

28 nm HKMG-Based Current Limited FeFET Crossbar-Array for Inference Application

Sourav De[®], *Member, IEEE*, Franz Müller[®], Sunanda Thunder[®], Sukhrob Abdulazhanov[®], Nellie Laleni, Maximilian Lederer[®], Tarek Ali, Yannick Raffel[®], Stefan Dünkel, Shaown Mojumder, Alptekin Vardar, Sven Beyer, Konrad Seidel, and Thomas Kämpfe[®], *Member, IEEE*

Abstract—This article reports a novel ferroelectric fieldeffect transistor (FeFET)-based crossbar array cascaded with an external resistor. The external resistor is shunted with the column of the FeFET array, as a current limiter and reduces the impact of variations in drain current (I_d) , especially in a low threshold voltage (LVT) state. We have designed crossbar arrays of 8 \times 8 sizes and performed multiply-and-accumulate (MAC) operations. Furthermore, we have evaluated the performance of the current limited FeFET crossbar array in system-level applications. Finally, the system-level performance evaluation was done by neuromorphic simulation of the resistor-shunted FeFET crossbar array. The crossbar array achieved software-comparable inference accuracy (~97%) for National Institute of Standards and Technology (MNIST) datasets with multilayer perceptron (MLP) neural network, whereas the crossbar arrays built solely with FeFETs failed to learn, yielding only 9.8% accuracy.

Index Terms—Ferroelectric field-effect transistor (FeFET), ferroelectric memory, hafnium oxide (HfO₂), inference, neuromorphic.

I. INTRODUCTION

THE arrival of convolution neural networks (CNNs) or ConvNets changed the paradigm of computing architecture in the present era. Built by LeCun et al. [1], the early versions of CNN mostly found their application in banking. However, the necessity of massive datasets and computing resources sidelined CNNs from the mainstream research roadmap of computer vision and artificial intelligence

Manuscript received 6 October 2022; accepted 20 October 2022. Date of publication 4 November 2022; date of current version 30 November 2022. This work was supported in part by the European Union's Electronic Components and Systems for European Leadership (ECSEL) Joint Undertaking under Grant 826655, in part by the Project TEMPO and ANDANTE under Grant 876925, in part by the German Bundesministerium für Wirtschaft (BMWI), and in part by the State of Saxony in the frame of the "Important Project of Common European Interest (IPCEI)." The review of this brief was arranged by Editor F. Bonani. (Corresponding author: Sourav De.)

Sourav De, Franz Müller, Sunanda Thunder, Sukhrob Abdulazhanov, Nellie Laleni, Maximilian Lederer, Yannick Raffel, Shaown Mojumder, Alptekin Vardar, Konrad Seidel, and Thomas Kämpfe are with the Center Nanoelectronic Technologies, Fraunhofer IPMS, 01099 Dresden, Germany (e-mail: sourav.de@ipms.fraunhofer.de).

Tarek Ali, Stefan Dünkel, and Sven Beyer are with Integration Engineering, GlobalFoundries, 01109 Dresden, Germany.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2022.3216973.

Digital Object Identifier 10.1109/TED.2022.3216973

(AI). The researchers started to demonstrate interest again in AI with the advent of AlexNet amidst the availability of the ImageNet dataset [2], [3]. The colossal computing resources enabled the users to create much larger CNNs, which could perform more complex tasks that had not been possible before.

Amid such developments in the computing world, another evolution was brewing in the world of electronics. The rapid advancement toward the deeply scaled and dense technology nodes made the edge devices more available to the users, which increased the real-time data generated by internet search engines, social media, Internet of Things (IoT) devices, and smart devices by many folds in recent times. The need for realtime processing of this enormous amount of data produced by the end-user devices has mandated a change in the computing system. The latency and massive computing power required by conventional Von-Neumann computing architecture for processing such an enormous amount of real-time data make them ill-suited for such purposes. However, it is worth noting that the data processing centers furnished by high-performance graphics processing units (GPUs) or tensor processing units (TPUs) can run real-time data processing with much lower latency. But their power-hungry nature makes them inappropriate for end-user applications. Therefore, the need for lowpower and fast real-time data processing has directed the researchers toward an alternative route beyond standard Von-Neumann architecture [4], [5], [6], [7], [8].

The scientific community is pursuing non-Von-Neumann architectures by implementing deep neural networks (DNNs) or spiking neural networks (SNNs) for data-centric computing with higher energy efficiency and lower latency. Simultaneously, research on emerging non-volatile memory (eNVM) has also accelerated the implementation of in-memory computing (IMC) architectures. Hafnium oxide (HfO₂) based ferroelectric memories are of great interest among the scientific community amidst many potential candidates like resistive random access memory (RRAM), magnetic random access memory (MRAM), and phase change memories (PCMs). This trait can be attributed to the CMOS compatibility and scalability of HfO₂, which facilitates very large-scale integration (VLSI) of ferroelectric memories with the advanced CMOS process. The compatibility of ferroelectric memory with 28-nm high-k-metal-gate (HKMG) technology, FinFETs and thinfilm technology have further accelerated the system-level

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

integration of ferroelectric memories [9], [10], [11], [12], [13], [14], [15].

However, the primary issue of system-level integration of ferroelectric memories is the increasing variability with scaling. The poly-crystalline nature of HfO₂ and intrinsic defect sites, which act as charge-trapping sites, creates a severe issue in deeply scaled ferroelectric field-effect transistors (FeFETs). The charge trapping sites may capture electrons or holes from the channel side (CS) or gate side (GS), leading to infidelity in program-erase (WRITE) operation. Quintessentially, ferroelectricity in HfO₂ is a crystal structure-dependent property. The non-centrosymmetric Pca₂₁ orthorhombic phase is responsible for ferroelectricity in HfO₂. Therefore, harnessing the ferroelectric orthorhombic phase is essential for the stable operation of ferroelectric memory. There have been several attempts from a semiconductor process perspective to stabilize the ferroelectric orthorhombic phase. Despite adopting several stabilization processes, atomic-layer-deposited (ALD) or physical vapor-deposited (PVD) HfO₂ films show variability in significantly scaled devices [16], [17], [18], [19], [20], [21].

In our previous work, we demonstrated a very effective way to reduce the device variations in the READ–WRITE operation of 28-nm HKMG FeFETs devices by shunting a resistor with the drain terminal (1F–1R) [22], [23]. We observed that the large variation in the drain current (I_d) is mostly associated with the low threshold voltage (LVT)-state current, which creates erroneous output during analog-to-digital conversion after accumulation. Therefore, we suggested to build a 1F–1R structure to compensate the variation in I_d of LVT-state. We observed sufficiently high resistance (2 M Ω), the I_d variation was strongly suppressed. Furthermore, the I_d variation originating from the random distribution of ferroelectric domains is reduced by the large operation window of V_{GS} .

This work, an extension of our previous work, focuses on evaluating the performance of 1F–1R structure-based crossbar array for neuromorphic computing applications. We have designed 8 × 8 crossbar arrays with FeFETs and the column was terminated with a resistive element. The resistive element acts as current limiter for the crossbar array and reduces the variation of the bitline (BL) current ($I_{\rm BL}$). Finally we have also evaluated the performance of the memory array as synaptic core. The system level performance demonstrates, ~97% accuracy for inference application.

II. EXPERIMENTS

The device considered in this work was fabricated in Global-Foundries', using their 28-nm HKMG technology on 300-mm wafers. The FeFETs were fabricated by integrating 8 nm silicon doped hafnium oxide (HfO₂) based ferroelectric layer with a \sim 1 nm silicon-di-oxide (SiO₂) interfacial layer in the gate-stack of a regular metal oxide semiconductor field-effect transistor (MOSFET). Fig. 1 shows the schematic illustration and the transmission electron microscopic (TEM) image of the FeFETs under consideration. The fabricated devices were programmed (WRITE) to non-overlapping binary states using 500-ns pulses at the gate terminal. Before conducting READ–WRITE operations, the FeFETs were cycled by



Fig. 1. TEM image of the 28-nm HKMG FeFET devices and schematic representation. After applying a positive and negative pulse, LVT, and high-threshold voltage (HVT) states are achieved.

50 consecutive wake-up pulses. Each wake-up pulse consists of one 4.5-V pulse followed by another -5 V pulse of 500 ns. The source, drain and bulk terminals were biased at 0 V during the WRITE operation. A non-disturbing direct current (dc) sweep with a step size of 100 mV was applied at the gate terminal for the READ operation while maintaining 100 mV at the drain terminal and 0 V at source and bulk. The WRITE-pulse applied at the gate terminal lines up the electrical dipoles in the ferroelectric layer according to their polarity, changes the surface charge density of the semiconductor layer, conductance of the channel (G_{ch}), and the threshold voltage (V_{th}). Quintessentially, for n-type FeFET a positive pulse at gate terminal sets the device to LVT-state and negative pulse at gate terminal sets the devices at HVT-state.

The characterization of single devices was followed by the characterization of 8×8 arrays. The layout, optical image of the array, and schematic representation of the mini-array are shown in Fig. 2(a) and (b), respectively. The gate terminals of FeFETs are connected row-wise in a single word-line (WL). The drains and sources are connected column-wise in bitline (BL) and source-line (SL). The WL receives inputs for the READ-WRITE operation. The BL is connected to the current limiter. The arrays were programmed row-wise through the direct access through word lines (WLs). The select lines (SLs) and BLs connected along the columns allows the read operation to be performed along the column. The transistors denoted by $M_{\rm SL}$ and $M_{\rm BL}$ are used as inhibit mode switches. The FeFETs are characterized using a PXI-Express system from National Instruments. The contacts of the memory array were controlled by controlled by the pin parametric measurement unit (PPMU) of NI PXIe-6570 and source measure unit (SMU) of NI PXIe-4143. SMUs were used for conduction programerase operation, while the SLs and BLs were biased at 0 V. The devices were allowed to de-trap for 2 s after programming. The read operation was conducted by a slowly varying voltage ramp with a step size of 100 mV at WL, while keeping BL and SL biased at 100 mV and 0 V, respectively. The bulk was also kept at 0 V. The read operation takes approximately 1 ms to complete.

The row-wise WRITE operation was carried out by applying a 4.5-V pulse of 500 ns at the WL, while the complete array was erased by applying a 5-V pulse of 40 μ s in the bulk. Fig. 2(d) and (e) shows the biasing scheme for row wise WRITE and bit-wise READ operation. The SL was connected to the resistive current limiter. The BLs were biased to 0.1 V through the wire named by V_{BL}^{PRG} . N_{SL} and N_{BL} transistors



Fig. 2. (a) Layout and (b) optical image of the fabricated memory array. (c) Schematic of the memory array. The operating voltages along BL, SL, and WL for carrying out (d) WRITE operation and (e) READ operation.

are turned on by applying 1.8 V at $V_{\rm SL}$ and $V_{\rm BL}$ to allow access to SL and BL. During the READ operation $V_{\rm INH}$ was biased at 0 V to turn-off the inhibit transistors denoted by $M_{\rm BL}$ and $M_{\rm SL}$. $M_{\rm BL}$ and $M_{\rm SL}$ were turned on during inhibit operation by applying 1.8 V at $V_{\rm INH}$. The inhibit operation on WL was carried out by applying -0.3 V, which enabled row-wise WRITE and cell-wise READ operation.

III. RESULTS AND DISCUSSION

A. Device Characterization

Fig. 3(a) shows the program-erase scheme used in this work. Fig. 3(b) and (d) displays the two-level READ-WRITE operations for 1F cell and 1F-1R cell, respectively. The LVT-state current (I_{LVT}) to HVT-state current (I_d^{HVT}) ratio has been reduced for 1F–1R synapses. However, high I_d^{LVT} also increases the variation in the I_d^{LVT} . The slightest variation in the $V_{\rm th}$ of any programmed state induces a significant variation related to the cell current of that state. Therefore, even though the $I_d^{\rm LVT}$ of the 1F–1R structure is reduced, embedding the current limiter reduces the standard deviation of the I_d^{LVT} significantly [Fig. 3(c) and (e)]. Further, the variability originating from the WL due to the random variation in ferroelectric domains is suppressed by a large operation window of the WRITE pulse. The drain was held at constant 0 V during WRITE operation to ensure low static power consumption.

The READ–WRITE characterization was followed by endurance and retention characterization. Quintessentially, the operation of front-end-of-line (FEoL) FeFETs with silicon channels is limited by the WRITE endurance up to 10^5 . This limitation on WRITE endurance makes online training of the neural network (NN). The WRITE-endurance characteristics has been described in Fig. 4(a), and we observed a MW with a stable behavior up to 10^3 cycles with an increase in degradation, which lead to a full closure of MW after 5×10^5 cycles. Fig. 4(b) shows stable data retention characteristics up to 10^4 s at 85 °C.



Fig. 3. (a) Schematic of the waveform. The devices were subjected to wake-up by 50 pulses before the READ–WRITE. (b)–(e) Transfer characteristics and CDF of $I_d^{\rm LVT}$ and $I_d^{\rm HVT}$ shows high variation in the $I_d^{\rm LVT}$ for regular FeFET cells. Although shunting a current limiter reduces the $I_d^{\rm LVT}$, the variation is also reduced significantly.



Fig. 4. (a) Endurance characteristics demonstrate stable endurance up to 10⁴ cycles. The memory window collapses around 5 \times 10⁵ cycles. (b) Retention characteristics measured at 85 °C demonstrates stable data retention characteristics up to 10⁴ s.

B. Array Characterization

Fig. 5(a) demonstrates the accumulated I_{BL} versus V_{WL} operation from a single column of memory array. The accumulation of the I_{BL} resembles the mathematical vector matrix multiplication or multiply-and-accumulate (MAC) operation. The MAC operations can be quintessentially conducted over the FeFET crossbar memory array demonstrated in this work. We can observe a negligible leakage current during I_{BL} accumulation. The leakage current with all transistors in the erased



Fig. 5. (a) MAC Operation shows accumulation of I_{BL} for various stages of cell activation. (b) Statistical plot of I_{BL} shows stable MAC operation in 20 crossbar arrays.



Fig. 6. (a) Architecture of MLP-based NN. (b) Simulation of NN predicts minimal loss of inference accuracy due to retention degradation over a time of 10 years.

state was lower than 1 nA. This ensures leakage free MAC operation. For statistical modeling, the MAC operation was performed over 20 segments across 300-mm wafers. Shows stable MAC operation over 20 different segments from the crossbar array with a non-overlapping variation in $I_{\rm BL}$.

C. Applications to In-Memory-Computing

We have performed a system-level simulation of handwritten digit recognition from the data set of "Modified National Institute of Standards and Technology (MNIST)" to quantify the efficacy of current limited FeFET-based crossbar array in multi-layer perceptron (MLP) based NNs as synaptic cores [24], [25]. Experimentally obtained device-to-device variation and retention degradation of I_{BL} have been modeled for NN simulation. The architectures and the layers of the NN are illustrated in Fig. 6(a).

The NN was trained offline for meeting the endurance criteria of FeFET synapses. The inference operation was conducted after the offline training of the NN was conducted. Quintessentially the online training or the retraining of the NN puts an excessive load on the hardware in terms of energy.

TABLE I BENCHMARKING TABLE

	VLSI, 2022 [26]	IEDM, 2021 [27]	IEDM, 2021 [28]	JEDS, 2021 [29]	This Work
Cell Type	1T-1C (FeFET)	1C (FeCAP)	1T-1C (FeRAM)	1T-1C (FeRAM)	1T-1C (FeFET)
Technology	180 nm	NA	130 nm	130 nm	28 nm
WRITE	6V	3V	2.5V	2.5V	4.5V
Retention	10 ⁴ s	$10^4 \mathrm{s}$	10 ⁴ s	$10^4 \mathrm{s}$	10 ⁴ s

Therefore, data retention capability becomes important to carry out inference operation without frequent retraining. The data retention measured up to 10^4 s at 85 °C has been extrapolated to estimate the data retention up to 10^8 s. The extrapolated data retention was used to gauge the inference accuracy for MNIST datasets with MLP NN for 10^8 s. The MLP-based NN achieves inference accuracy of over 97% initially and maintains inference accuracy above 95% for 10^8 s without being retrained.

IV. CONCLUSION

In this work, we demonstrate a high-performance current limited FeFET-based crossbar array. The externally shunted resistor acted as a current limiter and reduced the variation in $I_{\rm BL}$. The MAC operation conducted on 20 numbers of crossbar arrays demonstrates the stability of the operation. The high on-state resistance and low variation ensure error-free MAC operation using current limited FeFET arrays. Table I lists the comparison of this work with other state-of-the-art works.

REFERENCES

- Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, Jun. 2017, pp. 84–90, doi: 10.1145/3065386.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
- [4] T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations," *Frontiers Neurosci.*, vol. 10, p. 333, Jul. 2016, doi: 10.3389/fnins.2016.00333.
- [5] S. Ambrogio et al., "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60–67, Jun. 2018, doi: 10.1038/s41586-018-0180-5.
- [6] J.-H. Bae, S. Lim, B.-G. Park, and J.-H. Lee, "High-density and nearlinear synaptic device based on a reconfigurable gated Schottky diode," *IEEE Electron Device Lett.*, vol. 38, no. 8, pp. 1153–1156, Aug. 2017, doi: 10.1109/LED.2017.2713460.
- [7] C.-C. Chang et al., "Mitigating asymmetric nonlinear weight update effects in hardware neural network based on analog resistive synapse," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 116–124, Mar. 2018, doi: 10.1109/JETCAS.2017.2771529.
- [8] P.-Y. Chen et al., "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2015, pp. 194–199, doi: 10.1109/ICCAD.2015.7372570.
- [9] M. Trentzsch et al., "A 28 nm HKMG super low power embedded NVM technology based on ferroelectric FETs," in *IEDM Tech. Dig.*, Dec. 2016, p. 11, doi: 10.1109/IEDM.2016.7838397.

- [10] S. De et al., "READ-optimized 28 nm HKMG multibit FeFET synapses for inference-engine applications," *IEEE J. Electron Devices Soc.*, vol. 10, pp. 637–641, 2022, doi: 10.1109/JEDS.2022.3195119.
- [11] S. De et al., "Ultra-low power robust 3bit/cell Hf_{0.5}Zr_{0.5}O₂ ferroelectric finFET with high endurance for advanced computing-in-memory technology," in *Proc. Symp. VLSI Technol.*, 2021, pp. 1–2.
- [12] S. De et al., "Robust binary neural network operation from 233 K to 398 K via gate stack and bias optimization of ferroelectric FinFET synapses," *IEEE Electron Device Lett.*, vol. 42, no. 8, pp. 1144–1147, Aug. 2021, doi: 10.1109/LED.2021.3089621.
- [13] S. De et al., "Low-power vertically stacked one time programmable multi-bit IGZO-based BEOL compatible ferroelectric TFT memory devices with lifelong retention for monolithic 3D-inference engine applications," in *Proc. EAI Eur. Solid-State Devices Circuits Conf.*, 2022, pp. 1–20.
- [14] M. Jerry et al., "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *IEDM Tech. Dig.*, Dec. 2017, p. 6, doi: 10.1109/IEDM.2017.8268338.
- [15] S. Dutta et al., "Logic compatible high-performance ferroelectric transistor memory," *IEEE Electron Device Lett.*, vol. 43, no. 3, pp. 382–385, Mar. 2022, doi: 10.1109/LED.2022.3148669.
- [16] Y. Raffel et al., "Charge pumping and flicker noise-based defect characterization in ferroelectric FETs," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, Oct. 2020, pp. 1–4, doi: 10.1109/IIRW49815.2020. 9312851.
- [17] S. De, W.-X. Bu, B.-H. Qiu, C.-J. Su, Y.-J. Lee, and D. D. Lu, "Alleviation of charge trapping and flicker noise in HfZrO₂-based ferroelectric capacitors by thermal engineering," in *Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA)*, Apr. 2021, pp. 1–2, doi: 10.1109/VLSI-TSA51926.2021.9440091.
- [18] M. N. K. Alam et al., "On the characterization and separation of trapping and ferroelectric behavior in HfZrO FET," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 855–862, 2019, doi: 10.1109/JEDS.2019.2902953.
- [19] S. De et al., "Uniform crystal formation and electrical variability reduction in hafnium-oxide-based ferroelectric memory by thermal engineering," ACS Appl. Electron. Mater., vol. 3, no. 2, pp. 619–628, Feb. 2021, doi: 10.1021/acsaelm.0c00610.

- [20] M. Lederer et al., "Influence of annealing temperature on the structural and electrical properties of Si-doped ferroelectric hafnium oxide," ACS Appl. Electron. Mater., vol. 3, no. 9, pp. 4115–4120, Sep. 2021, doi: 10.1021/acsaelm.1c00590.
- [21] M. Lederer, D. Lehninger, T. Ali, and T. Kämpfe, "Review on the microstructure of ferroelectric hafnium oxides," *Phys. Status Solidi Rapid Res. Lett.*, vol. 16, Jul. 2022, Art. no. 2200168, doi: 10.1002/pssr.202200168.
- [22] T. Soliman et al., "FELIX: A ferroelectric FET based low power mixedsignal in-memory architecture for DNN acceleration," ACM Trans. Embedded Comput. Syst., vol. 21, no. 6, pp. 1–25, Nov. 2022, doi: 10.1145/3529760.
- [23] T. Soliman et al., "Ultra-low power flexible precision FeFET based analog in-memory computing," in *IEDM Tech. Dig.*, Dec. 2020, p. 29, doi: 10.1109/IEDM13553.2020.9372124.
- [24] A. Lu, X. Peng, W. Li, H. Jiang, and S. Yu, "NeuroSim simulator for compute-in-memory hardware accelerator: Validation and benchmark," *Frontiers Artif. Intell.*, vol. 4, Jun. 2021, Art. no. 659060.
- [25] X. Peng, S. Huang, Y. Luo, X. Sun, and S. Yu, "DNN+NeuroSim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies," in *IEDM Tech. Dig.*, Dec. 2019, p. 32.
- [26] K. Seidel et al., "Memory array demonstration of fully integrated 1T-1C FeFET concept with separated ferroelectric MFM device in interconnect layer," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Jun. 2022, pp. 355–356, doi: 10.1109/ VLSITechnologyandCir46769.2022.9830141.
- [27] Y.-C. Luo et al., "Experimental demonstration of non-volatile capacitive crossbar array for in-memory computing," in *IEDM Tech. Dig.*, Dec. 2021, pp. 1–4, doi: 10.1109/IEDM19574.2021.9720508.
- [28] T. Francois et al., "16kbit HfO₂: Si-based 1T-1C FeRAM arrays demonstrating high performance operation and solder reflow compatibility," in *IEDM Tech. Dig.*, Dec. 2021, p. 33, doi: 10.1109/IEDM19574.2021.9720640.
- [29] J. Okuno et al., "1T1C FeRAM memory array based on ferroelectric HZO with capacitor under bitline," *IEEE J. Electron Devices Soc.*, vol. 10, pp. 29–34, 2022, doi: 10.1109/JEDS.2021.3129279.