Efficient Large Scale Neural Network Acceleration With 3-D FeNOR-Based Computing-in-Memory Design

Yang Feng, *Graduate Student Member, IEEE*, Dong Zhang, Chen Sun, Zijie Zheng, Yue Chen, *Graduate Student Member, IEEE*, Qiwen Kong, *Graduate Student Member, IEEE*, Gan Liu, *Graduate Student Member, IEEE*, Xiaolin Wang, Yuye Kang, Kaizhen Han, Zuopu Zhou, Leming Jiao, *Graduate Student Member, IEEE*, Jixuan Wu, *Member, IEEE*, Jiezhi Chen, *Senior Member, IEEE*, and Xiao Gong

***Abstract*— In this work, we introduce and experimentally demonstrate a 3-D stacked ferroelectric NOR (FeNOR) memory, featuring a back-end-of-line (BEOL) zinc oxide (ZnO) channel, and a metal–ferroelectric–metal–insulator5 semiconductor (MFMIS) unit cell. The main contributions of this work are as follows: 1) enhanced memory window (MW) and high ON/OFF ratio: The MFMIS architecture in 3-D FeNOR enables a tunable and large MW (∼4 V), as well as an ON/OFF ratio (I_{on}/I_{off}) of six orders of magnitude; 2) low operation voltage and high endurance: The integration of ferroelectric materials allows for low operation voltages (∼4 V) and excellent endurance (10^7 cycles); 3) efficient neural network implementation: Leveraging the 3-D FeNOR structure, we further develop VGG-16 and ResNet-50 convolutional neural networks that achieve high prediction accuracy, decent area efficiency, and low power consumption. The emergence of 3-D FeNOR technology positions ferroelectric devices as a highly promising candidate for computing-in-memory (CIM) applications.**

***Index Terms*— 3-D structure, computing-in-memory (CIM), ferroelectric field-effect transistors (FeFETs), $Hf_xZr_{1-x}O_2$ (HZO), metal–ferroelectric–metal–insulator5 semiconductor (MFMIS), oxide semiconductor.**

Received 2 January 2025; revised 15 February 2025 and 14 March 2025; accepted 19 March 2025. Date of publication 8 April 2025; date of current version 7 May 2025. This work was supported in part by the Ministry of Education (Singapore) Tier 2 Academic Research under Grant MOE-T2EP50221-0008; in part by NSF 2346953; in part by China Key Research and Development Program under Grant 2023YFB4402500; and Grant 2023YFB4402400, and in part by the National Natural Science Foundation of China under Grant 62034006, Grant U23B2040, and Grant 92264201. The review of this article was arranged by Editor P.-Y. Du. *(Yang Feng and Dong Zhang contributed equally to this work.) (Corresponding authors: Jiezhi Chen; Xiao Gong.)*

Yang Feng is with the School of Information Science and Engineering, Shandong University, Qingdao 266100, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576.

Dong Zhang, Chen Sun, Zijie Zheng, Yue Chen, Qiwen Kong, Gan Liu, Xiaolin Wang, Yuye Kang, Kaizhen Han, Zuopu Zhou, Leming Jiao, and Xiao Gong are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 (e-mail: elegong@nus.edu.sg).

Jixuan Wu and Jiezhi Chen are with the School of Information Science and Engineering, Shandong University, Qingdao 266100, China (e-mail: chen.jiezhi@sdu.edu.cn).

Digital Object Identifier 10.1109/TED.2025.3554164

I. Introduction

COMPUTING-IN-MEMORY (CIM) has attracted considerable attention for its capability to support data- and computation-intensive workloads in artificial intelligence applications [1], [2], [3], [4], [5], [6], offering a promising solution to overcome the computing power bottleneck. Over the last decade, such CIM architectures have been constructed by various non-volatile memory technologies, including flash memory, resistive random access memory (RRAM), phase change memory (PCM), ferroelectric field-effect transistors (FeFETs), and so on. Among these, hafnium oxide (HfO_2) based FeFET stand out as a highly promising candidate, owing to several crucial advantages, such as low operation voltage, good endurance, fast operation speed, and compatibility with complementary metal–oxide–semiconductor (CMOS) technology [7], [8], [9], [10]. Fig. 1(a) and (b) compares FeFET with conventional NOR Flash memory, highlighting the device-level competitive advantages. Moreover, FeFET can be implemented in the 3-D architecture, offering substantial data storage capacity and enabling highly parallel computing capabilities. Leveraging on this compatibility, recently, 3-D ferroelectric NAND (FeNAND) array structures utilizing poly-Si or advanced oxide semiconductor channels have been demonstrated [11], [12]. These works show excellent performance and high-density integration, underscoring the potential of 3-D integrated ferroelectric memories in enabling CIM technologies.

However, despite considerable progress in 3-D ferroelectric memory, two fundamental challenges still persist, hindering its further advancements. The first challenge is the limited memory window (MW) in 3-D FeFET, resulting from weak ferroelectric switching. In prior 2-D FeFET studies, this problem could be solved by employing a metal-ferroelectric-metal-insulator-semiconductor (MFMIS) structure [13], [14], [15]. However, the complex 2-D MFMIS unit cell structure is not applicable in conventional 3-D memory designs. This necessitates the development of novel MFMIS unit cell structure in 3-D ferroelectric NOR (FeNOR) design to fully

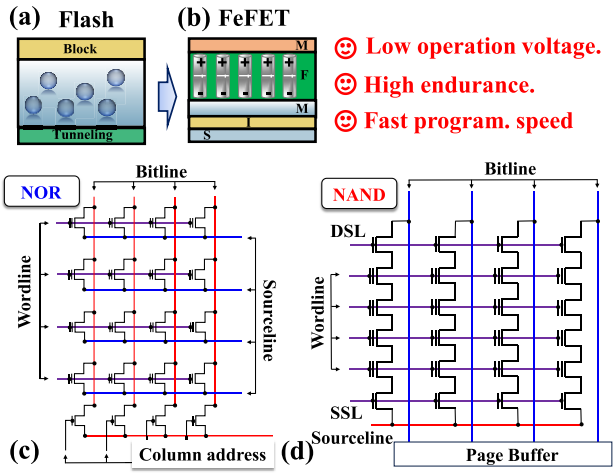


Fig. 1. (a) and (b) 3-D FeNOR offers lower power consumption, better reliability, and faster programming speed than 3-D NOR Flash. (c) and (d) Schematic illustartion of the NAND and NOR array. In comparison to NAND arrays, NOR arrays enable stronger parallel computing capabilities.

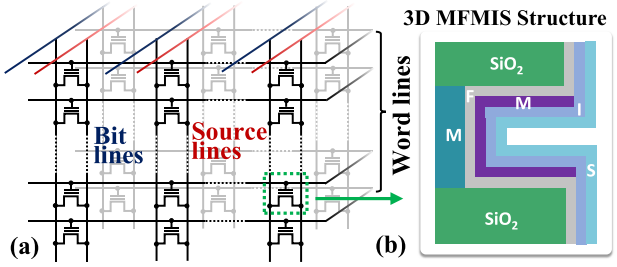


Fig. 2. (a) Schematic illustration of the 3-D FeNOR array, showcasing the array architecture and peripheral connections. (b) 3-D MFMIS unit cell proposed in this work, with a meticulously designed area ratio for MW enhancing.

unlock FeFET's potential. The second challenge is the lack of exploration into 3-D FeNOR architectures, although the development of the 3-D FeNOR is necessary and significant to CIM advancement. Compared to 3-D FeNAND, 3-D FeNOR is better suited for CIM applications due to its capability for parallel computing, which is enabled by utilizing multiple source lines as multiple outputs, as shown in Fig. 1(c) and (d) [16], [17]. Therefore, advancing 3-D FeNOR technology is a crucial step to boost FeFET's competitiveness as a leading non-volatile memory for CIM applications.

To address these challenges, in this work, we propose a 3-D MFMIS structure for MW enhancement in 3-D FeNOR. Furthermore, building on this design, we further experimentally demonstrated the 3-D FeNOR and evaluate its feasibility for CIM applications. During the experiment process, the highlights are: 1) the key layers, including the channel, dielectric, and other critical components, are deposited using atomic layer deposition (ALD), ensuring precise thickness control and exceptional uniformity and 2) all processes are conducted at temperatures $\leq$450 °C, ensuring back-end-of-line (BEOL) compatibility. The experimental results of our fabricated 3-D FeNOR showcase the fast and energy-efficient program operation, a large MW, outstanding ON/OFF ratio, and good endurance.

Initial results of this work are previously reported in [18], and in this version, we provide a more detailed and comprehensive analysis: we assess the 3-D FeNOR-based CIM systems through an in-depth analysis of key metrics, including noise-related accuracy, area efficiency, and power consumption. In summary, this work paves the way for the development of 3-D FeFET memory architectures capable of meeting the demands of next-generation high performance CIM applications.

II. Device Structure and Fabrication

A. FeNOR With 3-D MFMIS Structure

The schematic of the 3-D FeNOR array in this work is illustrated in Fig. 2(a). In this plot, each unit cell is connected to both the source line and bitline, enabling random access capability of the memory array. A detailed view of the unit cell in this NOR-type array is provided in Fig. 2(b), in which the 3-D side-fin MFMIS configuration is clearly illustrated. Moreover, the corresponding physical model of the 3-D FeNOR array incorporating the 3-D MFMIS structure is provided in Fig. 3(a). As depicted, the bitline and source line are oriented vertically, while the word line is planar. A more detailed description of this model can be found in [19]. Additionally, the X–Z view of 3-D FeNOR is presented in Fig. 3(b). Fig. 3(c) shows the planar view of 3-D FeNOR along the X- and Y-axis, clearly illustrating the channel length definition. Here, it should be pointed out our as-grown ZnO channel exhibits a positive V_{th}. Fig. 4(a) displays the I_D–V_G curve of the planar ZnO FeFET, indicating this positive V_{th}. The small MW here is attributed to the weak-erase issue. This issue is effectively addressed in our 3-D FeNOR by leveraging the 3-D MFMIS structure. In addition, the intrinsic positive V_{th} of ZnO channel is also reflected in the following I_D–V_G loop of the 3-D FeNOR, where the amplitude of the positive V_{th} is significantly higher than that of the negative V_{th}, which will be discussed in detail later. Fig. 4(b) illustrates the schematic of the FeNOR unit cell, where the channel is non-conductive in regions without gate control. This non-conductivity ensures electrical isolation between individual cells within the 3-D FeNOR array, effectively eliminating leakage and interference. Prior to the subsequent discussion, it is necessary to elucidate the mechanism of MW boosting within the MFMIS structure. In this MFMIS structure [see Fig. 2(b)], the area of the gate-controlled ferroelectric layer (denoted as A_{FE}) over the area of the floating-gate-controlled channel layer (denoted as A_{Mos}) is defined as the area ratio (A_{FE}/A_{Mos}). This area ratio can solve the longstanding weak-erase issue by adjusting the distribution of the applied gate voltage across the ferroelectric layer [20]. More specifically, a smaller area ratio—indicating a relatively reduced ferroelectric capacitance area—results in a larger fraction of the applied voltage being distributed across the ferroelectric layer. This increased voltage drop enhances the electric field within the ferroelectric layer, thereby promoting more effective ferroelectric switching. Consequently, this improved switching behavior leads to an expansion of the MW in FeFET devices. This is why a decent MW is obtained in this work.

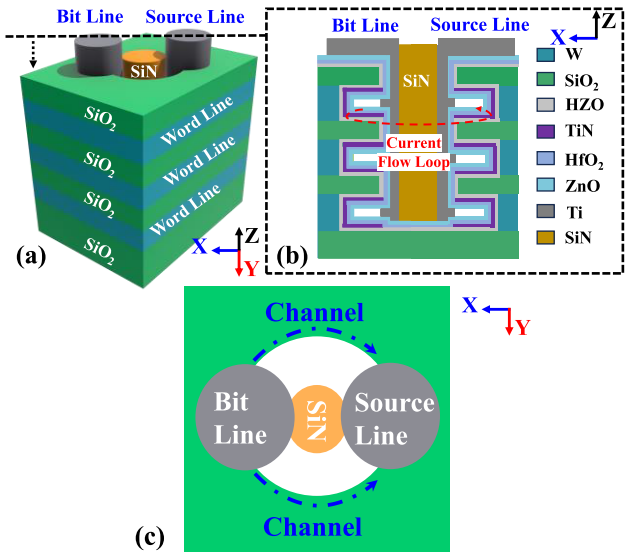


Fig. 3. (a) Schematic of the 3-D FeNOR array featuring vertical source and drain electrodes, along with planar gate electrodes. (b) X–Z view of the 3-D FeNOR array. (c) X–Y view of 3-D FeNOR.

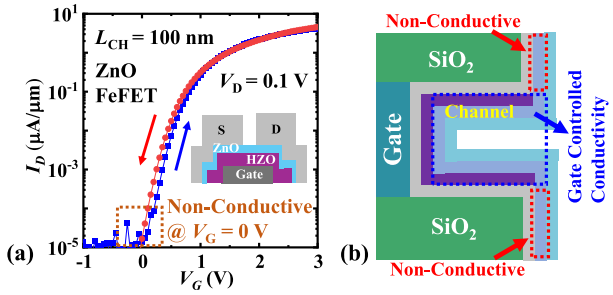


Fig. 4. (a) I_D–V_G curve of the planar ZnO FeFET, demonstrating a positive V_{th}. (b) Schematic of the FeNOR unit cell, where the positive V_{th} effectively suppresses leakage paths between adjacent cells.

B. Device Fabrication

Fig. 5(a) illustrates the key process steps for fabricating the 3-D FeNOR with the 3-D MFMIS unit cell. Fig. 5(b)–(g) shows the fabrication schematic for better understanding. The process in this work begins with the deposition of a 200 nm silicon dioxide (SiO_2) layer on the Si substrate, using plasma-enhanced chemical vapor deposition (PECVD). After this, a 100 nm-thick tungsten (W) layer is deposited by sputtering, followed by a 50 nm-thick SiO_2 layer via PECVD. This process is repeated for three cycles to construct the superlattice structure. Next, the gates and channel regions are dry-etched to expose them. Afterward, the sample is immersed in the tungsten etchant to form the side-fin structure. Subsequently, a 7 nm HZO layer is deposited using ALD with a 1:1 cycle ratio of Hf to Zr at 280 °C, followed by the deposition of a 7 nm TiN layer at 350 °C. After this, a post-metal annealing (PMA) process is carried out at the 450 °C to crystallize the HZO layer. Next, dry etching is performed to etch the side-wall TiN, forming three separate cells. Subsequently, a 10 nm HfO_2 layer and a 15 nm ZnO channel layer are deposited by ALD without breaking the vacuum. Following this, 100 nm Ti is deposited and lifted off to form the source/drain metal.

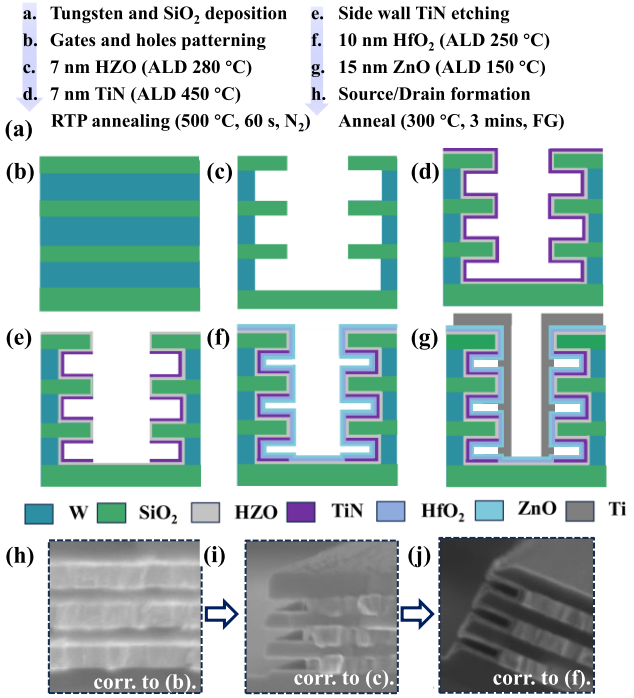


Fig. 5. (a) Key process steps. (b)–(g) Shcematic illustartion of fabricating the 3-D FeNOR featuring a 3-D MFMIS structure and oxide semiconductor channel with the detailed fabrication process flow. There are several key process: 1) HZO, TiN, HfO_2, and ZnO are grown with good conformality using ALD; 2) the holes of side-fin are formed with wet etching; and 3) the TiN layer is etched using dry etching. (h)–(j) SEM images after critical steps, showing a more direct and visual perspective of the fabrication processes.

Here, the sputtered thick source and drain layer are crucial for forming continuous 3-D contact with the channel layer. The channel length is defined as 1 μm. Finally, the ZnO channel is activated in the forming gas ambient at 300 °C for 3 min. During the abovementioned fabrication process, the maximum processing temperature is 450 °C, ensuring BEOL compatibility.

Moreover, to provide a direct and visual perspective, the scanning electron microscope (SEM) images of key fabrication steps are provided and shown in Fig. 5(h)–(j), including: 1) the stacked layers of SiO_2 and W; 2) the side-fin structure after vertical dry etching and lateral wet etching; and 3) the fabricated FeNOR array featuring the MFMIS structure. Furthermore, the zoomed-in view transmission electron microscope (TEM) images of the channel region in 3-D FeNOR are depicted in Fig. 6(a) and (b), confirming the thickness of each layer. From this, the area ratio can be determined: the hole has a width of 200 nm and a height of 100 nm, therefore the area ratio can be calculated as 1:5. Moreover, the energy dispersive X-ray spectroscopy (EDX) mapping of main elements (W, Zn, Hf, Zr, Si, Ti, and N) is illustrated in Fig. 6(c)–(i), clearly showing the uniform element distribution and MFMIS structure.

III. Characterizations of Proposed 3-D FeNOR

A. FeNOR Characterization

The transfer characteristics (I_D–V_G) of the fabricated 3-D FeNOR under varying gate voltages are presented in

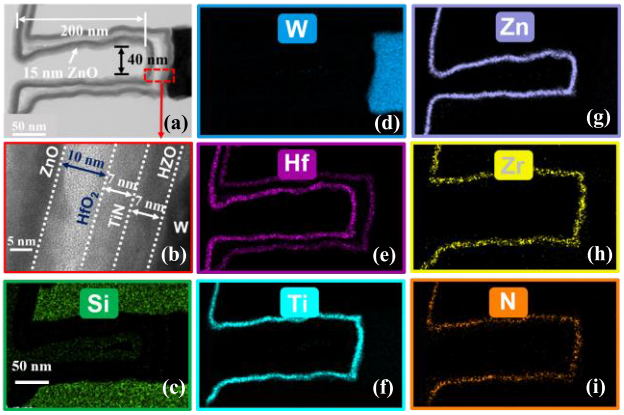


Fig. 6. (a) and (b) Zoomed-in view TEM image of the unit cell in the 3-D FeNOR array. (c)–(i) Corresponding EDX mapping, highlighting the distribution of the main elements and clearly showcasing the 3-D side-fin MFMIS structure.

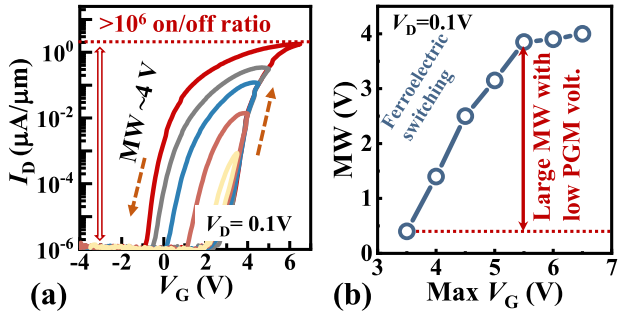


Fig. 7. I_D–V_G loops of the 3-D FeNOR unit cell by applying various V_G, showcasing over 10^6 ON/OFF ratio and MW (~4 V). (b) MW initially shows a significant increase as V_G increases, before eventually reaching a saturation point.

Fig. 7(a). Here, the V_D is fixed at 0.1 V. As depicted, the counterclockwise hysteresis induced by ferroelectric switching is obtained [21], [22]. In addition, it is observed that the MW expands with an increase in the maximum applied gate voltage. A more detailed explanation of this relationship is provided in Fig. 7(b). As plotted, initially, the MW grows as the voltage increases due to the enhanced ferroelectric switching at higher gate voltages. However, as the voltage continues to rise, the MW gradually saturates. Here, it is important to point out that this large MW is also influenced by the charge injection effect [23], [24], [25], particularly at high gate voltages when V_G exceeds 4.5 V. In contrast, in the low V_G region, where the charge injection effect is weak, the V_{th} changing is primarily attributed to ferroelectric switching. Finally, at a gate voltage of 6 V, the MW reaches 4 V, with an ON/OFF current ratio exceeding 10^6.

This large MW enables us to optimize memory performance based on specific application requirements. For instance, in online training tasks, energy efficiency is a significant criterion due to frequent data search and programming operations. In such scenarios, a narrow MW with low-power operation could enhance energy efficiency. Conversely, in high-precision CIM computing applications, such as solving partial differential equations, the primary concern is ensuring multi-state reliability. In this case, a wider MW can improve sensing margins across multiple states. In this study, we focus more on neural network training process. Therefore, an MW of ~2.5 V is selected and further analyzed in the following discussion.

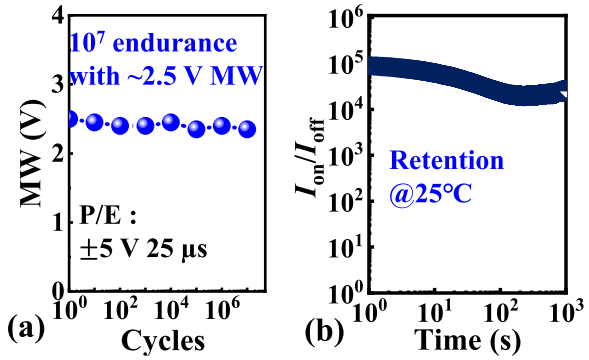


Fig. 8. (a) Under pulse amplitude of 5 V, the FeNOR demonstrates an good endurance. (b) Retention test of the 3-D FeNOR at room temperature indicates the non-volatile memory characteristics.

The reliability characteristics of 3-D FeNOR are investigated. First, endurance test is conducted at room temperature. During the endurance test, repeated pulses are applied to induce electrical stress, followed by the measurement of a single I_D–V_G loop for 3-D FeNOR. In this measurement process, the maximum gate voltage amplitude applied is 4.5 V, corresponding to a MW of 2.5 V. The results are shown in Fig. 8(a). Here, the 3-D FeNOR initially exhibits a MW of 2.5 V, and after 10^7 cycles, the MW remains stable. In further endurance testing, hard breakdown occurs before any evidence of ferroelectric fatigue. This demonstrates 3-D FeNOR's excellent endurance and stability across electric field cycles. Besides, the retention characteristics are also evaluated at room temperature, and the result is depicted in Fig. 8(b). As shown, after applying a program pulse (6 V, 1 μs) and an erase pulse (−6 V, 1 μs), the ON/OFF ratio remains above 10^4 even after 10^3 s, confirming robust non-volatile memory characteristics. In addition, it is worth noting that the reliability of the ZnO channel deteriorates at higher temperatures. However, forming gas annealing can be utilized to enhance its reliability. Further details on this can be found in our previous works [26]. Moreover, the long-term potentiation (LTP) and long-term depression (LTD) characteristics under non-identical pulse measurements are evaluated to assess the 3-D FeNOR's potential for CIM applications. The results are shown in Fig. 9. As illustrated, the conductance exhibits good linearity: with a fixed pulsewidth of 500 ns, increasing the LTP pulses from 4 to 6 V leads to a consistent increase in the conductance of the 3-D FeNOR, while decreasing the LTD pulses from −4 to −6 V leads to a consistent decrease in conductance. This good linearity in pulse-conductance response indicates 3-D FeNOR's capability for multi-bit storage, supporting its potential application in neural network system training [27], [28].

B. Array Level Characterization and Benchmarking

In addition to the single cell characteristic, the array level operation for 3-D FeNOR is also investigated. Fig. 10 illustrates the schematic and logic table for the array-level

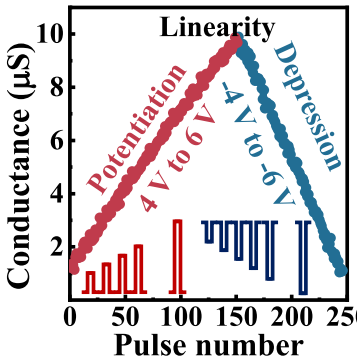
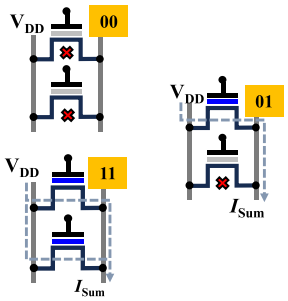


Fig. 9. Potentiation and depression characteristics of the 3-D FeNOR, indicating decent linearity for computing in memory applications.



	00	01	11
Cell1	OFF	On	On
Cell2	OFF	OFF	On
I_{Sum} (a.u.)	0	1	2

Fig. 10. Schematic and logic table for FeNOR array operation. In the NOR-structured memory array, the I_{Sum} is proportional to the number of active ON-state cells.

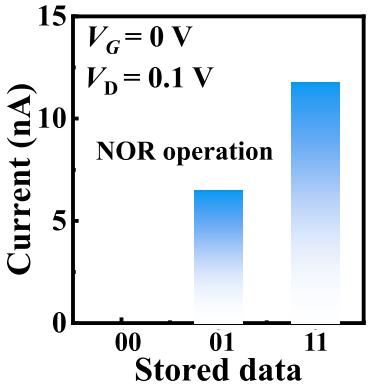


Fig. 11. Array-level logic operations of the 3-D FeNOR measured with varying storage values, showing the proportional relationship between I_{Sum} and ON-state cells. This notable linearity demonstrates the 3-D FeNOR's capability for array-level CIM applications.

operation involving two 3-D FeNOR cells. In this context, "00" represents both cells are off, "01" indicates one cell is on while the other is off, and "11" signifies both cells are on. The results are shown in Fig. 11. Here, it is observed that the summed current (I_{Sum}) increases linearly with the number of programmed ON-state cells, confirming the effective functionality of the NOR type array. This decent array-level linearity lays a solid foundation for large-scale CIM systems at the array level.

Finally, Fig. 12 presents a comparison of the main metrics between 3-D FeNOR, 3-D FeNAND, and 3-D NOR Flash [5], [11], [12], [17], [18], [23], [29], [30], [42]. As illustrated,

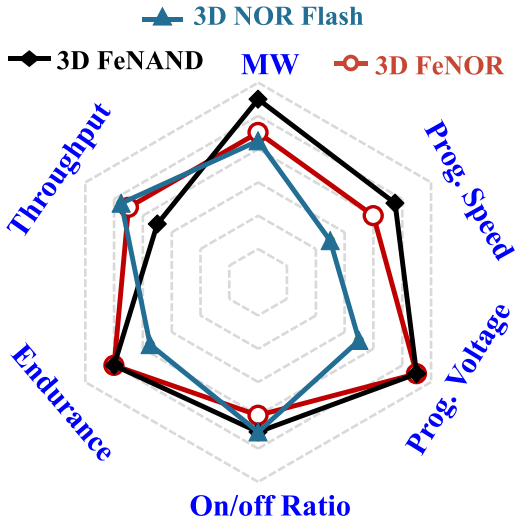


Fig. 12. Comparison of key metrics among 3-D FeNAND, 3-D FeNOR, and 3-D NOR Flash. The 3-D FeNOR exhibits good endurance, a decent MW, and low operating voltage, making it a strong candidate for CIM architectures.

the 3-D FeNOR exhibits several notable advantages over traditional 3-D NOR Flash, including faster programming speed, lower programming voltage, and enhanced endurance. These benefits are primarily due to the inherent advantages of ferroelectric switching. Additionally, when compared to the 3-D FeNAND structure, the 3-D FeNOR demonstrates a higher throughput. The reason for this higher throughput is discussed previously. In summary, these superior characteristics make the 3-D FeNOR a promising candidate for future CIM technologies.

IV. CIM Simulations and Analysis

We employ large-scale convolutional neural networks (VGG-16 and ResNet-50) to evaluate the performance of the 3-D FeNOR-based neural network system [31]. To efficiently manage the computational complexity of these networks, we adopt a large-scale matrix expansion approach, scaling the memory array size up to 4096 × 4096. This strategy significantly reduces dependence on analog-to-digital converters (ADCs) and other peripheral circuits, thereby improving both throughput and energy efficiency [32]. Furthermore, the CIFAR-10 dataset is utilized to assess the stability of convolutional neural networks integrated with the 3-D FeNOR architecture, demonstrating its potential for robust AI acceleration [33].

For neural network deployment, we conducted simulations based on device performance derived from realistic testing. Specifically, we extracted the non-linearity, and conductance fluctuation characteristics from real devices. Additionally, to simulate a realistic CIM circuit environment, we incorporated various circuit-level noise factors, including device-to-device variations, long-term retention degradation, and *RC* delay. These noise components were generated following the Monte Carlo principle to ensure statistical reliability.

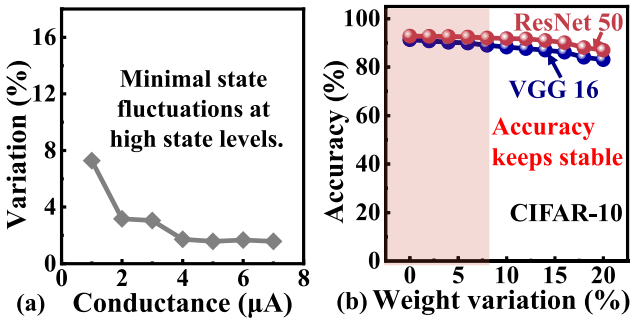


Fig. 13. (a) Fluctuations for different conductance levels extracted and averaged from ten devices. The state with lower current shows larger fluctuation. (b) Relationship between prediction accuracy and weight variation in the VGG-16 and ResNet-50, implemented using our proposed 3-D FeNOR architecture.

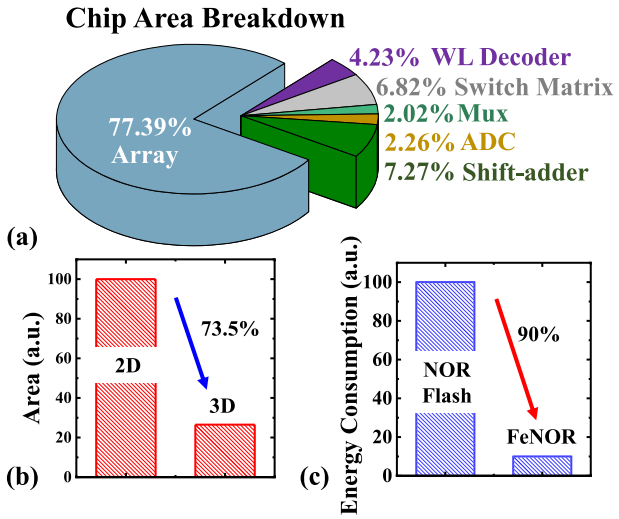


Fig. 14. (a) Area breakdown in a 2-D CIM circuit with an array size of 4096 × 4096, highlighting that the memory array occupies the majority of the area. (b) By utilizing a 3-D structure, the memory array area is reduced by 73.5%. (c) In comparison to a NOR Flash-based CIM circuit, the 3-D FeNOR achieves a 90% improvement in energy efficiency.

Prior to training, the relationship between weight magnitude and weight variation (conductance change) in 3-D FeNOR is first investigated, where the conductance ranges from 0 to 7 μs. All eight states are involved in the training process. In this case, the weight variation is measured following target weight programming. The corresponding results are presented in Fig. 13(a). Here, it can be observed that the I_D variation decreases as the memory weight magnitude increases. This could be attributed to the fact that a larger memory weight corresponds to stronger ferroelectric switching, which, in turn, enhances the weight's stability. Subsequently, this relationship between weight magnitude and weight variation is incorporated into the simulations, and the corresponding simulation results are displayed in Fig. 13(b). As depicted, high prediction accuracy (∼90%) under near-ideal conditions has been achieved. Additionally, it is observed that recognition accuracy remains stable despite fluctuations in memory weight: even with a maximum weight variation of 10%, accuracy exceeds 85%. This robustness meets the requirements for neural network training, highlighting the feasibility of the FeNOR-based neural network.

Furthermore, the total area of a conventional CIM circuit, including the 2-D memory array and associated peripheral circuitry, has been evaluated. For the circuit level evaluation in this work, we utilize the NeuroSim platform [34]. The results are presented in Fig. 14(a), indicating that when the memory size is scaled up to 4096 × 4096, the memory array occupies an impressive 77.39% of the total chip area. More seriously, this proportion increases as the memory size scales up. Such plight highlights a significant challenge faced by large-scale array-based CIM scheme: the memory array becomes the dominant factor determining the total area in 2-D array-based large-scale CIM chips. This large CIM array significantly increases the overall CIM macro area, raising two major concerns. First, the expanded memory array may limit the area scaling in system-level monolithic 3-D-integrated circuits, where the CIM macro and other functional layers are stacked in a layer-by-layer configuration. The second concern is that signal delay and degradation could become more pronounced in large-scale memory arrays due to higher wire resistance and increased parasitic capacitance. Nevertheless, this challenge can be effectively addressed through the adoption of our proposed 3-D FeNOR architecture. However, it is worth to point that the transition to 3-D CIM circuit introduces increased metal routing complexity, necessitating further optimization in data flow and peripheral circuit design.

To clearly illustrate the performance enhancement, the systematic evaluation has been conducted. The area evaluation result is presented in Fig. 14(b). As illustrated, in comparison to the conventional 2-D architecture, the 3-D FeNOR-based CIM design offers a substantial improvement in area efficiency, achieving a notable 73.5% reduction in the memory array footprint. In addition, Fig. 14(c) presents a comparison of the energy consumption of the memory arrays between the conventional NOR Flash and the proposed 3-D FeNOR architecture. As depicted, in macro level, the 3-D FeNOR achieves an excellent 90% improvement in energy consumption. This improvement can be attributed to two key factors. First, the proposed 3-D FeNOR operates at a lower programming voltage and requires a shorter programming time than conventional NOR Flash memory. Second, the operating voltage of the proposed 3-D FeNOR is 0.1 V, which is significantly lower than the typical drain voltage (∼1 V) used in NOR Flash CIM applications [35], [36]. Moreover, the benchmark of this work with other CIM schemes is presented in Table I. Leveraging the 3-D structure and the low operation voltage, 30.12 TOPS/W has been achieved in this work. Our proposed 3-D FeNOR demonstrates competitive performance relative to other reported works [37], [38], [39], [40], [41], [42], [43], [44].

Finally, it is worth noting that in future 3-D memory architecture-based CIM applications, the transition from conventional 2-D data flow to a fully integrated 3-D data flow in practical neural network deployments presents both opportunities and challenges. To enhance the efficiency of

TABLE I
BENCHMARK WITH OTHER WORKS

	Memory Type	3D Device	On/Off Ratio	Bit/Cell	Energy Efficiency (TOPS/W)
[37]	RRAM	Yes	10^2	3	-
[38]		No	-	4	76.25
[39]	PCM	Yes	10^3	3	-
[40]		No	-	4	20
[41]	Fe-NAND	No	10^4	2	8.08
[42]		Yes	10^3	7	-
[43]	NOR Flash	Yes	10^7	2	-
[44]		No	10^6	3	37.9
This work	**Fe-NOR**	**Yes**	**10^6**	**3**	**30.12**

3-D data flow, several optimization strategies could be implemented: 1) by employing a hierarchical data mapping strategy [39], computation can be efficiently distributed across 3-D memory arrays, significantly reducing memory access latency. Additionally, optimized routing techniques, such as vertical interconnects and through-silicon vias (TSVs), can be leveraged to minimize communication overhead between different layers, thereby improving overall system performance and 2) to fully release the computing potential of 3-D memory-based computing architectures, a more precise and well-organized sub-data division approach is required. For instance, advanced image segmentation strategies can facilitate higher levels of parallelism in data processing, ensuring efficient computation across multiple layers in the 3-D memory stack.

V. CONCLUSION

In this study, we introduced and experimentally demonstrated a 3-D BEOL-compatible MFMIS structure FeNOR-type memories featuring a ZnO channel, aimed at high-density neural network applications. The 3-D FeNOR proposed in this work achieves outstanding electrical performance, including a large MW of over 4 V, good ON/OFF ratio (10^6), and high endurance of 10^7 cycles. In addition, a high recognition accuracy of ∼90% is achieved in VGG-16 and ResNet-50 model. Our work advances NOR-type memory technologies, showcasing significant potential for compute-in-memory applications.

REFERENCES

[1] C. Li et al., "Analogue signal and image processing with large memristor crossbars," *Nature Electron.*, vol. 1, no. 1, pp. 52–59, Dec. 2018, doi: 10.1038/s41928-017-0002-z.

[2] P. Yao et al., "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, Jan. 2020, doi: 10.1038/s41586-020-1942-4.

[3] J. H. Shin, Y. J. Jeong, M. A. Zidan, Q. Wang, and W. D. Lu, "Hardware acceleration of simulated annealing of spin glass by RRAM crossbar array," in *IEDM Tech. Dig.*, Dec. 2018, pp. 3.3.1–3.3.4, doi: 10.1109/IEDM.2018.8614698.

[4] B. Yan, M. Liu, Y. Chen, K. Chakrabarty, and H. Li, "On designing efficient and reliable nonvolatile memory-based computing-in-memory accelerators," in *IEDM Tech. Dig.*, Dec. 2019, pp. 14.5.1–14.5.4, doi: 10.1109/IEDM19573.2019.8993562.

[5] D. Zhang et al., "Fast Fourier transform (FFT) using flash arrays for noise signal processing," *IEEE Electron Device Lett.*, vol. 43, no. 8, pp. 1207–1210, Aug. 2022, doi: 10.1109/LED.2022.3183111.

[6] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, no. 7567, pp. 47–55, Sep. 2015, doi: 10.1038/nature14956.

[7] Z. Zhou et al., "Experimental demonstration of an inversion-type ferroelectric capacitive memory and its 1 kbit crossbar array featuring high CHCS/CLCS, fast speed, and long retention," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 357–358, doi: 10.1109/VLSITechnologyandCir46769.2022.9830291.

[8] C. Sun et al., "First demonstration of BEOL-compatible ferroelectric TCAM featuring a-IGZO fe-TFTs with large memory window of 2.9 V, scaled channel length of 40 nm, and high endurance of 108 cycles," in *Proc. Symp. VLSI Technol.*, Jun. 2021, pp. 1–2.

[9] G. Kim et al., "High performance ferroelectric field-effect transistors for large memory-window, high-reliability, high-speed 3D vertical NAND flash memory," *J. Mater. Chem. C*, vol. 10, no. 26, pp. 9802–9812, Jul. 2022, doi: 10.1039/D2TC01608G.

[10] S. Migita, H. Ota, and A. Toriumi, "Design points of ferroelectric field-effect transistors for memory and logic applications as investigated by metal-ferroelectric-metal–insulator–semiconductor gate stack structures using $Hf_{0.5}Zr_{0.5}O_2$ films," *Jpn. J. Appl. Phys.*, vol. 58, Aug. 2019, Art. no. SLLB06, doi: 10.7567/1347-4065/ab389b.

[11] M.-K. Kim, I.-J. Kim, and J.-S. Lee, "CMOS-compatible ferroelectric NAND flash memory for high-density, low-power, and high-speed three-dimensional memory," *Sci. Adv.*, vol. 7, no. 3, Jan. 2021, Art. no. eabe1341, doi: 10.1126/sciadv.abe1341.

[12] K. Florent et al., "Vertical ferroelectric HfO_2 FET based on 3-D NAND architecture: Towards dense low-power memory," in *IEDM Tech. Dig.*, Dec. 2018, pp. 2.5.1–2.5.4, doi: 10.1109/IEDM.2018.8614710.

[13] Z. Zheng et al., "Boosting the memory window of the BEOL-compatible MFMIS ferroelectric/anti-ferroelectric FETs by charge injection," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 389–390, doi: 10.1109/VLSITechnologyandCir46769.2022.9830466.

[14] C. Sun et al., "Temperature-dependent operation of InGaZnO ferroelectric thin-film transistors with a metal-ferroelectric-metal-insulator-semiconductor structure," *IEEE Electron Device Lett.*, vol. 42, no. 12, pp. 1786–1789, Dec. 2021, doi: 10.1109/LED.2021.3121677.

[15] Y. Qin et al., "Understanding the memory window of ferroelectric FET and demonstration of 4.8-V memory window with 20-nm HfO_2," *IEEE Trans. Electron Devices*, vol. 71, no. 8, pp. 4655–4663, Aug. 2024, doi: 10.1109/TED.2024.3418942.

[16] C. Lee, S. H. Baek, and K. H. Park, "A hybrid flash file system based on NOR and NAND flash memories for embedded devices," *IEEE Trans. Comput.*, vol. 57, no. 7, pp. 1002–1008, Jul. 2008, doi: 10.1109/TC.2008.14.

[17] D. Zhang et al., "Implementation of image compression by using high-precision in-memory computing scheme based on NOR flash memory," *IEEE Electron Device Lett.*, vol. 42, no. 11, pp. 1603–1606, Nov. 2021, doi: 10.1109/LED.2021.3114407.

[18] Y. Feng et al., "First demonstration of BEOL-compatible 3D vertical FeNOR," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2024, pp. 1–2, doi: 10.1109/VLSITechnologyandCir46783.2024.10631352.

[19] H.-T. Lue et al., "3D AND: A 3D stackable flash memory architecture to realize high-density and fast-read 3D NOR flash and storage-class memory," in *IEDM Tech. Dig.*, Dec. 2020, pp. 6.4.1–6.4.4, doi: 10.1109/IEDM13553.2020.9372101.

[20] Z. Zheng et al., "BEOL-compatible MFMIS ferroelectric/anti-ferroelectric FETs—Part II: Mechanism with load line analysis and scaling strategy," *IEEE Trans. Electron Devices*, vol. 71, no. 9, pp. 5325–5331, Sep. 2024, doi: 10.1109/TED.2024.3421184.

[21] K. Ni et al., "Critical role of interlayer in $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric FET nonvolatile memory performance," *IEEE Trans. Electron Devices*, vol. 65, no. 6, pp. 2461–2469, Jun. 2018, doi: 10.1109/TED.2018.2829122.

[22] Z. Zhao et al., "A large window nonvolatile transistor memory for high-density and low-power vertical NAND storage enabled by ferroelectric charge pumping," *IEEE Electron Device Lett.*, vol. 45, no. 12, pp. 2554–2556, Dec. 2024, doi: 10.1109/LED.2024.3477510.

[23] G. Kim et al., "In-depth analysis of the Hafnia ferroelectrics as a key enabler for low voltage & QLC 3D VNAND beyond 1K layers: Experimental demonstration and modeling," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2024, pp. 1–2, doi: 10.1109/VLSITechnologyandCir46783.2024.10631559.

[24] S. Yoo et al., "Highly enhanced memory window of 17.8 V in ferroelectric FET with IGZO channel via introduction of intermediate oxygen-deficient channel and gate interlayer," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Honolulu, HI, USA, Jun. 2024, pp. 1–2, doi: 10.1109/vlsitechnologyandcir46783.2024.10631534.

[25] S. Yoon et al., "QLC programmable 3D ferroelectric NAND flash memory by memory window expansion using cell stack engineering," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2023, pp. 1–2, doi: 10.23919/VLSITECHNOLOGYANDCIR57934.2023.10185294.

[26] Q. Kong et al., "Back-end-of-line-compatible fin-gate ZnO ferroelectric field-effect transistors," *IEEE Trans. Electron Devices*, vol. 70, no. 4, pp. 2059–2066, Apr. 2023, doi: 10.1109/TED.2023.3242852.

[27] C. Sun et al., "Novel a-IGZO anti-ferroelectric FET LIF neuron with co-integrated ferroelectric FET synapse for spiking neural networks," in *IEDM Tech. Dig.*, Dec. 2022, pp. 2.1.1–2.1.4, doi: 10.1109/IEDM45625.2022.10019526.

[28] T. Soliman et al., "First demonstration of in-memory computing crossbar using multi-level cell FeFET," *Nature Commun.*, vol. 14, no. 1, p. 6348, Oct. 2023, doi: 10.1038/s41467-023-42110-y.

[29] P. Venkatesan et al., "Disturb and its mitigation in ferroelectric field-effect transistors with large memory window for NAND flash applications," *IEEE Electron Device Lett.*, vol. 45, no. 12, pp. 2367–2370, Dec. 2024, doi: 10.1109/LED.2024.3467210.

[30] H. Joh et al., "Oxide channel ferroelectric NAND device with source-tied covering metal structure: Wide memory window (14.3 V), reliable retention (> 10 Years) and disturbance immunity ($\Delta V_{th} \leq 0.1V$) for QLC operation," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2024, pp. 1–4, doi: 10.1109/iedm50854.2024.10873376.

[31] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification," in *Proc. Int. Conf. Disruptive Technol. Multi-Disciplinary Res. Appl. (CENTCON)*, vol. 1, Bengaluru, India, Nov. 2021, pp. 96–99, doi: 10.1109/CENTCON52345.2021.9687944.

[32] Y. Feng et al., "A novel array programming scheme for large matrix processing in flash-based computing-in-memory (CIM) with ultrahigh bit density," *IEEE Trans. Electron Devices*, vol. 70, no. 2, pp. 461–467, Feb. 2023, doi: 10.1109/TED.2022.3227529.

[33] X. Zhang, "The AlexNet, LeNet-5 and VGG NET applied to CIFAR-10," in *Proc. 2nd Int. Conf. Big Data Artif. Intell. Softw. Eng. (ICBASE)*, Sep. 2021, pp. 414–419, doi: 10.1109/ICBASE53849.2021.00083.

[34] P.-Y. Chen et al., "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 12, pp. 3067–3080, Dec. 2018, doi: 10.1109/TCAD.2018.2789723.

[35] Y. Feng et al., "Design-technology co-optimizations (DTCO) for general-purpose computing in-memory based on 55nm NOR flash technology," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2021, pp. 12.1.1–12.1.4.

[36] H.-T. Lue, T.-H. Hsu, C. Lo, T.-H. Yeh, K.-C. Wang, and C.-Y. Lu, "Investigation of methods that greatly improve 3D NOR flash to either gain superb retention or become DRAM-like with high endurance (> 1G cycling) and high write-bandwidth (> 4Gb/s)," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2021, pp. 10.3.1–10.3.4.

[37] M. Xie et al., "Monolithic 3D integration of 2D transistors and vertical RRAMs in 1T–4R structure for high-density memory," *Nature Commun.*, vol. 14, no. 1, p. 5952, Sep. 2023, doi: 10.1038/s41467-023-41736-2.

[38] C. Mu et al., "A 28nm 76.25TOPS/W RRAM/SRAM-collaborative CIM fine-tuning accelerator with RRAM-endurance/latency-aware weight allocation for CNN and transformer," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2024, pp. 1–3, doi: 10.1109/A-SSCC60305.2024.10848596.

[39] N. Hur et al., "Ultralow-power programmable 3D vertical phase-change memory with heater-all-around configuration," *Small Methods*, vol. 2024, Art. no. 2401381, doi: 10.1002/smtd.202401381.

[40] S. Ambrogio et al., "An analog-AI chip for energy-efficient speech recognition and transcription," *Nature*, vol. 620, no. 7975, pp. 768–775, Aug. 2023, doi: 10.1038/s41586-023-06337-5.

[41] I.-J. Kim, M.-K. Kim, and J.-S. Lee, "Highly-scaled and fully-integrated 3-dimensional ferroelectric transistor array for hardware implementation of neural networks," *Nature Commun.*, vol. 14, no. 1, p. 504, Jan. 2023, doi: 10.1038/s41467-023-36270-0.

[42] Z. Zhao et al., "In-situ encrypted NAND FeFET array for secure storage and compute-in-memory," in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4, doi: 10.1109/IEDM45741.2023.10413774.

[43] M.-L. Wei et al., "Analog computing in memory (CIM) technique for general matrix multiplication (GEMM) to support deep neural network (DNN) and cosine similarity search computing using 3D AND-type NOR flash devices," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2022, pp. 33.3.1–33.3.4, doi: 10.1109/IEDM45625.2022.10019495.

[44] D. Kwon et al., "Reconfigurable neuromorphic computing block through integration of flash synapse arrays and super-steep neurons," *Sci. Adv.*, vol. 9, no. 29, Jul. 2023, Art. no. eadg9123, doi: 10.1126/sciadv.adg9123.