Monolithic 3D Integration of High Endurance Multi-Bit Ferroelectric FET for Accelerating Compute-In-Memory S. Dutta1\*, H. Ye1\*, W. Chakraborty1, Y.-C. Luo2, M. San Jose1, B. Grisafe1, A. Khanna1, I. Lightcap1, S. Shinde1, S. Yu2and S. Datta1 1University of Notre Dame, Notre Dame, IN 46556, USA; 2Georgia Institute of Technology, Atlanta, GA, USA \*Equal contribution; Email: sdutta4@nd.edu

**Abstract:** We demonstrate, for the first time, monolithic 3D (M3D) integration of back-end-of-line (BEOL) compatible Hf0.5Zr0.5O2 (HZO) ferroelectric FET (FeFET) with front-end-of-line (FEOL) high-k/metal gate (HKMG) Si-NMOS. We use low thermal budget (<4000C) processing to integrate HZO with 1% Tungsten (W)-doped amorphous In2O3 (IWO) semiconducting oxide channel and demonstrate high remnant polarization charge density *2PR*, of 40!"/$%&with reliable switching characteristics. We report (a) read memory window of 0.45V in ultra-scaled 20nm channel length IWO FeFET, (b) write speed of 100ns, and (c) write endurance >108 cycle. We further demonstrate a 2bit/cell synaptic weight cell with well separated conductance states. System-level analysis of compute-in-memory (CIM) accelerators for performing inference on CIFAR-10 image dataset using VGG-8 model shows that 22nm BEOL FeFET achieves 3× higher energy-efficiency than 7nm SRAM while occupying a smaller memory array area due to area folding enabled by M3D architecture.

**I. Introduction**   
Billions of connected edge devices produce zettabytes of data per year that need to be transformed into actionable information. This has created an unprecedented demand for data-centric computing. With on-chip memory limited by SRAM size, there is an extraordinary volume of data traffic between processor and off-chip memory that adds to energy and latency. Compute-in-memory (CIM) is a promising approach to overcome memory bottleneck where compute is moved closer to the data residing in the memory. Embedded non-volatile memory (eNVM) technologies such as FeFET [1], RRAM [2] and STT-MRAM [3] are candidates for storing the weight matrix on-chip and for performing vector-matrix-multiplication (VMM) *in-situ* to accelerate deep neural network (DNN) inference. Conventional CIM architecture co-locates eNVM and access transistors in the FEOL along with CMOS logic and periphery like MUX, analog-to-digital converter (ADC) in the FEOL. M3D CIM architecture targets placement of the memory array in the BEOL with CMOS periphery under the array (CUA) (Fig. 1(a)). Such M3D CIM accelerator provides significant area, energy and latency advantage [4]. Here, we demonstrate a BEOL compatible FeFET that is monolithically integrated on top of FEOL silicon NMOS to realize high-density 3D synaptic array (Fig. 1(b)). While 1T-1RRAM and 1T-1MRAM require large FEOL access transistors resulting in 30-60F2 bit cell area [2-3], M3D 1T-1FeFET bit cell occupies only 15F2, providing a huge density advantage. Additionally, the low write and read energy of BEOL FeFET provides an energy-efficiency benefit when compared to SRAM and other eNVM technologies (Fig. 1(c)). In this work, we demonstrate BEOL compatible low thermal

budget processing and integration of HZO with amorphous oxide semiconductor (AOS) channel. IWO is preferred as the AOS channel material for FeFET owing to its high field-effect electron-mobility[5,6]under low thermal budget processing compared to poly-Si, absence of unwanted low-k interfacial layer formed between the Si interface and the HZO. Recently HZO has been integrated with BEOL compatible IGZO channel [7], however, IWO offers better improved VT stability due to higher oxygen-bond dissociation energy of W that acts as both stabilizer and electron donor [8]. We experimentally demonstrate 2bit/cell synaptic characteristics of IWO FeFET with well separated conductance states and scalability down to 20nm for providing performance boost for CIM accelerators.

**II. Fabrication Process**   
The key fabrication steps for monolithic integration of BEOL FeFET with FEOL Si-NMOS are shown in Fig. 2(a). FEOL Si-NMOS fabrication comprises source/drain (S/D) ion implantation, junction activation, thermal ALD of 5nm HfO2 followed by 20nm thick W deposition as patterned gate metal. 100nm thick PECVD SiO2 is deposited to serve as interlayer dielectric (ILD). Following via pattern and etch, Ti/Al contact metallization and anneal completes the FEOL process. BEOL processing of FeFET starts with patterned W back gate formation and plasma enhanced ALD (PEALD) of 10nm thick HZO at 2500C. For strain-induced stabilization of ferroelectric orthorhombic phase in HZO, W sacrificial capping layer (SCL) is sputter-deposited followed by 3500C anneal for 300s in N2 and subsequent W removal. Next, 1% by weight W-doped amorphous In2O3 (IWO) channel of 3nm thickness was sputtered in the presence of 0.02Pa excess O2 at room temperature, followed by liftoff. 30 nm thick Pd was patterned as S/D electrodes, followed by a 1500C N2 anneal for contact resistance improvement. Next, 5nm thick HfO2 was deposited using thermal ALD as top-gate dielectric at 1200C, followed by patterning Pd top electrode. The top gate ensures effective application of electric field across HZO compared to floating-body scenario [7]. Fig. 2(b) shows the top view false-colored SEM image of the FEOL Si-NMOS and the BEOL HZO FeFET. Cross-sectional STEM of the highlighted region is shown in Fig. 2(c) showing the fully processed 2-tier M3D integrated structure. Fig. 2(d) shows the STEM image and EDS elemental mapping of the 500nm LG FEOL Si-NMOS. Similar STEM image and EDS elemental mapping of the 20nm LG BEOL FeFET is shown in Fig. 2(e). The impact of the 3D stacking and the associated process temperature on the bottom Si-NMOS is characterized by comparing the ID-VG characteristics pre and post 3D integration (Fig.2(f)) showing matched transfer characteristics.

**III. Low Thermal Budget Processing of HZO**   
Conventional processing of Zr-doped ferroelectric HfO2 involves crystallization anneal with thermal budget larger than

5000C. Here, we report on optimization of BEOL compatible low thermal budget processing of HZO. Large area (80!%´80!%) W/HZO/W metal-ferroelectric-metal (MFM) capacitors were fabricated using plasma ALD, followed by annealing at various temperatures ranging from 3000C to 4000C and different anneal durations of 30s and 300s. Fig. 3(a) shows the remnant polarization (PR) measured using DC sweep P-E loop measurements for various anneal conditions. While lower thermal budget leads to lowering of PR, higher ferroelectricity with PR=20!"/$%& is still achievable at 3500C with 300s anneal which is comparable with our baseline process of 5000C anneal for 30s. Fig. 3(b) shows representative P-E loops measured for 3500C anneal for 30s and 300s and compared with the baseline 5000C 30s anneal. We also analyzed the evolution of the ferroelectric orthorhombic phase for low thermal budget processing by performing Grazing-Incidence X-Ray Diffraction (GI-XRD) as shown in Fig. 3(c). The appearance of o(111) peak around 30.50 and o(200) peak around 35.50 confirms stabilized ferroelectric phase in BEOL compatible HZO. We measured the switching speed of the MFM capacitor using squared-shaped program and erase pulses. Fig. 3(d) shows excellent switching capability with partial polarization switching occurring within 50ns - suitable for multi-bit weight cell. The low thermal budget MFM capacitors showed >107 endurance cycle and retention **>**103 seconds, similar to that of baseline processing as shown in Fig. 3(e) and (f), respectively.

**IV. Low Thermal Budget Processing of IWO FeFET** To enable BEOL compatible FeFET, we integrate HZO with IWO as AOS channel material. Fig. 4(a) highlights a key fabrication step of introducing an additional W SCL during annealing to stabilize ferroelectricity in HZO. This is confirmed by P-E loop measurements performed on 80!%×80!% W/HZO/IWO/W capacitors exhibiting higher PR with SCL compared to that without SCL (Fig. 4(b)). The switching speed of the two samples with and without SCL is shown in Fig. 4(c, d) showing partial polarization switching within 50ns in the integrated structure with SCL.

**V. Characterization of IWO FeFET**   
We characterize the weight cell operation of IWO FeFET using pulsed ID-VG and single pulse program and erase measurements on a long channel (LG=1!m) device. For all measurements, gate voltage was applied to the bottom gate while the top gate was kept grounded to ensure effective application of electric field across HZO [7]. Fig. 5(a) shows a large memory window (MW) of 1.2V measured from pulsed ID-VG using a wide VG sweep from -2V to +5V. We also measured a MW of 0.45V using a single program and erase pulse of ±5V and 100!s (Fig. 5(b)). Similar measurements are reported in Fig. 6(a, b) for ultra-scaled FeFET of LG=20nm showing MW of 0.45V and 0.2V measured using pulsed ID-VG measurement with -2V to +5V VG sweep and using program/erase pulse of ±5V and 100!s. The dependence of the MW on the program pulse amplitude and pulse width is shown in Fig. 7(a). Fast write speed of 100ns is achieved while maintaining a MW of 0.3V. Note that the write speed of the fabricated IWO FeFET is limited by the parasitic capacitances

and with optimized device structure sub-10ns switching is possible. High endurance (>108 cycle) of the IWO FeFET with 3-orders of magnitude improvement over conventional Si-FeFET [1] is shown in Fig. 7(b) which can be attributed to the absence of low-k interfacial layer between IWO channel and HZO ferroelectric. A memory retention **>**103 seconds is shown in Fig. 7(c). We demonstrate the multi-bit program capability of IWO FeFET using partial polarization switching. Fig. 8(a, b) shows four tightly distributed distinct ID levels (corresponding to 2bit per cell) achieved by tuning the programming pulses. For writing bits ‘00’, ‘01’, ‘10’ and ‘11’, program/erase pulses of -5V, +3V, +4V and +5V were applied with pulse widths of 1!s. The distribution in ID due to cycle-to-cycle variation was measured for 100 program/erase cycles.

**VI. BEOL FeFET CIM Accelerator Benchmark** We benchmark the BEOL IWO FeFET with SRAM and other eNVM devices (FEOL FeFET[1], 2-bit RRAM [2], STT-MRAM[3]) for 8-bit inferencing VGG-8 model on CIFAR-10 dataset using DNN+NeuroSim framework [9] (Fig. 9(a)), which is a widely used benchmark simulator developed for evaluating the CIM accelerators that include all the necessary peripheral logic. We developed the cell structure of the 1T-1BEOL FeFET in a pseudo-crossbar array (Fig. 9(b, c)) to realize key operations such as VMM. The FEOL 1T serves as selection transistor to avoid disturb in the write operations in pseudo-crossbar array [10]. Thanks to the small current needed to write/read BEOL FeFET, a compact BEOL FeFET cell area of 15F2 assuming minimum W/L. On the other hand, 1T1R cell area is generally in the range of 30F2-60F2 due to RRAM/STT-MRAM’s large write current. The SRAM based accelerators are evaluated at both 7nm and 22nm node, and the other eNVM based accelerators were evaluated at 22nm node. Fig. 9(d) summarizes the benchmarking results. BEOL FeFET CIM array occupies smallest memory array area due to minimum size FEOL access transistor under the BEOL FeFET in M3D. BEOL FeFET also achieves more than 3× improvement in energy-efficiency than 7nm SRAM due to the large RON (4(Ω) of BEOL FeFET limiting the read dynamic energy.

**VII. Conclusion**   
We demonstrated monolithic 3D integration of BEOL compatible FeFET with FEOL Si-NMOS. We show low thermal budget (<4000C) processing and integration of HZO with W doped In2O3 amorphous oxide semiconductor. We report 0.45V memory window in 20nm channel length IWO FeFET with fast write speed of 100ns, >108 cycle write endurance and >103 seconds memory retention. We also demonstrate 2bit/cell synaptic weight cell that provides 3× improvement in energy-efficiency than 7nm SRAM and density advantage due to compact cell area of 15F2 when benchmarked against other eNVM technologies for CIM accelerators.

**References**   
[1] K. Ni, et al, IEDM 2018. [2] P. Jain, et al, ISSCC 2019. [3] Y. Kim et al, VLSI 2011. [4] S. Datta et al, IEEE Micro 2019. [5] K. Nomura et al, Nature 2004. [6] W. Chakraborty, et al, VLSI 2020. [7] F. Mo et al, VLSI 2019. [8] H. Li. et al, EDL 2013. [9] X. Peng et al, IEDM 2019. [10] K. Ni. et al, EDL 2018.

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













