

# Logic Compatible High-Performance Ferroelectric Transistor Memory

Sourav Dutta<sup>1\*</sup>, Huacheng Ye<sup>1</sup>, Abhishek Khanna<sup>1</sup>, Yuan-Chun Luo<sup>2</sup>, Lillian Pentecost<sup>3</sup>, Akif A. Khandker<sup>1</sup>, Wriddhi Chakraborty<sup>1</sup>, Gu-Yeon Wei<sup>3</sup>, David Brooks<sup>3</sup>, Michael Niemier<sup>5</sup>, Xiaobo Sharon Hu<sup>5</sup>, Shimeng Yu<sup>2</sup>, Kai Ni<sup>4</sup> and Suman Datta<sup>1\*</sup>

<sup>1</sup>Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

<sup>2</sup>School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

<sup>3</sup>School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138

<sup>4</sup>Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623 USA

<sup>5</sup>Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

\* Corresponding author: [sdutta4@nd.edu](mailto:sdutta4@nd.edu), [sdatta@nd.edu](mailto:sdatta@nd.edu)

## ABSTRACT

Silicon ferroelectric field-effect transistors (FeFETs) with low-k interfacial layer (IL) between ferroelectric gate stack and silicon channel suffers from high write voltage, limited write endurance and large read-after-write latency due to early IL breakdown and charge trapping and detrapping at the interface. We demonstrate low voltage, high speed memory operation with high write endurance using an IL-free back-end-of-line (BEOL) compatible FeFET. We fabricate IL-free FeFETs with 28nm channel length and 126nm width under a thermal budget  $<400^{\circ}\text{C}$  by integrating 5nm thick  $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$  gate stack with amorphous Indium Tungsten Oxide (IWO) semiconductor channel. We report 1.2V memory window and read current window of  $10^5$  for program and erase, write latency of 20ns with  $\pm 2\text{V}$  write pulses, read-after-write latency  $<200\text{ns}$ , write endurance cycles exceeding  $5 \times 10^{10}$  and 2-bit/cell programming capability. Array-level analysis establishes IL-free BEOL FeFET as a promising candidate for logic-compatible high-performance on-chip buffer memory and multi-bit weight cell for compute-in-memory accelerators.

**KEYWORD:** Ferroelectric memory, FeFET, interfacial layer (IL), logic compatible, BEOL, monolithic 3D, HZO, IWO, endurance, multi-bit per cell, global buffer, compute-in-memory.

## INTRODUCTION

The paradigm of computing is constantly evolving to enable the vision of edge artificial intelligence (e-AI) where trillions of connected smart devices are pervasively being integrated into our daily life. These devices will constantly measure physical parameters and process them locally to provide actionable intelligence in real-time [1]. Data-centric local computing at the point of data acquisition can minimize bandwidth and latency issues associated with communication and enable critical applications that require sensing, processing and feedback in rapid succession. Today's primary candidate for on-chip memory - static random-access memory (SRAM) is volatile, suffers from large footprint area ( $120\text{-}150F^2$ ) and high standby leakage power. The embedded dynamic random-access memory (eDRAM) is also volatile and requires frequent refresh, thus incurring high power dissipation. Even with the continued advancement in transistor scaling, leading to improvement in processor speed and energy-efficiency, the performance of the overall system is limited by the extraordinary volume of data traffic between the processor and the off-chip memory. New architectural innovations such as compute-in-memory (CIM) promises an advantage in energy and latency by moving compute closer to the data residing in the memory block. The holy grail of CIM is high bandwidth, low-latency access to on-chip embedded non-volatile memory (eNVM) that is programmable with logic-compatible voltages, has the footprint of a single transistor, exhibits infinite endurance and is BEOL compatible for integrating on top of high-performance advanced logic to realize the true potential of M3D-CIM[2], [3].

Harnessing the polar orthorhombic phase  $\text{Pca}_2\text{I}$  in thin films of doped Hafnium Oxide ( $\text{HfO}_2$ ), one can construct a one transistor non-volatile ferroelectric memory (FeFET) that can serve as a logic-compatible, high-performance embedded non-volatile memory [2]. Recent demonstrations of

robust memory operation in FeFETs with high density integration in relatively scaled CMOS technology nodes (28nm planar bulk [4] and 22nm FD-SOI [5]) has provided legitimacy of FeFET as a contender for eNVM with best-in-class energy efficiency, latency and footprint area for CIM applications standing next to spin-transfer-torque magnetic random-access memory (STT-MRAM), resistive random-access memory (RRAM) and phase change memory (PCM) [2], [6]–[8]. However, FeFETs still face several key roadblocks such as logic-compatible programming, write endurance, read-after-write latency and BEOL-compatible low temperature processing [9]–[13] for integration in the BEOL.

Fig. 1(a) shows the schematic of a silicon FeFET with an undesired interfacial layer (IL) with low-dielectric constant ( $k=3.9$  for  $\text{SiO}_2$ ) between the silicon channel and ferroelectric HZO gate stack. The limited write endurance in the conventional FeFET predominantly arises due to the presence of the low- $k$  IL. As elucidated in the energy band diagram (see supplementary section **S1** for details) in Fig. 1(c, i), during the write operation of a conventional FeFET, most of the programming voltage in the gate stack drops across the IL which may reach as high as  $E_{\text{IL}}=10\text{MV/cm}$  [2], [14]. The high electric-field stresses the IL along with continuous charge trapping and trap generation at the either the interface between the IL and silicon or the FE-IL interface that tend to limit the write endurance in the range of  $10^4$ – $10^6$  [9]–[12]. It is to be noted that, the ferroelectric HZO in a stand-alone metal-ferroelectric-metal (MFM) capacitor withstands a write endurance of over  $10^{10}$  cycles (see supplementary section **S2** for endurance measurements on MFM capacitors) [14], [15]. Thus, improving the write endurance of FeFET requires us to either re-explore the known techniques of IL engineering in HKMG technology such as reducing the IL thickness through scavenging [16] and increasing the dielectric constant of IL [17], or re-imagine

novel ways to build an IL-free FeFET with high write endurance. The latter can be achieved by fabricating a back-gated, channel-last FeFET where an oxide semiconductor channel is grown directly on top of the ferroelectric HZO gate oxide, analogous to the fabrication of thin-film transistors (TFTs) used commercially in flat-panel display application [8], [18]–[22]. Fig. 1(b) shows a schematic of an IL-free FeFET along with the calculated band diagram in Fig. 1(c, ii). The absence of IL allows lowering the write voltage to logic-compatible voltage levels and significantly reduces the charge trapping, and trap generation at the interface between FE and channel, thereby improving the write endurance.

The absence of IL also reduces the read-after-write latency of the FeFET which is predominantly due to electron and hole trapping at the interface between FE-IL and IL-silicon, respectively. Prior work on conventional FeFET have reported a large read-after-write latency due to charge trapping at the interface of HZO/IL and IL/silicon during the write operation associated with ferroelectric switching. During programming with positive gate voltage, electrons get trapped primarily at the HZO/IL interface. As shown in the energy band diagram in Fig. 1(c, iii), during the read operation, a V-shaped energy potential is formed at the HZO/IL interface. The large read-after-write latency is set by the time required for the electrons to de-trap via tunneling through the IL which has been measured to be in the range of 10ms. Similar scenario also arises for the erase operation where the negative gate voltage favors the hole trapping primarily at the interface of IL and silicon channel, requiring close to  $100\mu\text{s}$  time delay to de-trap [9], [14]. On the other hand, as shown in the band diagram in Fig. 1(c, iv), the absence of IL allows fast de-trapping, resulting in low read-after-write latency.

Here, we present a novel IL-free BEOL compatible back-gated, channel last FeFET with ultra-scaled dimensions of  $L_g=28\text{nm}$  and  $W=126\text{nm}$ . The FeFET is fabricated by depositing a 5nm thick amorphous Tungsten-doped Indium Oxide (IWO) semiconductor channel material directly on top of 5nm ferroelectric Zirconium-doped  $\text{HfO}_2$  ( $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ ) gate oxide, thus avoiding the creation of the undesired low-k IL. Prior to IWO thin film deposition, the  $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$  film undergoes a crystallization anneal at  $400^0\text{C}$  for 300s using a sacrificial capping layer (SCL) to stabilize the polar orthorhombic phase (see supplementary section **S2** for details). The entire fabrication is done under a low thermal budget of  $400^0\text{C}$ . The details are described in the **Methods** section. We choose IWO as the n-type amorphous metal oxide semiconductor (AOS) channel material since it provides a higher field-effect electron-mobility of  $20 \text{ cm}^2/\text{V}\cdot\text{s}$  compared to other contemporary AOS materials such as IGZO ( $\sim 10 \text{ cm}^2/\text{V}\cdot\text{s}$ ) under low thermal budget processing. Furthermore, IWO has a higher threshold voltage ( $V_T$ ) stability compared to IGZO due to higher oxygen-bond dissociation energy of Tungsten that acts as both stabilizer and electron donor [3], [23], [24].

## RESULTS

Fig. 1(d) shows a schematic of the fabricated BEOL FeFET consisting of a Tungsten (W) local back-gate, 5nm HZO as the ferroelectric gate oxide, 5nm IWO as the semiconductor channel and Palladium (Pd) as the source and drain contacts. Fig. 1(e) shows the top view false-colored SEM image of the fabricated device and a zoomed-in image highlighting the ultra-scaled device dimensions of  $L_g = 28\text{nm}$  and width =  $126\text{nm}$ . Fig. 1(f) shows a cross-sectional TEM of the FeFET, confirming the ultra-scaled channel length and thickness of individual layers. A high-resolution (HR) TEM cross-section image of the gate-source overlap region shows the distinct amorphous nature of the IWO channel material as opposed to the crystalline nature of the HZO

layer. The elemental composition is further confirmed from the EDS elemental mapping of the fabricated device as shown in Fig. 1(g).

The BEOL FeFET operates as a junction-less FET with the gate-to-drain and gate-to-source overlap regions playing a crucial role in the memory operation of the device. As shown in the HRTEM image in Fig. 1(f), the overlap regions resemble a metal-ferroelectric-semiconductor-metal (MFSM) structure where most of the applied electric field during the program and erase operation is concentrated and induces effective polarization switching in the HZO layer. In addition, there exists fringing electric fields inside the channel region away from the source and drain contacts that result in polarization switching in the HZO layer underneath the channel. This implies that the read voltage memory window will monotonically increase with decreasing channel length in such a back-gated FeFET device geometry [18]. During the programming state, the polarization charge in HZO can accumulate a large channel charge in IWO, thus turning on the transistor. In the erase state, the channel charge in IWO is depleted by the polarization charge in HZO, thus turning off the transistor.

Unlike perovskite-based ferroelectrics, the polar orthorhombic phase is retained in ultra-thin layers of doped  $\text{HfO}_2$  up to 1 nm thickness [25] (see supplementary section **S2** for characterization of 5nm thick MFM capacitors). This allows lowering of the programming voltage of HZO-based FeFETs from  $\pm 4\text{V}$  for 10nm thickness [26] to  $\pm 3.3\text{V}$  for 4.5nm [12]. However, in Si FeFETs, most of the programming voltage in the gate stack drops across the low-k IL between the HZO and silicon. On the other hand, an IL-free FeFET enables further lowering of the programming voltage for the same ferroelectric thickness. Fig. 2(a) shows the dual-sweep DC  $\text{Id}-\text{Vg}$

characteristic of the BEOL FeFET with  $W/L = 126\text{nm}/28\text{nm}$  and  $V_{DS} = 0.2\text{V}$ . A memory window (MW) over  $1.2\text{V}$  is obtained with a sweep of  $\pm 2\text{V}$ . We report a read current  $I_{LVT} = 2\mu\text{A}$  for the program (or low  $V_T$ ) state at read voltage  $V_{read} = -0.2\text{V}$  and a read current  $I_{HVT} \sim 10\text{pA}$  for the erase (or high  $V_T$ ) state which provides a high read current window of  $I_{LVT}/I_{HVT} > 10^5$ . Fig. 2(b) shows the measured subthreshold slope (SS) for the dual sweep. In the forward sweep, we measure a SS  $\sim 95\text{mV/dec}$  while a steep slope of SS  $\sim 4\text{mV/dec}$  is measured in the reverse sweep owing to the ferroelectric switching kinetics [27]. Such a steep-slope, also known as transient negative capacitance (NC) effect, arises due to the difference between the rate of change of polarization switching and rate of change of screening or compensating free charge [28]–[30].

Next, we characterize the fast read and write operation of the BEOL FeFET memory. In most studies, the memory characteristics of a FeFET is determined in terms of threshold voltage ( $V_T$ ) and MW measurement. In real applications such as compute-in-memory, the actual current being read out of the memory device (after program and erase) will play the critical role of determining the overall array-level efficiency such as energy and latency. Hence, here we adopt a one-shot fast read-out scheme of BEOL FeFET by applying a short read pulse to the gate of the FeFET following the write (program or erase) pulse. For either program or erase, we apply write pulses of amplitude  $\pm 2\text{V}$  and pulse-width of  $50\text{ns}$  at the gate of the FeFET. During the write operation, both the source and drain of the FeFET are grounded by applying  $V_{DS} = 0\text{V}$ . Following the write operation, we apply  $V_{read} = -0.2\text{V}$  to the gate of FeFET while a  $V_{DS} = 0.2\text{V}$  is applied across the source and drain terminals to measure the drain current  $I_D$ . Figs. 2(c, d) show the time-domain response of the FeFET during the program and erase operation. For the programming scenario, we see a sharp increase in read current  $I_D$  over  $1\mu\text{A}$  during the read-out, whereas, for the erase case, the read

current  $I_D$  remains close to 1nA. Note that a key characteristic of the IL-free BEOL FeFET is the ability to read the memory state almost instantaneously after the write operation without the need for the charge trapped at the FE-IL interface during write phase to detrap before read operation can commence. In contrast to the conventional Si FeFET that in some cases require as high as almost 10ms of read-after-write latency [9], [14], we demonstrate fast read-after-write operation with a latency of 200ns (limited by the pulsed I-V measurement setup). We also measure the cycle-to-cycle variation by performing repeated program and erase operations with  $\pm 2V$  and 50ns pulses. Fig. 2(e) shows the distribution of the measured  $I_D$  for 140 repeated cycles of program and erase, showing a clear separation of memory states. We quantified the switching speed of the BEOL FeFET in terms of the write pulse amplitude and pulse width required to obtain a  $I_{LVT} > 1\mu A$ ,  $I_{HVT} < 10nA$  and  $\Delta I_D > 1\mu A$ . Fig. 2(f) shows the exponential dependence of the required pulse width on the applied voltage. We obtain symmetrical switching behavior for both program and erase operation with write latency as small as 20ns for  $\pm 2V$ .

Next, we investigate the write endurance characteristics of the BEOL-compatible IL-free FeFET. We use fast bipolar write pulses of  $\pm 2V$  and 20ns for endurance cycling with periodic reading of the memory state ( $I_{LVT}$  and  $I_{HVT}$  at  $V_{read}$  for program and erase) and DC  $I_d$ - $V_g$  measurement to monitor the device characteristics. Fig. 3(a) shows the readout current for program and erase states as a function of endurance cycles. Post  $5 \times 10^{10}$  write cycles, the FeFET maintains an  $I_{LVT} > 1\mu A$ , while the  $I_{HVT}$  gradually increases from 10pA up to 100nA. Still, after  $5 \times 10^{10}$  cycles, the device maintains a clear separation of the read current level for the program and erase states with a read current window of  $I_{LVT}/I_{HVT} > 10$  as seen in Fig. 3(b). Fig. 3(c) shows the DC  $I_d$ - $V_g$  characteristics of the device measured periodically post endurance cycling. In contrast to rapid degradation and

permanent failure exhibited by Si FeFETs after  $10^4$ - $10^6$  cycles [9]–[13] due to catastrophic breakdown of the IL, we see a graceful degradation in the case of the IL-free IWO FeFET with increasing subthreshold slope and a shift in the threshold voltage  $V_T$ . This can be correlated to an increase in the generation of interface trap density ( $N_{it}$ ) at the HZO/IWO interface (see supplementary section S4 for details). The anti-clockwise hysteresis remains preserved in the FeFET, exhibiting robust ferroelectricity in the 5nm HZO gate oxide with no sudden breakdown observed whatsoever. We also characterized the read endurance of the BEOL FeFET. After programming (or erasing) the device with a positive (or negative) write pulse, we apply continuous read pulses with  $V_{read} \sim -0.2V$ . We witness no read disturb and degradation in  $I_{LVT}$  and  $I_{HVT}$  as seen in Fig. 3(d). Fig. 3(e) shows the excellent memory retention characteristic of the BEOL FeFET, measured at  $25^0C$  for  $10^3$ s.

Finally, we demonstrate the multi-bit per cell programming characteristic of the FeFET that allow storage of multi-bit weight matrix on-chip for performing fast and energy-efficient vector-matrix multiplication *in situ* to accelerate deep neural network (DNN) inference workloads. We utilize the voltage-driven partial polarization switching in the ferroelectric HZO layer to implement the multi-bit programming capability. As shown in Fig. 4, we use the amplitude modulation scheme. Following an erase operation with -2V, programming pulses with increasing amplitudes of 1V, 1.3V, 1.7V and 2V are applied to gradually modulate the conductance state of the device and program it to four distinct states representing bits ‘00’, ‘01’, ‘10’ and ‘11’, respectively. Fig. 4 shows the four tightly distributed current levels measured over 50 programming cycles, clearly exhibiting a 2bit/cell memory characteristic.

## DISCUSSION

The back-gated, channel last process for fabricating FeFETs provides an effective pathway to constructing an IL-free BEOL compatible logic-compatible, high-performance ferroelectric memory that mitigates several challenges faced by conventional Si FeFETs. In Table 1(a), we enumerate a comprehensive device-level comparison between the state-of-the-art results reported on FEOL silicon FeFETs [5], [17], the IL-free BEOL FeFET and other contemporary BEOL FeFET embodiments [19]–[21]. We report the best-in-class device performance with a low operating voltage of  $\pm 2V$ , high-speed read and write operation and endurance cycling exceeding  $5 \times 10^{10}$  write cycles in an extremely scaled device geometry of  $W^*L=0.0035\mu m^2$  with a ferroelectric gate oxide of 5nm thickness. Additional thinning of the ferroelectric gate dielectric to 3nm will further reduce the programming voltage down to 1.5V. This makes the IL-free BEOL FeFET a promising CMOS logic compatible high-performance on-chip buffer memory and multi-bit weight cell for future compute-in-memory accelerators.

We perform system-level benchmarking to compare the performance of various eNVMs and SRAM for storing multi-bit weight matrix on-chip to realize key operations such as vector-matrix-multiplication during inference. We use an 8-bit inference workload using the VGG-8 model on CIFAR-10 dataset. We use the DNN+NeuroSim framework [31] which is a widely used to benchmark and evaluate CIM accelerators (see supplementary section **S5** for details). We use a cell structure consisting of 1T-1BEOL FeFET in a pseudo-crossbar array [8]. Table 1(b) shows the comparison between IL-free BEOL FeFET, FEOL FeFET [6], RRAM [32], STT-MRAM [33] and SRAM. We report the best-in-class energy efficiency of 255.3 TOPs/W in the densest footprint

area of  $119.9\text{mm}^2$  for a single CIM array which provides a 10x improvement in energy-efficiency and 4x improvement in area compared to SRAM.

In this benchmarking exercise, we assume a 128KB SRAM-based on-chip global buffer memory (scratchpad memory) which can fit the maximum size of activation for VGG-8 on CIFAR-10 dataset with a single data transfer from off-chip DRAM. Larger datasets such as ImageNet would require a much denser on-chip global buffer memory. Furthermore, in the benchmarking, the weights are assumed to be stored on-chip in the synaptic array prior to the inference task. For lightweight smart edge devices with area-constraint design, this may not hold and only a part of the weight matrix can be stored on-chip [34]. For example, more powerful DNN models for image processing such as ResNet [35] have around 12-30 MB of parameters that would require either frequent, costly data movement from off-chip DRAM to load weights per inference or building denser on-chip global buffer memories. We show that the IL-free FeFET can serve as an incredibly dense on-chip storage for parameters and thus enable larger DNN models to be efficiently executed on the CIM accelerator without fetching DNN weights from the DRAM during inference. This in turn can further maximize inference efficiency [36]. Table 1(c) shows the performance comparison between SRAM, STT-MRAM [37] and IL-free BEOL FeFET when used as a 32MB global buffer memory to support a ResNet50 DNN model. The array-level characteristics are computed using NVSim framework [38]. Compared to SRAM, IL-free FeFET increases the storage density by over 6x with similar read/write access energy and latency. IL-free FeFET also significantly improves the memory density over iso-capacity when compared to STT-MRAM, a prime candidate for scratchpad and last-level-cache memories.

## METHODS

**Device Fabrication.** We fabricated FeFETs starting with a patterned W local back gate formation. We use plasma enhanced atomic layer deposition (PEALD) to deposit 5nm thick HZO at a deposition temperature of 250<sup>0</sup>C. For strain-induced stabilization of ferroelectric orthorhombic phase in HZO, we sputter-deposit W as a sacrificial capping layer (SCL). We perform rapid thermal anneal (RTA) at 400<sup>0</sup>C for 300s in N<sub>2</sub>. Subsequently, W SCL is removed used etching. Next, 1% by weight W-doped amorphous In<sub>2</sub>O<sub>3</sub> (IWO) channel of 5nm thickness was deposited using RF magnetron sputtering in the presence of 0.02Pa excess O<sub>2</sub> at room temperature. This is followed by patterning and liftoff of the active IWO channel. Next, 30 nm thick Pd was patterned as source and drain electrodes, followed by a 150<sup>0</sup>C N<sub>2</sub> anneal for contact resistance improvement. The channel conductivity significantly depends on the thickness of IWO and O<sub>2</sub> partial pressure used during deposition of IWO. As such, we also fabricated non-ferroelectric IWO transistors with 5nm HfO<sub>2</sub> gate stack for comparison. Supplementary section **S3** shows the measured transfer characteristics and dependence of the threshold voltage (V<sub>T</sub>) of the baseline transistors fabricated with different IWO channel thickness and different O<sub>2</sub> partial pressures.

**Device Characterization.** All electrical characterizations are performed using Keithley 4200-SCS probe station, Agilent 81150A pulse generator and Agilent DSO9104A oscilloscope.

## AUTHOR CONTRIBUTIONS

S. Dutta and S. Datta conceived the idea. S. Dutta, A. Khanna., A. Khandker and W. C. performed the experiments. H. Y. performed the fabrication. B. R. supervised the ALD process. K. N. performed the TCAD simulations. Y. L. and S. Y. performed the benchmarking for inference

workload using DNN+Neurosim. L. P. G. W. and D. B. performed the benchmarking on global buffer memory using NVSim framework. M. N and X. S. H. participated in useful discussion. S. Dutta and S. Datta wrote the manuscript.

## ACKNOWLEDGEMENT

This work was supported by ASCENT, one of six centers in JUMP, sponsored by DARPA and the Semiconductor Research Corporation (SRC), and IMPACT center in nCORE, sponsored by SRC.

## REFERENCE

- [1] A. Keshavarzi, K. Ni, W. Van Den Hoek, S. Datta, and A. Raychowdhury, “FerroElectronics for Edge Intelligence,” *IEEE Micro*, 2020.
- [2] A. I. Khan, A. Keshavarzi, and S. Datta, “The future of ferroelectric field-effect transistor technology,” *Nature Electronics*. 2020.
- [3] S. Datta, S. Dutta, B. Grisafe, J. Smith, S. Srinivasa, and H. Ye, “Back-End-Of-Line Compatible Transistors for Monolithic 3D Integration,” *IEEE Micro*, 2019.
- [4] M. Trentzsch *et al.*, “A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2017.
- [5] S. Dünkel *et al.*, “A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2018.
- [6] M. Jerry *et al.*, “Ferroelectric FET analog synapse for acceleration of deep neural network training,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2018.

- [7] M. Jerry *et al.*, “A ferroelectric field effect transistor based synaptic weight cell,” *J. Phys. D. Appl. Phys.*, 2018.
- [8] S. Dutta *et al.*, “Monolithic 3D integration of high endurance multi-bit ferroelectric FET for accelerating compute-in-memory,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2020.
- [9] E. Yurchuk *et al.*, “Charge-Trapping Phenomena in HfO<sub>2</sub>-Based FeFET-Type Nonvolatile Memories,” *IEEE Trans. Electron Devices*, 2016.
- [10] H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, “Ferroelectric FETs With 20-nm-Thick HfO<sub>2</sub> Layer for Large Memory Window and High Performance,” *IEEE Trans. Electron Devices*, 2019.
- [11] T. Ali *et al.*, “High Endurance Ferroelectric Hafnium Oxide-Based FeFET Memory Without Retention Penalty,” *IEEE Trans. Electron Devices*, 2018.
- [12] A. J. Tan *et al.*, “Hot Electrons as the Dominant Source of Degradation for Sub-5nm HZO FeFETs,” in *Digest of Technical Papers - Symposium on VLSI Technology*, 2020.
- [13] N. Gong and T. P. Ma, “A Study of Endurance Issues in HfO<sub>2</sub>-Based Ferroelectric Field Effect Transistors: Charge Trapping and Trap Generation,” *IEEE Electron Device Lett.*, 2018.
- [14] K. Ni *et al.*, “Critical Role of Interlayer in Hf<sub>0.5</sub>Zr<sub>0.5</sub>O<sub>2</sub> Ferroelectric FET Nonvolatile Memory Performance,” *IEEE Trans. Electron Devices*, 2018.
- [15] M. G. Kozodaev *et al.*, “Mitigating wakeup effect and improving endurance of ferroelectric HfO<sub>2</sub>-ZrO<sub>2</sub> thin films by careful La-doping,” *J. Appl. Phys.*, 2019.
- [16] T. Ando, “Ultimate Scaling of High-κ Gate Dielectrics: Higher-κ or Interfacial Layer Scavenging?,” *Materials (Basel)*., 2012.

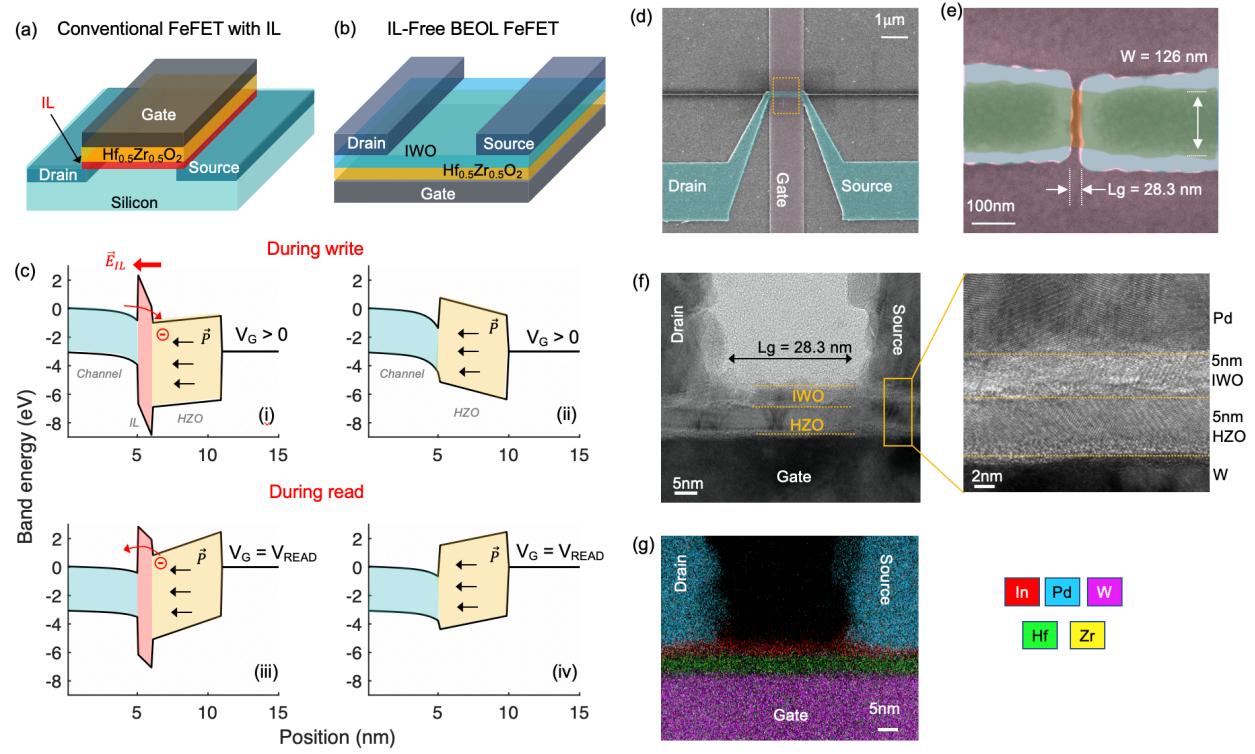
- [17] A. J. Tan, Y.-H. Liao, L.-C. Wang, J.-H. Bae, C. Hu, and S. Salahuddin, “Ferroelectric HfO<sub>2</sub> Memory Transistors with High-\$\kappa\$ Interfacial Layer and Write Endurance Exceeding \$10^{10}\$ Cycles,” *arXiv Prepr. arXiv2103.08806*, 2021.
- [18] A. A. Sharma *et al.*, “High speed memory operation in channel-last, back-gated ferroelectric transistors,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2020.
- [19] F. Mo *et al.*, “Experimental Demonstration of Ferroelectric HfO<sub>2</sub> FET with Ultrathin-body IGZO for High-Density and Low-Power Memory Application,” in *Digest of Technical Papers - Symposium on VLSI Technology*, 2019.
- [20] F. Mo *et al.*, “Low-Voltage Operating Ferroelectric FET with Ultrathin IGZO Channel for High-Density Memory Application,” *IEEE J. Electron Devices Soc.*, 2020.
- [21] M. Halter *et al.*, “Back-End, CMOS-Compatible Ferroelectric Field-Effect Transistor for Synaptic Weights,” *ACS Appl. Mater. Interfaces*, 2020.
- [22] M. K. Kim and J. S. Lee, “Ferroelectric Analog Synaptic Transistors,” *Nano Lett.*, 2019.
- [23] W. Chakraborty, B. Grisafe, H. Ye, I. Lightcap, K. Ni, and S. Datta, “BEOL Compatible Dual-Gate Ultra Thin-Body W-Doped Indium-Oxide Transistor with Ion=370uA/um. SS = 73mV/dec and Ion/Ioffratio > 4\*10<sup>9</sup>,” in *Digest of Technical Papers - Symposium on VLSI Technology*, 2020.
- [24] W. Chakraborty, H. Ye, B. Grisafe, I. Lightcap, and S. Datta, “Low Thermal Budget (<250 °c) Dual-Gate Amorphous Indium Tungsten Oxide (IWO) Thin-Film Transistor for Monolithic 3-D Integration,” *IEEE Trans. Electron Devices*, 2020.
- [25] S. S. Cheema *et al.*, “Enhanced ferroelectricity in ultrathin films grown directly on silicon,” *Nature*, 2020.
- [26] S. Dünkel *et al.*, “A FeFET based super-low-power ultra-fast embedded NVM technology

for 22nm FDSOI and beyond,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2018.

- [27] M. Jerry, J. A. Smith, K. Ni, A. Saha, S. Gupta, and S. Datta, “Insinhts on the DC characterization of ferroelectric field-effect-transistors,” in *Device Research Conference - Conference Digest, DRC*, 2018.
- [28] P. Sharma, J. Zhang, K. Ni, and S. Datta, “Time-Resolved Measurement of Negative Capacitance,” *IEEE Electron Device Lett.*, 2018.
- [29] J. Gomez, S. Dutta, K. Ni, S. Joshi, and S. Datta, “Steep Slope Ferroelectric Field Effect Transistor,” in *2019 Electron Devices Technology and Manufacturing Conference (EDTM)*, 2019, pp. 59–61.
- [30] J. Gomez *et al.*, “Hysteresis-free negative capacitance in the multi-domain scenario for logic applications,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2019.
- [31] 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 *Technical Digest - International Electron Devices Meeting, IEDM*, 2019.
- [32] P. Jain *et al.*, “13.2 A 3.6Mb 10.1Mb/mm<sup>2</sup> Embedded Non-Volatile ReRAM Macro in 22nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0.5V with Sensing Time of 5ns at 0.7V,” in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, 2019.
- [33] J. G. Alzate *et al.*, “2 MB Array-Level Demonstration of STT-MRAM Process and Performance Towards L4 Cache Applications,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2019.
- [34] A. Lu, X. Peng, Y. Luo, and S. Yu, “Benchmark of the Compute-in-Memory-Based DNN

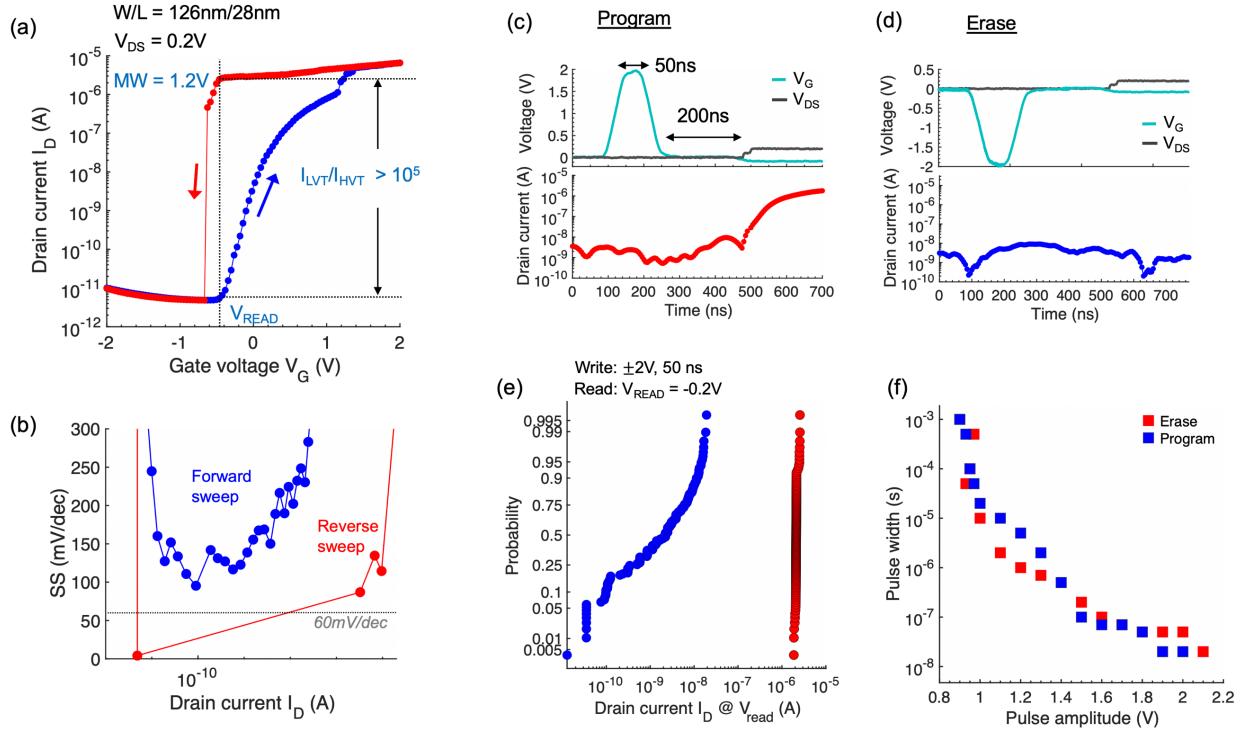
- Accelerator with Area Constraint,” *IEEE Trans. Very Large Scale Integr. Syst.*, 2020.
- [35] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.
- [36] L. Pentecost *et al.*, “MaxNVM: Maximizing DNN storage density and inference efficiency with sparse encoding and error mitigation,” in *Proceedings of the Annual International Symposium on Microarchitecture, MICRO*, 2019.
- [37] L. Wei *et al.*, “13.3 A 7Mb STT-MRAM in 22FFL FinFET Technology with 4ns Read Sensing Time at 0.9V Using Write-Verify-Write Scheme and Offset-Cancellation Sensing Technique,” in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, 2019.
- [38] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, “NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory,” *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, 2012.

## Figures



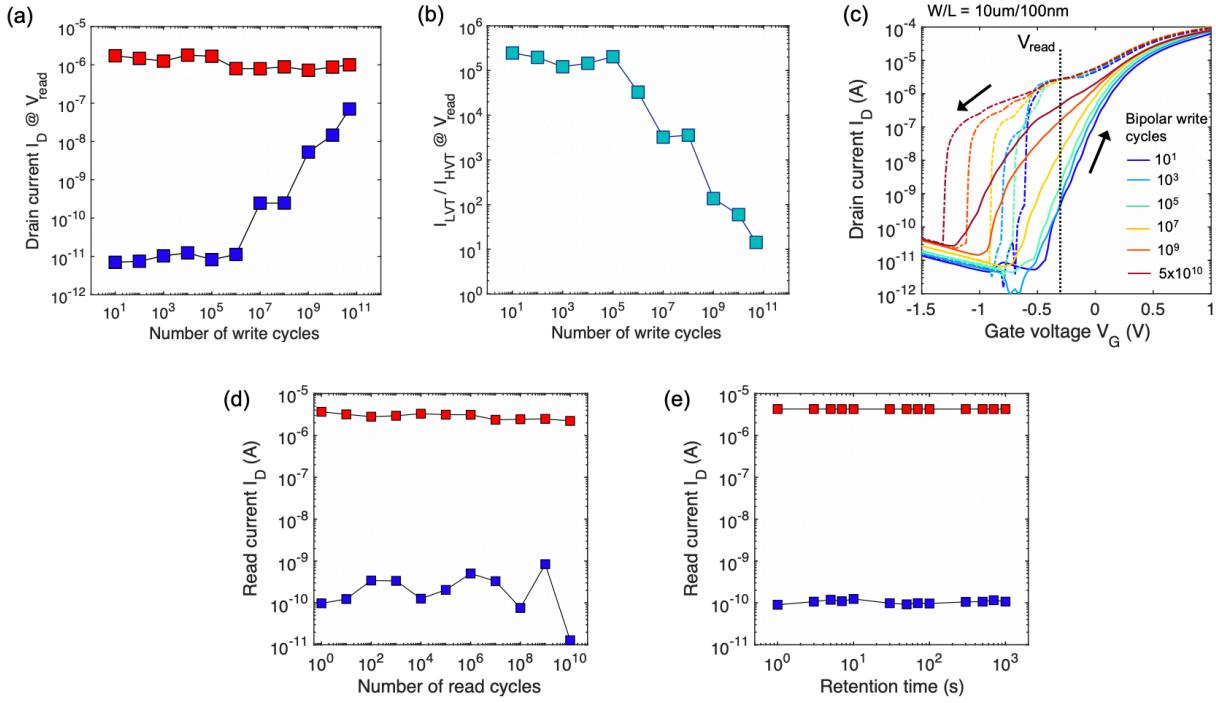
**Figure 1.** IL-free BEOL FeFET. (a) Schematic of a conventional FeFET with a low-k interfacial layer (IL) and (b) IL-free FeFET. (c) Calculated energy band-diagrams for the two structures during write and read operations. The high electric-field ( $E_{IL}$ ) stress across the IL along with continuous charge trapping and trap generation at the interface tend to limit the write endurance of conventional FeFET. The presence of IL also sets a large charge de-trapping time constant that increases the read latency in conventional FeFET. The absence of IL allows lowering the write voltage and significantly reduces the interfacial charge trapping, thereby improving the write endurance and read latency. (d) Schematic of the fabricated device consisting of a local Tungsten (W) back-gate, 5nm  $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$  as the ferroelectric gate dielectric, 5nm Tungsten (W)-doped amorphous  $\text{In}_2\text{O}_3$  (IWO) as the semiconductor channel and Palladium (Pd) as the source and drain contacts. (e) Top-view false colored SEM image of the fabricated device showing the ultra-scaled

device dimensions of  $L_g = 28.3\text{nm}$  and width =  $126\text{nm}$ . (f) Cross-sectional TEM image confirms the ultra-scaled channel length and thickness of individual layers. High-resolution TEM image of the gate-to-source overlap region shows the distinct amorphous nature of the IWO channel material as opposed to the crystalline nature of the HZO layer. (g) Elemental composition of the fabricated device shown in the EDS elemental mapping image.



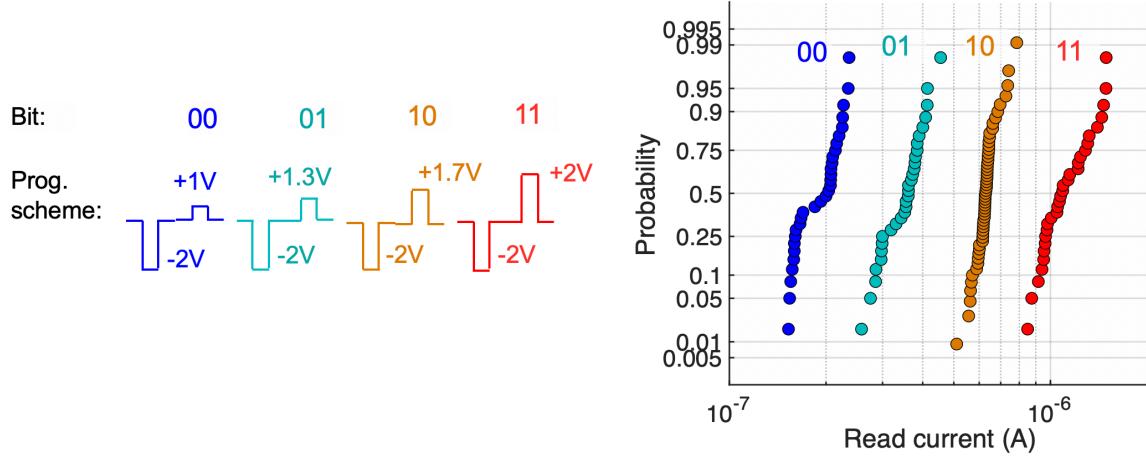
**Figure 2.** Memory characterization of the ultra-scaled BEOL FeFET with  $W/L=126\text{nm}/28\text{nm}$ . (a) Dual sweep DC  $Id-V_g$  characterization with  $\pm 2\text{V}$  shows a memory window (MW) of  $1.2\text{V}$ . The device exhibits a high read current  $I_{LVT} = 2\mu\text{A}$  for the program state and a low read current  $I_{HVT} \sim 10\text{pA}$  for the erase state with read current window of  $I_{LVT}/I_{HVT} > 10^5$ . (b) Measured subthreshold slope (SS) with an  $SS=95\text{mV/dec}$  for the forward sweep and a steep slope of  $SS=4\text{mV/dec}$  for the reverse sweep. (c, d) Time-domain response of the device during program and erase operation. A write voltage of  $\pm 2\text{V}$  and  $50\text{ns}$  is applied to the gate of FeFET to program or erase, followed by a fast read out using a  $V_{read} = -0.2\text{V}$  at the gate and  $V_{DS} = 0.2\text{V}$  with a  $200\text{ns}$  read-after-write latency. (e) Distribution of the measured  $I_D$  for 140 repeated program and erase cycles, showing a clear separation of memory states upon cycling. (f) Switching speed, defined as the write pulse amplitude and pulse width required to obtain a  $I_{LVT} > 1\mu\text{A}$ ,  $I_{HVT} < 10\text{nA}$  and read current window

$\Delta I_D > 1\mu A$ . An exponential dependence of the required pulse width on the applied voltage is seen write latency as small as 20ns for  $\pm 2V$ .



**Figure 3.** Endurance characterization of the BEOL FeFET. (a) Bipolar stress of  $\pm 2\text{V}$  and 20ns is applied to the device. The measured drain current for program and erase states as a function of endurance cycles shows that even after  $5 \times 10^{10}$  write cycles, the device maintains a  $I_{LVT} > 1\mu\text{A}$  while the  $I_{HVT}$  slowly increases from 10pA up to 100nA. (b) Measured read current window of the device with endurance cycling. Even after  $5 \times 10^{10}$  cycles, the device maintains a clear separation of current level for the program and erase states with  $I_{LVT}/I_{HVT} > 10$ . (c) DC  $I_d$ - $V_g$  characteristics of the device measured periodically post endurance cycling. A graceful degradation with increasing subthreshold slope and a shift in the threshold voltage  $V_T$  is seen. However, the anti-clockwise hysteresis is retained in the BEOL FeFET, exhibiting robust ferroelectric switching in the 5nm HZO gate oxide with no sudden breakdown and permanent device failure. (d) Read endurance characterization of the BEOL FeFET. (a) After programming (or erasing) the device

with a positive (or negative) write pulse, continuous read pulses with  $V_{\text{read}} \sim -0.2\text{V}$  are applied to readout the memory state. No read disturb and degradation in  $I_{\text{LVT}}$  and  $I_{\text{HVT}}$  is seen. (e) Measured memory retention of the BEOL FeFET at  $25^{\circ}\text{C}$ , showing excellent retention characteristics with no degradation.



**Figure 4.** Multi-bit per cell characteristics of the FeFET that will allow storing multi-bit weight matrix on-chip. We use the amplitude modulation scheme to induce voltage-driven partial polarization switching in the ferroelectric HZO layer and implement multi-bit programming capability. Following an erase operation with -2V, programming pulses with increasing amplitudes of 1V, 1.3V, 1.7V and 2V are applied to gradually modulate the conductance state of the device and program it to four distinct states representing bits ‘00’, ‘01’, ‘10’ and ‘11’, respectively. A 2bit/cell memory characteristic with four tightly distributed current levels is clearly seen for repeated programming of the device over 50 times.

(a) Device Level Benchmarking				
	Conventional FeFET with IL	IL-Free BEOL FeFET		
	[5, 16]	[18, 19]	[20]	This Work
Channel Material	Silicon	IGZO	WO <sub>x</sub>	IWO
Ferroelectric	4.5nm HZO	15nm HZO	10nm HZO	5nm HZO
Maximum process temperature	500°C	500°C	375°C	400°C
Scaling (W/L)	80nm/20nm	30μm/10μm	20μm/5μm	126nm/ 28nm
Subthreshold slope (SS)	100mV/dec	90 mV/dec	/	95mV/dec
Memory Window	1.5V	0.5V	1V	1.2V
I <sub>LVT</sub> /I <sub>HVT</sub> ratio	10 <sup>5</sup>	10 <sup>5</sup>	/	>10 <sup>5</sup>
Programming voltage	±3V	±3V	±3.6V	±2V
Write latency	250ns	500 us	5us	20ns
Write endurance	>10 <sup>10</sup>	>10 <sup>8</sup>	/	>10 <sup>10</sup>
Read latency	80us	/	/	200ns
Retention	>10 <sup>4</sup> s	>4x10 <sup>3</sup> s	>10 <sup>4</sup> s	>10 <sup>3</sup>
Multi-bit/Cell	1	1	4	2

(b) Inference Benchmarking with VGG-8 model on CIFAR-10					
	SRAM	RRAM	STT-MRAM	Conventional FeFET	IL-Free BEOL FeFET
Bit/Cell	1	2	1	2	2
Cell area (F2)	600	60	100	26	15
Single CIM array area (mm <sup>2</sup> )	481.69	475.79	792.99	206.18	118.95
L-by-L Dynamic Energy (uJ)	46.28	21.94	79.98	7.67	4.74
L-by-L latency (ms)	1.3	0.76	1.31	0.7	0.67
Energy Efficiency (TOPs/W)	25.4	55.90	15.34	158.83	255.33
Throughput (TOPs)	0.95	1.61	0.94	1.76	1.83

(c) Benchmarking 32MB Global Buffer Memory for ResNet50			
	SRAM	STT-MRAM	IL-Free BEOL FeFET
Global buffer area (mm <sup>2</sup> )	21.68	9.92	3.09
Memory density (MB/mm <sup>2</sup> )	1.48	3.23	10.3
Write energy (pJ)	8.2	55.8	39.1
Write latency (ns)	13	5.9	21.7
Read energy (pJ)	4.9	29.0	9.0
Read latency (ns)	13	5.8	21.3

**Table 1.** Comprehensive benchmarking of the IL-free BEOL FeFET. (a) Device-level comparison of the IL-free BEOL FeFET demonstrated in the work with the state-of-the-art results reported on FEOL silicon FeFETs and other contemporary BEOL FeFET proposals. Overall, BEOL FeFET exhibits the best-in-class device performance with a low operating voltage, high-speed memory operation and high write endurance in a highly scaled device geometry. (b) System-level benchmarking using the DNN+NeuroSim framework [31] to compare the performance of various eNVMs and SRAM for storing multi-bit weight matrix on-chip. We use an 8-bit inference

workload using the VGG-8 model on CIFAR-10 dataset. BEOL FeFET exhibits the best-in-class energy efficiency in the densest footprint area with 10x improvement in energy-efficiency and 4x improvement in area compared to SRAM. (c) Array-level performance comparison performed using NVSim framework [37] for a 32MB global buffer memory used to support a ResNet50 DNN model. Compared to SRAM, IL-free FeFET increases the storage density by over 6x with similar read/write access energy and latency.