_{\text{write}}$ is supposed to be 1.4V, 1.276V and 1.166V at 300K, 330K and 360K respectively. Therefore, compared to 300K, the V$_{\text{write}}$ reduction at 330K/360K can reach 8.9%/16.7%. Thanks to the V$_{\text{write}}$ reduction, the write energy (E$_{\text{write}}$) can be lowered from 2.57 fJ/bit at 300K to 2.24/1.74 fJ/bit at 330K/360K, which means that the E$_{\text{write}}$ reduction can be 13%/32.5% at 330K/360K (Fig. 6(b)).

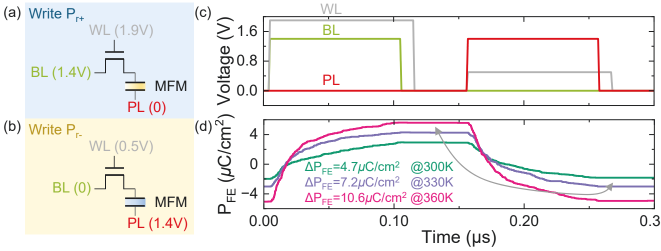

Fig. 5. The schematics and applied voltages of circuit simulation for write (a) positive P_r (i.e., data '1') and (b) negative P_r (i.e., data '0'). (c) The waveform of applied voltages for 1.4 V programming. (d) The simulated P_{FE} waveform at 300K, 330K and 360K.

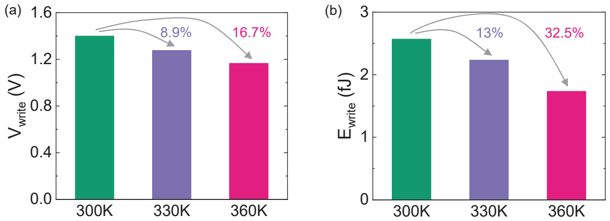

Fig. 6. Compared to 300K, (a) the V_{write} reduction at 330K/360K can reach 8.9%/16.7%, and (b) the E_{write} reduction can reach 330K/360K is 13%/32.5% with similar MW.

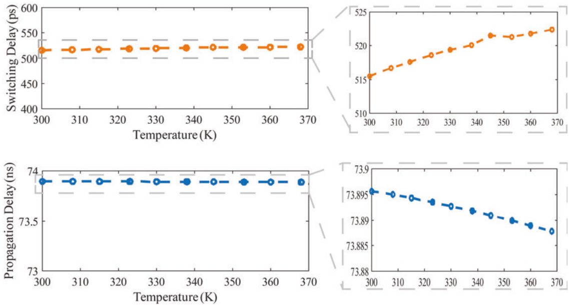

Fig. 7. Effect of temperature on switching delay and propagation delay.

C. Case Study

A study has been developed to gain insight into the impact of temperature fluctuations in a hardware accelerator with ferroelectric memory. A deep neural network (DNN) is employed for this purpose as 3D memory technologies provide significant advantages for these data-centric workloads. A Yolo DNN model [26] is selected as the workload for machine learning. This benchmark offers a rigorous evaluation of DNN efficiency, replicating fundamental neural network operations. The structural attributes of the accelerator are modified from Google Tensor Processing Unit [4]. It comprises a processing unit, buffer memory, and peripheral components, which include an input-output unit, memory management unit, and control unit. A 256 by 256 systolic array matrix-multiply unit within the DNN accelerator functions as the processing core. Three 64 MiB FeRAM buffers are employed for holding input, filter weights and activations, and accumulators.

The DNN model must be mapped to a systolic array to assess the thermal characteristics of a DNN accelerator running a specific workload. Assuming an output-stationary data flow i.e. each processing element computes a pixel of the output feature map by accumulating inputs and weights provided each cycle, we use SCALE-Sim [27] tool to get cycle-accurate simulation for systolic array-based CNNs. It provides insights into compute cycles, systolic array utilization, and memory accesses per layer, which is crucial for estimating power traces for thermal analysis.

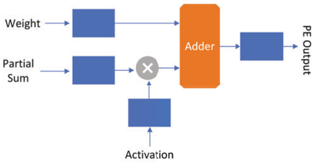

Fig. 8. Block level diagram of processing element [14].

Heat sink material	Aluminium 6061
Heat sink thermal conductivity	167 W/(m-K)
Heat sink specific heat capacity	8.9×10^5 J/(m^{3}K)
Heat sink thickness	6.9 mm
Heat sink side	25 mm
Heat spreader thickness	3 mm
Ambient temperature	300K

TABLE I
CONFIGURATION FOR THERMAL SIMULATION.

The CMOS-based processing element Fig. 8 from [14] is used to derive power traces for the processing core. The average dynamic and static power consumption per processing element is approximately 370μW and 13μW, respectively, at a supply voltage of 0.9V and a clock frequency of 700MHz. These values are utilized to calculate the total energy dissipated during the execution of each DNN layer, which, divided by the total number of cycles, gives us a power trace for the processing element. To get the power trace for memory modules, we multiply the read and write energy per access per bit by the number of bytes read and written obtained from SCALE-Sim, respectively.

We use hotspot [28] for thermal analysis. The 2D floor plan, resembling a conventional 2D technology similar to TPU [4], is depicted in Fig. 9 (a). For the 3D stack, apart from the buffer memories, we have a 4GB on-chip main memory similar to NVDRAM [19] with a chip density of 0.42Gb/mm^2. The floor plan for the monolithic 3D stack with stack 0 close to the heat sink is shown in Fig. 9 (b). Tab. I shows the cross-sectional information of the chip, ambient temperature and heat sink parameters for the simulation.

Fig. 9(c) illustrates the steady-state temperatures in Kelvin, of our 2D floor plan. The processing unit and the FeRAM buffers have steady-state temperatures of 314.7K and 314.1K, respectively. The on-chip memory access energy, representing the energy consumed during data transfer between the buffer and systolic array for a single inference cycle, is 65.47mJ.

Transitioning from a 2D to a 3D floor plan introduces significant benefits in terms of latency and throughput for hardware accelerators. The 3D architecture integrates on-chip main memory more effectively, considerably reducing data transfer distances compared to traditional 2D layouts. The steady-state temperatures in Kelvin of our 3D-floor plan are shown in Fig. 9 (d). The processing unit and the buffer memory on the bottom stack close to the heat sink reach a steady state temperature of 352.18K and 350K, respectively. The FeRAM buffers on stack one and the on-chip main memory operate at

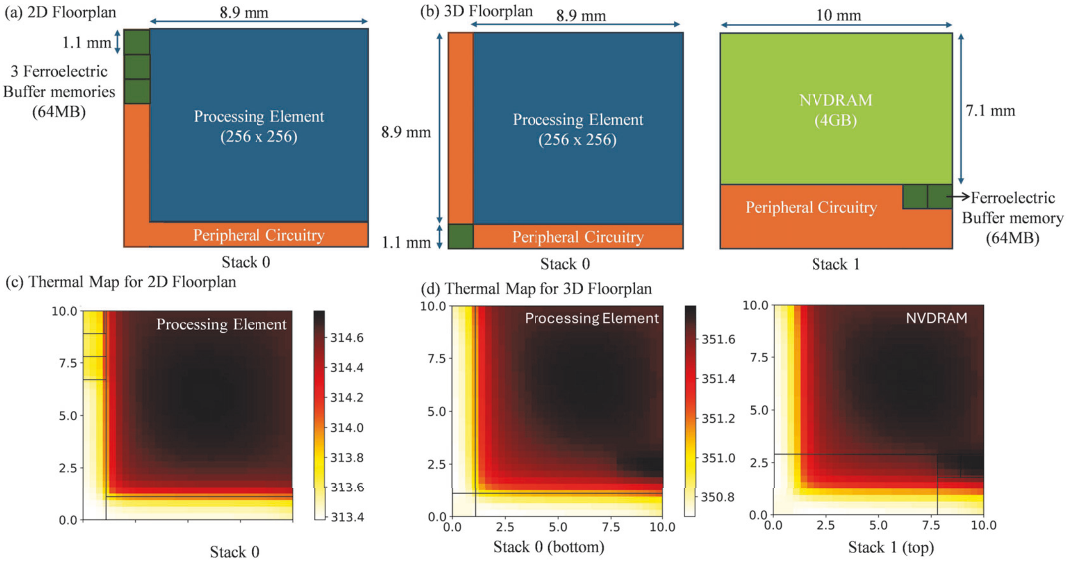

Fig. 9. (a) Floor plan of 2D Stack: $100mm^2$ 2D floor plan with $1.21mm^2$ ferroelectric buffer memories with the proposed design. (b) Floor plan of 3D Stack: In addition to the buffer memory, we also have an on-chip main memory. It is similar to nvdram [19] with the proposed model (c) Steady state thermal map of 2D and (d) 3D Stack with maximum temperature 314.78 and 351.77 respectively.

351K steady-state temperature.

These temperatures are well below the rated operating condition of NVDRAM at 368.15K, where the bit error ratio (BER) is 3×10^{-11} [19]. We also investigate the effect of temperature variations on switching and propagation delays by the circuit-level simulation implemented in Cadence Spectre Simulator. The readout circuitry for measuring propagation delay is provided in the supplementary material. Our simulation show that the propagation delay remains largely unaffected, while the switching delay only increases by about 1.5% (refer Fig. 7).

Thus, at 351K, we can reduce the access energy per bit from 2.57 fJ to 1.86 fJ. As a result, the memory access energy for one deep neural network inference cycle on our 3D hardware accelerator reduces from 92.04 mJ to 66.63 mJ, i.e., 27.56 % less memory access energy.

The HotSpot tool [28], while useful for thermal estimations, may not achieve the precision of direct measurements obtained through experimental methods using thermal sensors. This inaccuracy is attributed to its inability to account for the limited number of thermal sensors available on the chip. To address the discrepancies in experimental and estimated temperature profiles, methods described by Zhang et al. can be adopted [29]. Depending on the grid size, placement and density of on-chip thermal sensors, and workload characteristics, the error in the on-chip thermal profile ranges from 0.01K to 1.05K. Even when considering a temperature discrepancy of 1.05K for conservative estimations, we still observe a reduction in memory access energy by 26.83%.

In the hardware accelerator architecture being examined in the use case, the compute unit inherently induces a rise in temperature due to its operational demands. This temperature increase is a characteristic challenge independent of memory technology. As we observe, adopting FeRAMs facilitated with the proposed thermal-dependent ferroelectric switching as memory modules within these accelerators offers a promising approach to effectively leveraging the elevated thermal conditions to reduce the energy consumption associated with memory accesses. Additionally, transitioning from 2D to 3D memory can significantly reduce the energy required for on-chip memory transfers, even when employing an on-chip main memory. Furthermore, the presence of on-chip main memory also reduces inference latency. Furthermore, the integration of higher-density ferroelectric memories can reduce the physical footprint of the 3D hardware accelerator. However, in this paper, we focus exclusively on demonstrating performance gains in terms of access energy and plan to address area optimization in future work.

V. Conclusion

Our study has systematically analyzed the temperature-dependent switching behavior of HfO_2-based FeRAM within a 3D memory architecture, emphasizing its application in neural network hardware accelerators. Experiments and simulations reveal that by dynamically adjusting write voltages in response to temperature variations, we can significantly reduce the access energy per bit while maintaining the reliability and performance of these memory systems under elevated temperatures. This work contributes to the theoretical understanding of ferroelectric behavior under varying thermal conditions and provides practical insights for developing more efficient and thermally resilient memory solutions in next-generation hardware accelerators. Our findings underscore the importance of incorporating temperature-aware strategies in designing memory architectures for advanced computing systems, which will be critical in maintaining system robustness and energy efficiency with ever-increasing computational loads. In the future, we would like to study the thermal profile of these memory systems and leverage temperature-dependent switching behavior at more granularity, such as a bank, channel, or sub-array level.

Acknowledgment

Research on ferroelectric switching dynamics and associated thermal simulations is supported by the center for 3D Ferroelectric Microelectronics (3DFeM), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Energy Frontier Research Centers program under Award Number DE-SC0021118. Architectural and neural network effort is supported by the National Science Foundation under the awards 1955815, A Technology-Architecture-Algorithm Co-Design Exploration of Scalable Spiking Neural Networks (SNNs), Award 2346953, High performance ferroelectric memory for in-memory computing and Award 2008365, Leveraging Monolithic 3D for Architectural Innovations.

References

[1] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From words to watts: Benchmarking the energy costs of large language model inference," *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–9, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263620702

[2] S. Rokicki, E. Rohou, and S. Derrien, "Hardware-accelerated dynamic binary translation," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. IEEE, 2017, pp. 1062–1067.

[3] A. Devulapally, M. F. F. Khan, S. Advani, and V. Narayanan, "Multi-modal fusion of event and rgb for monocular depth estimation using a unified transformer-based architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 2081–2089.

[4] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[5] J. Wang, M. Ge, B. Ding, Q. Xu, S. Chen, and Y. Kang, "Nicepim: Design space exploration for processing-in-memory dnn accelerators with 3-d stacked-dram," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 5, pp. 1456–1469, 2024.

[6] S. Singh, A. Sarma, N. Jao, A. Pattnaik, S. Lu, K. Yang, A. Sengupta, V. Narayanan, and C. R. Das, "Nebula: A neuromorphic spin-based ultra-low power architecture for snns and anns," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture*, 2020, pp. 363–376.

[7] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "Tetris: Scalable and efficient neural network acceleration with 3d memory," in *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, 2017, pp. 751–764.

[8] H. Wen and W. Zhang, "Exploiting gpu with 3d stacked memory to boost performance for data-intensive applications," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*, 2018, pp. 1–6.

[9] B. Akin, J. C. Hoe, and F. Franchetti, "Hamlet: Hardware accelerated memory layout transform within 3d-stacked dram," in *2014 IEEE High Performance Extreme Computing Conference (HPEC)*, 2014, pp. 1–6.

[10] P. Shukla, V. F. Pavlidis, E. Salman, and A. K. Coskun, "Tread-m3d: Temperature-aware dnn accelerators for monolithic 3-d mobile systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 12, pp. 4350–4363, 2023.

[11] T. D. Richardson and Y. Xie, "Evaluation of thermal-aware design techniques for microprocessors." [Online]. Available: https://ieeexplore.ieee.org/document/1611250/

[12] E. Rotem, R. Ginosar, A. Mendelson, and U. Weiser, "Power and thermal constraints of modern system-on-a-chip computer," 09 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026269215002086

[13] X. Liu, M. Zhou, T. S. Rosing, and J. Zhao, "Hr 3 am: A heat resilient design for rram-based neuromorphic computing," in *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2019, pp. 1–6.

[14] A. Abdurrob, E. Salman, and J. Lombardi, "Thermal integrity of reram-based near-memory computing in 3d integrated dnn accelerators," in *2023 IEEE 36th International System-on-Chip Conference (SOCC)*, 2023, pp. 1–6.

[15] L. Yang, R. M. Radway, Y.-H. Chen, T. F. Wu, H. Liu, E. Ansari, V. Chandra, S. Mitra, and E. Beigné, "Three-dimensional stacked neural network accelerator architectures for ar/vr applications," *IEEE Micro*, vol. 42, no. 6, pp. 116–124, 2022.

[16] G. Zervakis, I. Anagnostopoulos, S. Salamin, O. Spantidi, I. Roman-Ballesteros, J. Henkel, and H. Amrouch, "Thermal-aware design for approximate dnn accelerators," *IEEE Transactions on Computers*, vol. 71, no. 10, pp. 2687–2697, 2022.

[17] T. Francois, L. Grenouillet, J. Coignus, P. Blaise, C. Carabasse, N. Vaxelaire, T. Magis, F. Aussenac, V. Loup, C. Pellissier *et al.*, "Demonstration of beol-compatible ferroelectric hf 0.5 zr 0.5 o 2 scaled feram co-integrated with 130nm cmos for embedded nvm applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2019, pp. 15–7.

[18] J. Okuno, T. Kunihiro, K. Konishi, H. Maemura, Y. Shuto, F. Sugaya, M. Materano, T. Ali, K. Kuehnel, K. Seidel *et al.*, "Soc compatible 1t1c feram memory array based on ferroelectric hf0. 5zr0. 5o2," in *2020 IEEE Symposium on VLSI Technology*. IEEE, 2020, pp. 1–2.

[19] N. Ramaswamy, A. Calderoni, J. Zahurak, G. Servalli, A. Chavan, S. Chhajed, M. Balakrishnan, M. Fischer, M. Hollander, D. P. Ettisserry, A. Liao, K. Karda, M. Jerry, M. Mariani, A. Visconti, B. R. Cook, B. D. Cook, D. Mills, A. Torsi, C. Mouli, E. Byers, M. Helm, S. Pawlowski, S. Shiratake, and N. Chandrasekaran, "Nvdram: A 32gb dual layer 3d stacked non-volatile ferroelectric memory with near-dram performance for demanding ai workloads," in *2023 International Electron Devices Meeting (IEDM)*, 2023, pp. 1–4.

[20] H. Chen, L. Tang, L. Liu, Y. Chen, H. Luo, X. Yuan, and D. Zhang, "Temperature dependent polarization-switching behavior in $\text{hf}_{0.5}\text{zr}_{0.5}\text{o}_2$ ferroelectric film," *Materialia*, vol. 14, p. 100919, 2020.

[21] J. Hur, Y.-C. Luo, Z. Wang, S. Lombardo, A. I. Khan, and S. Yu, "Characterizing ferroelectric properties of hf0.5zr0.5o2 from deep-cryogenic temperature (4 k) to 400 k," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 7, no. 2, pp. 168–174, 2021.

[22] T. Ali, K. Kühnel, M. Czernohorsky, C. Mart, M. Rudolph, B. Pätzold, M. Lederer, R. Olivo, D. Lehninger, F. Müller *et al.*, "A study on the temperature-dependent operation of fluorite-structure-based ferroelectric hfo 2 memory fefet: A temperature-modulated operation," *IEEE Transactions on Electron Devices*, vol. 67, no. 7, pp. 2793–2799, 2020.

[23] C. Alessandri, P. Pandey, A. Abusleme, and A. Seabaugh, "Monte carlo simulation of switching dynamics in polycrystalline ferroelectric capacitors," *IEEE Transactions on Electron Devices*, vol. 66, no. 8, pp. 3527–3534, 2019.

[24] S. Deng, G. Yin, W. Chakraborty, S. Dutta, S. Datta, X. Li, and K. Ni, "A comprehensive model for ferroelectric fet capturing the key behaviors: Scalability, variation, stochasticity, and accumulation," in *2020 IEEE symposium on VLSI technology*. IEEE, 2020, pp. 1–2.

[25] A. K. Tagantsev, I. Stolichnov, N. Setter, J. S. Cross, and M. Tsukada, "Non-kolmogorov-avrami switching kinetics in ferroelectric thin films," *Physical Review B*, vol. 66, no. 21, p. 214109, 2002.

[26] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[27] A. Samajdar, J. M. Joseph, Y. Zhu, P. Whatmough, M. Mattina, and T. Krishna, "A systematic methodology for characterizing scalability of dnn accelerators using scale-sim," in *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2020, pp. 58–68.

[28] J.-H. Han, R. E. West, K. Skadron, and M. R. Stan, "Thermal simulation of processing-in-memory devices using hotspot 7.0," in *2021 27th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC)*. IEEE, 2021, pp. 1–5.

[29] Y. Zhang, A. Srivastava, and M. Zahran, "On-chip sensor-driven efficient thermal profile estimation algorithms," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 15, no. 3, jun 2010. [Online]. Available: https://doi.org/10.1145/1754405.1754410