Ferroelectric FET Analog Synapse for Acceleration of Deep Neural Network Training

Matthew Jerry[1], Pai-Yu Chen[2], Jianchi Zhang[1], Pankaj Sharma[1], Kai Ni[1], Shimeng Yu[2] and Suman Datta[1]
[1]Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
[2]School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA
Phone: (574)-631-5480; Fax: (574)-631-4393; email: mjerry@nd.edu

***Abstract*—** The memory requirement of at-scale deep neural networks (DNN) dictate that synaptic weight values be stored and updated in off-chip memory such as DRAM, limiting the energy efficiency and training time. Monolithic cross-bar / pseudo cross-bar arrays with analog non-volatile memories capable of storing and updating weights on-chip offer the possibility of accelerating DNN training. Here, we harness the dynamics of voltage controlled partial polarization switching in ferroelectric-FETs (FeFET) to demonstrate such an analog synapse. We develop a transient Presiach model that accurately predicts minor loop trajectories and remnant polarization charge (P_r) for arbitrary pulse width, voltage, and history. We experimentally demonstrate a 5-bit FeFET synapse with symmetric potentiation and depression characteristics, and a 45x tunable range in conductance with 75ns update pulse. A circuit macro-model is used to evaluate and benchmark on-chip learning performance (area, latency, energy, accuracy) of FeFET synaptic core revealing a 10^3 to 10^6 acceleration in online learning latency over multi-state RRAM based analog synapses.

I. Introduction

Deep neural networks have demonstrated success in cognitive tasks such as speech and image recognition (Fig. 1(a)). However, the energy consumption and training time of at-scale deep neural networks are limited by off-chip memory (DRAM) access bottleneck owing to the large memory requirements of the weight matrices. For a fully connected DNN, significant acceleration in training can be achieved by minimizing data movement by utilizing on-chip storage and performing weight updates at the same node, where all the nodes are all connected together in an array. The ability to expand on-chip SRAM cache sizes to accommodate training of larger networks is limited due to $150F^2$ (F is the smallest patterned feature) cell size and motivates the development of an area efficient high speed analog synapse capable of on-chip learning. Device requirements for accelerating the training of DNNs include ±1V, 1 nanosecond potentiation and depression programming pulses, a symmetric and linear conductance response with ≥32 conductance states (≥5-bit), and a G_{max}/G_{min} ratio of >10 [1][8]. Emerging non-volatile memories such resistive random access memory (RRAM) and phase change memory (PCM) are potential candidates owing to their small cell size ($4F^2$) and the ability to program the cells with multiple intermediate states. However, achieving symmetric potentiation and depression characteristics, with nanosecond pulse widths, and sufficient G_{max}/G_{min} ratios in RRAM/PCM has not been realized (Fig. 2). In the case of filamentary RRAM, the lack of symmetry in the characteristics [2] results in the utilization of fewer conductance states, and often quenching $G_{max}/G_{min}<10$ in addition, which negatively impacts the system performance. While interfacial RRAM and RRAM exploiting multiple weak filaments [3]–[5] exhibit increased symmetry between potentiation and depression, the program pulse widths are as large as 10ms due to the slow diffusion process of oxygen vacancies during weak programming. Therefore, training with a modest 1M images from the MNIST database on such devices would take years (>$5.6x10^7$s) on such devices (Fig. 17). In this work, we harness electric-field controlled partial polarization switching in ALD ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ (HZO) [6] to demonstrate a FeFET based analog synapse (Fig. 1(c)). The FeFET synapse exhibits, highly symmetric conductance values for potentiation and depression, in a pseudo-crossbar array, for 75ns pulses with progressive amplitude (Fig. 1(b)). Therefore, enabling high speed training of networks with high online learning accuracy (90%), in a scaled cell footprint.

II. Metal-Ferroelectric-Metal Capacitor Fabrication and Modeling

The FeFET synapse utilizes multi-domain polarization switching dynamics in ferroelectric HZO thin films to gradually tune the threshold voltage (V_T) of the underlying channel, and consequently its drain-to-source conductance, by the application of short voltage pulses to the gate. First, we measure and model the dynamics of the metal-ferroelectric-metal (MFM) capacitors (TiN/10nm HZO/TiN). The multi-domain effects are modeled using a Presiach theory of hysteretic switching where the MFM capacitor response is assumed to be an aggregate response of a distribution of individual domains with discrete and deterministic coercive fields (E_c^+ and E_c^-) (Fig. 3). The dynamic response of each domain is then computed by rescaling the applied voltage amplitude with an experimentally calibrated transfer function, that is dependent on the applied pulse width. The scaled effective voltage (V_{EFF}) is then applied to the static model to compute the net polarization (P) response from a write pulse of arbitrary duration and amplitude [7] (Fig. 3). Fig. 4(a-c) show that the model captures both the history and the memory wipeout of the minor loop trajectories of the MFM device.

We study the effect of multiple pulse schemes on the remnant polarization to optimize the number of accessible polarization states (Fig. 6(a)-(c)). The remnant polarization charge (P_r) is calculated by integrating the transient current (Fig. 5). The experimental and simulation results are in excellent agreement as shown in Fig. 6(a)-(c). Pulse schemes

1 and 2 exhibit regions of near linear P_r change, but the number of available states is limited (6-12 states) (Fig. 6(a)-(b). While scheme 3 (Fig. 6(c)) accesses >50 stable intermediate polarization states as it directly samples from the assumed distribution (Fig. 3) of E_c^+ and E_c^-. Therefore, scheme 3 results in symmetric and sigmoidal potentiation and depression characteristics. The simulated trajectory of the polarization (1→2→3) in response to three example potentiation pulses from pulse schemes 2 and 3 are shown in Fig. 7(a)-(b), highlighting the enhanced control of ΔP_r in scheme 3. Point 1 indicates the initial P_r state while Point 3 is the new stable P_r state. Next, we evaluate the channel conductance of a FeFET based on the polarization states of Scheme 3 (Fig. 6(c)). The FeFET characteristics are simulated by computing the surface potential $\boldsymbol{\psi_s}$ based on charge sharing between the capacitors in an MFIS stack using the Pr values from Fig. 6(c) and experimentally extracted FET parameters. The I_{DS}-V_{GS} characteristics are shown in Fig. 8(a)-(b). The simulated channel conductance (G_{ds}) as a function of pulse number for potentiation and depression results in a symmetric conductance response with G_{max}/G_{min} ratio of 69 (Fig. 9) highlighting the potential of the FeFET as an analog synapse candidate.

III. FeFET Analog Synapse Fabrication, Characterization, and Analysis

Fig. 10 summarizes the fabrication process flow of n-channel FeFETs. The ferroelectric gate stack consists of 10nm thick ALD HZO deposited on p-Si with 0.8nm thick interfacial SiO_2 layer (confirmed by TEM), capped by ALD TiN layer, followed by a 600°C anneal. This gives rise to multiple ferroelectric domains within a nanocrystaline structure of HZO. The conductance behavior in response to pulse schemes 1, 2, 3 are shown in Fig. 11(a)-(c). The shape of the measured FeFET channel conductance (G_{ds}) vs pulse number curves mirrors that of the P_r vs pulse number response, as shown previously in Fig. 6(a)-(c). This confirms that the programming sequence is deterministically switching a fraction of the total FE domains with each pulse. Scheme 3 exhibits the highest number of programmed states, 32 (5-bit), as it encompasses a more uniform sampling from the domain coercive field (E_c^+ and E_c^-) distributions compared to schemes 1 and 2. Pulse widths are limited to 75ns currently constrained by the experimental FeFET dimensions and will scale with geometry. The G_{max}/G_{min} ratios are extracted for each scheme and are compared with other reported values in Fig. 12(a). The ratio directly correlates with the neural network accuracy, where G_{max}/G_{min}>10 is required to achieve accuracies of >80% [8]. FeFET achieves a ratio of 47× near the ideal design target of 50× [8]. Fig. 12(b) shows the extracted non-linearity (as described in [8]) and corresponding asymmetry parameters for each pulse scheme. The equal sign (+) and near zero value of, α_p=1.75 α_d=1.46, indicated the linearity, and lead to a near ideal asymmetry value of 0.29 (ideal asymmetry=0, implying perfectly symmetric characteristics) highlighting the symmetric behavior of the FeFET between potentiation and depression using scheme 3. The corresponding transient I_{DS}-V_{GS} characteristics for all 32 potentiation and depression states are shown in Fig. 13.

IV. FeFET Synaptic Core Macro Model & Bencharmking

We benchmark the experimental FeFET with SRAM and other analog RRAM synaptic devices using a 2-layer multilayer perceptron (MLP) neural network (Fig. 14(a)) with the support of a circuit-level macro model, NeuroSim [8], by estimating the chip area, latency, dynamic energy and leakage power. We develop the cell structure of FeFET in a pseudo-crossbar array to realize key operations such as vector-matrix product (weighted sum Fig. 14(b)) and weight update (Fig. 15). The FeFET synaptic core is simulated with the supporting peripheral circuits (Fig. 16). The benchmarking highlights the increased learning accuracy (>90%) and faster speed (75ns) of FeFET than reported RRAMs for a training cycle comprising of 1M images from MNIST database (Fig. 17). Further, FeFET synaptic cores achieve a 10× reduction in area and >30× reduction of leakage power compared to a 6-bit SRAM cache with similar accuracy.

V. Conclusion

In conclusion, we experimentally demonstrate a FeFET analog synapse based on partial polarization switching, for acceleration of on-chip learning in deep neural networks. A transient Presiach model quantitatively captures the dynamics of voltage controlled partial polarization switching in 10nm HZO films. The fabricated FeFET synapse exhibits symmetric 5-bit potentiation and depression characteristics, resulting in 90% accuracy for image recognition after training on the MNIST database. Further, the 75ns experimental programing pulse width improves training time on 1M images by 1000× compared to demonstrated RRAM devices while maintaining a 10× area advantage over SRAM.

VI. Acknlowledgements

This project was supported by the National Science Foundation under grant 1640081 and 1552687, and the Nanoelectronics Research Corporation (NERC), a wholly-owned subsidiary of the Semiconductor Research Corporation (SRC), through Extremely Energy Efficient Collective Electronics (EXCEL), an SRC-NRI Nanoelectronics Research Initiative under Research Task IDs 2698.001. The authors thank useful discussion with Wilfried Haensch of IBM Research, Yorktown Heights.

VII. References

[1] T. Gokmen, *et al.*, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations," *Front. Neurosci.*, 10, 1–13, 2016.

[2] J. Woo, *et al.*, "Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Lett.*, 37, 8, 994–997, 2016.

[3] L. Gao, *et al.*, "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology*, 26, 45, 455204, 2015.

[4] S. Park, *et al.*, "Neuromorphic speech systems using advanced ReRAM-based synapse," *International Electron Devices Meeting*, 2013.

[5] S. H. Jo, *et al.*, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, 10, 4, 1297–1301, 2010.

[6] S. Oh, *et al.*, "HfZrOx -based Ferroelectric Synapse Device with 32 levels of Conductance States for Neuromorphic Applications," *IEEE Electron Devices Lett.*, 99, 732–735, 2017.

[7] J. Chow, *et al.*, "A voltage-dependent switching-time (VDST) model of ferroelectric capacitors for low-voltage FeRAM circuits," *Symp. VLSI Circuits. Dig. Tech. Pap.*, 2004.

[8] S. Yu, *et al.*, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," *International Electron Devices Meeting, IEDM*, 2016.

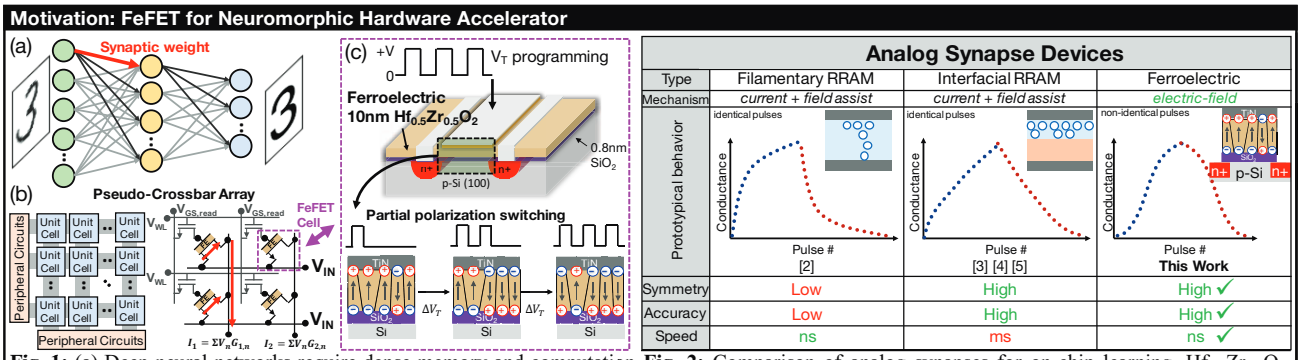

Fig. 1: (a) Deep neural networks require dense memory and computation of inner-dot products. (b) Structure of FeFET psuedo-crossbar array. (c) Principle of analog synapse operation where partial polarization switching results in gradual programming of the channel conductance (G_{ds}).

Fig. 2: Comparison of analog synapses for on-chip learning. $Hf_{0.5}Zr_{0.5}O_2$ (HZO) FeFET based analog synapse exhibits the desired characteristics of high speed electric-field controlled switching and symmetric potentiation and depression, allowing fast training of high accuracy neural networks.

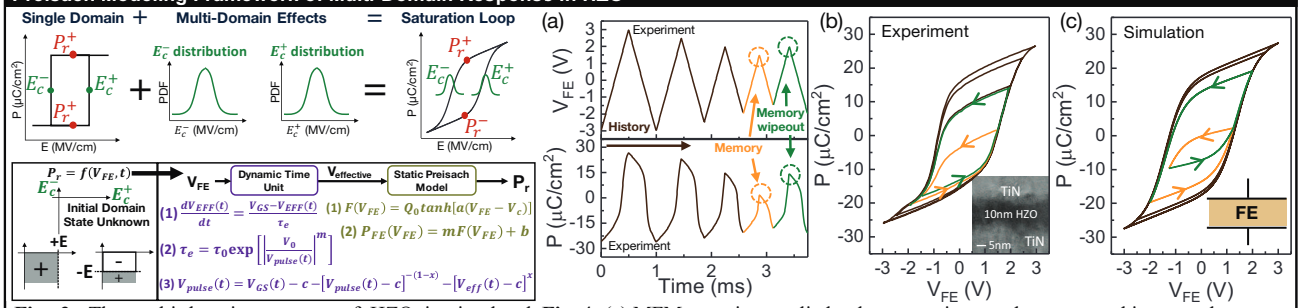

Fig. 3: The multi-domain response of HZO is simulated using a dynamic Preisach model which accurately captures the transient trajectory and polarization of the ferroelectric for arbitrary voltage amplitudes and pulse widths as in [7].

Fig. 4: (a) MFM capacitor applied voltage transient used to measure history and memory wipeout of previous minor loop trajectories. This is observed in experiment by the reduction in the slope (dP/dV) of minor loops near E_c (b). (c) The simulated response of the ferroelectric accurately captures the history and trajectory of the minor loops.

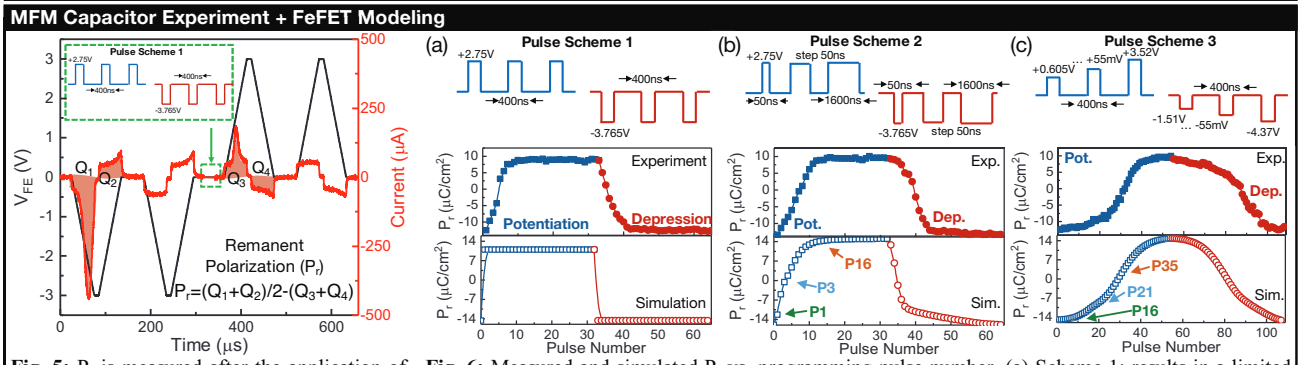

Fig. 5: P_r is measured after the application of potentiation/depression pulses, where P_r is calculated by integrating the transient current according to the equation in the lower right.

Fig. 6: Measured and simulated P_r vs. programming pulse number. (a) Scheme 1: results in a limited number of polarization states with asymmetric response. (b) Scheme 2: modulation of the pulse width improves upon the characteristics at the cost of increased latency. (c) Scheme 3: exhibits the greatest number of states with symmetric response due to optimal sampling of the $E_c^{\pm}$ distributions.

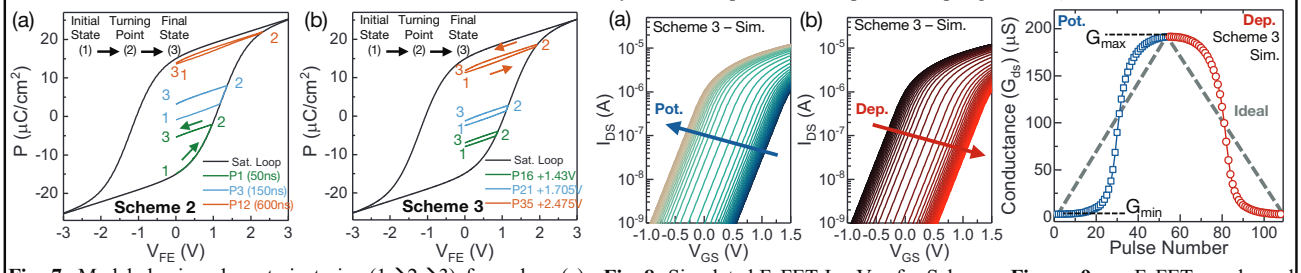

Fig. 7: Modeled minor loop trajectories (1→2→3) for pulse (a) Scheme 2 and (b) Scheme 3. In scheme 2 the increase in P_r (1→3) is non-equal for each pulse number (P#). While scheme 3 each pulse increases P_r (1→3) similarly, creating a more linear response.

Fig. 8: Simulated FeFET I_{DS}-V_{GS} for Scheme 3 polarization values. (a) Potentiation shows gradual decrease in V_T (b) depression gradual increase in V_T. Vds=50mV.

Fig. 9: FeFET channel conductance (G_{ds}) vs. pulse number is symmetric and G_{max}/G_{min}=69. $V_{gs,read}$=1.2V.

Synaptic FeFET Characterization

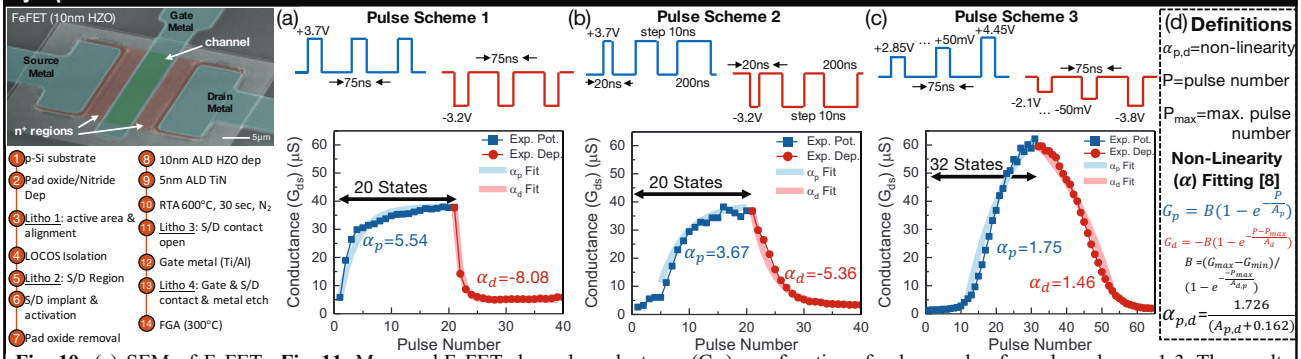

Fig. 10: (a) SEM of FeFET with 10nm ALD $Hf_{0.5}Zr_{0.5}O_2$ and 0.8nm interfacial SiO_2 gate stack. (b) Process flow.

Fig. 11: Measured FeFET channel conductance (G_{ds}) as a function of pulse number for pulse schemes 1-3. The results match the shape of the polarization response in Fig. 6. (a) Pulse schemes 1 and (b) 2 result in low G_{max}/G_{min}, high non-linearity ($\alpha_{p,d}$), and high asymmetry ($|\alpha_p - \alpha_d|$). (c) Scheme 3 exhibits G_{max}/G_{min}=45 and symmetric characteristics, ideally suited for on-chip learning. (d) Definitions of fitting equations for non-linearity ($\alpha_{p,d}$), asymmetry extraction.

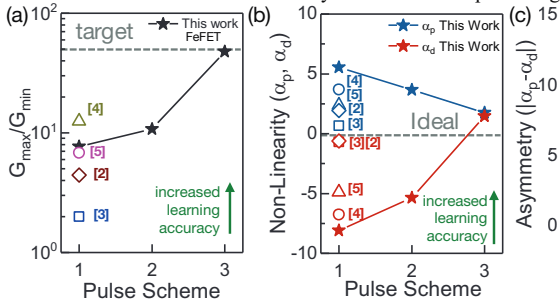

Fig. 12: (a) Comparison of G_{max}/G_{min} ratios as a function of pulse scheme (target=50x) [8]. The non-linearity (ideal $\alpha_{p,d}$=0) and asymmetry (ideal=0) of the potentiation and depression characteristics are shown in (b) and (c). FeFET scheme 3 achieves near ideal values of G_{max}/G_{min}=45, α_p=1.75, α_d=1.46, and asymmetry of 0.29.

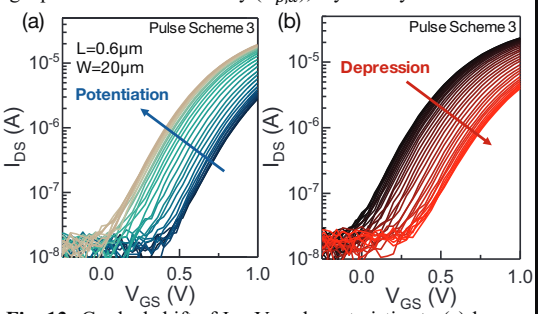

Fig. 13: Gradual shift of I_{DS}-V_{GS} characteristics to (a) lower V_T states with successive potentiation programming and (b) higher V_T states with successive depression programming. Charge trapping reduces the total V_T swing.

Device and System Benchmarking

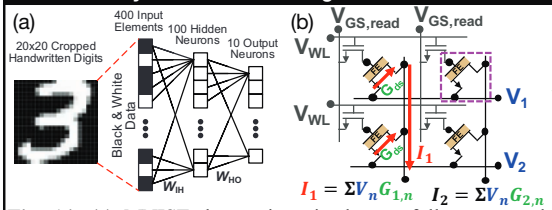

Fig. 14: (a) MNIST dataset is trained on a fully connected multi-layer perceptron using NeuroSim [8]. (b) FeFET synapse read scheme where the pseudo-crossbar array calculates the product of the input vector (V_n) and weight matrix ($G_{ds(n,m)}$).

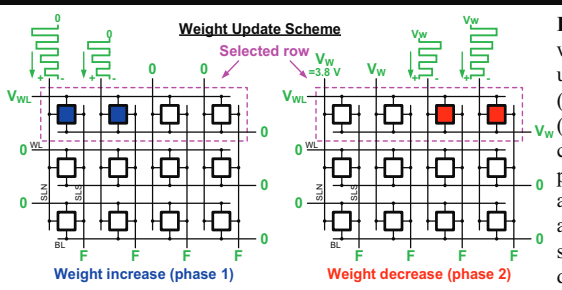

Fig. 15: Row-by-row weight update scheme used to potentiate (*left*) and depress (*right*) the FeFET conductance within a pseudo-crossbar array. Weight updates are calculated via stochastic gradient descent [8].

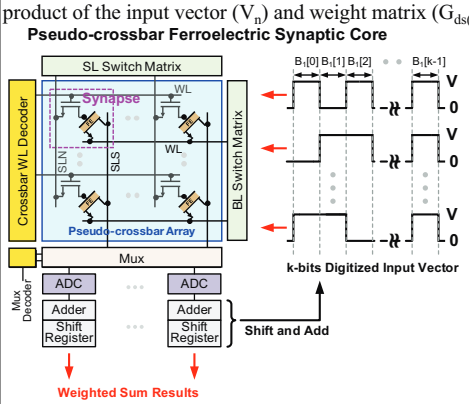

Fig. 16: Schematic of FeFET synaptic core macro including peripherals used to benchmark system performance of FeFET synapse against other reported analog RRAM synapses.

Analog eNVM type	TaO_x/TiO_2 [3]	PCMO [4]	Ag:a-Si [5]	AlO_x/HfO_2 [2]	FeFET (This Work)	6-bit SRAM
# of conductance states	102	50	97	40	32	--
Nonlinearity (weight increase/decrease)	0.66/-0.69	3.68/-6.76	2.4/-4.88	1.94/-0.61	1.75/1.46	--
Asymmetry	1.35	10.44	7.28	2.55	0.29	
R_{ON}	5 MΩ	23 MΩ	26 MΩ	16.9 kΩ	559.28 kΩ	
ON/OFF ratio	2	6.84	12.5	4.43	45	--
Weight increase pulse	3V/40ms	-2V/1ms	3.2V/300μs	0.9V/100μs	3.65V (avg.)/75ns	--
Weight decrease pulse	-3V/10ms	2V/1ms	-2.8V/300μs	-1V/100μs	-2.95V (avg.)/75ns	--
Weight update cycle-to-cycle variation (σ)	<1%	<1%	3.5%	5%	<0.5%	--
Accuracy for online learning	~10%	~10%	~73%	~41%	~90%	~94%
Area	1,071.3 μm²	1,071.3 μm²	1,072.0 μm²	3,657.2 μm²	1,190.4 μm²	10,311 μm²
Latency for online learning (1M images)	1132 years ($3.57x10^{10}$ s)	22.19 years ($7.00x10^{8}$ s)	13.3 years ($4.20x10^{8}$ s)	1.77 years ($5.60x10^{7}$ s)	9.33 hours (3.36E4 s)	7.76 s
Energy for online learning (1M images)	65.86 mJ	29.4 mJ	87.94 mJ	150 mJ	98.01 mJ	6.98 mJ
Leakage power	35.29 μW	35.29 μW	35.29 μW	35.29 μW	35.29 μW	1.1 mW

Fig. 17: Benchmark of system level performance (including peripheral circuits) on 1M images from the MNIST database. FeFET based synapse achieves the highest network accuracy and training speed compared to other analog RRAM. 6-bit SRAM maintains a speed and accuracy advantage but increases chip area 10x compared to FeFET, creating difficulty to scale the network size with SRAM.