Experimental Demonstration of a Ferroelectric HfO_2-Based Content Addressable Memory Cell

Ava J. Tan, Korok Chatterjee, Jiuren Zhou, Daewoong Kwon, Yu-Hung Liao, Suraj Cheema, Chenming Hu, and Sayeef Salahuddin

***Abstract*— In this work, we present an experimental demonstration of a content addressable memory (CAM) cell based on ferroelectric HfO_2 field effect transistors (FeFETs). Our proposed ferroelectric CAM (FeCAM) utilizes a CMOS-compatible ferroelectric material, hafnium zirconium oxide (HZO), as the gate dielectric. We discuss operation of the FeCAM cell and propose a suitable architecture to realize in-memory computation as well as single clock cycle content-driven search. In addition, the HZO FeFET is analyzed for its intrinsic memory characteristic, and design considerations are identified for improving device and therefore projected system-level performance. Our results indicate that FeCAM is well-suited to accommodate demanding modern computational needs by sealing the gaps between conventional memory, logic, and continued device scaling.**

***Index Terms*— Ferroelectrics, hafnium zirconium oxide, content addressable memory, ferroelectric memory, FeFET, logic-in-memory.**

I. Introduction

As COMPUTING continues to evolve with an ever-increasing emphasis on data processing for cognitive functions such as learning, recognition, and synthesis, so do the requirements on underlying hardware. Though dimensional scaling has driven the semiconductor industry since its inception, innovation in memory technologies has become perhaps the most crucial aspect in recent years [1]. Solutions to tackle the "von-Neumann bottleneck" via specialized ASICs [2], [3] and emerging memory technologies [4]–[6] are actively being explored. To minimize latency and additional heat generated by data movement within the traditional Von-Neumann architecture, a wiser approach may be to distribute memory elements alongside logic units. This approach, when combined with nonvolatile memories (NVMs), enables logic-in-memory functionality with substantially lowered energy consumption.

Ternary content addressable memories (TCAMs) are specialized memory solutions that allow for fast parallel lookup/search of data, but typically at the expense of power consumption. Memory cells are searched by *content*, rather than by address, to determine the location and existence of a partial or exact match. Traditionally used for high-speed data processing for IP filters, network routers, etc., TCAMs are recently being explored for neural network acceleration and pattern-matching in data-intensive applications [7].

Naturally, research interest in TCAMs has gravitated towards the development of emerging memories to improve upon existing CAMs, which typically have low area efficiencies and utilize many transistors [8]–[11]. Of these solutions, ferroelectric FET solutions stand out as they provide nonvolatility, reconfigurability, and the ability to achieve high density due to their scalability and ease of integration with current CMOS processes [12], [13]. Furthermore, they have been shown to successfully implement logic functions using a small number of transistors [11]. When compared against magnetic tunnel junction (MTJ), resistive RAM (ReRAM), and even SRAM-based TCAMs, FeFET-based TCAMs show superior array-level performance characteristics. A 4T-2FeFET cell design, for example, occupies 58% of the equivalent 16T SRAM-based TCAM design, and a 64-row array boasts an energy-delay product which is 7.5X and 149X better than similar MTJ- and ReRAM-based designs, respectively [14]. Additional improvement on these metrics are anticipated from the design investigated in this work as our fundamental CAM cell design requires just two FeFETs acting as both the selector and memory element.

In this work, we experimentally demonstrate a HZO-based two-FeFET CAM cell for parallel search and in-memory computing. Our CAM cell shows competitive programming characteristics, achieving a 0.54V memory window with ±3.5V microsecond-order write pulses and 10-year projected retention of the XNOR function at 85 °C.

II. Experiment

Fig. 1 details the CAM cell geometry, the gate oxide stack and corresponding TEM. After thinning down the device active layer on 150 mm SOI substrates through a combination of wet oxidation and HF etching to 30 nm in thickness, the active regions are defined and the ferroelectric gate oxide is deposited via atomic layer deposition (ALD). Once the tungsten gates are defined, the source/drain regions are implanted with As+ ions, thereby completing the self-aligned gate-first process. The activation of the implanted source/drain regions and the crystallization of the ferroelectric HfO_2 layer are carried out simultaneously with one rapid thermal annealing (RTA) step

Manuscript received December 13, 2019; revised December 23, 2019; accepted December 28, 2019. Date of publication December 31, 2019; date of current version January 27, 2020. This work was supported in part by the Energy Efficient Computing: from Devices to Architecture (E2CDA) Energy Efficient Learning Machines (ENIGMA) Project and in part by the University of California Multicampus Research Programs and Initiatives (UC MRPI) Initiative. The work of Ava J. Tan was supported by the National Defense Science and Engineering Graduate Fellowship. The review of this letter was arranged by Editor D. Ha. *(Corresponding author: Ava J. Tan.)*

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: ava@eecs.berkeley.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LED.2019.2963300

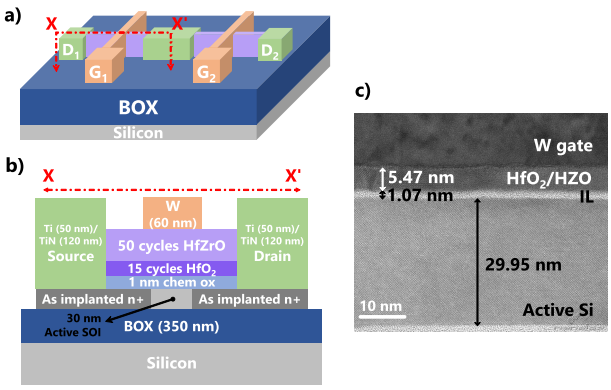

Fig. 1. (a) Tilted 3D view of CAM cell schematic on thin SOI substrate. (d) 2D cross-section across channel of one FeFET, showing the gate stack composition and active device layer. (c) TEM (tunneling electron microscopy) image of the gate stack of a fabricated CAM cell device with W/L = 1/1 μm.

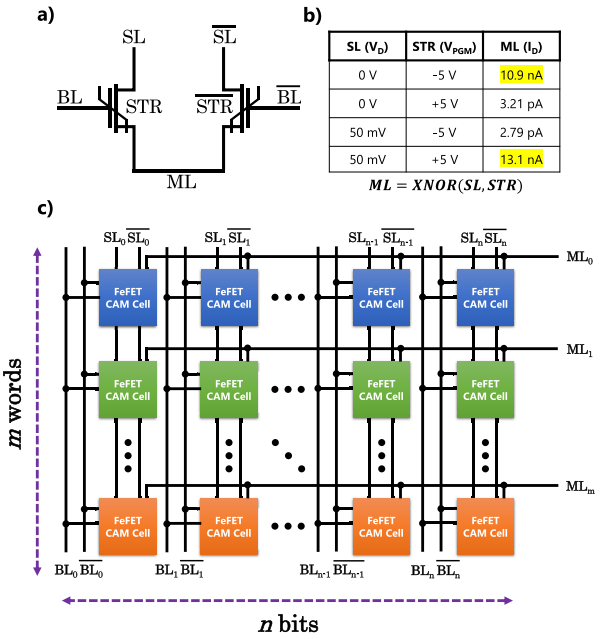

Fig. 2. (a) Circuit representation of a single CAM cell, which consists of two FeFETs with their sources shorted together. (b) The operation of the single CAM cell can be described by the XNOR Boolean equation and associated truth table. (c) Proposed CAM cell array of dimension $m \times n$ for parallel searching. Peripheral circuitry (not shown) senses and compares current levels on the MLs to determine closest match to input data.

at 500 °C for 30 seconds. The devices are then isolated with a thick 350 nm interlayer dielectric of SiO_2, and an additional metal deposition/etch step is utilized to make contact to the source/drain/gate regions.

Based on Fig. 1c, the HfO_2 layer is roughly 1.3 nm thick, and the HZO layer is 4.2 nm thick. The ratio of cycles of Hf to Zr is chosen as 4:1 as previously investigated in [15].

III. Results and Discussion

The intrinsic FeCAM demonstrated in this work consists of two FeFETs with their sources tied together, as shown in Fig. 2a. Though similar architectures have been investigated using other emerging memories (i.e, ReRAM [16]), the FeFET boasts an additional advantage of combining the memory and selector element within a single device, thereby significantly reducing area footprint.

Within a single FeCAM cell, a bit and its complement are stored by applying voltage pulses of opposite polarity to the gates of each respective transistor. During the search phase, the select lines for each FeCAM cell are probed with the corresponding data to be matched. In the case of a mismatch, neither transistor in the cell will conduct high current through the match line, as one FeFET will be in the OFF state while the other will be programmed in the "0" or low-current state. In the case of a match, either the STR or $\overline{STR}$ FeFET will conduct high current to the match line upon select line activation. In this manner, the FeCAM cell implements the XNOR operation, as outlined in Fig. 2b. Fig. 2b also contains experimental data of match line current readout for the cases of a match or a mismatch. Note that the same cell can be utilized as a ternary FeCAM by writing a "1" or high-current state to both transistors to force a match when a "don't care" bit is required.

A full example FeCAM array architecture is shown in Fig. 2c. Notably, such an array could be constructed in 3D following a BEOL process with a polysilicon channel, for example. n total FeCAM cells cascaded together to share a match line form a word in an array of m words, thus enabling single clock cycle determination of a match, which enables Hamming distance calculations for cognitive computing schemes (e.g, hyperdimensional computing [17]). Additionally, the FeCAM cell allows parallel search in hardware, thus speeding up deep neural network/convolutional neural network implementation. Further, an FeCAM crossbar array permits distributed weight updates and multiply-accumulate operations to be performed in-memory, thereby drastically reducing energy and time associated with shuttling weight data to the CPU.

To understand projected FeCAM performance we characterize representative single FeFET devices and well as the FeFETs in a fabricated FeCAM cell. Figs. 3a and 3b illustrate the change in $\pm V_T$, defined at a constant current criteria of $I_D = 10$ pA, at varying programming voltages and durations, for a single FeFET memory device. The inset of Fig. 3a shows a typical $I_D V_G$ after applying programming pulses of $+5.5V$, 25 ns, and $-5.5V$, 35 ns. Applying a voltage pulse of positive polarity shifts the V_T to the left, and applying a voltage pulse of negative polarity shifts the V_T to the right. The memory window widens as duration and/or amplitude is increased until charge trapping within the oxide inhibits further shifts in $\pm V_T$. The same program/erase data for the FeFETs which comprise a CAM cell are shown in Figs. 3c and 3d, with the inset of Fig. 3c again showing both CAM cell devices' $I_D V_G$s after applying pulses of $+5.5V$, 25 ns, and $-5.5V$, 35 ns.

Though Fig. 3 illustrates the measurable PGM/ERS speeds of the FeFETs, it should be noted that at the lower limit of pulse speeds tested, significant oscillation of the gate voltage pulse was observed as shown in Fig. 4. Furthermore the peak voltage values seen by the device were degraded as the pulse widths were reduced below 50 ns. These effects are due to the non-negligible contribution of extrinsic device impedances at shorter programming pulse durations. For this reason, long channel devices ($L_g \geq 250$ nm) were used to enable robust PGM/ERS testing, as going down to very small L_g ($\sim$ 50 nm) requires much thinner SOI, thereby significantly increasing series resistance and requiring longer

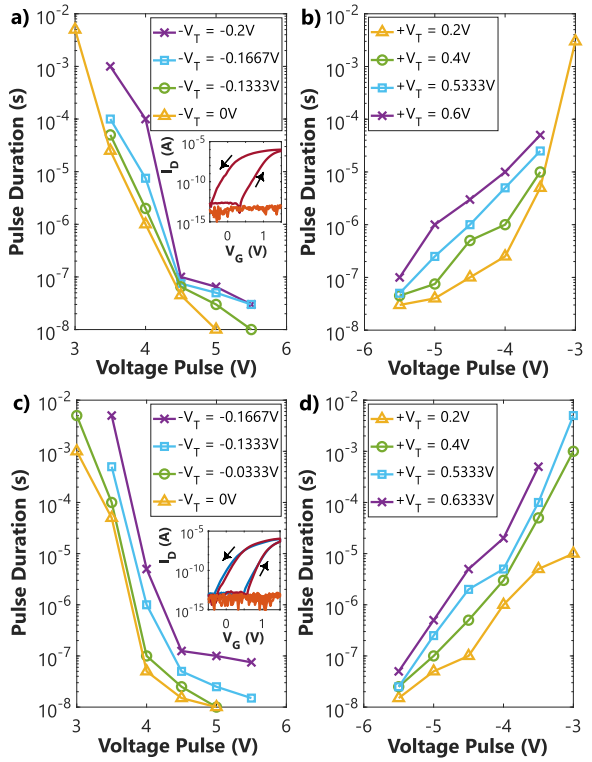

Fig. 3. (a) Curves of constant $-V_T$ obtained at varying pulse durations/amplitudes to illustrate the voltage/time tradeoff for a single FeFET with W/L = 500/250 nm. Inset includes $I_D V_G$ characteristics. (b) Curves of constant $+V_T$ obtained at varying pulse durations/amplitudes for the single FeFET. (c) Curves of constant $-V_T$ obtained at varying pulse durations/amplitudes to illustrate the voltage/time tradeoff for the CAM FeFETs with W/L = 1/1 μm. Inset includes $I_D V_G$ characteristics. (d) Curves of constant $+V_T$ obtained at varying pulse durations/amplitudes for the CAM FeFET.

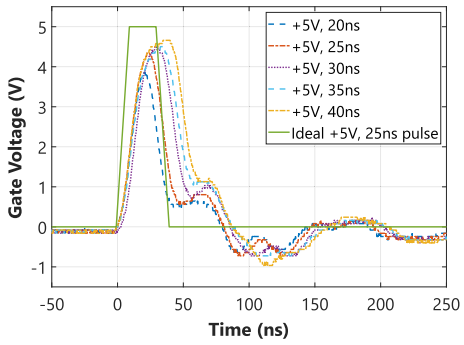

Fig. 4. The applied ideal pulse of +5V, 25 ns with 10 ns rise/fall times plotted against the actual voltage waveforms measured at the gate of a transistor with W/L = 500/250 nm.

programming pulses. On the other hand, achieving a full self-aligned, gate-last process is beyond the capabilities of most university laboratories.

Retention testing of the FeFET CAM cell is carried out at 85 °C, as shown in Fig. 5. In this particular test, the matchline current (I_{ML}) from a CAM cell is measured as function of time for all possible states (a miss, a match, or a forced match in the case of a "don't care" query). After 24 hours, the I_{on}/I_{off} ratio degrades roughly 30%, and there remains two orders of magnitude of current level separation after 10 years.

Initial endurance testing (Fig. 6) shows that a gradual shrinking of the memory window occurs during the FeFET's

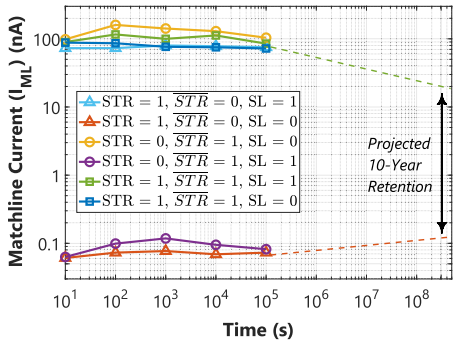

Fig. 5. Retention testing at 85 °C for a CAM cell with device dimensions W/L = 1/1 μm. PGM/ERS conditions are −5.5V, 35 ns and +5.5V, 25 ns respectively.

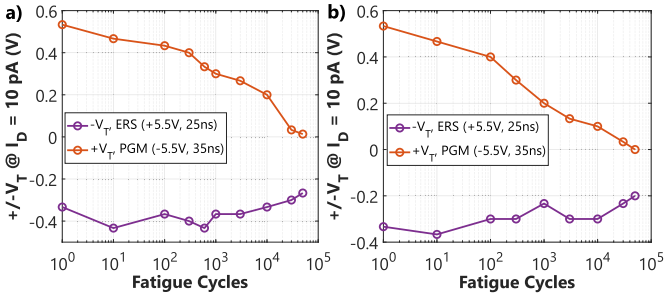

Fig. 6. (a) MW as a function of fatigue cycles for a single device with W/L = 500/250 nm at a constant current criteria of 10 pA. (b) MW as a function of fatigue cycles for a CAM cell FeFET with W/L = 1/1 μm at a constant current criteria of 10 pA.

lifetime. Alternating pulses of +5.5V, 25 ns and −5.5V, 35 ns with a wait time of 1.5 seconds between each pulse are applied at the gate of the device, and $I_D V_G$'s are taken periodically throughout the endurance test to record the evolution of $+/-V_T$ at 10 pA. Fig. 6a shows endurance cycling for a single FeFET memory and Fig. 6b shows endurance cycling for the CAM cell FeFET.

Previous reports, typically on much thicker FE layers, suggest comparable endurance properties on the order of $10^4 - 10^5$ cycles [18], [19]. We suspect that charge trapping/interface charge generation at the porous ~1 nm chemical oxide interface limits our device endurance. Though a thinner IL was used to in order to scale the FeFET gate stack, additional engineering will be necessary to improve the chemical oxide quality to reduce trap-assisted tunneling. Most existing reports on FE-HfO_2 memories utilize thicker SiO_2 or SiON IL's [15], [20], [21] to improve the endurance, but several recent works have established the importance of further engineering the IL for improved breakdown properties, operating voltage, retention time and cycling endurance [22]–[25].

IV. Conclusion

We have demonstrated an SOI FeCAM cell based ferroelectric HZO which shows large memory window characteristics at competitive programming voltages and speeds. Complete cell operation and elevated temperature retention studies with ternary in-memory XNOR functionality were demonstrated. Simple cell structure, scalability, and fast read speed essentially determined by the channel mobility show that FeCAMs are a promising memory technology for both emerging computing and current computing needs.

REFERENCES

[1] S. De Boer, "Memory technology: The core to enable future computing systems," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 3–6, doi: 10.1109/vlsit.2018.8510707.

[2] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, May 2014, doi: 10.1109/JPROC.2014.2313565.

[3] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014, doi: 10.1126/science.1254642.

[4] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. L. Gallo, K. Moon, J. Woo, H. Hwang, and Y. Leblebici, "Neuromorphic computing using non-volatile memory," *Adv. Phys., X*, vol. 2, no. 1, pp. 89–124, 2017, doi: 10.1080/23746149.2016.1259585.

[5] S. Mueller, "Ferroelectric HfO_2 and its impact on the memory landscape," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2018, pp. 1–4, doi: 10.1109/imw.2018.8388831.

[6] N. K. Upadhyay, H. Jiang, Z. Wang, S. Asapu, Q. Xia, and J. J. Yang, "Emerging memory devices for neuromorphic computing," *Adv. Mater. Technol.*, vol. 4, no. 4, Apr. 2019, Art. no. 1800589, doi: 10.1002/admt.201800589.

[7] M. Imani, D. Peroni, Y. Kim, A. Rahimi, and T. Rosing, "Efficient neural network acceleration on GPGPU using content addressable memory," in *Proc. Design, Autom. Test Eur. Conf. Exhibit. (DATE)*, Mar. 2017, pp. 1026–1031, doi: 10.23919/date.2017.7927141.

[8] R. Yang, H. Li, K. K. H. Smithe, T. R. Kim, K. Okabe, E. Pop, J. A. Fan, and H.-S. P. Wong, "Ternary content-addressable memory with MoS_2 transistors for massively parallel data search," *Nature Electron.*, vol. 2, no. 3, pp. 108–114, Mar. 2019, doi: 10.1038/s41928-019-0220-7.

[9] B. Chen, Y. Zhang, W. Liu, S. Xu, R. Cheng, R. Zhang, and Y. Zhao, "Ge–based asymmetric RRAM enable $8F^2$ content addressable memory," *IEEE Electron Device Lett.*, vol. 39, no. 9, pp. 1294–1297, Sep. 2018, doi: 10.1109/led.2018.2856537.

[10] M. Imani, D. Peroni, A. Rahimi, and T. S. Rosing, "Resistive CAM acceleration for tunable approximate computing," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 2, pp. 271–280, Apr. 2019, doi: 10.1109/tetc.2016.2642057.

[11] E. T. Breyer, H. Mulaosmanovic, S. Slesazeck, T. Mikolajick, and T. Mikolajick, "Demonstration of versatile nonvolatile logic gates in 28 nm HKMG FeFET technology," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5, doi: 10.1109/iscas.2018.8351408.

[12] A. Aziz, E. T. Breyer, A. Chen, X. Chen, S. Datta, S. K. Gupta, M. Hoffmann, X. S. Hu, A. Ionescu, M. Jerry, T. Mikolajick, H. Mulaosmanovic, K. Ni, M. Niemier, I. O'Connor, A. Saha, S. Slesazeck, S. K. Thirumala, and X. Yin, "Computing with ferroelectric FETs: Devices, models, systems, and applications," in *Proc. Design, Autom. Test Eur. Conf. Exhibit. (DATE)*, Mar. 2018, pp. 1289–1298, doi: 10.23919/date.2018.8342213.

[13] X. Yin, A. Aziz, J. Nahas, S. Datta, S. Gupta, M. Niemier, and X. S. Hu, "Exploiting ferroelectric FETs for low-power non-volatile logic-in-memory circuits," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2016, pp. 1–8, doi: 10.1145/2966986.2967037.

[14] X. Yin, M. Niemier, and X. S. Hu, "Design and benchmarking of ferroelectric FET based TCAM," in *Proc. Conf. Design, Autom. Test Eur. (DATE)*, Leuven, Belgium: European Design and Automation Association, Mar. 2017, pp. 1448–1453. [Online]. Available: http://dl.acm.org/citation.cfm?id=3130379.3130719

[15] K. Chatterjee, S. Kim, G. Karbasian, A. J. Tan, A. K. Yadav, A. I. Khan, C. Hu, and S. Salahuddin, "Self-aligned, gate last, FDSOI, ferroelectric gate memory device with 5.5-nm $Hf_{0.8}Zr_{0.2}O_2$, high endurance and breakdown recovery," *IEEE Electron Device Lett.*, vol. 38, no. 10, pp. 1379–1382, Oct. 2017, doi: 10.1109/led.2017.2748992 .

[16] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, Y. Hayakawa, K. Tsuji, S. Yoneda, A. Himeno, K. Shimakawa, T. Takagi, T. Mikawa, and K. Aono, "An 8 MB multi-layered cross-point ReRAM macro with 443 MB/s write throughput," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 178–185, Jan. 2013, doi: 10.1109/jssc.2012.2215121.

[17] P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cognit. Comput.*, vol. 1, no. 2, pp. 139–159, Jun. 2009, doi: 10.1007/s12559-009-9009-8.

[18] J. Müller, T. S. Böscke, S. Müller, E. Yurchuk, P. Polakowski, J. Paul, D. Martin, T. Schenk, K. Khullar, A. Kersch, W. Weinreich, S. Riedel, K. Seidel, A. Kumar, T. M. Arruda, S. V. Kalinin, T. Schlösser, R. Boschke, R. van Bentum, U. Schröder, and T. Mikolajick, "Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories," in *IEDM Tech. Dig.*, Dec. 2013, pp. 10.8.1–10.8.4, doi: 10.1109/iedm.2013.6724605.

[19] S. Dünkel, M. Trentzsch, R. Richter, P. Moll, C. Fuchs, O. Gehring, M. Majer, S. Wittek, B. Müller, T. Melde, H. Mulaosmanovic, S. Slesazeck, S. Müller, J. Ocker, M. Noack, D. Löhr, P. Polakowski, J. Müller, T. Mikolajick, J. Höntschel, B. Rice, J. Pellerin, and S. Beyer, "A FeFET based super-low-power ultra-fast embedded NVM technology for 22 nm FDSOI and beyond," in *IEDM Tech. Dig.*, Dec. 2017, pp. 19.7.1–19.7.4, doi: 10.1109/iedm.2017.8268425.

[20] H. Mulaosmanovic, J. Ocker, S. Müller, U. Schroeder, J. Müller, P. Polakowski, S. Flachowsky, R. van Bentum, T. Mikolajick, and S. Slesazeck, "Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors," *ACS Appl. Mater. Interfaces*, vol. 9, no. 4, pp. 3792–3798, 2017, doi: 10.1021/acsami.6b13866.

[21] E. Yurchuk, J. Müller, S. Müller, J. Paul, M. Pešic, R. van Bentum, U. Schroeder, and T. Mikolajick, "Charge-trapping phenomena in HfO_2-based FeFET-type nonvolatile memories," *IEEE Trans. Electron Devices*, vol. 63, no. 9, pp. 3501–3507, Sep. 2016, doi: 10.1109/TED.2016.2588439.

[22] A. J. Tan, A. K. Yadav, K. Chatterjee, D. Kwon, S. Kim, C. Hu, and S. Salahuddin, "A nitrided interfacial oxide for interface state improvement in hafnium zirconium oxide-based ferroelectric transistor technology," *IEEE Electron Device Lett.*, vol. 39, no. 1, pp. 95–98, Jan. 2018, doi: 10.1109/LED.2017.2772791.

[23] 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 $Hf_{0.5}Zr_{0.5}O_2$ferroelectric FET nonvolatile memory performance," *IEEE Trans. Electron Devices*, vol. 65, no. 6, pp. 2461–2469, Jun. 2018, doi: 10.1109/TED.2018.2829122.

[24] T. Ali, P. Polakowski, S. Riedel, T. Büttner, T. Kämpfe, M. Rudolph, B. Pätzold, K. Seidel, D. Lähr, 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," *IEEE Trans. Electron Devices*, vol. 65, no. 9, pp. 3769–3774, Sep. 2018, doi: 10.1109/ted.2018.2856818.

[25] M. Pešiü, V. D. Lecce, D. Pramanik, and L. Larcher, "Multiscale modeling of ferroelectric memories: Insights into performances and reliability," in *Proc. Int. Conf. Simulation Semiconductor Process. Devices (SISPAD)*, Sep. 2018, pp. 111–114, doi: 10.1109/SISPAD.2018.8551722.