BEOL Compatible Superlattice FerroFET-based High Precision Analog Weight Cell with Superior Linearity and Symmetry

Khandker Akif Aabrar1, Jorge Gomez1, Sharadindu Gopal Kirtania1, Matthew San Jose1, Yandong Luo3, Priyankka G. Ravikumar2, Prasanna Venkatesan Ravindran3, Huacheng Ye1, Sanjukta Banerjee1, Sourav Dutta1, Asif Islam Khan3, Shimeng Yu3, Suman Datta1   
1University of Notre Dame, Notre Dame, IN 46545, USA; 2Indian Institute of Technology Palakkad, Kerala 678623,India; 3Georgia Institute of Technology, Atlanta, GA 30332, USA; Email:kaabrar@nd.edu

***Abstract*—**Off-chip DRAM memory accesses limit the energy efficiency and training time of state-of-the-art deep neural networks (DNN). Compute-in-memory (CIM) accelerators leveraging pseudo-crossbar arrays and on-chip weight storage have emerged as alternatives to GPUs for fast and efficient training. However, this comes at the cost of reduced training accuracy due to weight cell non-idealities such as: low bit precision, nonlinearity, asymmetry, low Gmax /Gmin ratio, and slow programming speed. Here, we engineer the ferroelectric domain structure in a carefully designed superlattice (SL) ferroelectric(FE)/dielectric(DE) stack, to experimentally demonstrate high precision FEFET analog weight cells with excellent linearity and symmetry during potentiation and depression. We demonstrate switching speed as low as 100 ns in the SL-based ferroelectric capacitor (FECAP), with no degradation in either retention or endurance. We integrate the SL FE/DE/FE with a back-end-of-line (BEOL) compatible Indium Tungsten Oxide transistors, to demonstrate 128 stable conductance states with improved linearity and symmetry. System-level analysis of SL-FEFET based CIM accelerators show an excellent 94.1% online learning accuracy without degrading any other performance parameter, with potential for monolithic 3D integration.

**I.**  **INTRODUCTION**   
 Deep neural networks have achieved remarkable breakthroughs in image analysis and pattern recognition with extreme accuracy. Training these networks to perform these tasks involves constantly updating the values of high-dimensional synaptic weight matrices. As the weights are usually stored in off-chip DRAM, the energy consumption and training time of the network rapidly increases with the complexity of the input data set, with data transfers dictating most of the execution time and consumed energy. Therefore, the energy efficiency and training time of DNNs can be substantially improved by storing weights on-chip in a compute-in-memory configuration. The device requirements of an on-chip analog weight cell consist of a wide dynamic range of conductance states with a symmetric, linear update profile during potentiation and depression. Such characteristics impact the learning accuracy of a network. Previous work has demonstrated a 5-bit analog weight cell implemented with a compact silicon FET with a ferroelectric gate stack of CMOS-compatible Zr-doped HfO2 (HZO) [5]. It is possible to increase the number of conductance states through domain engineering of the ferroelectric gate stack. By inserting a dielectric layer within the ferroelectric HZO, the domain density within the gate stack increases not only to accommodate more polarization and,

hence, conductance states, but also to enhance the symmetry and linearity of the resulting profile. While the related 2T1F embedded memory cell also has high linearity[8], its large area footprint hampers its retention time, when compared to a compact 1T design utilizing an engineered ferroelectric film.

**II.DOMAIN ENGINEERING THROUGH SUPERLATTICE**  **FEDESIGN**   
In a ferroelectric (FE) film, the total free energy of the system is the sum of its electrostatic energy and domain wall energy. This dictates the number of domains that exist in its equilibrium state. As the depolarization field increases with an applied voltage, the FE film tends to decompose into a higher number of domains to minimize the overall free energy of the system. This opens up an opportunity for domain-engineering in the FE film to stabilize a large number of domains, and hence more polarization states, through the modulation of the depolarization screening field. To explore this notion, we fabricate structures consisting of FE/DE/FE superlattices with 5nm HZO FE films, and a DE (HfO2) spacer of variable thickness. Fig. 2(a) shows that the energy minima of the system, calculated using [9], moves towards a configuration with higher domain density as we increase the DE thickness until it saturates around 5nm (Fig. 2 (b)). This gives us the optimized, domain-engineered superlattice (SL) stack with three components: FE(5nm)/DE(5nm)/FE(5nm). Both FE constituents contain a higher number of domains than a standalone FE film, which in turn support a higher number of stable partial polarization states. Furthermore, a thicker dielectric enables a higher depolarization field, permitting an improved control of the polarization switching through the applied voltage, which will tilt the PV loops of SL FE/DE/FE depending on the thickness of the ferroelectric.

**III.FECAPFABRICATION AND CHARACTERIZATION** Fig. 3 (a) shows the fabrication process flow for the ferroelectric capacitor (FECAP) device. The SL stack is deposited on a tungsten-sputtered P+ Si substrate through plasma-enhanced atomic-layer deposition (PEALD) followed by patterning of a sputtered tungsten top electrode. A 400oC rapid-thermal anneal (RTA) is performed in N2 ambient for 60 seconds to stabilize the orthorhombic phase within HZO, which gives rise to ferroelectricity. Both the TEM and EDX, shown in Figs. 3(b) and 3(c) respectively, highlight clearly distinguishable layers of HZO/HfO2/HZO within the superlattice. We also fabricate a control FECAP with 10nm HZO. As evident from the measured P-V loops (Fig. 4 (a)-(d)) of the FECAPs, increasing the DE thickness stabilizes more minor loops which can be attributed to a higher number of domains with wider coercive field distribution (Ec) distribution. This is also seen

in the transient switching current vs applied electric field characteristics of the FECAPs in Fig. 4(e), where the polarization peaks are broadened for SL FECAPs with thicker DE – a signature of a larger number of domains. To study the switching behavior of the FECAPs, we use positive-up and negative-down (PUND) measurement. The polarization charge for any given pulse scheme is extracted from the difference between the charges associated with the first negative pulse and the first positive pulse as shown in Fig. 5(a). The contour map in Fig 5(b) shows excellent switching capability, with up to 100 ns switching speed for the SL FECAP. For characterizing the potentiation/depression (pot/dep) profile of the FECAPs shown in Fig. 5 (c-f), amplitude-modulated, 100 ns wide pulses are employed. Whereas the control sample and the SL with the thinnest DE show less polarization states and nonlinear behavior due to abrupt changes in polarization with pulse number (PN), a gradual and linear potentiation/depression profile is obtained for the SL FECAPs with 5 nm and 7 nm DE yielding a total of 128 polarization states. The endurance and retention characterization of the 5 nm DE SL FECAP shows high endurance (>108 cycles) and a high retention (>103 seconds) of the partial polarization states as seen in Fig.6 (a)-(b).

**IV.BEOLSLFEFET ANALOG SYNAPSE FABRICATION &**  **CHARACTERIZATION**   
We utilize the optimized SL(5/5/5) as gate stack to fabricate BEOL-compatible superlattice FerroFET (SL FEFET) that comprises an Indium Tungsten Oxide (IWO) channel. A conventional FEFET with 10nm HZO gate stack is fabricated for comparison. The process flow in Fig. 7(a) summarizes the key fabrication steps, all of which involve processing temperatures of 400oC or less. The TEM and EDX, in Figs. 7(b) and (c), respectively, confirm the integrity of the SL gate stack. The ID-VG characteristics show, at VDS= 50mV, a record memory window (MW) of ~5.7V and Ion/Ioff ~105for the SL FEFET in Fig. 7(d), and MW of 1.45V, and Ion/Ioff~102for the 10nm HZO FEFET in Fig. 7(e). For analog weight cell characterization of the FEFETs, an amplitude-modulated pulse scheme was used with

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| an | incremental(decremental) | voltage | of | 10mV | for |

potentiation(depression). The starting voltage level for the pulse train was chosen for the optimum conductance behavior. The control FEFET shows a nonlinear conductance profile having a maximum of 70 analog states with nonlinearity αp=6.81and αd=-6.51 in Fig 8(a). On the other hand, in Fig. 8(b), the SL FEFET with the optimized gate stack shows much higher linearity and greater symmetry of the conductance states for the pot/dep (αp=-0.7, αd=-1.56 and asymmetry |αp – αd|=0.86) and exhibits 128 analog states. The SL FEFET also shows a high Gmax/Gmin ratio of ~280. The ID-VG characteristics in Figs. 9(a) and (b) for the SL FeFET show the gradual change in the drain current due to VT shift caused by the pot/dep program pulses. We perform endurance and retention characterization of the SL FEFET which shows no on/off current degradation after 1010cycles, and retention of the intermediate conductance states after 103 seconds, seen in Figs. 10 and 11 respectively.

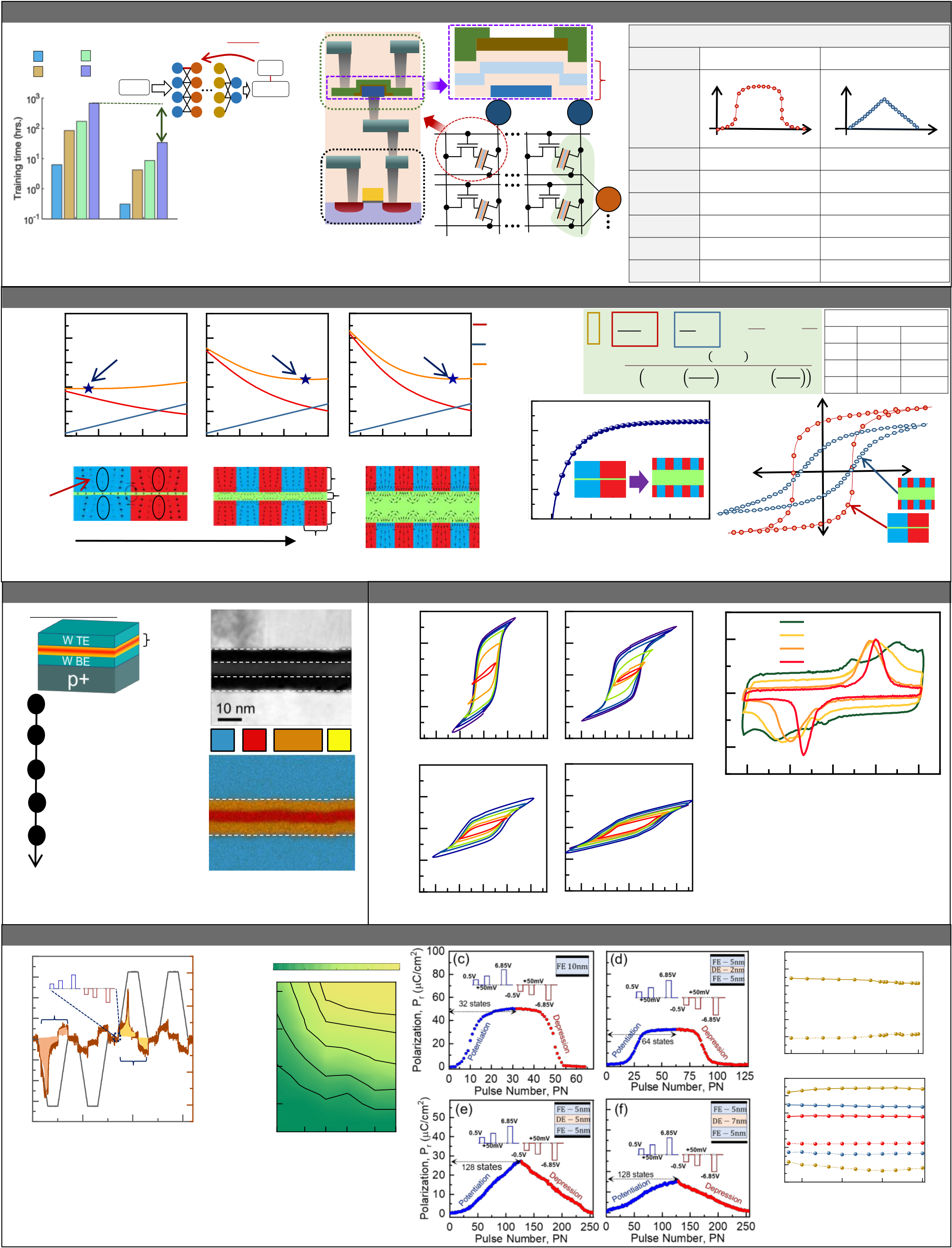
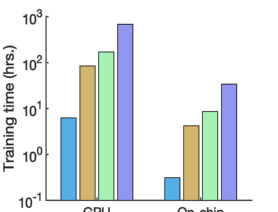
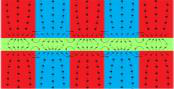
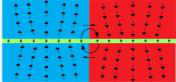
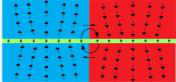
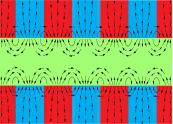
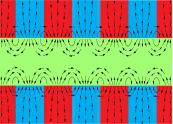
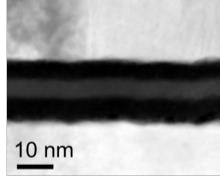
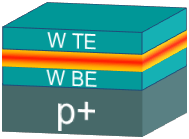
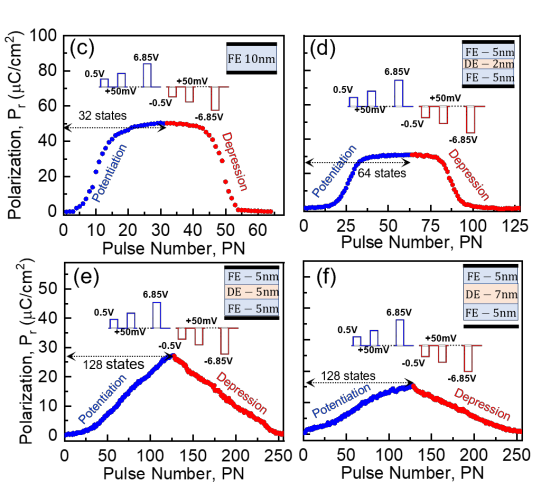
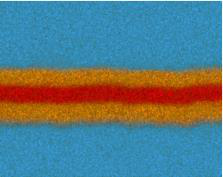
**V.**  **SLFEFETMODELING AND BENCHMARKING**  A compact model of the SL FEFET is developed considering an SL FECAP model coupled to the gate of an

IWO transistor model as shown in Fig. 12 (a). The SL FECAP model is based on the Landau equivalent Positive Feedback model [10] integrated with the gate stack of the semi-empirical physics based Virtual Source model for IWO transistors [11]. Both models: SL FECAP and the IWO transistor, are implemented on SPICE compatible Verilog-A. First, we calibrate the SL FECAP model to the experimental results obtained from the 5nm/5nm/5nm SL FECAP structures, that present an acceptable trade off between linearity, number of states and Pr, as shown in Fig. 12 (b). We couple the calibrated SL FECAP model to the gate stack of the IWO transistor model. The complete model captures the experimental conductance vs number of pulses (Fig. 12 (c)). Starting from the calibrated model, we explore the SL FEFET performance when the gate length is reduced to 50 nm, shown in Fig. 12 (c), which shows a conductance improvement of ~2.5x. The SPICE compatible compact model of SL FEFET captures the potentiation and depression characteristics of the SL FEFET as shown in Fig. 12 (d),(e). We use the results to obtain and benchmark the performance of the SL FEFET based CIM accelerator.Cycle accurate accelerator level simulations are performed with MLP+NeuroSim simulator. A two layer MLP (400 ×100 ×10) is utilized to classify MNIST dataset (Fig.13 (a)). 1T-1 SL FEFET bit cell is assumed for the analog weight device. Fig. 13 (b) shows the SL FEFET array design. During reading, the input is applied to the read word line (RWL) and the partial sum is obtained along the vertical source line which is then converted to 400 digital values using an ADC. During the weight update, the write line (WL) of the selected row is turned on and the programming voltage is applied to the gate of FEFET through bit line (BL). Thanks to the high number of conductance states, excellent linearity and symmetry, and high Gmax/Gmin ratio, the SL FEFET achieves an excellent training accuracy (94.1%) compared to other non-volatile memories (NVMs) while simultaneously maintaining a competitive chip area and latency, as shown in Fig. 13 (c).

**VI.CONCLUSIONS**   
We experimentally demonstrate that, through SL FE domain engineering, it is possible to increase the number of analog states by 4x and linearity by 2x over a conventional FE. We experimentally demonstrate that introduction of a SL layer does not have any negative impact on the retention, endurance or switching speed. We further demonstrate that the improved characteristics of the SL FECAP map reliably in experiments to the BEOL compatible SL FEFET. Finally, the system level benchmark shows that it is possible to achieve 94.1% accuracy in online training with SL-FEFET as analog weight cell without significant trade off in chip area and latency. The training accuracy is markedly higher than any other emerging non-volatile memory.

**References:** [1] S.H. Jo, et al., Nano Letters 2010. [2] W.Wu, et al., VLSI 2018. [3] S. Park, et al., IEDM 2013. [4] J. Woo, et al., IEDM 2016. [5] M. Jerry, et al., IEDM 2017. [6] M.K. Kim, et al., Nano Letters 2019. [7] W. Chung et al., IEDM 2019. [8] X. Sun et al. IEDM 2018 [9] A. Kopal et al., Ferroelectrics 1999. [10] J Gomez et al., IEDM 2019 [11] H. Ye, et al., IEDM 2020.

**Acknowledgement:** This work was supported in part by ASCENT, one of six centers in JUMP, sponsored by DARPA and the Semiconductor Research Corporation (SRC).



**Motivation: Accelerate DNN training with Compute-in-Memory (CIM) + Monolithic 3D (M3D) Analog Ferroelectric FET**

(a)

Backpropagation   
 dError (b) Source

IWO − Channel Drain (c) Conventional vs Superlattice Ferroelectric (FE)

Text Images 𝑤𝑖𝑗 dwij HZO Type Conventional FE Superlattice FE BEOL FEOL   
 Superlattice (SL)   
Audio Videos Error HfO2

HZO

Prototypical Input Output S G D Back Gate

Behavior   
 GDS   
 GDS   
N1 Nn

× 20 Towards On-Chip CIM

Learning: NPulses NPulses

✓ High Density Symmetry High High

✓

✓  
 High Speed

Weight Symmetry   
 G+

+ Linearity Low High

GPU + Ext. DRAM On Chip CIM  
 ✓ Gradual Weight

Change   
 n+   
 G

Si n+

G− −  
 M1 Speed

Analog States   
 High

Low   
 High

High

**Fig.1** (a) Single-GPU with external DRAM training times across different DNN applications; on-chip implementation reduces the training

time drastically, (b) Monolithic 3D CIM provides area, energy and latency advantage by placing high-density embedded memory in a BEOL Endurance High High

pseudo-crossbar array. (c) Superlattice (SL) FEFET improves multiple aspects of conventional FE such as: linearity, number of analog states ΔG Change Nonuniform Uniform

and the conductance update uniformity.

**Domain Engineering though Superlattice (SL) FE Design**

(a)   
 0.5

0.4

0.3

0.2

0.1   
 77 Domains/μm   
 Emin:   
 tDE = 2nm

~249 Domains/μm   
 Emin ∶  
 tDE = 5nm

~259 Domains/μm   
 Emin ∶  
 tDE = 7nm

(b)   
Domain Wall Energy

Total Energy   
Electrostatic Energy

300 300   
 300

250   
 S = ෍  
 F = 8ωS

n=1  
∞  
 π3ε0

n3εD coth πntD   
 P0 2 + ktF

ω P0

sin2n

2𝑤   
 2 tD

+ εF coth πntF   
π 2

(c)  
′ = tDE

εD   
 tF

2𝑤  
′ = tFE

εF

P   
 εFE

εDE

P0

k   
 Parameters

0.7   
32

30

25 μC/cm2

m2/F  
−

−Energy Density (J/m2)   
 # of Domains / mm # of Domains / mm # of Domains / mm   
 0 250 250

100 200 300 100 200 300 100 200 300 200

Domain / mm Domain / mm Domain / mm 200 200

Dipole   
 +  
 − tFE

tDE   
 150 150   
 150 V

(or domain) + −

tFE 100 100 100

Increasing Dielectric Thickness 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5

w Dielectric Thickness tD (nm)

**Fig.2** (a) FE/DE/FE stack reaches equilibrium by minimizing the total free energy; the domain density (#/um) is a result of competition between the electrostatic energy and the domain wall energy; (b) the

ferroelectric decomposes into more domains as the thickness of the dielectric layer is increased; (c) thicker dielectric enables higher depolarization fields which tilts the PV loops.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SL FECAP Fabrication** | | | | | | | | **SL FECAP Characterization** | | | | | | | 0 | 1.5V to | | 1 | (e) | | 7 nm | | | | | |
| Process Flow | | SL | | (b) | W TE | | | 40 | | (a) | | | (b) | |
| (a) | | 20 | |
| 5 nm | | | | | |
| 2 nm | | | | | |
| 5 nm HZO (FE) | | | | | | | | Current (a.u.)   0  -1 |
| 0 | | 1.5V to  4V | | | | | 0 nm | | | | | |
| 5 nm HfO2 (DE) | | | | | | | |
| 5 nm HZO (FE) | | | | | | | | 6V | |
| 1 | p+ Si substrate | W BE | | | | | | P (mC/cm2) | -20 | (c) | 10 nm | | | (d) |
| 5-2-5 | |
| -40 |
| 2 | Sputter W back | W | | | Hf | Hf/Zr | Zr |
| electrode (BE) |
| 3 | 250oC PEALD SL: | | (c) | | | | | 40 |
| HZO / HfO2 / HZO | | | | | | | | 20 | 3.5V to | | -4 | | -2 0 2  Electric Field (MV/cm) | | | | | | 4 |
| 4 | Pattern sputtered W | | | | | | |
| top electrode (TE) | | SL | | | | | | 0 | | 3V to | | | | | **Fig.4** | (a) | P-V | characteristics | | | for | FECAP | with |
| 5 | 60s RTA in N2 at  400oC | | | | | | | conventional 10 nm FE. (b) P-V characteristics for SL | | | | | | | | |
| FECAP with 2 nm DE, (c) 5 nm DE and (d) 7 nm | | | | | | | | |
| -20 | | 6.5V  5-5-5 | | | | | 8V | |
| DE. A thicker DE results in more skewing in the P-V | | | | | | | | |
| **Fig.3** (a) Process flow for the SL FECAP fabrication with schematic. (b) | | | | | | | | 5-7-5 | |
| characteristics. | | | | | | | | |
| -40 | | -6 -3 | 0 | 3 | 6 -6 -3  Voltage (V) | | (e) Measured I-V characteristics showing stretched | | | | | | | | |
| Cross-sectional TEM image and (c) STEM-EDX elemental map of fabricated | | | | | | | | 3 | 6 |
| polarization | | current | | peaks | with | | increasing | DE |
| FECAP with the FE(5nm) /DE(5nm)/FE(5nm) SL stack. | | | | | | | |
| thickness, from 0 nm to 7 nm. | | | | | | | | |

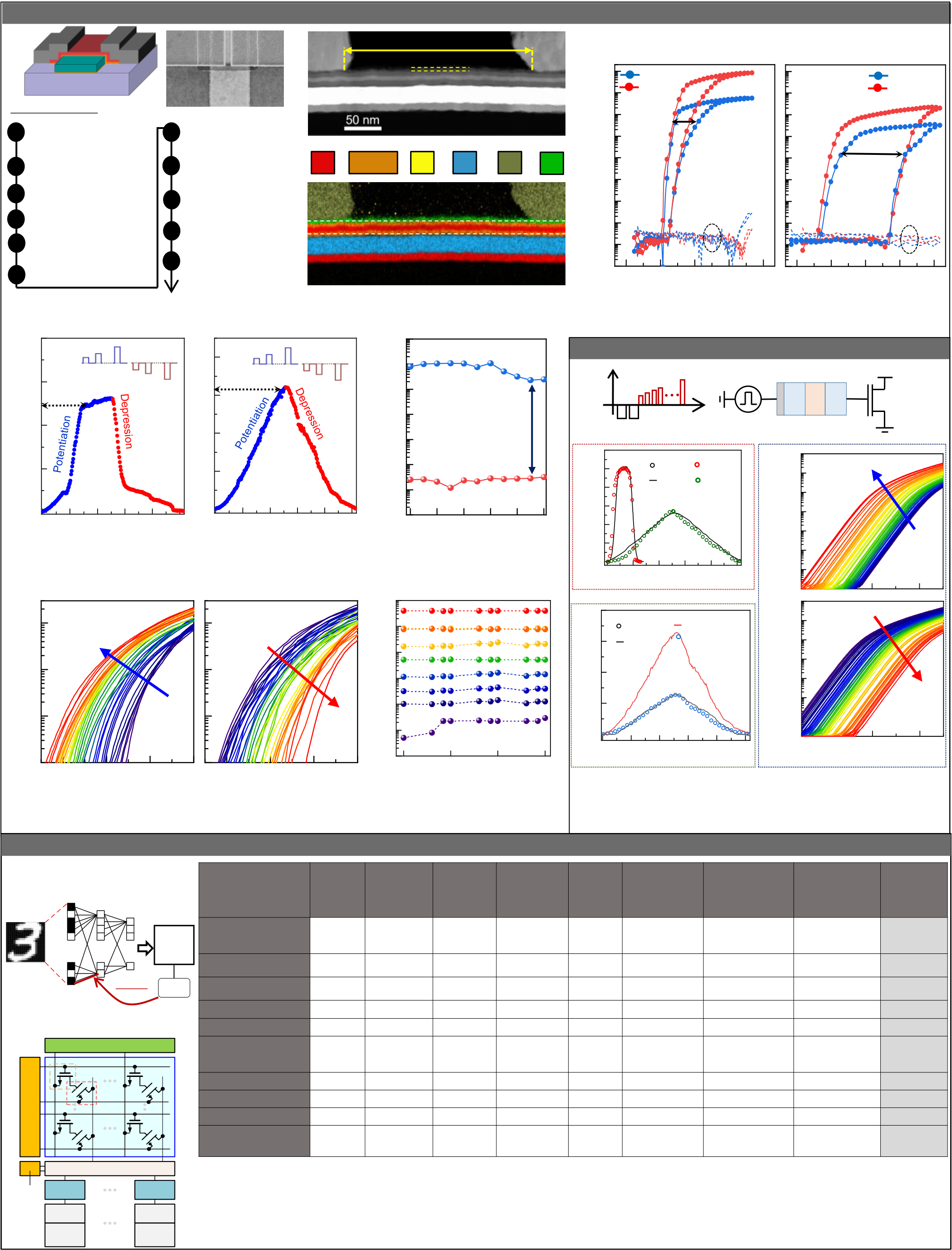
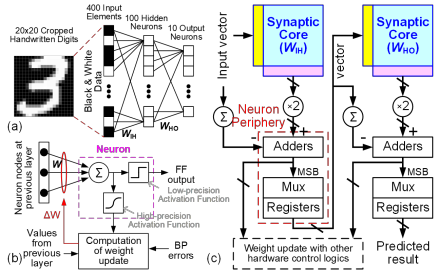
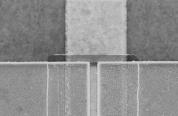
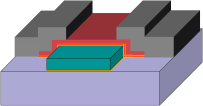
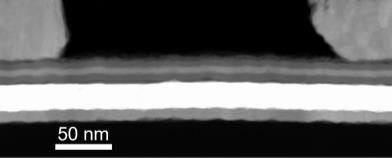
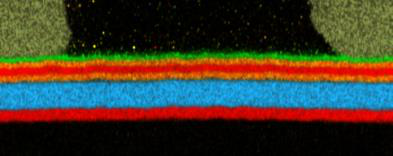
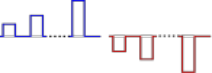
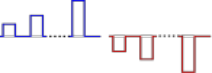
**SL FECAP Characterization (Potentiation, Depression, Switching Speed, Endurance, Retention)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8 | (a) | Programming pulse | | 400 | | (b) | 0% | ΔPr/Pr | | | 75% | 100% | 30   20  Pr (mC/cm2)   10   0  -10 | | (a) | Endurance pulse: | 108 |
| 6 | 25% | | 50% |
| 4 | Q1 | | Q2 | | Current (mA) 200  0  -200 | Pulse Amplitude (V) | 8 | | 90% | | | |
| 2 VFE (V)   0  -2 | 6 | | 75% | | | | +/-5V, 1us |
| 50% | | | | -20 | | |
| 4 | | -30   100 | | | 102 104 106  Number of Cycles |
| 25% | | | |
| -4 | 2 | | 10% | | | | 30  Pr (mC/cm2)   20   10   0  -10 | (b) | |
| -6 | Pr = Q1 – Q2 | | -400 |
| 25oC |
| -8 | | | | | | | | | | | | |
| 0 | 50 100  Time (µs) | | | 150 | | Pulse Width (ns) | | | | | | |
| **Fig.5** (a) PUND measurement showing the extraction technique of switching polarization; | | | | | | | | | | | | | -20 | | |

(b) Switching speed characteristics of SL MFM: showing switching speed up to 100ns. (c)-(f) Measured Pr vs. programming pulse number (PN) for FE/DE/FE structures with increasing dielectric thickness ranging from 0nm to 7nm. By increasing the DE thickness the number of states and linearity increases at the cost of reducing Pr, defining the design space for this system.

|  |  |  |  |
| --- | --- | --- | --- |
| -30 | 10-6 | 10-3 100  Retention time (s) | 103 |

**Fig.6**(a) Measured endurance (>108cycles) and (b) retention (>103seconds) with no degradation in Pr for various states.



**BEOL-compatible Superlattice (SL) FEFET with Indium Tungsten Oxide (IWO) Channel: Fabrication + Characterization**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (a) | | IWO | | Source | | Drain | (b) | Source | | Lg ~ 250 nm | | Pd | Drain | | (d) | IWO FEFET (10nm HZO) | | | | | (e) | IWO SL FEFET (5-5-5) | | | | |
| S | | BG | D | IWO channel | |
| BG | | | (c) | Hf | Hf/Zr | SL | | 10-4 | | | VD : 50mV  VD : 1V | | | | | LG = 250nm | | | VD : 50mV VD : 1V | | |
| SiO2 | | | |
| W BG | |
| Process Flow | | | |
| Etch Stop Layer | |
| 7 | 60 sec RTA in N2 | | MW~2V | | | | |
| SiO2 | | In | 10-6 ID (A)   10-8 | |
| 1 | 90nm SiO2 on p-Si substrate | | |
| MW~5.7V | | | | | |
| at 400oC | | | Ion/Ioff~102 | | | | |
| Zr | W |
| 2 | 250oC Thermal ALD HfO2 etch stop layer | | | 8 | Wet etch sacrificial layer | | LG = 250nm | | | | |
| Ion/Ioff~105 | | | | | |
| 3 | Sputter W film | | | 9 | Pattern sputtered IWO channel | | 50 nm | | SL | | IG | | | | |
| 10-10 | | | IG | | | | | |
| 4 | Pattern and etch W BG | | |
| 10Pattern evaporated Pd S/D pads | | |
| 5 | 250oC PEALD SL: HZO / HfO2 / HZO | | |
| 10-12 | | |
| 11 | 10 min RTA in N2 at 150oC | |
| 6 | Sputter sacrificial W film | | | -6 | -3 | 0  VG (V) | 3 | 6 | -6 | | -3 | 0  VG (V) | 3 | 6 |

**Fig.7**: (a) Key process flow steps for the BEOL-compatible IWO SL FEFET (b) Cross-sectional TEM and (c) STEM-EDX elemental map of the IWO FEFET with the FE (5nm)/DE (5nm)/FE(5nm) SL FEFET (d)-(e) ID-VG characteristics of the IWO FEFET and SL FEFET; the SL FEFET shows a large memory window and higher on/off current ratio compared to the FEFET.

200

(a) 1.5V +10mV 2.77V

-10mV 8 (b) 3.5V +10mV 4.77V

-10mV

10-6   
 VDS = 50 mV, VG = 0 V

150  
 -0.5V

-1.77V 128 states -0.5V

-1.77V   
 Forward sweep

70 states 6

GDS (mS)   
GDS (mS) ID (A)   
 10-8

ID (A) 100

4 4 orders

𝛼𝑝: 𝛼𝑑: 𝛼𝑝: 𝛼𝑑:

50 6.81 -6.51 2 -0.7 -1.56 10-10

Reverse sweep

0

0 50 100 150 200 250   
 0

0 50 100 150 200 250 10-12

1001021041061081010

Pulse Number, PN Pulse Number, PN Number of Cycles

**Fig.8**: Conductance vs pulse number for the BEOL IWO FEFET with (a) **Fig.10**: High endurance (>108) cycles of

10nm FE and (b) SL gate stack. Integrating the latter gate stack reduces 𝛼𝑝 SL FEFET showing no memory window

and 𝛼𝑑, leading to improved linearity and symmetry in the conductance degradation, with distinct Ion-Ioff states

response.

10-7 (a) VDS = 50 mV (b) VDS = 50 mV

10-6

Potentiation Depression

10-8   
 ID (A) 10-8

10-9

10-10

Measured at 25o

10-10

-5

VG (V)  
-4 -3 -2 1

VG (V)   
2 3 4 100

Retention time (s)   
 101   
 C

102 103

**Fig.9**: ID-VG characteristics of the SL FEFET shifted by programming pulses: **Fig.11**: High retention(103s) in the BEOL

(a) potentiation shows gradual decrease in VT ,(b) depression shows gradual SL IWO-FET observed for multiple

increase in VT with successive programming pulses. analog states, each with an associated ID.

**System Benchmarking (Prof Yu)**

**Compact Model for IWO SL-FEFET**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (a)   Vapp | | | | | | | Vdd | | | | | |
| FE DE | | FE | | IWO FET | |
| time | | | | | | Vapp |
| SL FEFET | | | | | |
| (b) | SL − based FECAP calibration | | | | | |
| 60 | | | | | | | (d)  10-6 | Potentiation | | | | |
| Polarization, Pr (mC/cm2) | 50 | Exp. | | 10nm | | |
| 40 | Sim. | | . 5/5/5 | | |
| VDS = 50mV | | | | | |
| 30 | 10-8 ID (A)  10-10 | ΔV = 10mV | | | | |
| 20 |
| 10 |
| 0 |
| 0 | | 50 | 100 | 150 | 200 | 250 | 10-12 | LG = 50nm | | | | |
| Pulse Number, PN | | | | | | |
| (c) | SL FEFET calibration | | | | | |
| (e)   10-6 | Depression VDS = 50mV  ΔV = 10mV | | | | |
| 20 | | | | | | |
| Exp. | | | | LG = 50nm | | |
| 15  Conductance (mS)   10  5 | | Sim. | | LG = 250nm | | |
| 10-8  ID (A)  10-10 | | | | | |
| 0 | | | | | | | 10-12 | 0 | | LG = 50nm | | |
| 1  VG (V) | | 2 |
| 0 | | 50 100 150 200  Pulse Number, PN | | | | 250 |

**Fig.12**: (a) Verilog-A based simulation setup and pulse scheme. (b) Pr as a function of pulse number, showing good agreement between modelled and experimental

FECAP data. (c) Model calibration for FEFET performance projection of the SL gate

stack FEFET based on conductance response. (d) ID-VG characteristics of Verilog-A based model with successive potentiation and (e) successive depression.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (a) Training MLP network on | | | | | | | | (c) | **Devices** | **RRAM Ag:** | | **RRAM** | **RRAM** | **AlOx/HfO2 [4] FEFET [5] FETFT IGZO [6] pFEFET Ge NW [7] Hybrid 2 CMOS T** | | | | | | | **SL FEFET** |
| MNIST dataset  **400 Input** | | | | | | | | |
| **Elements** | | | |  |  | | --- | --- | | **s** | **100 Hidden** | | | | | | | **a-Si [1]** | | **TaOx/HfOx[2]** | **PCMO [3]** | **(this work)** |
| **Black &**  **White Data** | | | **Neurons**  4 | | | | | | **# of conductance** | 97 | | 128 | 50 | 40 | | 32 | 64 | 320/256 | | 64 | **128** |
| **states** |
| **Nonlinearity: αp, αd** | 2.4/-4.88 | | 0.04/-0.63 | 3.68/-6.76 | 1.94/-0.61 | | 1.75/1.46 | -0.8/-0.69 | 1.2/-1.75 | | 0.5/0.5 | **-0.7/-1.56** |
| wij | | | | dError | | | Error | |
| dwij | | | **Asymmetry,|αp-αd|** | 7.28 | | 0.67 | 10.44 | 2.55 | | 0.29 | 0.11 | 2.95 | | 0 | **0.86** |
| Backpropagation  (b) SL FEFET-based synaptic core  BL Switch Matrix | | | | | | | | | **RON** | 26 MΩ | | 100 kΩ | 23 MΩ | 16 kΩ | | 559.28 kΩ 5.4M(0.185μS) | | | 5kΩ (~200μS) | 559.28 kΩ | **61 kΩ** |
| **Gmax/Gmin** | 12.5 | | 10 | 6.84 | 4.43 | | 45 | 14.4 | 500 | | 45 | **285.7** |
| **Online learning** | 72% | | 80% | 33% | 20% | | 88% | 91.5% | 85.8% | | 94.3% | **94.1%** |
| WL/RS Switch Matrix | WL  Access  Transistor | | | | | SL | | | **accuracy** |
| **Chip Area** | 6292µm2 | | 8663µm2 | 6292.4µm2 | | 21760µm2 | 7032.6µm26.23×108 μm2 | | 21,442μm2 | | 232, 320μm2 | **11,508μm2** |
| RWL | | | | | **Latency (125 epochs)** | | 31997s | 10.15s | 12218s | 470.42s | | 2.73s | 1.97×106s | 108.1.1s | | 0.174s | **38.34s** |
| FeFET | | | | | **Energy** | 13.44mJ | | 4.01mJ | 2.53mJ | 15.26mJ | | 1.9mJ | 9.83mJ | 94.99mJ | | 3.24mJ | **17.99mJ** |
| BL | | | | |
| **FeFET Array** | | | | | **TOPS/W (ops in 125** | | 9.15 | 30.67 | 48.62 | 8.06 | | 64.7 | 0.0125 | 1.29 | | 37.96 | **6.83** |
| **epochs)** | |
| Mux | | | | | | | | | **Fig.13**: We perform a system benchmark with MLP+NeuroSim simulator. (a) Two-layer MLP network (400×400×10) is utilized to classify MNIST dataset in which 125 training epochs were considered. (b) Synaptic array design: 1T-1FEFET bit cell is assumed for the FEFET synaptic device with Lg = 50nm and W=100nm. During read, the input is applied at the read word line (RWL), and the partial sum is obtained along the vertical source line. It is then converted to digital values using an ADC. During the weight update, the WL of the selected row is turned on and the programming voltage is applied to the gate of FEFET through BL. (c) Benchmarking results show the BEOL SL FEFET providing high online learning accuracy thanks to the high number of conductance states, low nonlinearity, and low asymmetry. This is achieved without excessive penalties in latency and area. The high energy consumption is due to the relative low value of RON. | | | | | | | | | | | | |
| Mux  Decoder | | ADC | ADC | | | | | |
| Adder | Adder | | | | | |
| Shift  Register | | | | | Shift  Register | | | |