High-Peformance BEOL-Compatible Atomic-Layer-Deposited In_2O_3 Fe-FETs Enabled by Channel Length Scaling down to 7 nm: Achieving Performance Enhancement with Large Memory Window of 2.2 V, Long Retention > 10 years and High Endurance > 10^8 Cycles

Z. Lin[1], M. Si[1], Y.-C. Luo[2], X. Lyu[1], A. Charnas[1], Z. Chen[1], Z. Yu[3], W. Tsai[4], P. C. McIntyre[4], R. Kanjolia[6], M. Moinpour[5], S. Yu[2], P. D. Ye[1,*]
[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA, *Email: yep@purdue.edu
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[3]Department of Electrical Engineering, [4]Department of Materials Science, Stanford University, Stanford, CA, USA
[5]EMD Electronics, San Jose, CA, USA [6]EMD Electronics, Boston, MA, USA

***Abstract*—**In this work, we report ultra-scaled Fe-FETs with channel length down to 7 nm enabled by atomically thin In_2O_3 channels and ferroelectric hafnium zirconium oxide grown by atomic layer deposition (ALD) as back-end-of-line (BEOL) compatible non-volatile memory devices for monolithic 3D integration and in-memory computing applications. High performance ALD In_2O_3 short-channel Fe-FETs are achieved, exhibiting a large memory window of 2.2 V, long retention > 10 years, and high endurance greater than 10^8 cycles. It is found that the memory characteristics of ALD In_2O_3 Fe-FETs are enhanced significantly by channel length scaling due to charge balance requirements at the ferroelectric/dielectric (FE/DE) interface and the insufficient positive charge supply at long channel lengths. Therefore, for wide bandgap oxide semiconductors that lack ambipolar carriers, to scale the channel length of the Fe-FET is the key to achieve high performance devices. The aggressive scaling of Fe-FET enables its integration with logic periphery at the leading edge node (e.g. 7 nm), yielding 3~10× system-level benefits over 22 nm Fe-FET design and 7 nm SRAM design, respectively.

I. Introduction

Oxide semiconductors, such as In_2O_3 [1-5], Sn-doped and W-doped In_2O_3 (ITO and IWO) [6-9], Indium-Gallium-Zinc-Oxide (IGZO) [10, 11], are promising channel materials for back-end-of-line (BEOL) compatible transistors toward monolithic 3D integration [12]. In particular, the outlook for ALD In_2O_3 is very positive due to simultaneous featuring low thermal budget of 225 °C, atomically smooth surfaces, highly controllable thickness down to sub-nanometer, wafer scale homogeneity and conformality, and high electron mobility over 100 $cm^2/V \cdot s$ [1, 5]. High-performance ALD In_2O_3 transistors have been demonstrated with drive current over 2 mA/μm with enhancement-mode operation and low subthreshold swing (SS) down to 63.8 mV/dec [3]. The conformal capability of ALD on side walls, deep trenches, and other 3D structures enables tremendous new opportunities for BEOL device processes and integration [5].

To realize the monolithic 3D integration of in-memory computing architecture, as shown in Fig. 1, both logic devices and non-volatile memory (NVM) devices are required. Ferroelectric field-effect transistors (Fe-FETs) with ALD FE hafnium oxide (HfO_2), such as hafnium zirconium oxide (HZO) [13], as the ferroelectric gate insulator and oxide semiconductors as the semiconducting channel is one of the most promising NVM device candidates due to the fully BEOL-compatible device fabrication process with a low thermal budget below 400 °C, and the superior performance of ALD HfO_2 such as scaling and speed [14-16]. Meanwhile, although FE HfO_2 based Fe-FETs with endurance > 10^8 cycles were reported [17-19], endurance on the level of only 10^4-10^6 [20-22] or below is commonly observed. How to understand such discrepancies and how to improve the endurance and other memory characteristics of HfO_2 based Fe-FETs [23-26] is crucial for embedded NVM and in-memory computing.

In this work, ultra-scaled Fe-FETs with channel length down (L_{ch}) to 7 nm are demonstrated, as BEOL-compatible NVM devices for monolithic 3D integration, enabled by the ALD growth of atomically thin In_2O_3 channels and FE HZO. It is found that the memory characteristics of ALD In_2O_3 Fe-FETs are enhanced significantly by channel length scaling due to charge balance requirements at the FE/DE interface and the insufficient positive charge supply at long channel lengths. For wide bandgap semiconductors such as In_2O_3 with a bandgap of around 3 eV and ultra-thin body, it is hard to realize strong inversion and hence there is a lack of ambipolar carriers. To address this challenge, high-performance ALD In_2O_3 Fe-FETs with channel length down to 7 nm are demonstrated, exhibiting large memory window (MW) of 2.2 V, long retention > 10 years, and high endurance > 10^8 cycles.

II. Experimental

Fig. 2 illustrates a schematic diagram of an In_2O_3 Fe-FET. The gate stack includes 40 nm W as the gate metal, 8 nm HZO and 1 or 2 nm Al_2O_3 as the FE gate stack, 1.5 nm In_2O_3 as the semiconducting channel and 40 nm Ni as the source/drain (S/D) contacts. Fig. 3 presents a false-colored scanning electron microscopy (SEM) image of a fabricated In_2O_3 transistor, capturing the In_2O_3 channel, W gate electrode and Ni S/D electrodes. Fig. 4 shows a high-resolution transmission electron microscopy (HRTEM) image with energy-dispersive x-ray spectroscopy (EDX) mapping of a representative In_2O_3 transistor with L_{ch} of 7 nm, highlighting O/Ni/In/Hf elements.

Fig. 5 shows the fabrication process flow of the In_2O_3 Fe-FETs. The device fabrication process started with solvent cleaning of p+ Si substrate. 10 nm Al_2O_3 was then deposited by ALD at 175 °C as an etch stop layer for 40nm W gate metal isolation, by CF_4/Ar ICP dry etching. HZO/Al_2O_3 stack were then deposited by ALD at 200 °C. $[(CH_3)_2N]_4Hf$ (TDMAHf), $[(CH_3)_2N]_4Zr$ (TDMAZr), $(CH_3)_3Al$ (TMA) and H_2O were used as the Hf, Zr, Al, and O precursors, respectively. Then, devices were annealed by rapid thermal annealing (RTA) at 350 °C in an N_2 environment for 10 min. In_2O_3 thin films with thickness of 1.5 nm were then deposited by ALD at 225 °C, with $(CH_3)_3In$ (TMIn) and H_2O used as In and O precursors. Channel isolation was done by wet etching of In_2O_3 using concentrated hydrochloric acid and BCl_3/Ar dry etching. 40 nm Ni was patterned by e-beam lithography and deposited by e-beam evaporation as S/D ohmic contacts in two steps to avoid the lift-off difficulties of short channel devices. Finally, devices were annealed by RTA at 300 °C in O_2 for 4 min. annihilating fabrication defects.

III. Results and Discussion

Fig. 6 shows the P-V characteristics at different voltage ranges of capacitors with (a) W/8 nm HZO/1 nm Al_2O_3/Ni structure and

(b) W/8 nm HZO/1 nm Al_2O_3/1.5 nm In_2O_3/Ni structure. The metal/FE/DE/metal (M/FE/DE/M) structure and metal/FE/DE/semiconductor/metal (M/FE/DE/S/M) structure give similar remnant polarization, indicating a high-quality oxide/semiconductor interface. Fig. 7 presents the I_D-V_{GS} characteristics of an In_2O_3 Fe-FET with L_{ch} of 7 nm and with 1 nm Al_2O_3 as an interfacial DE layer, showing a clear ferroelectric hysteresis loop with a large MW of 2.2 V. The gate leakage current of the device is below the detection limit, as also shown in Fig. 7. The large MW is mainly attributed to the highly scaled channel length in the Fe-FETs *because the polarization switching in short channel and long channel is very different*. Fig. 8 shows the threshold voltage (V_T) versus L_{ch} of the ALD In_2O_3 Fe-FETs after erase and program, where the V_T is measured from transfer curve at V_{DS}=0.1 V at an I_D of 10^{-6} A/μm. The corresponding L_{ch}-dependent MW is shown in Fig. 9, where the MW increases significantly by L_{ch} scaling below 100 nm down to 7 nm. Such L_{ch} dependence suggests a stronger polarization switching process in shorter channel devices, which is also captured by the SS characteristics, as shown in Fig. 10, with deep sub-60 mV/dec at room temperature and smaller SS achieved in shorter channel devices.

The L_{ch} dependence of memory characteristics is because in a Fe-FET, especially with a wide bandgap semiconducting ultra-thin channel without ambipolar carriers, the electrostatic potential has a 2D distribution, which cannot be approximated as a 1D case like a FE capacitor [25, 26]. As a result, the polarization density in the FE gate insulator also has a space distribution in the lateral direction because FE insulator in the middle of the channel in a long channel device can only be partially polarized and time-dependent, as shown in Fig. 11. Thus, to satisfy the charge balance condition [22-25], the FE gate insulator near the source/drain region can be fully switched while FE gate insulator is only partially switched in the middle of the device, because positive charge can only be supplied from the source/drain metal contacts. Fig. 12 shows calculated electrostatic potential distributions of a wide-bandgap n-type transistor at V_{GS} of (a) -10 V and (b) 10 V by TCAD simulation, using similar method and structure as in Ref. [25]. As we can see, at positive V_{GS}, the electrostatic potential at the semiconductor surface is uniform while at negative V_{GS}, the electrostatic potential at the semiconductor surface is non-uniform, so that in the middle of the channel, voltage across the gate insulator is also smaller than the source/drain region, which also leads to less polarization switching in the middle of the channel. Therefore, considering the above fundamental switching process in Fe-FETs, scaling the channel length of the Fe-FET can improve the MW of the device due to the reduction of partial polarization switching.

Fig. 13 presents the I_D-V_{GS} characteristics of an In_2O_3 Fe-FET with L_{ch} of 10 nm and with 2 nm Al_2O_3 as the DE layer, showing a clear ferroelectric hysteresis loop with a MW of 1.4 V. The reduced MW compared to device with 1 nm Al_2O_3 as DE layer can be well described by the charge balance mechanism in Ref. [26]. Fig. 14 shows the MW versus L_{ch} of the ALD In_2O_3 Fe-FETs with 2 nm Al_2O_3 as DE layer, where increasing MW with L_{ch} scaling is also observed. Fig. 15 shows that the multilevel capability (MLC) of 4 states (00, 01, 10, 11) are obtained using pulsing condition on a 7nm Fe-FET with 1 nm Al_2O_3 DE layer as shown in Fig. 7.

Fig. 16 illustrates the pulse sequence for (a) retention and (b) endurance testing used in this work. Retention and endurance measurements were performed on ALD In_2O_3 Fe-FETs with 1 nm Al_2O_3 as DE layer. The pulse width of both erase and program processes are 200 μs with an amplitude of 3.5 V in retention measurements while pulse width of both erase and program processes are 500 ns with amplitude of 2.2 V in endurance measurements unless otherwise specified. Fig. 17 shows the retention characteristics of a short-channel In_2O_3 Fe-FET with L_{ch} of 7 nm with extracted retention time > 10 years. Fig. 18 shows the on/off ratio versus retention time for device with L_{ch} of 7 nm and 0.8 μm, showing that the retention performance is also improved by channel length scaling because the full polarization state is more stable than partial polarization state. Fig. 19 shows I_D of an In_2O_3 Fe-FET with L_{ch} of 20 nm at V_{DS}=0.1 V and V_{GS}=0 V in polarization up and down states versus different pulse width down to 500 ns, suggesting that 500 ns is enough to obtain a sufficiently large MW. Fig. 20 presents the endurance performance of an In_2O_3 Fe-FET with L_{ch} of 50 nm at V_{DS}=0.1 V. A high endurance > 10^8 cycles is achieved. Further process optimization on the gate stack can suppress the V_T shift due to charge trapping and trap generation. Such high endurance performance is also contributed to by the L_{ch} scaling, as shown in Fig. 21, indicating that device operation with partial polarization states also degrades the endurance performance.

To demonstrate the benefits of aggressive L_{ch} scaling on the system-level, a widely used in-memory computing benchmark simulator DNN+NeuroSim [27] is used to compare the BEOL Fe-FETs using the 7nm experimental data as shown in Fig. 15 and a 7nm node process with 7nm SRAM and other state-of-the-art NVM at 22nm. Table 1 shows that using the 7nm BEOL Fe-FET reported in this work, the energy efficiency is improved ~3~10× over 22 nm BEOL Fe-FET design at 22nm and SRAM design at 7nm.

IV. Conclusion

In conclusion, high-performance BEOL-compatible ultra-scaled Fe-FETs with channel lengths down to 7 nm are demonstrated with a large memory window of 2.2 V, long retention > 10 years, and high endurance > 10^8 cycles, which is promising for monolithic 3D integration and in-memory computing applications. The memory characteristics of ALD In_2O_3 Fe-FETs are found to be enhanced significantly by channel length scaling due to charge balance requirements at FE/DE interface and the insufficient positive charge supply mid-channel in long channel length devices.

ACKNOWLEDGMENT

The work is supported by SRC nCore IMPACT Center, AFOSR, and SRC/DARPA JUMP ASCENT Center.

References

[1] M. Si et al., Nano Lett., vol. 21, p. 500, 2021. [2] M. Si et al., IEEE EDL, vol. 42, p. 184, 2021. [3] M. Si et al., IEEE TED, vol. 68, p. 1075, 2021. [4] A. Charnas et al., Appl. Phys. Lett., vol. 118, p. 052107, 2021. [5] M. Si et al., VLSIT, p. T2-4, 2021. [6] S. Li et al., Nat. Mater., vol. 18, p. 1091, 2019. [7] W. Chakraborty et al., VLSIT, p. TH2.1, 2020. [8] M. Si et al., ACS Nano, vol. 14, p. 11542, 2020. [9] M. Si et al., IEEE TED, vol. 68, pp. 3195, 2021. [10] J. Wu et al., VLSIT, p. THL 2.4, 2020. [11] S. Samanta et al., p. TH2.3, 2020. [12] S. Datta et al., IEEE Micro, vol. 39, p. 8, 2019. [13] J. Muller et al., Nano Lett., vol. 12, p. 4318, 2012. [14] X. Lyu et al., in VLSIT, p. T44, 2019. [15] M. Si et al., Appl. Phys. Lett., vol. 115, p. 072107, 2019. [16] X. Lyu et al., IEDM, p. 342, 2019. [17] S. Dutta et al., in IEDM, p. 801, 2020. [18] A. Sharma et al., IEDM, p. 391, 2020. [19] A.J. Tan et al., EDL, 2021. [20] J. Muller et al., VLSIT, p. 25, 2012. [21] E. Yurchuk et al., IEEE TED, vol. 61, p. 3699, 2014. [22] K. Ni et al., IEEE TED, vol. 65, p. 2461, 2018. [23] M. Si et al., ACS Appl. Electron. Mater., vol. 1, p. 745, 2019. [24] K. Toprasertpong et al., IEDM, p. 570, 2019. [25] M. Si et al., IEEE JEDS, vol. 8, p. 846, 2020. [26] M. Si et al., arXiv:2105.12892, 2021. [27] X. Peng et al. IEDM, 2019 https://github.com/neurosim.

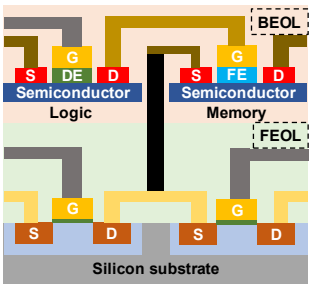

Fig. 1. Illustration of BEOL-compatible MOSFETs and Fe-FETs for monolithic 3D integration and in-memory computing applications.

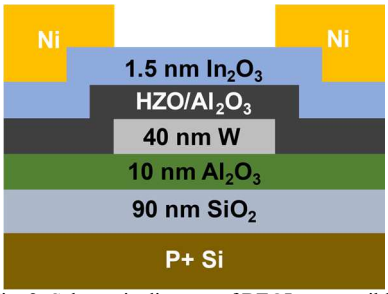

Fig. 2. Schematic diagram of BEOL-compatible In_2O_3 Fe-FETs. 8 nm HZO/x nm Al_2O_3 (x=1, 2) is used as the gate insulator.

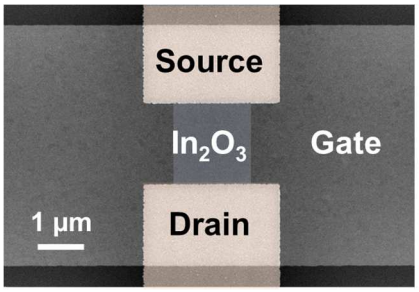

Fig. 3. False-colored SEM image of a fabricated In_2O_3 Fe-FET.

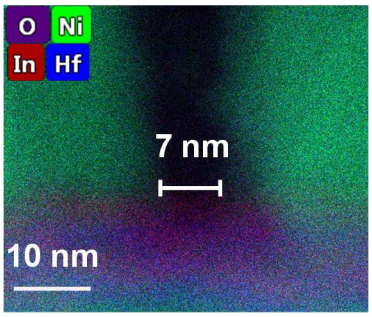

Fig. 4. HRTEM cross-section image of an ultra-short-channel In_2O_3 Fe-FET with EDX elemental mapping, highlighting channel length of 7 nm.

- Solvent clean of SiO_2/Si substrate
- ALD 10 nm Al_2O_3 at 175 °C
- 40 nm W sputtering
- W gate isolation by CF_4/Ar ICP dry etching
- ALD ferroelectric gate stack at 200 °C
  - 8 nm HZO/x nm Al_2O_3 (x=1, 2)
- RTA 350 °C in N_2 for 10 mins
- ALD 1.5 nm In_2O_3 at 225 °C
- Channel isolation by HCl wet etching
- S/D contact formation
  - Pattern by e-beam lithography in two steps
  - 40 nm Ni e-beam evaporation
- Channel isolation by BCl_3/Ar dry etching
  - Dry etching of HZO/Al_2O_3 on gate pad
- RTA 300 °C in O_2 for 4 mins

Fig. 5. Fabrication process flow of the In_2O_3 Fe-FETs.

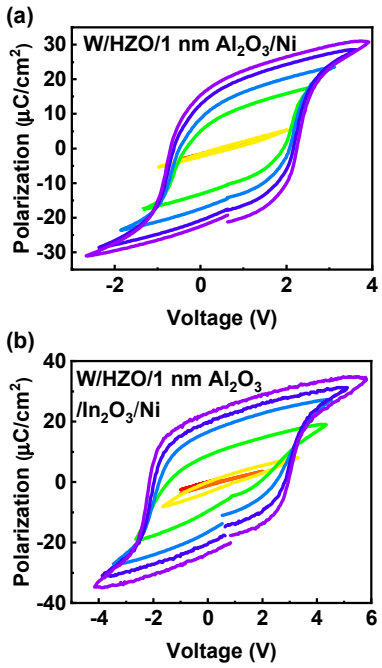

Fig. 6. P-V loops of representative capacitors with (a) W/8 nm HZO/1 nm Al_2O_3/Ni and (b) W/8 nm HZO/1 nm Al_2O_3 /1.5 nm In_2O_3/Ni.

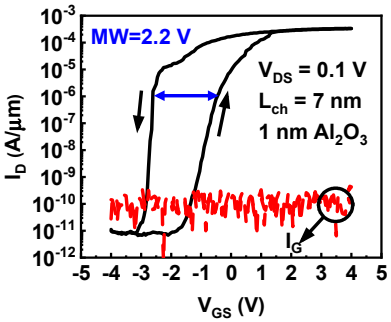

Fig. 7. I_D-V_{GS} characteristics of an ALD In_2O_3 Fe-FET with channel length of 7 nm, Al_2O_3 of 1 nm at V_{DS}=0.1 V, and with a memory window of 2.2 V. Al_2O_3 capping is important here since ALD of In_2O_3 will degrade HZO ferroelectricity if directly grown on HZO without an Al_2O_3 layer.

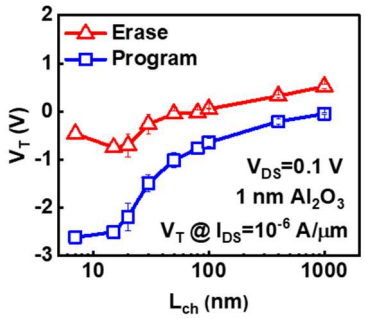

Fig. 8. Threshold voltages after erase and program versus channel length of ALD In_2O_3 Fe-FETs with Al_2O_3 of 1 nm at V_{DS}=0.1 V. V_T is extracted at I_{DS} of 10^{-6} A/μm. Sweep range is ± 2.5 V, except for L_{ch} of 7/15 nm, which is ± 4 V because of the large negative V_T.

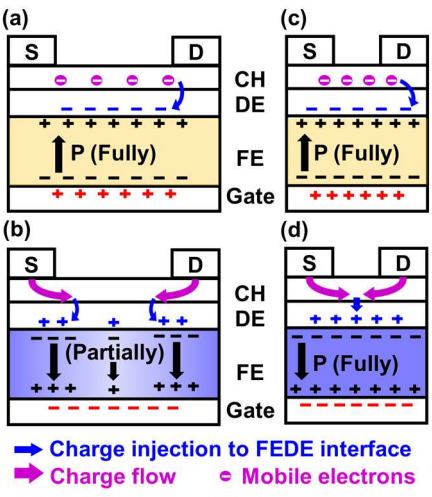

Fig. 11. Charge distribution and polarization states of a typical long-channel Fe-FET at (a) positive bias and (b) negative bias. Charge distribution and polarization states of a typical short-channel Fe-FET at (c) positive bias and (d) negative bias.

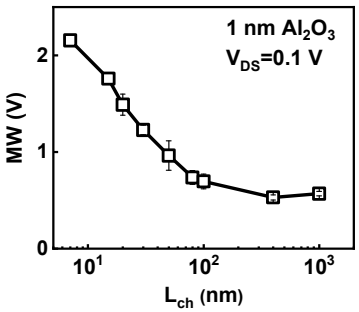

Fig. 9. Memory window versus channel length of ALD In_2O_3 Fe-FETs with Al_2O_3 of 1 nm at V_{DS}=0.1 V. Memory window is calculated as ΔV_T in Fig. 8.

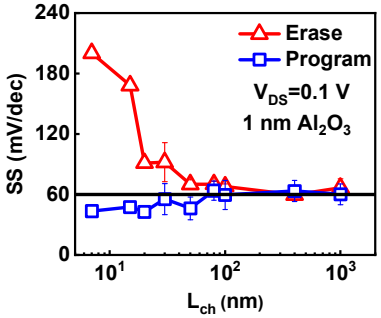

Fig. 10. Subthreshold slope versus channel length of ALD In_2O_3 Fe-FETs with Al_2O_3 of 1 nm at V_{DS}=0.1 V.

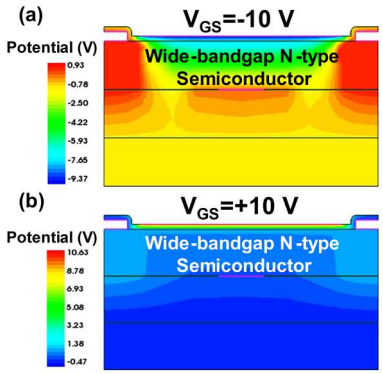

Fig. 12. TCAD simulated electrostatic potential distribution of a wide-bandgap n-type transistor at V_{GS} of (a) -10 V and (b) 10 V.

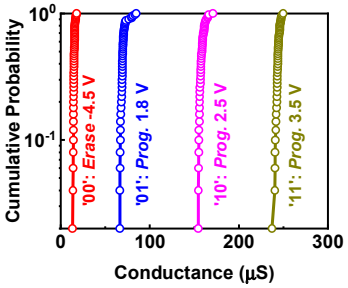

Fig. 15. Distribution of four well distributed conductance levels measured, demonstrating the capability toward 2-bit inference.

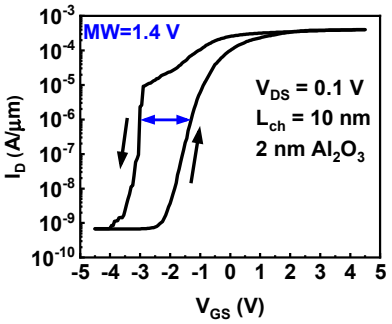

Fig. 13. I_D-V_{GS} characteristics of an ALD In_2O_3 Fe-FET with channel length of 10 nm, Al_2O_3 of 2 nm and V_{DS} of 0.1 V. Smaller breakdown electrical strength in 2nm Al_2O_3 compared to 1nm Al_2O_3 lowers the MW [26].

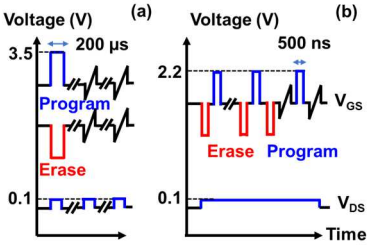

Fig. 16. (a) Pulse sequence for retention test used. Pulse width of both erase and program processes are 200 μs with amplitude of 3.5 V. (b) Pulse sequence for endurance test used. Pulse width of both erase and program processes are 500 ns with amplitude of 2.2 V.

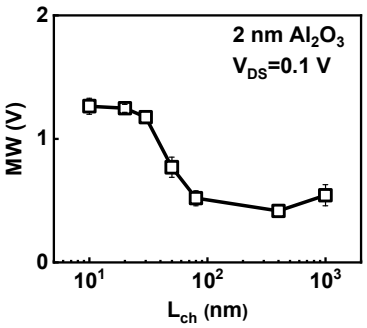

Fig. 14. Memory window versus channel length of ALD In_2O_3 Fe-FETs with Al_2O_3 of 2 nm at V_{DS}=0.1 V.

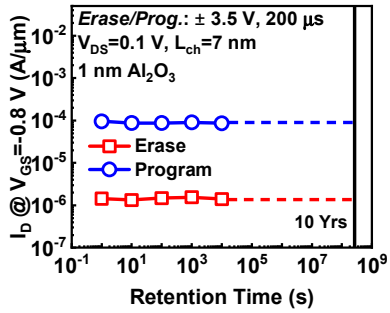

Fig. 17. Retention performance of a short-channel In_2O_3 Fe-FET with channel length of 7 nm at V_{DS}= 0.1 V.

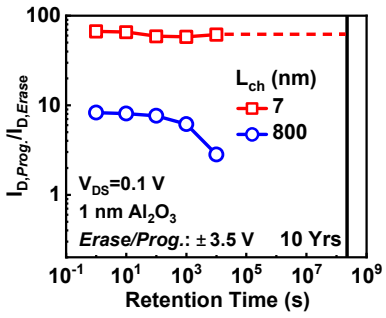

Fig. 18. On/off ratio versus retention time of In_2O_3 Fe-FETs with Al_2O_3 of 1 nm at V_{DS}=0.1 V. I_D is read at V_{GS} of -0.8 V and -0.2 V for device with channel length of 7 nm and 800 nm, respectively.

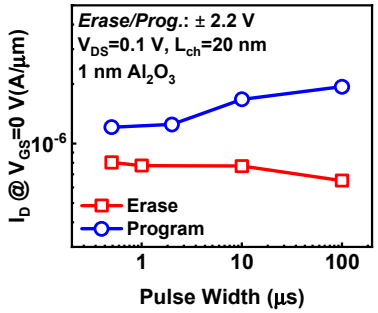

Fig. 19. I_D at V_{GS}=0 V versus pulse width down to 500 ns of an In_2O_3 Fe-FET with channel length of 20 nm, Al_2O_3 of 1 nm at V_{DS}=0.1 V.

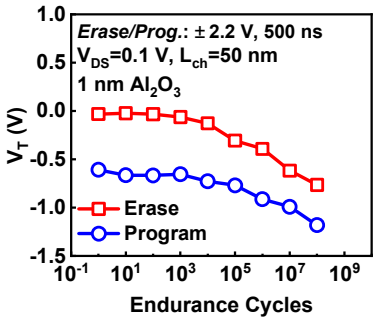

Fig. 20. Endurance performance of a short-channel In_2O_3 Fe-FET with channel length of 50 nm at V_{DS}=0.1 V.

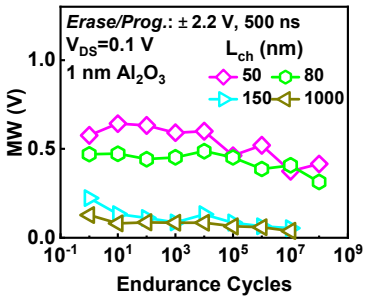

Fig. 21. Memory window versus endurance cycles of In_2O_3 Fe-FETs with Al_2O_3 of 1 nm at V_{DS}=0.1 V and at different channel lengths.

VGG-8 (8-bit activation; 8-bit weight) on CIFAR10 Inference, simulated by DNN+NeuroSim [27]						
Technology node (LSTP)	7nm	22nm				7nm
Device	SRAM	FEOL Fe-FET (Globalfoundries) [IEDM'17]	IWO Fe-FET (Notre Dame) [IEDM'20]	RRAM (Intel) [ISSCC'19]	STT MRAM (Intel) [ISSCC'19]	In_2O_3 Fe-FET (This work)
Bit/Cell	1	2	2	2	1	2
Ron (Ω)	/	67k	4M	6k	2.5k	245k
On/Off Ratio	/	100	4	17	2.8	17
Cell Area (F^2)	600	26	15	60	100	45
Chip Area (mm^2)	13.16	23.52	22.80	33.08	99.22	2.94
Energy Efficiency (TOPS/W)	25.40	57.49	71.04	25.80	7.57	248.55
Throughput (TOPS)	0.95	1.31	1.31	0.98	0.59	1.80
Compute Efficiency ($GOPS/mm^2$)	72.24	55.52	57.61	29.50	5.91	611.41

Subarray size = 128*128; 5-bit SAR ADC; F=7/22nm for normalizing cell area, doesn't indicate physical feature size.

Table 1. Benchmark for in-memory computing to demonstrate benefits of aggressive scaling of Fe-FET.