Ferroelectric FET based Non-Volatile Analog Synaptic Weight Cell

Matthew Jerry, Sourav Dutta, Kai Ni, Jianchi Zhang, Pankaj Sharma, Suman Datta
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
Email: mjerry@nd.edu

Arman Kazemi, X Sharon Hu, Michael Niemier
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556

Pai-Yu Chen
School of Electrical, Computer, and Energy Engineering
Arizona State University
Tempe, AZ 85287

Shimeng Yu
School of Electrical and Computer Engineering
Gerogia Institute of Technology,
Atlanta, GA 30332

***Abstract*— Dense analog synaptic crossbar arrays are a promising candidate for neuromorphic hardware accelerators due to the ability to mitigate data movement by performing in-situ vector-matrix products and weight updates within the storage array itself. However, many analog weight storage cells suffer from long latencies or low dynamic ranges, limiting the achievable performance. In this work, we demonstrate that the voltage-controlled partial polarization switching dynamics in ferroelectric-field-effect transistors (FeFET) can be harnessed to enable a 32 state non-volatile analog synaptic weight cell with large dynamic range (67×) and low latency weight updates (50 ns) for an amplitude modulated pulse scheme.**

Keywords—neuromorphic computing, analog synapse, ferroelectric, field-effect-transistor

III. INTRODUCTION

The confluence of steadily increasing computing power and the availability of large datasets has enabled deep neural networks (DNNs) and convolutional neural networks (CNNs) to perform complex cognitive tasks such as language translation, image recognition, and computer vision with unprecedented accuracy [1]. However, the computational demands of larger datasets, deeper networks, as well as energy constraints of mobile and edge devices pose challenges for current hardware systems in terms of data movement and memory technology. Large networks typically exceed the size of available on-chip static random access memory (SRAM) caches and expanding their size is limited by the large cell area (100–200F^2, where F is the smallest patterned feature). Therefore, off-chip high-bandwidth memory such as dynamic random-access memory (DRAM) is often used for storing network parameters but comes at the expense of lower energy-efficiency and longer latency compared to on-chip solutions owing to the von-Neumann bottleneck.

This has spurred investigation of new architectures [2], [3] and devices[4], [5], which can both accelerate and increase the energy-efficiency of training and inference machine learning tasks by reducing data movement. One such device oriented approach involves the development of a non-volatile analog synaptic memory device that can be densely integrated within crossbar or pseudo-crossbar arrays such that the weight values of fully connected layers can be directly stored in the non-volatile weight storage cells. Thereby, it enables analog vector-matrix multiplication and weight updates to be performed within the storage array itself, reducing the latency and energy cost of data movement between logic and off-chip memory. Additionally, a multi- state analog synaptic weight cell would enable significantly denser on-chip storage due to not only the advantage in cell size (4–24F^2 compared to 100–200F^2 for SRAM), but also a reduction in the total number of cells required to store individual weights. However, in order to realize an acceleration in DNN or CNN training, the analog synaptic memory element requirements include >5-bit conductance levels per cell [6], with a G_{max}/G_{min} ratio >50× [6], where the conductance levels are linearly spaced and modulated (potentiation and depression) by fast identical voltage pulses of 1 nanosecond duration within range of on-chip voltage levels (<1.5 V), all with minimal variation [6].

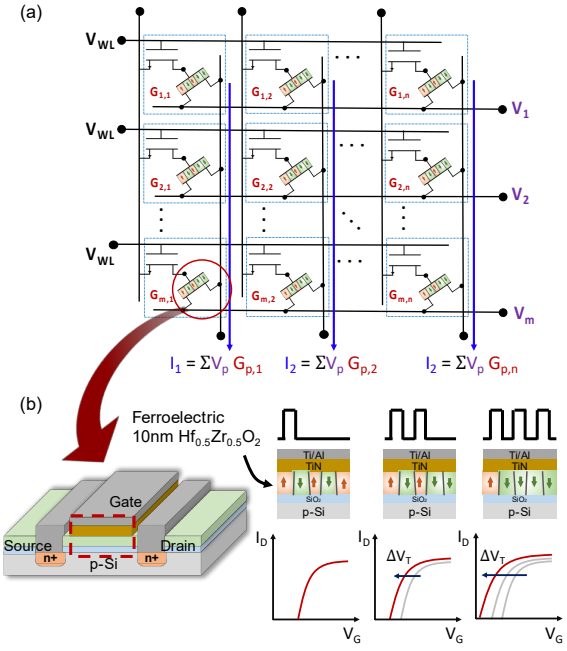

Figure 1. (a) Ferroelectric field-effect transistor (FeFET) pseudo-crossbar array enabling analog vector-matrix multiplication and row-wise parallel weight updates. Each synaptic weight cell utilizes an access transistor in addition FeFET storage device to reduce disturbance effects. (b) Storage of analog conductance values within the FeFET result from partial polarization switching within the ferroelectric gate oxide ($Hf_{0.5}Zr_{0.5}O_2$). Changes in the net polarization charge shift the transistor threshold voltage, and therefore the channel conductance (for a fixed gate read voltage).

IV. Results and Discussions

Till date, such an analog synaptic memory element has not yet been realized with various material systems including phase change memory (PCM), resistive random-access memory (RRAM), and ferroelectrics currently being explored as solutions. PCM, although attractive due to their multi-state analog capabilities and $4F^2$ cell size, exhibits an abrupt reset characteristic, while oxygen vacancy based RRAM devices often suffer from cycle-to-cycle variation and small G_{max}/G_{min} ratios. Further, slow write times demonstrated in the range of μs to ms can result in training times of several years [6], when training on a modest one million images from the MNIST database. In this work, we demonstrate the potential use of electric-field controlled partial polarization switching in ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ (HZO) [6] to demonstrate a ferro-electric field-effect-transistor (FeFET) based analog synapse (figure 1(b)). The FeFET synapse combines the ability to modulate the ferroelectric polarization charge using high speed weight update pulses with a large dynamic range (G_{max}/G_{min}) due to the underlying metal-oxide-semiconductor field-effect transistor (MOSFET) (figure 1(b)). The FeFET synapse can be integrated into pseudo-crossbar arrays suit- able for parallel row-wise weight update and column-wise weighted sum of the individual FeFET conductance's (figure 1(a)). The experimentally demonstrated FeFET synapse exhibits a 67× G_{max}/G_{min} ratio using 50 ns amplitude modulated weight updates (figure 2(a-d)) [7]. When benchmarked using a circuit-level macro model, NeuroSim+ [6], the FeFET synapse achieves an enhanced accuracy (~90%) and orders of magnitude lower latency (0.876s) compared to demonstrated RRAM devices for training on one million images from the MNIST database.

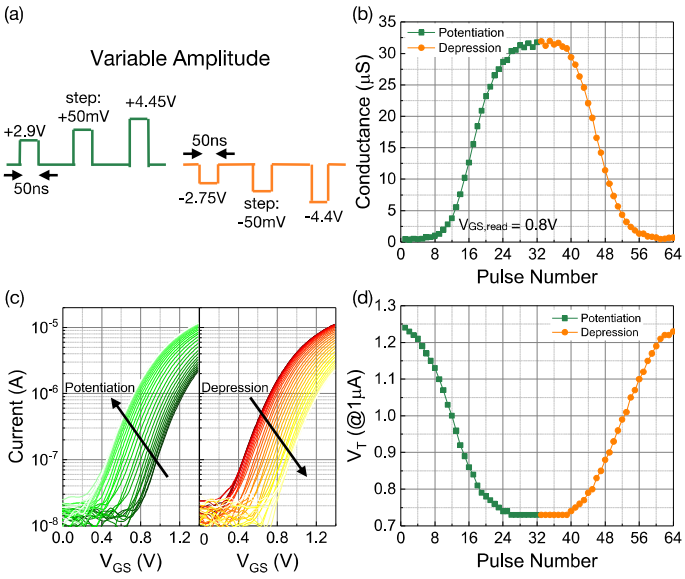

Figure 2. (a-d) Variable amplitude pulse scheme results in a symmetric response with the largest dynamic range (G_{max}/G_{min}) compared to identical pulse and variable pulse widths schemes as it accesses the full FeFET memory window.

III. Conclusion

We experimentally demonstrate these dynamics can be harnessed for developing FeFET analog synaptic memory in neural network hardware accelerators. The fabricated FeFET synapse exhibits 32 analog states that can be modulated symmetrically (potentiation and depression) using a variable amplitude pulse scheme with 50 ns pulse widths over a large dynamic range of G_{max}/G_{min} = 67×. A system level benchmarking demonstrates that such an analog synaptic weight cells can be densely integrated within the configuration of pseudo-crossbar arrays for building neuromorphic hardware with accelerated learning capability and high inference accuracy.

Acknowledgment

This work was supported in part by ASCENT, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA under task number 2776.038. Additionally, 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.

References

[1] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nat. Methods*, vol. 13, no. 1, pp. 35–35, Dec. 2015.

[2] T. Gokmen and Y. Vlasov, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," *Front. Neurosci*., vol. 10, no. JUL, pp. 1–13, Jul. 2016.

[3] T. Gokmen, M. Onen, and W. Haensch, "Training Deep Convolutional Neural Networks with Resistive Cross-Point Devices," *Front. Neurosci*., vol. 11, no. October, pp. 1–22, Oct. 2017.

[4] M. Jerry, P. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, and S. Datta, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *IEEE International Electron Devices Meeting (IEDM)*, 2017, p. 6.2.1-6.2.4.

[5] S. Kim, T. Gokmen, H. Lee, and W. E. Haensch, "Analog CMOS-based resistive processing unit for deep neural network training," in *IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, no. i, pp. 422–425.

[6] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," in *IEEE International Electron Devices Meeting (IEDM), 2017*, p. 6.1.1-6.1.4.

[7] M. Jerry, S. Dutta, A. Kazemi, K. Ni, J. Zhang, P.-Y. Chen, P. Sharma, S. Yu, X. S. Hu, M. Niemier, and S. Datta, "A ferroelectric field effect transistor based synaptic weight cell," *J. Phys. D. Appl. Phys.,* vol. 51, no. 43, p. 434001, Oct. 2018.