A 2-Transistor-2-Capacitor Ferroelectric Edge Compute-in-Memory Scheme with Disturb-Free Inference and High Endurance

Xiaoyang Ma, Shan Deng, Juejian Wu, Zijian Zhao, David Lehninger, Tarek Ali, Konrad Seidel, Sourav De, Xiyu He, Yiming Chen, Huazhong Yang, Vijaykrishnan Narayanan, Suman Datta, Thomas Kämpfe, Qing Luo, Kai Ni, and Xueqing Li

***Abstract*—This paper proposes C^{2}FeRAM, a 2T2C/cell ferroelectric compute-in-memory (CiM) scheme for energy-efficient and high-reliability edge inference and transfer learning. With certain area overhead, C^{2}FeRAM achieves the following highlights: (i) compared with FeFET/FeMFET, it achieves disturb-free CiM and much higher write endurance (equal to FeRAM), leading to >100x inference time with <1% accuracy drop for VGG8 in CIFAR-10 dataset, along with the enhanced endurance for weight updates, e.g., CiM-based transfer learning; (ii) compared with 1T1C FeRAM inference cache, the achieved disturb-free feature and CiM capability in C^{2}FeRAM lead to improvements of 4x energy, 200x speed, and 3.2e5x life cycles. Such benefits highlight an intriguing solution for future intelligent edge AI.**

***Index Terms*—Ferroelectric memories, FeFET, FeRAM, compute-in-memory (CiM), endurance, read disturb.**

I. Introduction

COMPUTE-in-memory (CiM) based on ferroelectric memories are being actively exploited for high-speed and energy-efficient edge intelligence [1][2]. This has been increasingly promising with the rapid ferroelectric device development progress, including ferroelectric capacitors, transistors, and tunneling junctions in [3][4][5]. However, the existing ferroelectric devices are still facing the challenge of reliability, as illustrated in Fig. 1(a). For ferroelectric random access memory (FeRAM), the destructive read limits the application of continuous inference with CiM [6]; For ferroelectric FET (FeFET) and ferroelectric-metal FET (FeMFET), the low write endurance raises lifetime concerns in data-intensive training with frequent updates [7].

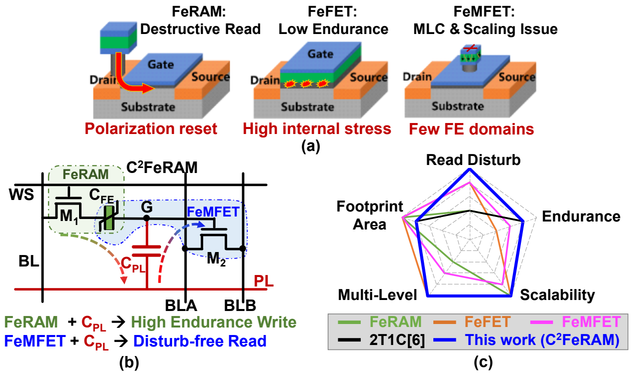

Fig. 1. (a) Challenges of ferroelectric devices; (b) cell scheme and read/write of proposed C^{2}FeRAM; (c) ferroelectric devices comparison.

Different mechanisms cause these limitations in edge CiM acceleration. The destructive read results from the same charging/discharging path shared by write/read. By integrating the ferroelectric (FE) layer to the gate of MOSFET, the FeFET supports non-destructive read through the MOSFET channel but still suffers from reliability issues, e.g., high write voltage and low endurance. As reported in [8][9], the ferroelectric layer fatigue and the gate stack deterioration lead to the endurance degradation of the HfO$_2$-based FeFET. In some devices, gate stack deterioration is the critical cause of device degradation [10][11][12]. For FeMFET, although it is possible to achieve lower write voltage and higher endurance with a small A_{FE}/A_{MOS} ratio, its endurance is still lower than that of FeRAM [13]. In addition, a small A_{FE}/A_{MOS} ratio raises scaling difficulty and susceptibility to disturbances. The prior work in [14] using in-cell amplification tries to combine FeRAM and FeMFET for higher read sensitivity, but the read destruction issue remains unsolved, which limits its application using CiM approaches.

To overcome these challenges, this letter proposes C^{2}FeRAM, a ferroelectric CiM scheme that achieves both disturb-free read and high write endurance. As illustrated in Fig. 1(b), it consists of one capacitor, one ferroelectric capacitor (FeCap), and two transistors. The disturb-free read also enables the multi-level cell (MLC) for further memory density

Manuscript received 14 March 2023; revised 18 April 2023; accepted 1 May 2023. Date of publication XXX XXX 2023; date of current version XXX XXX 2023. This work is supported in part by NSFC (U21B2020, 92264204), in part by the KDT JU under Grant Agreement N°101007321, in part by NSF (2008365, 2132918), and in part by the Tsinghua - Daimler JISM. Corresponding authors are Xueqing Li and Kai Ni.

X. Ma, J. Wu, X. He, Y. Chen, H. Yang, and X. Li are with BNRist, EE, Tsinghua University, Beijing 100084, China (e-mail: xueqingli@tsinghua.edu.cn).

S. Deng, Z. Zhao, and K. Ni are with The Department of Microsystems Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA (e-mail: kai.ni@rit.edu).

D. Lehninger, K. Seidel, S. De, and T. Kämpfe are with the Center Nanoelectronic Technologies, Fraunhofer IPMS, Dresden, Germany.

T. Ali is with GlobalFoundries, 01109 Dresden, Germany.

V. Narayanan is with The Department of Computer Science and Engineering, Penn State University, University Park, PA 16802, USA.

S. Datta is with The School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

Q. Luo is with The Institute of Microelectronics Chinese Academy of Science, Beijing 100029, China.

Color versions of one or more figures in this letter are available at https://doi.org/XX.XXXX/LEDXXXX.

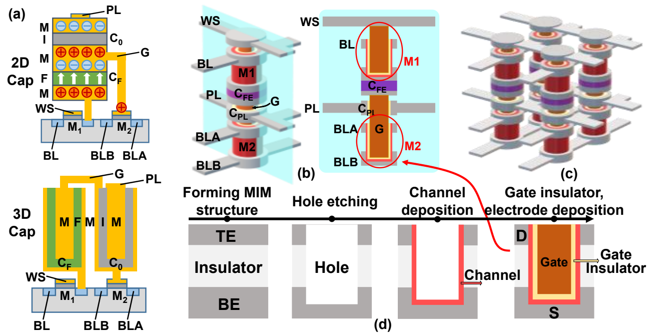

Fig. 2. (a) Proposed structure with 2D and 3D capacitor; (b) CAA structure and its cross-section of the C^{2}FeRAM cell; (c) array example of CAA C^{2}FeRAM (d) brief process sequence of the CAA MOSFET [17].

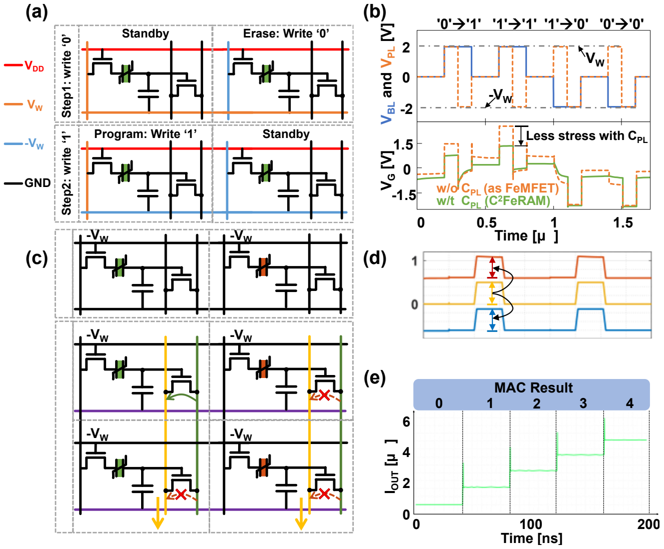

Fig. 3. (a) two-step-write weight update scheme with bias setting, (b) write transient signals and reduced internal gate stress with 5fF C_{PL} and 3~7fF C_{FE}, (c) current summation read scheme, (d) transient read signals showing non-destructive read, (e) bitline read currents for different accumulation results.

improvement. In addition, without the constraint of the small A_{FE}/A_{MOS} area ratio, the proposed structure can scale down with advanced process. The comparison between existing ferroelectric devices is shown in Fig. 1(c), which highlights the advantages of C^{2}FeRAM. Section II and III will present the operating mechanism and experimental results, respectively.

II. Device Mechanism

The FE layer in both FeMFET and FeFET couple with the MOSFET in read and write. Therefore, the MOSFET gate voltage can hardly be manipulated. In the proposed C^{2}FeRAM, the M$_2$ gate voltage (node G) can be controlled by the plate line (PL) through the coupling of the plate capacitor C_{PL}. The direct control over the internal gate voltage enables disturb-free read and high-endurance write.

To reduce the footprint overhead, stacking the capacitors on top of the transistors is an effective approach [15]. Fig. 2(a) illustrates two integration approaches. The planar C_{FE} and C_{PL} in Fig. 2(a) is compatible with the existing CMOS process. The structure of the cylinder cap in Fig. 2(a) has been demonstrated by [16] with 1Xnm DRAM technology. Moreover, a potential structure with channel-all-around (CAA) transistors [17], in which C_{FE} is a planar capacitor, and C_{PL} is a cylinder capacitor,

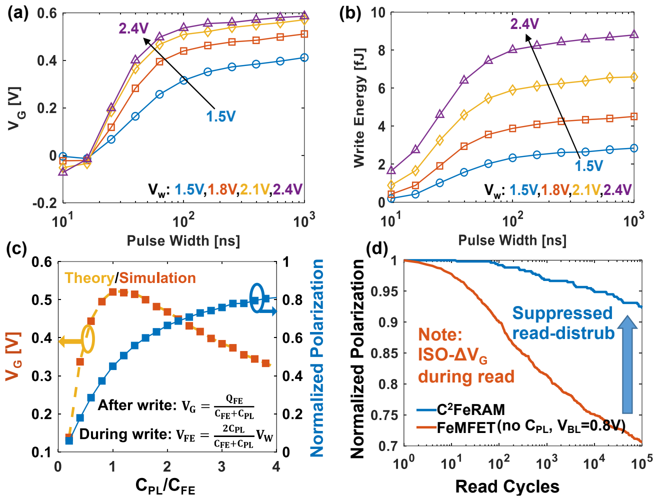

Fig. 4. C^{2}FeRAM optimization and simulation: (a), (b) internal voltage V_G and write energy vs pulse width; (c) dynamic range optimization; (d) improved read disturb.

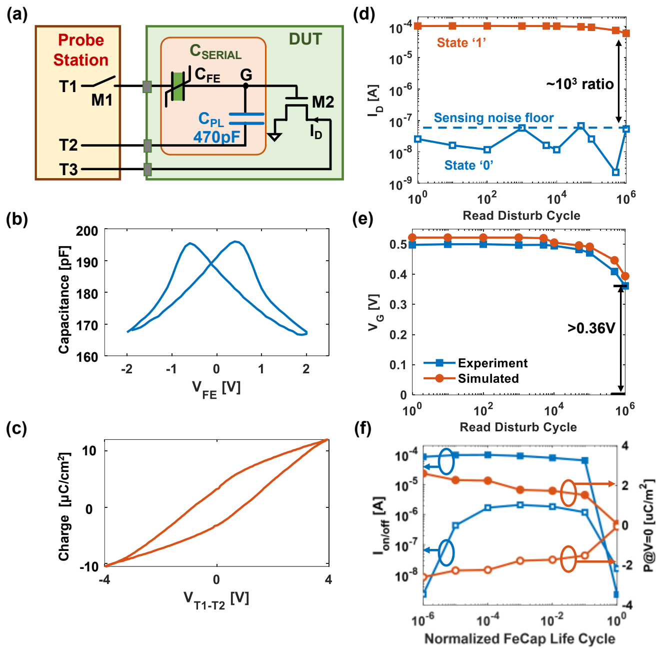

Fig. 5. Experiments: (a) setup with discreet C_{FE}, C_{PL} and MOSFET on a probe station; (b) C_{FE} vs V_{FE}; (c) Q_{SERIAL} vs V_{T1-T2}; (d) read-disturb measurement showing ~10^3 I_{ON}/I_{OFF} after 10^6 cycles; (e) Remnant V_G of state '1' degradation; (f) measured endurance as good as FeCAP.

could achieve higher density, as illustrated in Fig. 2(b), (c). Like [17], a possible brief process sequence of the two CAA NMOS FETs is illustrated in Fig. 2(d). Firstly, an MIM structure is formed. A hole is then etched on the MIM structure for the vertical channel. The channel, gate insulator, and gate electrode are deposited in the hole in sequential.

The compact 3D cell structure could lead to interference with adjacent cells due to coupling in a memory array. To alleviate the interference of coupling, the odd rows and even columns can be activated in a time-interleaved style. In one phase, once the odd PLs are activated, that adjacent even PLs will be grounded. In turn, the odd PLs and the even PLs will be grounded and activated in next phase, respectively.

In the C^{2}FeRAM, the FeRAM-style write operation is adopted. Additionally, the gate voltage stress of M$_2$ during programming can be lowered through PL by C_{PL} coupling. As

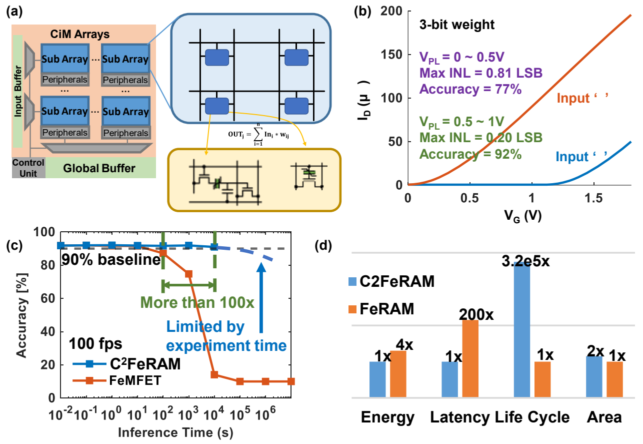

Fig. 6. VGG8 inference application benchmark for CIFAR-10 dataset: (a) benchmark architecture; (b) optimizing V_{PL} range for better linearity between I_D and V_{PL}; (c) CiM accuracy based on experimental results, C^2FeRAM reduces refresh rate more than 100x; (d) normalized energy, latency, and life cycle improvement with C^2FeRAM over FeRAM.

the gate stress is mitigated, the C^2FeRAM can achieve the endurance comparable to FeRAM. Fig. 3(a), (b) shows the low-V_G-stress high-endurance weight update during learning, including the biasing schemes and transient snapshots. Sub-10fF is a typical value of cell capacitance in 1y nm DRAM cell [18]. Therefore, the transient snapshots are simulated with the model in [19], with C_{PL}=5fF, 30nm x 30nm C_{FE} between 3fF and 7fF. While the conventional write scheme of FeRAM or FeMFET is applicable, the proposed write scheme achieves low-V_G-stress by applying an opposite voltage bias at the plateline PL. For example, to switch to positive polarization, $+V_W$ and $–V_W$ are applied to BL and PL, respectively, leading to sufficient write voltage across C_{FE} but a small M_2 gate stress voltage V_G. The row-level write is carried out by performing erasing (write ‘0’) and programming (write ‘1’) sequentially. In practice, the write voltage in BL and PL could be set as needed; C_{PL} could be optimized to achieve the maximum difference between settled V_G of states ‘1’ and ‘0’, as V_G is modulated by both the total capacitance at node G and the C_{PL}/C_{FE} ratio. For the unselected cells, the WS will be set to $-V_W$ so that C_{FE} will not be disturbed by the write operations to the selected rows.

For the read operation, the FeMFET-style non-destructive read is performed through M_2. Fig. 3(c)-(e) shows disturb-free read and CiM operations with ultra-low disturbance. With C^2FeRAM, a traditional crossbar sensing scheme is supported by setting V_{PL} as the input to access the cells in the selected rows, with the scheme in (c), waveforms in (d), and bitline currents for different accumulation results in (e). Excitingly, raising V_{PL} to set M_2 in linear region improves the linearity without the worry of C_{FE} state disturbance, as this does not add extra voltage stress to the floating C_{FE} with M_1 turned off. Practically, an optimized V_{PL} could be achieved for the balance between high computing accuracy and low power consumption during inference. To deal with leaky V_G, a restore could be performed as in FeRAM. For the CiM operation, the summation of cell currents within an array can be affected by the device-to-device variation. For binary cells, the impact of variation is insignificant. For MLC cells, a verify operation after writing the trained weights will effectively alleviate the impact of variation.

Write amplitude, pulse width, and the ratio between C_{FE} and C_{PL} could all affect the internal voltage V_G, which indicates the dynamic range of a C^2FeRAM memory cell. Fig. 4(a), (b) shows simulated fJ-level write energy and settled V_G vs 10-100ns write pulse duration, for a varying write amplitude of 1.5V-2.4V. A larger V_W and longer pulse width would result in larger V_G as well as greater write energy. Meanwhile, the ratio between C_{FE} and C_{PL} has a more complex relationship with V_G, where V_G is determined by both the C_{FE} polarization charge and the total capacitance of C_{FE} and C_{PL} in parallel. Under a given V_w, a larger C_{PL}/C_{FE} would result in a larger voltage across C_{FE} and thus a greater ferroelectric polarization, but the total capacitance of C_{PL} and C_{FE} would also be greater. Therefore, a maximized V_G could be achieved when C_{PL} and C_{FE} are nearly matched, as shown in Fig. 4(c). Fig. 4(d) shows the simulated CiM stability characteristics with matched C_{FE} and C_{PL}. Thanks to the floating C_{FE} scheme, even with over 10^5 times read, the polarization degradation is <7%. In contrast, the degradation of FeMFET after 10^5 times read is close to 30%.

III. Experimental Results

We assembled in-house discrete components to evaluate C^2FeRAM. Fig. 5(a) shows the experiment setting, with C_{FE} measured between 165pF and 195pF in Fig. 5(b), and the hysteric C_{FE} in series with a 470pF C_{PL} in Fig. 5(c). With V_W=2V, Fig. 5(d-e) shows that after 10^6 CiM read, the dynamic ‘0/1’ I_D sensing range with V_D=50mV is well around 10^3, and the remnant V_G for ‘1’ is still over 0.36V. In the experiment, sensing ‘0’ is limited by the noise floor sensitivity, and sensing ‘1’ is limited by M_2 I_D-V_G saturation. Moreover, Fig. 5(f) shows that the C^2FeRAM write endurance is as good as the standalone FeCap. Although the achieved write endurance is currently limited to 10^8 due to a large in-house device size above 2,500μm², it could be significantly improved with size scaling.

We have also evaluated the C^2FeRAM-based CiM arrays for VGG8 inference on the CIFAR-10 dataset with the architecture shown in Fig. 6(a). The simulation is carried out with NeuroSim [20]. Without disturbing the ferroelectric state, Fig. 6(b) shows that V_{PL} optimization in C^2FeRAM CiM achieves high linearity. Fig. 6(c) shows that, owing to the disturb-free capability, C^2FeRAM CiM successfully achieves <1% accuracy degradation after 10^6 CNN inferences, while the baseline FeMFET array degrades by 77%. In comparison with the FeRAM cache solution, Fig. 6(d) shows the energy, latency, and lifetime improvements up to 4x, 200x, and 3.2e5x, respectively, because C^2FeRAM does not need an extra recover operation after each inference cycle.

IV. Conclusions

This letter proposes C^2FeRAM, a 2T2C/cell CiM scheme that achieves disturb-free CiM and high write endurance comparable with FeRAM. The compact 3D cell structures are also proposed to alleviate the footprint overhead. Application-level benchmarks for VGG8 in CIFAR-10 dataset have demonstrated C^2FeRAM CiM achieving >100x inference time with <1% accuracy drop compared with FeMFET CiM, and 4x energy, 200x speed, and 3.2e5x life cycles over FeRAM cache.

REFERENCES

[1] A. I. Khan, A. Keshavarzi, and S. Datta, “The future of ferroelectric field-effect transistor technology,” Nat Electron, vol. 3, no. 10, Art. no. 10, Oct. 2020, doi: 10.1038/s41928-020-00492-7.

[2] A. Keshavarzi, K. Ni, W. Van Den Hoek, S. Datta, and A. Raychowdhury, “FerroElectronics for Edge Intelligence,” in IEEE Micro, vol. 40, no. 6, pp. 33-48, 1 Nov.-Dec. 2020, doi: 10.1109/MM.2020.3026667.

[3] M. I. Popovici, J. Bizindavyi, P. Favia, S. Clima, Md. Nur K. Alam, R.K. Ramachandran, A.M. Walke, U. Celano, A. Leonhardt, S. Mukherjee, O. Richard, A. Illiberi, M. Givens, R. Delhougne, J. Van Houdt, and G. Sankar Kar, "High performance La-doped HZO based ferroelectric capacitors by interfacial engineering," 2022 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2022, pp. 6.4.1-6.4.4, doi: 10.1109/IEDM45625.2022.10019525.

[4] A. A. Sharma, B. Doyle, H. J. Yoo, I-C. Tung, J. Kavalieros, M. V. Metz, M. Reshotko, P. Majhi, T. Brown-Heft, Y-J. Chen, and V. H. Le, "High Speed Memory Operation in Channel-Last, Back-gated Ferroelectric Transistors," 2020 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2020, pp. 18.5.1-18.5.4, doi: 10.1109/IEDM13553.2020.9371940.

[5] H. -L. Chiang, J.-F. Wang, K. -H. Lin, C.-H Nien, J. -J. Wu, K.-Y. Hsiang, C. -P. Chuu, Y.-W. Chen, X. W. Zhang, C. W. Liu, T. Wang, C. -C. Wang, M.-H. Chang, C.-S. Chang, and T. C. Chen, "Interfacial-Layer Design for Hf1-xZrxO2-Based FTJ Devices: From Atom to Array," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 361-362, doi: 10.1109/VLSITechnologyandCir46769.2022.9830462.

[6] D. Takashima, “Overview of FeRAMs: Trends and perspectives,” in 2011 11th Annual Non-Volatile Memory Technology Symposium Proceeding, 2011, pp. 1–6, doi: 10.1109/NVMTS.2011.6137107.

[7] M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck, J. Ocker, M. Noack, J. Müller, P. Polakowski, J. Schreiter, S. Beyer, T. Mikolajick, and B. Rice, “A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs,” in 2016 IEEE International Electron Devices Meeting (IEDM), 2016, p. 11.5.1-11.5.4, doi: 10.1109/IEDM.2016.7838397.

[8] E. Yurchuk, J. Müller, S. Müller, J. Paul, M. Peši´c, R. van Bentum, U. Schroeder, and T. Mikolajick, “Charge-Trapping Phenomena in HfO2-Based FeFET-Type Nonvolatile Memories,” in IEEE Transactions on Electron Devices, vol. 63, no. 9, pp. 3501-3507, Sept. 2016, doi: 10.1109/TED.2016.2588439.

[9] K. Ni, P. Sharma, J. Zhang, M. Jerry, J. A. Smith, K. Tapily, R. Clark, S. Mahapatra, and S. Datta, “Critical Role of Interlayer in Hf0.5Zr0.5O2 Ferroelectric FET Nonvolatile Memory Performance,” in IEEE Transactions on Electron Devices, vol. 65, no. 6, pp. 2461-2469, June 2018, doi: 10.1109/TED.2018.2829122.

[10] N. Gong and T. -P. Ma, “A Study of Endurance Issues in HfO2-Based Ferroelectric Field Effect Transistors: Charge Trapping and Trap Generation,” in IEEE Electron Device Letters, vol. 39, no. 1, pp. 15-18, Jan. 2018, doi: 10.1109/LED.2017.2776263.

[11] H. Zhou, J. Ocker, A. Padovani, M. Pesic, M. Trentzsch, S. Dünkel, H. Mulaosmanovic, S. Slesazeck, L. Larcher, S. Beyer, S. Müller, and T. Mikolajick, “Application and Benefits of Target Programming Algorithms for Ferroelectric HfO2 Transistors,” 2020 IEEE International Electron Devices Meeting (IEDM), 2020, pp. 18.6.1-18.6.4, doi: 10.1109/IEDM13553.2020.9371975.

[12] T. Ali, P. Polakowski, S. Riedel, T. Büttner, T. Kämpfe, M. Rudolph, B. Pätzold, K. Seidel, D. Löhr, R. Hoffmann, M. Czernohorsky, K. Seidel, D. Löhr, R. Hoffmann, M. Czernohorsky, K. Kühnel, P. Steinke, J. Calvo, K. Zimmermann, and J. Müller, “High Endurance Ferroelectric Hafnium Oxide-Based FeFET Memory Without Retention Penalty,” in IEEE Transactions on Electron Devices, vol. 65, no. 9, pp. 3769-3774, Sept. 2018, doi: 10.1109/TED.2018.2856818.

[13] K. Ni, J. A. Smith, B. Grisafe, T. Rakshit, B. Obradovic, J. A. Kittl, M. Rodder, and S. Datta, “SoC Logic Compatible Multi-Bit FeMFET Weight Cell for Neuromorphic Applications,” in 2018 IEEE International Electron Devices Meeting (IEDM), Dec. 2018, p. 13.2.1-13.2.4, doi: 10.1109/IEDM.2018.8614496.

[14] S. Slesazeck, V. Havel, E. Breyer, H. Mulaosmanovic, M. Hoffmann, B. Max, S. Duenkel, and T. Mikolajick, “Uniting The Trinity of Ferroelectric HfO2 Memory Devices in a Single Memory Cell,” in 2019 IEEE 11th International Memory Workshop (IMW), 2019, pp. 1–4, doi: 10.1109/IMW.2019.8739742.

[15] S.-C. Chang, N. Haratipour, S. Shivaraman, T. L Brown-Heft, J. Peck, C.-C. Lin, I-C. Tung, D. R Merrill, H. Liu, C.-Y. Lin, F. Hamzaoglu, M. V Metz, I. A Young, J. Kavalieros, and U. E Avci, “Anti-ferroelectric HfxZr1-xO2 Capacitors for High-density 3-D Embedded-DRAM,” in 2020 IEEE International Electron Devices Meeting (IEDM), 2020, p. 28.1.1-28.1.4, doi: 10.1109/IEDM13553.2020.9372011.

[16] M. Sung, K. Rho, J. Kim, J. Cheon, K. Choi, D. Kim, H. Em, G. Park, J. Woo. Y. Lee, J. Ko, M. Kim, G. Lee, S. W. Ryu, D. S. Sheen, Y. Joo, S. Kim, C. H. Cho, M-H. Na, and J. Kim, "Low Voltage and High Speed 1Xnm 1T1C FE-RAM with Ultra-Thin 5nm HZO," 2021 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2021, pp. 33.3.1-33.3.4, doi: 10.1109/IEDM19574.2021.9720545.

[17] X. Duan, K. Huang, J. Feng, J. Niu, H. Qin, S. Yin, G. Jiao, D. Leonelli，X. Zhao, W. Jing, Z. Wang, Q. Chen, X. Chuai, C. Lu, W. Wang, G. Yang, D. Geng, L. Li, and M. Liu, “Novel Vertical Channel-All-Around(CAA) IGZO FETs for 2T0C DRAM with High Density beyond 4F^2 by Monolithic Stacking,” 2021 IEEE International Electron Devices Meeting (IEDM), 2021, pp. 10.5.1-10.5.4, doi: 10.1109/IEDM19574.2021.9720682.

[18] A. Spessot and H. Oh, “1T-1C Dynamic Random Access Memory Status, Challenges, and Prospects,” in IEEE Transactions on Electron Devices, vol. 67, no. 4, pp. 1382-1393, April 2020, doi: 10.1109/TED.2020.2963911.

[19] S. Deng, G. Yin, W. Chakraborty. S. Dutta, S. Datta, X. Li, K. Ni, “A Comprehensive Model for Ferroelectric FET Capturing the Key Behaviors: Scalability, Variation, Stochasticity, and Accumulation,” in 2020 IEEE Symposium on VLSI Technology, Jun. 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265014.

[20] 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 2019 IEEE International Electron Devices Meeting (IEDM), 2019, p. 32.5.1-32.5.4, doi: 10.1109/IEDM19573.2019.8993491.