A FeFET Based Processing-In-Memory Architecture for Solving Distributed Least-Square Optimizations

Insik Yoon[1], Muya Chang[1], Kai Ni[2], Matthew Jerry[2], Samantak Gangopadhyay[1], Gus Smith[3], Tomer Hamam[1], Vijaykrishnan Narayanan[3], Justin Romberg[1], Shih-Lien Lu[4] , Suman Datta[2] and Arijit Raychowdhury[1]
[1]*Georgia Institute of Technology, Atlanta, GA 30318, USA,* [2]*University of Notre Dame, Notre Dame, IN 46556 USA*
[3]*Pennsylvania State University, State College, PA 16801 USA,* [4]*TSMC, Taiwan*
Email: iyoon@gatech.edu /Phone: (678)643-5264

Introduction: HfO_2 based ferroelectric FET (FeFET) has recently received great interest for its application in nonvolatile memory (NVM) [1]. Unlike conventional perovskite based ferroelectric materials, HfO_2 is CMOS compatible and retains ferroelectricity for thin film with thickness around 10 nm. Therefore, successful integration of ferroelectric HfO_2 into advanced CMOS technology makes this technology highly promising for NVM [1]. Moreover, by tuning the portion of switched ferroelectric domain, a FeFET can exhibit multiple intermediate states, which enables its application as an analog conductance in mixed-signal in-memory computing. Currently, such architectures have been applied to neuromorphic computing [2,3]. In this paper, we present a processing-in-memory (PIM) architecture with FeFETs and demonstrate how this can be used to solve a new class of optimization problems, in particular, distributed least square minimization.

Device Operation and Modeling: The operation of FeFET is illustrated in Fig. 1. A positive/negative gate pulse erase/program the device by setting the ferroelectric polarization to point toward the channel/gate, which would set the device in low/high V_{TH} state, respectively. The corresponding I_{DS}-V_{GS} characteristics for the two states are shown in Fig. 1(c). The separation of the two states indicate the memory window (MW), which is around 1.0 V for 10 nm HfO_2 FeFET [1]. By applying a read pulse, V_R, the two states can be differentiated by sensing the read current. The operation of FeFET as an analog conductance is different from the binary memory, in that a series of weak pulses are applied to set the device in desired state [2]. Out of the various pulse schemes proposed in [3] to tune the state, we use the pulse-amplitude modulation scheme (Fig. 2). Fig. 2(e) shows the applied pulse waveform. After each pulse, the percentage of switched ferroelectric domains is modified. The device states are shown in Fig. 2 (a)-(d). The device I_{DS}-V_{GS} corresponding to different states are shown in Fig. (f), which shows the intermediate states. The different states could be sensed by applying a read pulse, V_R, the corresponding drain-to-source conductance, G_{DS}, can be sensed. Fig. 2(g) shows the ideal G_{DS} as a function of applied pulse numbers. G_{DS} increases/decreases linearly and symmetrically with pulse numbers during potentiation/depression, respectively. The modeling framework for FeFET is shown in Fig. 3(a), which is composed of two sub-components: (1) baseline 45nm MOSFET, which is models $Q_{MOS}(V_{MOS})$ in BSIM 4, and (2) the ferroelectric which models $Q_{FE}(V_{FE})$ using the dynamic Preisach model. Fig. 4(a) and (b) show the measured and simulated I_D-V_G characteristics under potentiation and Fig. 5 shows FeFET channel conductance vs. the number of pulses from both simulations and experiments.

Architecture and Application: In Fig. 6, we present the cell schematic of a differential FeFET memory cell that computes MAC on positive and negative numbers. One number is stored as a differential value of conductance in two FeFETs and the other number is applied as a differential input voltage to BL1 and BL2. As WL is asserted, differential current flows to the SL and an ADC digitizes the result. The vector is mapped to input voltage in BL and numbers in matrix are mapped to conductance in FeFET array. Fig.7 describes the end-to-end and vertically integrated model from device to architecture. We use this architecture to demonstrate distributed least-square optimization (with wide applications to signal processing, machine learning etc.) which solves an inverse problem of the form $Az=B$ where A is the Grammian matrix of the basis and B is the observation vector [4]. We show that this problem can be mapped to iterative and distributed PIM where each unit is shown in Fig. 8. In Fig. 9, we compare compute latency and energy dissipation from a compute unit with FeFET memory vs. DAC resolution and number of bits per FeFET cell. The figure also shows the design-space exploration of ADC, DAC bit resolution and number of bits per FeFET cell in terms of latency and energy dissipation. At a system level, this results in *21X* (*3X*) improvement in energy-efficiency (performance) compared to an SRAM+ALU architecture. We use the proposed algorithm and architecture for solving a classic problem of signal reconstruction from non-uniform samples: (1) Signal reconstruction from 1D EEG Signals and (2) Recovery of CT Images used in medical imaging in Fig. 10. We note more than 35dB PSNR after signal reconstruction.

Conclusion: In this paper, we present a FeFET based processing in memory architecture and demonstrate its application to solving distributed least squares optimizations.

References:

[1] H. Mulaosmanovic et al., *VLSI Tech. Dig.,* 2017.
[2] S. Oh, et al., *IEEE Electron Device Letters*, pp. 732-735, 2017.
[3] M. Jerry, et al., *IEDM*, 2017.
[4] Ferreira et. al., *Signal Processing*, 1994.

978-1-5386-3028-0/18/$31.00 ©2018 IEEE

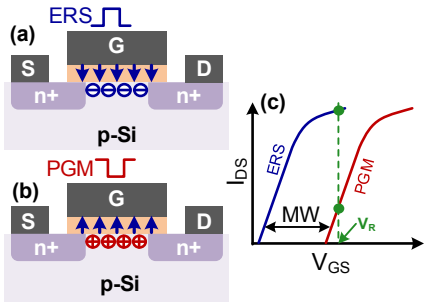

Fig. 1. (a) Erase and **(b)** program operation of binary FeFET memory. (c) Device I_{DS}-V_{GS} characteristics after the program/erase. The separation of the two curves is the memory window.

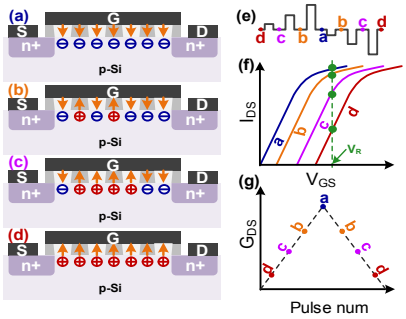

Fig. 2. (a) – (d) FeEFT states w.r.t ferroelectric domain switching. **(e)** pulse amplitude modulation scheme. **(f)** I_{DS}-V_{GS} characteristics after each pulse. **(g)** measured w.r.t applied pulse number

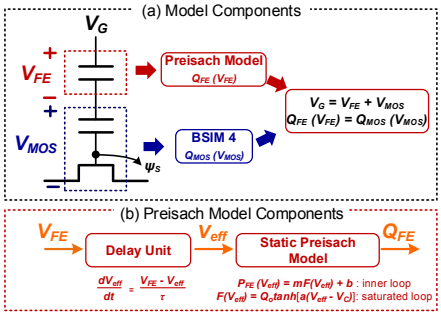

Fig. 3. (a) Components of compact model for FeFET and **(b)** the dynamic Preisach model components which is composed of a delay unit and the static Preisach model.

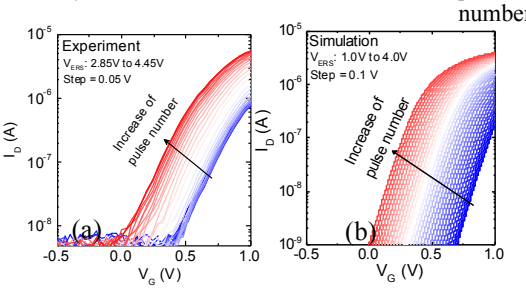

Fig. 4. (a) simulated **(b)** measured I_{DS}-V_{GS} characteristics of FerroFET w.r.t. increasing number of potentiation gate pulses

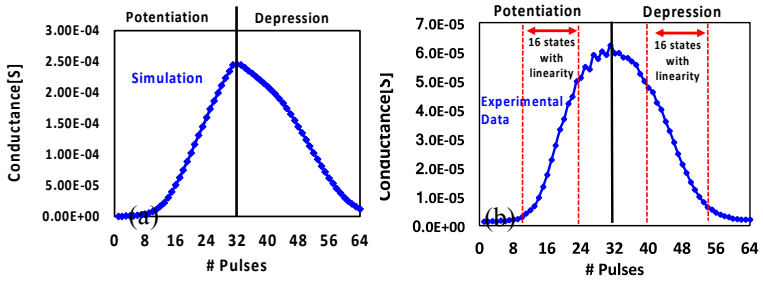

Fig. 5. FerroFET channel conductance (G_{DS}) as a function of pulse number from **(a)** simulation **(b)** experiment

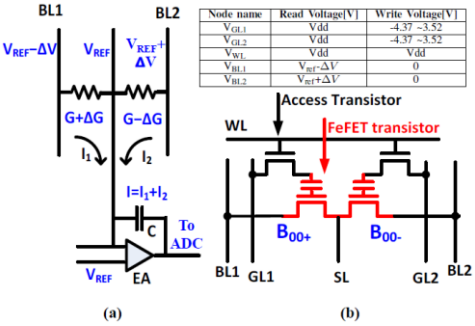

Fig. 6: FerroFET cell schematic **(a)** conceptual **(b)** transistor level implementation

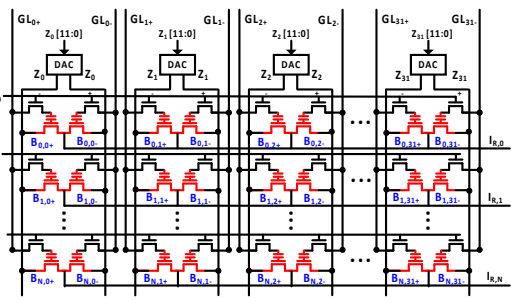

Fig. 6: FerroFET memory array schematic

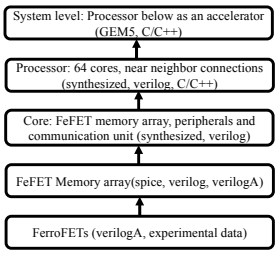

Fig. 7. Flow-chart of design hierarchy from device to system

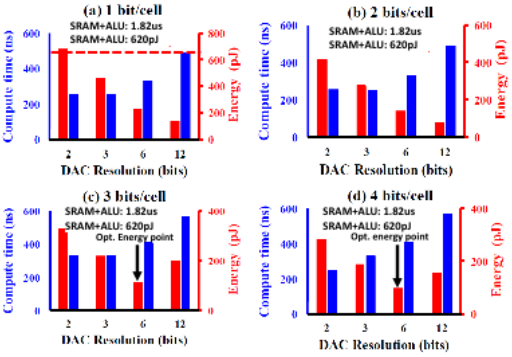

Fig. 9. Compute time and energy behavior of the compute unit versus DAC resolution and storage per FerroFET memory cell is **(a)** 1bit/cell **(b)** 2bits/cell **(c)** 3bits/cell and **(d)** 4bits/cell

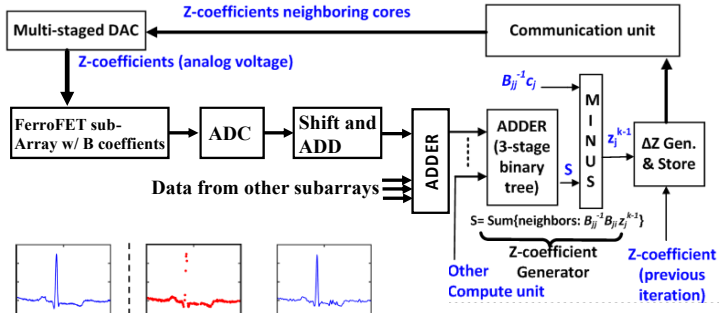

Fig. 9. Compute unit schematic for non-uniform sampling algorithm

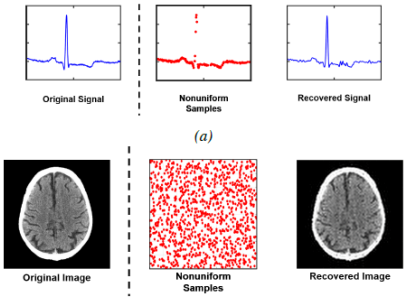

Fig. 10. Reconstruction steps. **(a)** 1D Example: Recovery of EEG Signal Profile**. (b)** 2D Example: Brain Computed Topography Recovery.

978-1-5386-3028-0/18/$31.00 ©2018 IEEE