Power-Efficient Clustering Using Programmable V_T FETs in Neuromorphic Architectures

S. Barve[1], N. Haehn[1], C. Socolik[1], Aaron Ruen[1], Joshua Mayersky[1], Amber Reed[2], Kevin Leedy[2], and R. Jha[1]
[1]University of Cincinnati, Cincinnati, OH, USA
[2] Air Force Research Laboratory, Wright Patterson, Ohio, USA

2023 International Electron Devices Meeting (IEDM) | 979-8-3503-2767-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEDM45741.2023.10413791

***Abstract*—**Existing Deep Neural Network (DNNs) implementations exceed power and time budgets of many applications. Additionally, DNN accelerators are supervised, highly specialized, and focus on multiply-and-accumulate operations. This work proposes a novel ultra-fast low power unsupervised neuromorphic architecture integrating programmable threshold voltage transistors for clustering using squared Euclidean distance.

I. Introduction

Many existing Deep Neural Network (DNN) accelerators focus on the multiply-and-accumulate (MAC) function [4]. In contrast, acceleration of Squared Euclidean Distance (SED) has seen comparatively little development [1-3, 7-8, 11]. Additionally, many Application Specific Integrated Circuits (ASICs) are supervised and highly specialized [3, 11]. This prevents deployment in dynamic or unexplored environments, as the labeled data may be nonexistent or unreliable. Existing ASICs are mostly co-developed with CMOS devices such as SRAMs or DRAMs, making scalability of the DNN dependent on the scalability of the device [1-3].

We present a novel neuromorphic architecture capable of performing analog in-memory SED computing via programmable threshold voltage (V_T) field-effect transistors (FET). The architecture can train and perform efficient data clustering at low power. A framework capable of emulating this architecture is presented. Finally, we demonstrate the feasibility of this architecture by integrating ferroelectric FETs (FeFETs) with experimental validation.

II. Neuromorphic SOFM Framework

A. In-memory Error Computing Crossbar

SED is among the most common measures of similarity between points as the shortest (straight line) distance between the points: $d(x, y) = (x - y)^2$[1, 8, 11]. It is at the center of supervised algorithms such as K-nearest neighbors and unsupervised algorithms such as K-means clustering [1, 7]. SED also produces a more symmetrical and unbiased measure of similarity in comparison to MAC (Fig. 1). While hardware accelerators for SED have been investigated, they tend to utilize digital CMOS technology or perform inaccurate computation of SED [6-7,11]. Meanwhile, programmable V_T FETs show promise for in-memory computation of SED since the long-channel saturation drain current I_{DS} is proportional to the SED between the gate bias V_{GS} and V_T of the device. For long-channel FETs, $I_{DS} = \frac{WC'_{ox}\mu}{2L}(V_{GS} - V_T)^2$, where W is the channel width, L is the channel length, C_{ox}' is the oxide capacitance per unit area, and μ is the charge carrier mobility. Exploiting this basic relationship, Fig. 2. shows the crossbar architecture of programmable V_T FET synapses for in-memory computation of SED input (V_{GS}) and weight (V_T) of the device. We developed two implementations of the FET synapses integrating either only n-channel FETs, or both n-channel (nFET) and p-channel FETs (pFET). The nFET only implementation (Fig. 3) can be used in cases where a pFET version of the device is not available or well-studied [1]. The nFET and pFET (Fig. 4) mitigate the overhead of reading V_T [2]. Therefore, users can write weights by programming the V_T of the FETs and applying inputs as V_{GS} for rapid low power analog in-memory computing of SED. We demonstrate and validate the framework's emulation of this circuit behavior in both Python and SPICE (Fig. 5). Clearly, the framework can accurately emulate the circuit behavior simulated in SPICE while performing near-ideal SED computation. More importantly, the framework allows users to integrate their own programmable V_T FET device models in the architecture. Capitalizing on SED acceleration, the framework is tailored for unsupervised clustering by implementing the self-organizing feature map algorithm (SOFM) discussed in the next section.

B. Neuromorphic SOFM

SOFM is an unsupervised neural network consisting of a neuron map that learns the topography of the input data via competitive learning. NeuroSOFM and DySON are low-power neuromorphic SOFM architectures integrating FeFETs in the in-memory SED computing crossbar [1-2]. The framework emulates both neuromorphic architectures. Fig. 6 demonstrates different FeFET models with varying number of V_T states. Fig. 7 demonstrates different FeFET models resulting from different programming schemes [1].

An advantage of these architectures is the neuron-sliceable or chip-sliceable implementation improving scalability as shown in Fig. 8. Therefore, the size and shape of neuron maps can be defined in the framework, as desired by application. Emulating the DySON architecture, the framework implements a growing SOFM which dynamically grows in response to the data. Growth allows users to emulate and test the architectures to identify the optimal size for their application (Fig. 6-7).

The best matching unit (BMU) is the neuron with highest similarity with the presented input. The BMU is often selected via competitive learning by exhaustively comparing neuron errors [1]. However, the framework enables users to select

between competitive BMU selection and thresholded BMU selection adopted from [2]. The thresholded BMU selection is done via a separate architecture, shown in Fig. 9. This threshold also acts as a filter for the SED produced in the crossbar accelerator.

C. Variability Testing

To test the impact of non-idealities, the framework contains parameters for adding variability to the V_T states of FETs via gaussian distribution. Overall variability to the crossbar can be applied via a gaussian distribution to the SED to produce an effective error (Fig. 10).

An additional user-defined parameter emulates Random stuck-at-faults (SAF) in neurons (Fig. 11). To account for non-idealities due to short-channel effects in scaled technologies, a user-defined parameter is added that reduces the I_{DS} (Fig. 12). The framework was tested using FeFET device models benchmarked against experimental data, discussed next.

III. Experimental Validation using FeFETs

Of several available candidates for programmable V_T FETs, (ex. floating gate, charge trap memory, etc.) FeFETs are promising options for the proposed clustering architecture [1-2]. In FeFETs, the V_T is modulated via polarization of the ferroelectric material integrated in the gate stack [1]. The proposed clustering architecture and framework was validated on experimental data obtained using Barium Titanate (BTO) Ferroelectric material and FeFETs based on BTO that can achieve multiple V_Ts and provide IDS proportional to the SED between V_{GS} and V_T.

A. BTO Electrical Testing for Multi-State Polarization

BTO is a promising ferroelectric material [9-10]. The programmable polarization characteristics of BTO were experimentally evaluated by fabricating ferroelectric capacitors (Fig. 13). BTO was deposited using pulse laser deposition (PLD) reported in [9-10]. We demonstrated multiple polarization states in BTO by multiple voltage sweeps (Fig. 14). Fig. 15 demonstrates the excellent retention, at room temperature (RT), of multiple polarization states in BTO that were measured using the technique reported in [5]. Endurance of BTO was measured by cycling it through PUND pulsing cycles (Fig.16). As evident from this data, BTO demonstrated high endurance, which is crucial for the proposed neuromorphic architecture for on-chip clustering.

B. Fabrication and Testing of BTO-based FeFET

FeFET devices were fabricated using BTO in back-gated configuration shown in Fig.17. On Nb:STO substrate BTO was deposited using PLD [10]. Thereafter, 70 nm amorphous IGZO channel material was deposited using PLD at RT. W was deposited using RF magnetron sputtering and patterned to form a source and drain. Fig. 18 (a) shows I_{DS} vs. V_{DS} and (b) shows I_{DS} vs. V_{GS} characteristics of these FeFETs. Fig. 19 shows that the I_{DS} is indeed proportional to the SED between the programmed V_T and the applied input V_{GS}. This experimentally validates the in-memory SED computation emulated in our developed framework integrating the FeFET device model.

IV. Architecture Scalability

The framework produces a SPICE netlist using models specified by the user. This SPICE netlist can be used for circuit level simulation for analyzing power and timing. We demonstrated the power consumption as the architecture is scaled (Fig. 20a). Comparing with existing SED computation methods (Fig. 20b) we observe that the in-memory SED accelerator is extremely power and time efficient.

V. Discussion and Summary

This work provides a framework to catalyze the development of ASIC to accelerate SED calculations and online training in unsupervised data-clustering neuromorphic architectures using programmable V_T FETs, even with non-idealities. The framework was evaluated using FeFET models benchmarked on experimental data. The proposed architecture shows potential to be power and time efficient. This work lays a solid foundation for extending the application of emerging memory devices beyond dot-product multiplication.

Acknowledgment

This work was funded by the National Science Foundation under award number ECCS-1926465, and Southwestern Council for Higher Education under agreement No. FA8650-19-2-9300. Joshua Mayersky, now affiliated with TEL Technology Center America, Albany, NY, USA, contributed to this work during graduate studies at the University of Cincinnati. Portions of sample fabrication were completed at the Cleanroom at the University of Cincinnati.

References

[1] S. Barve et al., "NeuroSOFM: A Neuromorphic Self-Organizing Feature Map Heterogeneously Integrating RRAM and FeFET," in IEEE JxCDC, doi: 10.1109/JXCDC.2021.3119489.

[2] S. Barve et al., "Dynamic Self-Organizing Neurons", IEEE TNNLS, Under Review

[3] T. J. Bailey, A. J. Ford, S. Barve, J. Wells and R. Jha, "Development of a Short-Term to Long-Term Supervised Spiking Neural Network Processor," in IEEE TVLSI, doi: 10.1109/TVLSI.2020.3013810.

[4] S. Fichtner et Al ; AlScN: A III-V semiconductor based ferroelectric. Journal of Applied Physics 21 March 2019; 125 (11): 114103. https://doi.org/10.1063/1.5084945

[5] V. Mikheev et Al., "Retention Model and Express Retention Test of Ferroelectric HfO2-Based Memory", Phys. Rev. Appl., vol. 18 Dec 2022

[6] C. Deotte. "Accelerating K-Nearest Neighbors 600x Using Rapids Cuml." NVIDIA Technical Blog, 21 Aug. 2022, developer.nvidia.com/blog/accelerating-k-nearest-neighbors-600x-using-rapids-cuml/.

[7] N. Singh et Al., Hardware Accelerator for Squared-euclidean Distance (2020). International Journal of Electrical Engineering and Technology, 11(3), 2020, pp. 186-193,

[8] M. M. Dahan, "Sub-Nanosecond Switching of Si:HfO2 Ferroelectric Field-Effect Transistor", Nano Letters 2023 23 (4), 1395-1400 ,DOI: 10.1021/acs.nanolett.2c04706

[9] J. Mayersky et. Al., Investigation and characterization of the annealing effects on the ferroelectric behavior of PLD BaTiO3. MRS Communications 11, 288–294 (2021). https://doi.org/10.1557/s43579-021-00030-2

[10] J. Mayersky et. Al.; True ferroelectric switching and trap characterization in BaTiO3/Nb:STO heterostructures. Appl. Phys. Lett. 5 September 2022; 121 (10): 102903. https://doi.org/10.1063/5.0097212

[11] R. Wang et al. Implementing in-situ self-organizing maps with memristor crossbar arrays for data mining and optimization. Nat Commun 13, 2289 (2022). https://doi.org/10.1038/s41467-022-29411-4

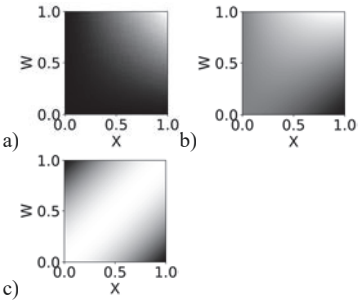

Fig. 1. Similarity measures darker areas mean higher (lower) error (similarity). a) MAC (dot-product similarity). b) SED computed via memristors [12]. c) SED computed via FET I_{DS}.

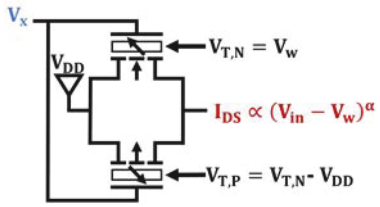

Fig. 4. Synapse implemented using n-channel and p-channel programmable V_T FETs. I_{DS} produced by synapse is proportional to SED between input and weight.

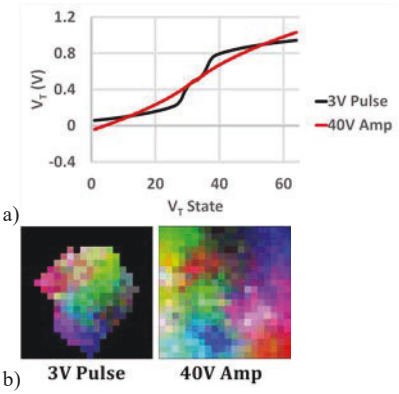

Fig. 7 a) V_T state distributions for 64 state FeFET device models programmed using 3V pulses and pulses of increasing amplitude from 0-40V. b) Emulation of 20 x 20 growing SOFM architecture using framework integrating the two device model programming schemes. Different distributions of V_T may require different architectural parameters.

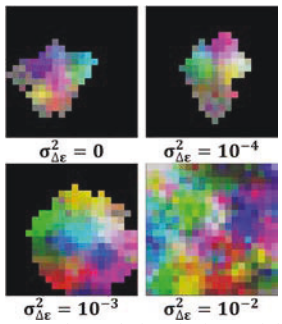

Fig. 10. Framework emulating device variability reflected in the variance induced in the in-memory computed error. Framework implemented 20x20 growing SOFM with increasing variability. Higher variability requires more neurons.

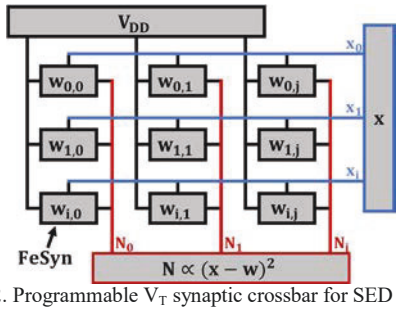

Fig. 2. Programmable V_T synaptic crossbar for SED computation acceleration. Each synapse computes pointwise SED, between input x and weight w, in parallel while each neuron N accumulates the error to produce a total error.

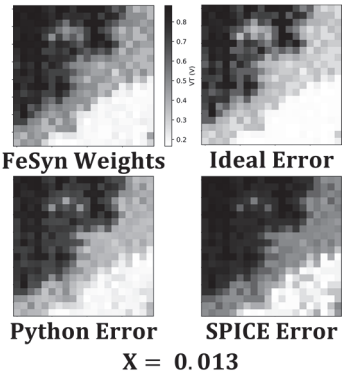

Fig. 5. Three identical 20x20 synaptic crossbars were simulated, with an input of 0.013, in both Python and SPICE using the BSIM4 model. Inputs were applied as a voltage input of 0.013V in Python and SPICE models of the architecture. Black indicates high error while white indicates low error.

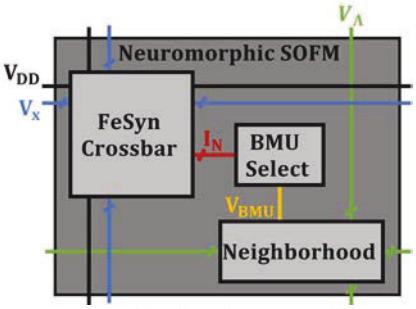

Fig. 8 Neuron or chip-slice of neuromorphic SOFM architecture. Each slice contains a crossbar for error computation, BMU selection, and neighborhood controller. Neighborhood controllers allow slices to interact with adjacent slices for weight update for clustering.

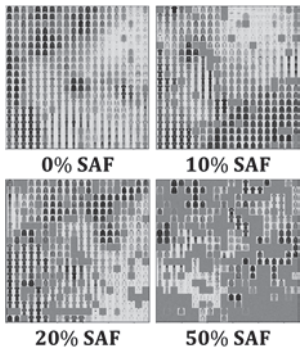

Fig. 11. Framework emulating 20 x 20 growing SOFM architecture trained on Fashion MNIST with 0-50% neurons with SAF. Neuron map clusters around the SAFs.

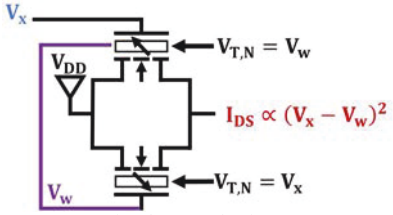

Fig. 3. Synapse implemented using two n-channel programmable V_T FETs. V_T of top FET must be read and applied as gate bias of bottom FET. The presented input must be programmed into the V_T of the Bottom FET. I_{DS} produced by synapse is proportional to SED between input and weight.

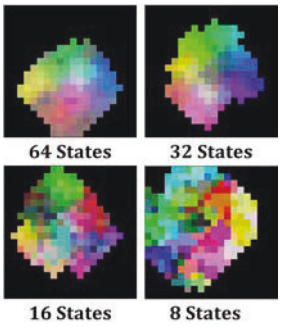

Fig. 6. FeFET models with 64, 32, 16, and 8 states were simulated using the framework with a 20 x 20 growing SOFM architecture trained on RGB colors. Clusters are still present with as little as 8 states albeit with lower quality. Number of neurons required to meet satisfiable error criteria is larger as number of states decreases; indicating users with low state devices may need larger maps.

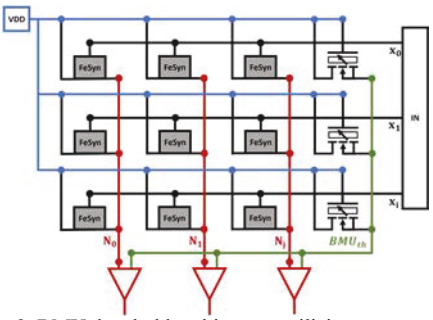

Fig. 9. BMU threshold architecture utilizing an additional reference neuron which produces a reference error using its programmed weights. This reference error thresholds neuron error indicating sufficiently low error (high similarity) between neuron weights and presented input.

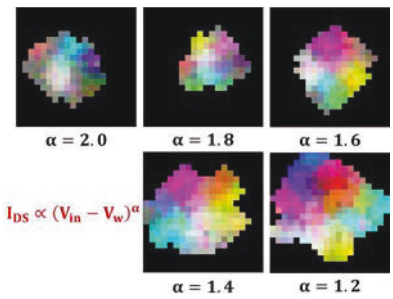

Fig. 12. Short-channel effects are emulated by having sub-squared non-linear I_{DS} modulated by user-defined parameter $1 < \alpha \leq 2$. Shorter channel lengths will result in a smaller α. Shorter channel lengths may require larger neuron maps.

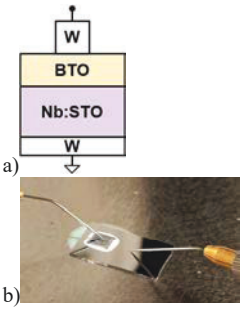

Fig. 13. a) Ferroelectric capacitor fabricated using BTO. b) Probing of ferroelectric capacitor wafer for measuring.

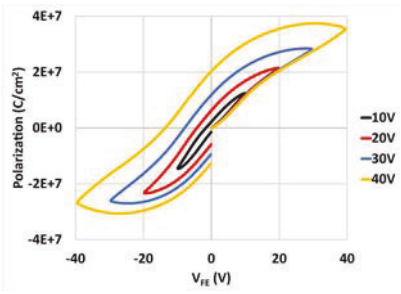

Fig. 14. Hysteresis of BTO capacitor with pulse amplitudes of 10-40V. Each hysteresis contains two remanent polarizations (up and down) resulting in 8 total polarization states.

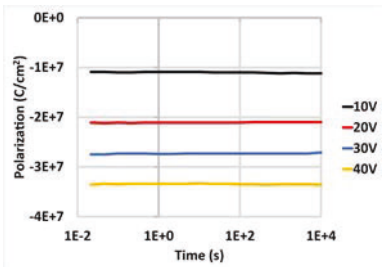

Fig. 15. Polarization state retention was tested using method described in [5]. Read polarization was calculated using method described in [6]. Polarization state is retained and shows little to no signs of decay.

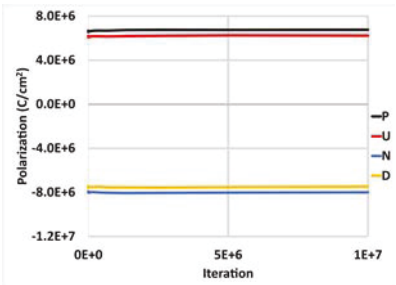

Fig. 16. Endurance of BTO tested using PUND (Positive Up Negative Down) sweeps followed by 9 segment fatigue pulse trains of alternating positive and negative pulses. Polarization at each point in PUND is consistent.

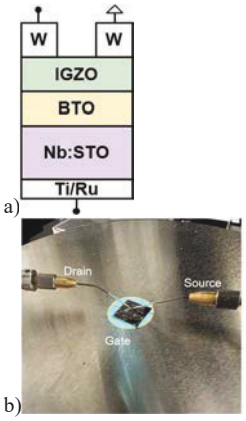

Fig. 17. a) FeFET fabricated by integrating BTO into gate stack of FET. b) Probing of FeFET wafer for measuring.

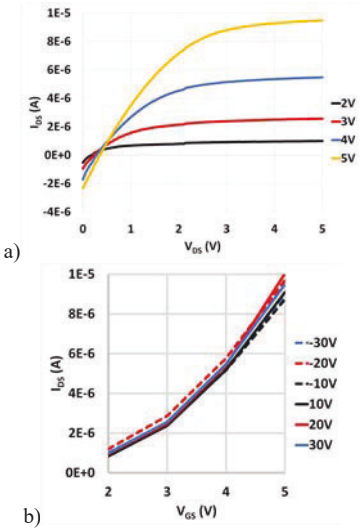

Fig. 18. a) V_{DS} vs I_{DS} while V_{GS} was varied from 2-5V. b) V_{GS} vs I_{DS} was measured while V_{DS} was maintained at 5V for FeFET programmed with pulses ranging from -30 to 30V. Each programming pulse amplitude results in a unique V_T as observed from the offset curves.

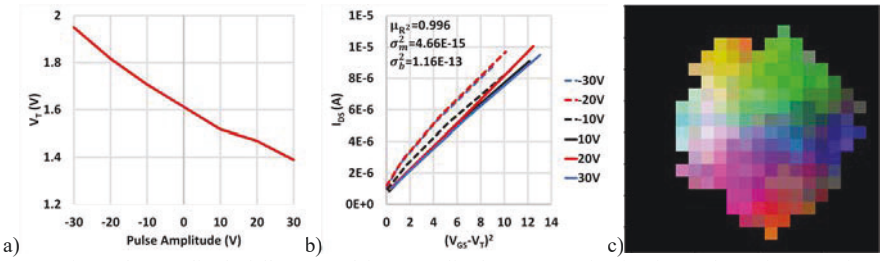

Fig. 19. a) V_T state for each programming pulse amplitude follows trend from amplitude programming in Fig. 6a. b) I_{DS} is proportional to SED computed by all FeFET states with some scaling and offset variance. c) 20x20 growing SOFM emulated by framework using error scaling and offset variance from measured data.

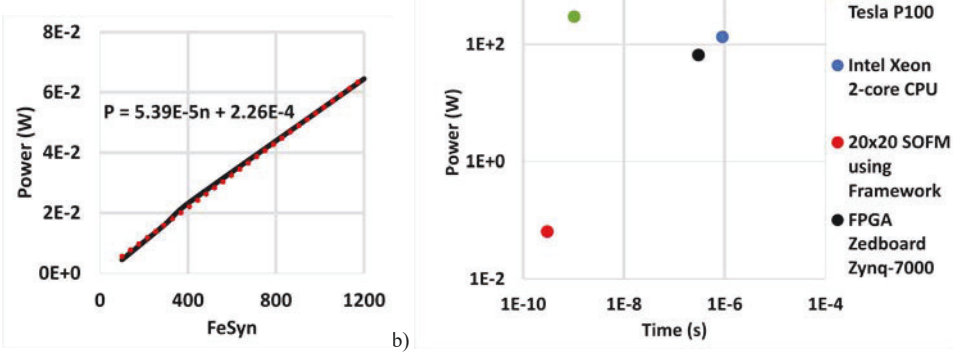

Fig. 20. Framework utilized 45nm BSIM4 FETs and state-of-the-art FeFET switching speed [9]. a) Power consumption per synapse using framework emulating varying crossbar sizes. b) In-memory SED computation accelerator compared to state-of-the-art CPU [7], GPU [7], and FPGA [8] implementations. Since I_{DS} computes the SED and each synapse computes the error in parallel, the computation speed is entirely limited by the switching speed of the FETs. Therefore, any time bottleneck of this architecture will be from peripheral circuitry.