2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS) | 978-1-6654-1913-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/AICAS51828.2021.9458437

Characterization of Drain Current Variations in FeFETs for PIM-based DNN Accelerators

Nathan Eli Miller*†, Zheng Wang*, Saurabh Dash*, Asif Islam Khan*, and Saibal Mukhopadhyay*
*School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
†Email: nathan.miller@gatech.edu

***Abstract*—We analyze the impact of drain current (I_{DS}) variation in 28 nm high-K metal-gate Ferroelectric FET devices on FeFET-based processing-in-memory (PIM) deep neural network (DNN) accelerators. Non-Normal variation in I_{DS} is observed due to repeated read operation on FeFET devices with different channel dimensions at various read frequencies. Device-circuit co-analysis using the measured current distribution shows a 1 to 3 percent accuracy degradation of an FeFET-based PIM platform when classifying the Fashion-MNIST dataset with the LeNET-5 DNN model. This accuracy drop can be fully recovered with variation-aware training methods, showing that individual FeFET device current variation over many read cycles is not prohibitive to the design of DNN accelerators.**

***Index Terms*—DNN Accelerator, FeFET, PIM**

I. Introduction

Ferroelectric FETs (FeFETs) based on silicon doped hafnium oxide (Si:HfO_2) have shown promise in a variety of applications for non von Neumann computing [1]–[9]. This includes applications such as logic-in-memory [5], reconfigurable computing [6], [7], coupled oscillators [8] and content-addressable memory [9]. One notable application for FeFETs is in the area of processing-in-memory (PIM) based machine learning (ML) accelerators [1]–[4]. The key function of a PIM accelerator is vector matrix multiplication (VMM), wherein the memory (FeFET crossbar array) stores the synapse matrix (weights), the input vector is applied from the rows, and the output is obtained from the columns. Some designs utilize FeFETs as analog synapses with the channel transconductance acting as an analog weight that can be tuned to achieve symmetric potentiation and depression characteristics [2], [3].

A different approach is to consider a multi-bit representation of the weights and use each FeFET in the crossbar to represent one bit of the weight [1]. Thus, a high or low current through the device represents a stored bit of '1' or '0', respectively [1]. The FeFET device structure includes a ferroelectric oxide layer of Si:HfO_2 in the gate stack that can be electrically polarized to two distinct states to produce high and low threshold voltages (Fig. 2). These V_{th} can be used to store logic high and logic low states in the device such that each FeFET stores '1' or '0' depending on the programmed V_{th}. With the V_{GS} as the input, the combination of different V_{th} (stored bit) and V_{GS} (input bit) states produces a logic AND function with the drain current I_{DS} as the output [1] (Fig. 1). This AND logic creates single-bit multiplication between the input bit V_{GS} and the weight bit V_{th}. The I_{DS} of each FeFET in each column are summed in the analog domain to enable multiply-and-accumulate operation for each bit (Fig. 1). The final output is obtained by digitizing the currents of each column, followed by a hierarchical shift-and-add logic to generate the final multi-bit result. Thus the design creates a VMM engine that can be used for deep neural network (DNN) acceleration [1].

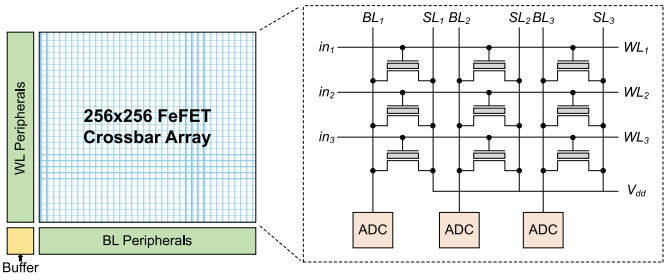
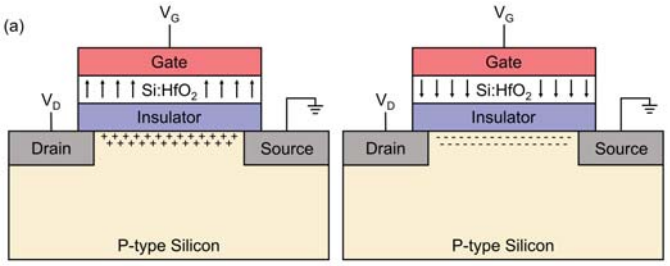


Fig. 1. A FeFET crossbar based PIM for DNN acceleration, similar to [1].

A key challenge in the design of FeFET based PIM systems is I_{DS} variation in devices within the crossbar leading to inaccuracies in PIM computation [1]. Although prior works have acknowledged this [1], [5], to the best of our knowledge, there are no measurement-based studies which characterize the impact of I_{DS} variation from individual FeFETs over time on the accuracy of PIM based ML Accelerators. Characterization of I_{DS} in FeFETs with various piezoelectric materials has been done for sources of variation such as retention time and V_{DS} and V_{GS} variation [10], [11], but by our knowledge characterization has not been performed on I_{DS} variation due to repeated read operation, nor has this characterization been applied to the accuracy of PIM-based DNN accelerators. This study characterizes variation in the I_{DS} of two 28 nm HKMG FeFET devices with differing channel dimensions due to repeated read operations at different frequencies. The measured I_{DS} distributions under different channel dimension and read frequency conditions are used to emulate performance variation in a PIM-based vector matrix multiplication (VMM) engine. Device-circuit co-analysis is used to estimate the potential drop in classification accuracy of a PIM-based DNN accelerator in implementing the LeNet-5 convolutional neural network to classify the Fashion-MNIST dataset [12].



(b)

Input (V_{GS})	Weight (V_{th})	Output (I_{DS})
0 (V_{GS} low)	0 (V_{th} high)	0 (I_{DS} low)
0 (V_{GS} low)	1 (V_{th} low)	0 (I_{DS} low)
1 (V_{GS} high)	0 (V_{th} high)	0 (I_{DS} low)
1 (V_{GS} high)	1 (V_{th} low)	1 (I_{DS} high)

Parameter	Logic Low	Logic High
V_{GS}	-1.0 V	0.0 V
$V_{th, D1}$	0.7 V	-0.1 V
$V_{th, D2}$	1.1 V	-0.5 V
I_{DS}	< 0.1 nA	> 1μ A

Fig. 2. FeFET device: (a) an FeFET device in its two polarization states, high V_{th} (left) and low V_{th} (right) and (b) summary of parameter values and bit-wise AND operation using FeFET.

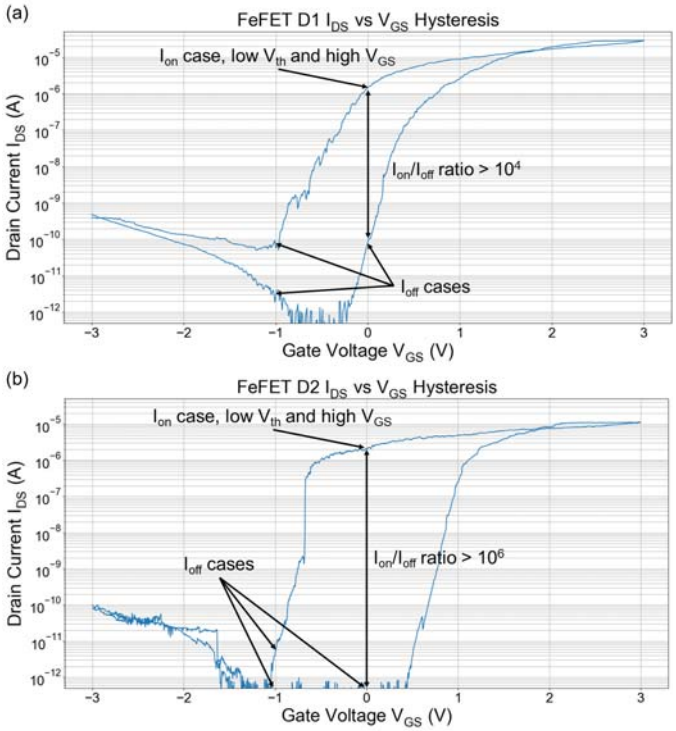


Fig. 3. Measured I_{DS}-V_{GS} hysteresis for 28 nm HKMG FeFET with transistor dimensions (a) 500 nm × 80 nm (Device 1, or D1) and (b) 80 nm × 34 nm (Device 2, D2). V_{DS} of 50 mV is applied and a voltage sweep is applied to V_G. Low V_{GS} of -1 V and high V_{GS} of 0 V are chosen to maximize I_{on}/I_{off} ratio. V_{th} is defined where I_{DS} = 1 μA.

II. Measured Device Characteristics

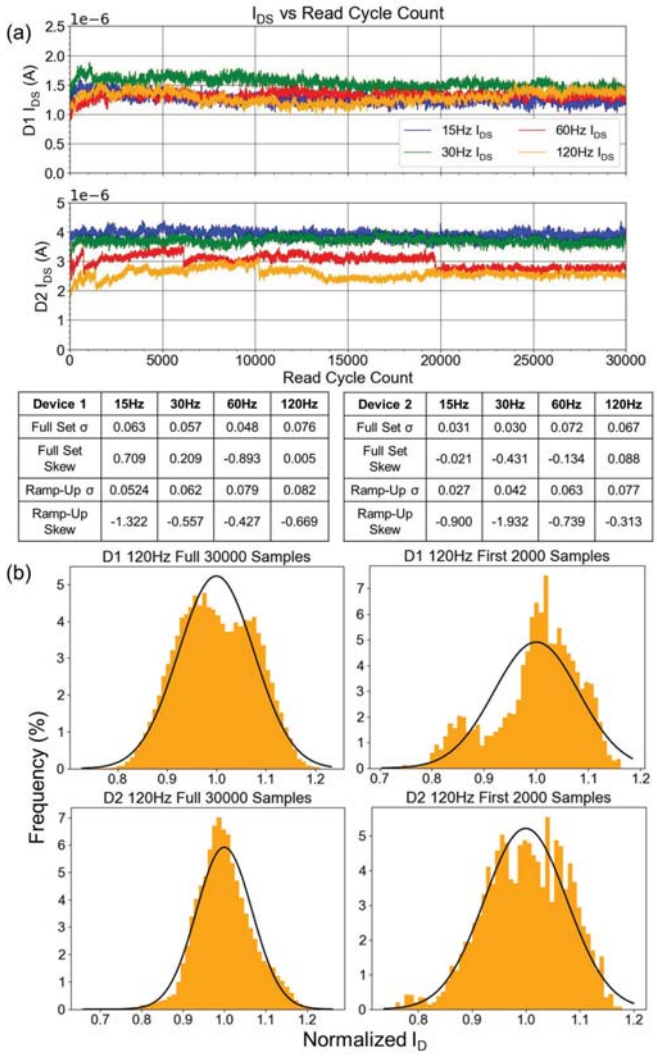
We characterize two 28 nm FeFET devices [13] with channel dimensions 500 nm × 80 nm (D1) and 80 nm × 34 nm (D2). Fig. 3 shows measured I_{DS}-V_{GS} hysteresis curves of D1 and D2. These hysteresis curves are measured by applying a voltage sweep to the gate (to implement a program cycle and an erase cycle), ground to the source and body and 50 mV to the drain to measure I_{DS}. Based on the measured hysteresis, V_{GS} of -1 V and 0 V are chosen for low and high logic inputs, respectively (Fig. 2b). The only combination of V_{th} and V_{GS} which will produce an output current of greater than 1 μA is the case where V_{th} and V_{GS} are logic 1. All other cases produce a current less than 0.1 nA for D1 and less than 0.01 nA for D2, resulting in an I_{on}/I_{off} ratio of greater than 10^4 in D1 and greater than 10^6 in D2. This large I_{on}/I_{off} ratio creates robustness in the full system to noise in I_{off}.

The logic 1 output current I_{on} occurs where V_{GS} and V_{th} are logic 1. I_{on} of the FeFET is measured by applying 50 mV to the drain, ground to the source and body, and a voltage pulse to the gate. The gate pulse waveform is generated with a base of -1 V and a pulse of 0 V at a pulse width of 10 μs. The I_{DS} measured at each 0 V pulse with V_{th} programmed to be logic 1 represents I_{on}. Measurement is taken over many read cycles after the device is programmed to a logic 1 V_{th} and is not reprogrammed between read cycles.

Measurements are taken for gate pulse frequencies of 15 Hz, 30 Hz, 60 Hz, and 120 Hz. As each device weight in the FeFET crossbar array is read once while processing an image, the different read frequencies represent different frame rates of the VMM engine. For example, in a design where the entire DNN can be loaded into the FeFET crossbar arrays, each layer is accessed individually, feeding into the next layer, and is not accessed again until the next image is passed. Thus the FeFETs are on but not read until the layer they are in is accessed during the next image pass, and a 15 Hz read frequency represents processing images at 15 frames per second (FPS).

Fig. 4 shows I_{DS} measurements for both devices for various read frequencies over 30000 read cycles, as well as standard deviation and skewness of the measurement datasets when normalized to mean 1 and fit to normal distributions. In each case, there is a noticeable "ramp up" period in the first 2000 cycles. We hypothesize that this trend could be caused by a parasitic capacitance present in the device or the measurement setup. To account for the ramp up period, we analyze PIM performance over the ramp up period via bootstrap sampling of the first 2000 samples, as well as over many read cycles via bootstrap sampling of the full dataset. We also observe abrupt changes in measured output current in some instances, such as in D2 measured at 60Hz at approximately 20000 cycles, which we attribute to PVT variations and V_{th} retention loss after repeated measurement. The original device characterization [13] shows no ferroelectric degradation until 10^{10} cycles, which we do not approach in this study, so this is ruled out as a potential cause of the abrupt current changes.

The original device characterization measures I_{DS} variation due to bipolar stress, though this test shows only the average I_{DS} rather than individual measurement results, and does not show ferroelectric breakdown until approximately 10^{10} cycles. The original characterization does not analyze variation due



Device 1	15Hz	30Hz	60Hz	120Hz
Full Set σ	0.063	0.057	0.048	0.076
Full Set Skew	0.709	0.209	-0.893	0.005
Ramp-Up σ	0.0524	0.062	0.079	0.082
Ramp-Up Skew	-1.322	-0.557	-0.427	-0.669

Device 2	15Hz	30Hz	60Hz	120Hz
Full Set σ	0.031	0.030	0.072	0.067
Full Set Skew	-0.021	-0.431	-0.134	0.088
Ramp-Up σ	0.027	0.042	0.063	0.077
Ramp-Up Skew	-0.900	-1.932	-0.739	-0.313

Fig. 4. (a) Measured I_{DS} for D1 and D2. The device is programmed once before the first measurement and is not reprogrammed. Distributions are normalized based on mean I_{DS} and fit to Gaussians of mean 1. Standard deviation and skewness of the full set and the first 2000 points are measured. (b) 120 Hz distributions are shown for D1 and D2 with bootstrap sampling of the full 30000 measurement set and the 2000 sample ramp up period.

to repeated read operations on single devices, as we perform here. In the same way that measuring many different devices tends to draw the distributions in the original device characterization close to Gaussian, we expect that characterizing more devices would help bring our measured distributions closer to Gaussian. Therefore, as we apply measurements from individual devices to many instances in a PIM architecture, our study is representative of a potential worst case as far as PIM classification accuracy is concerned, wherein all of the devices show non-Gaussian variation in individual I_{DS} measurements.

The logic 0 output current (I_{off}) is also measured for each device by applying -1V to the gate. In all cases, the measured I_{off} remains below 0.1 nA, and is often in the noise floor of the measurement system in the pA range. Thus, with the I_{on}/I_{off} ratio remaining larger than 10^4 in all cases, noise

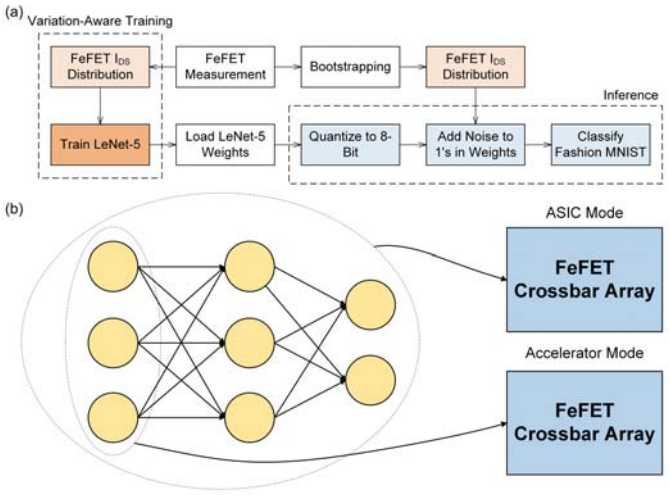


Fig. 5. Device-circuit co-analysis (a) overall analysis flow and (b) different operation modes for the PIM architecture. In ASIC Mode, the FeFET array is large enough to store the full DNN. In Accelerator Mode, one layer of the DNN is loaded into the array at a time and then overwritten by the next layer.

in I_{off} for the logic 0 outputs is considered negligible with respect to the full system. For this reason, we focus on the effects of noise in I_{on} on the accuracy of the PIM system.

III. Device Circuit Co-Analysis

We perform device-circuit co-simulation using PyTorch [14] to analyze the effect of FeFET I_{DS} variation on the PIM architecture's accuracy in classifying the Fashion-MNIST dataset using the LeNet-5 convolutional neural network model (Fig. 5a). The PIM architecture studied consists of many coupled 256 × 256 FeFET crossbar arrays [1]. We simulate the output of the PIM system considering bootstrap sampling of the measured I_{DS} distributions in two modes of PIM operation which we call ASIC Mode and Accelerator Mode (Fig. 5b).

A. ASIC Mode

First, we consider ASIC mode, where we assume the system of coupled FeFET crossbar arrays is large enough to store the full DNN weight matrix. The weights are written to the FeFET arrays once and streaming inputs (images) are used for inference. In this case, the FeFET V_{th} are programmed once and the I_{DS} is measured many times during inference without the V_{th} being rewritten in between measurements. Therefore, we perform bootstrapped sampling of the full 30000 measurement set to represent long-term variation in I_{DS} after the weights are programmed. The read frequency of the individual FeFET devices represents the frame rate of the system in processing entire images, such that an FeFET read frequency of 15 Hz represents a frame rate of 15 FPS.

B. Accelerator Mode

Second, we consider accelerator mode, where we assume the on-chip storage is not sufficient to store the weight matrix for the entire DNN and is time-multiplexed to compute a large network (similar to [1]). In this case, as the cells are frequently

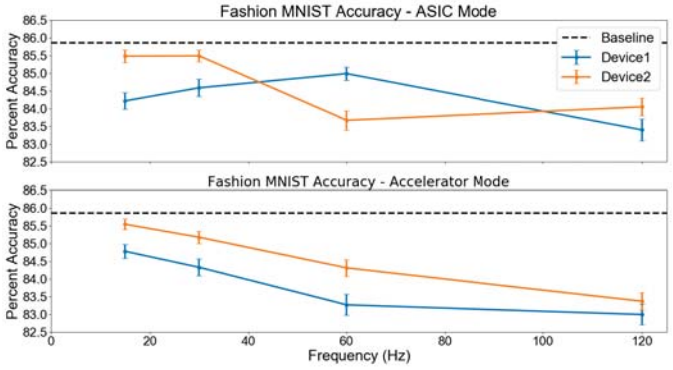


Fig. 6. Fashion-MNIST accuracy drops by up to 3 % from the baseline 8-bit quantization result of 85.86 % accuracy without variation-aware training.

rewritten during computation, we perform bootstrap sampling from the first 2000 samples of each dataset to study the effect of I_{DS} variation shortly after reprogramming (wherein I_{DS} is measured during the ramp up period described previously). In accelerator mode, the different frequency of read cycles represents different ratios of the size of the weight matrix to the total capacity of the crossbar. For example, processing a batch of 20 images through a 5 layer neural network in 1 second (for a 20 FPS output) in which only one layer is loaded into the FeFET crossbar at a time corresponds to a read frequency on each FeFET of 100 Hz (considering write time is in the ns range and is therefore negligible by comparison [13]). In this example, the first layer of the neural network would be programmed to the FeFET weights in the VMM engine, each of the 20 images in the batch would be passed through back to back, the weights would be rewritten for the next layer of the DNN, and so on.

C. Co-Analysis

To perform device-circuit co-analysis in PyTorch, each model parameter is quantized to 8 bits and noise is introduced to each logic 1 bit to represent variation in I_{on} during the VMM operation of the LeNet-5 network. The accumulation of noise leads to degradation of the PIM architecture's ability to classify the Fashion-MNIST dataset using LeNet-5. Fig. 6 demonstrates that the magnitude of degradation depends on the variance and skewness of the I_{DS} distributions, which depend on the read frequency and device dimensions. We observe accuracy reduction of up to 3 % from the baseline noiseless classification accuracy of 85.86 %, wherein larger read frequencies tend to result in worse classification accuracy.

In order to recover this accuracy loss, we use the measured noise distributions to add noise during training of the DNN (performing variation-aware training [15]). With this method, we observe full recovery of the accuracy drop back to the baseline of 85.86 %, negating the performance reduction caused by the variation in the FeFET I_{DS} measurements.

IV. Conclusion

In conclusion, measurements of individual 28 nm FeFET device output currents show non-Gaussian variation over many read cycles. Device-circuit co-analysis demonstrates that this variation in I_{DS} leads to only marginal (up to 3 %) drop in classification accuracy of an FeFET-based PIM architecture, even when performed on difficult datasets such as Fashion-MNIST. Additionally, we observe that the loss caused by I_{DS} variation can be fully recovered by variation-aware DNN training by using the measured variation to add noise during training of the DNN. Hence, we conclude that FeFET device current variation due to many read cycles is not preventative for our 28 nm FeFET-based DNN accelerator design.

Acknowledgment

This material is based on work supported by National Science Foundation (1810005). A.I.K. thanks GLOBALFOUNDRIES for providing FeFET technology wafers.

References

[1] Y. Long, et al., "A ferroelectric FET-based processing-in-memory architecture for DNN acceleration," in *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits,* vol. 5, no. 2, pp. 113–122, Dec. 2019, 10.1109/JXCDC.2019.2923745.

[2] M. Jerry, et al., "Ferroelectric FET analog synapse for acceleration of deep neural network training," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, Dec. 2017, pp. 6.2.1–6.2.4, 10.1109/IEDM.2017.8268338.

[3] H. Mulaosmanovic, et al., "Novel ferroelectric FET based synapse for neuromorphic systems," 2017 Symposium on VLSI Technology, Kyoto, June 2017, pp. T176–T177, 10.23919/VLSIT.2017.7998165.

[4] I. Yoon, et al., "Design space exploration of ferroelectric FET based processing-in-memory DNN accelerator" Aug. 2019, pp. 1–4, eprint, arXiv:1908.07942.

[5] E. T. Breyer, et al., "Compact FeFET circuit building blocks for fast and efficient nonvolatile logic-in-memory," in *IEEE Journal of the Electron Devices Society,* Apr. 2020, pp. 1–9, 10.1109/JEDS.2020.2987084.

[6] S. K. Thirumala, et al., "Reconfigurable ferroelectric transistor–Part I: Device design and operation," in *IEEE Transactions on Electron Devices,* vol. 66, no. 6, pp. 2771-2779, June 2019, 10.1109/TED.2019.2897960.

[7] N. Tasneem, et al., "On the possibility of dynamically tuning and collapsing the ferroelectric hysteresis/memory window in an asymmetric DG MOS device: a path to a reconfigurable logic-memory device," 2018 76th Device Research Conference (DRC), Santa Barbara, CA, June 2018, pp. 1–2, 10.1109/DRC.2018.8442250.

[8] Z. Wang, et al., "Ferroelectric Oscillators and Their Coupled Networks," in *IEEE Electron Device Letters,* vol. 38, no. 11, pp. 1614–1617, Nov. 2017, 10.1109/LED.2017.2754138.

[9] K. Ni, et al., "Ferroelectric ternary content-addressable memory for one-shot learning," in *Nature Electronics* 2, pp. 521–529, Nov. 2019, 10.1038/s41928-019-0321-3.

[10] C. Besleaga, et al., "Ferroelectric Field-Effect Transistors Based on PZT and IGZO," *IEEE Journal of the Electron Devices Society,* vol. 7, Jan. 2019, pp. 268–275, 10.1109/JEDS.2019.2895367.

[11] I. Katsouras, et al., "Controlling the on/off current ratio of ferroelectric field-effect transistors," *Sci Rep* 5, Article Number 12094, Jul. 2015, pp. 1–8, 10.1038/srep12094.

[12] H. Xiao, et al., "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," Aug. 2017, pp.1–6, eprint, arXiv:cs.LG/1708.07747.

[13] M. Trentzsch, et al., "A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 11.5.1–11.5.4, 10.1109/IEDM.2016.7838397.

[14] A. Paszke, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019, arXiv:1912.01703.

[15] Y. Long, et al., "Design of reliable DNN accelerator with un-reliable ReRAM," 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence, Italy, Mar. 2019, pp. 1769–1774, 10.23919/DATE.2019.8715178.