Journal of MATERIALS RESEARCH

Invited Paper
DOI:10.1557/s43578-021-00393-1

Impact of HKMG and FDSOI FeFET drain current variation in processing-in-memory architectures

Nathan Eli Miller[1,a], Zheng Wang[1], Saurabh Dash[1], Asif Islam Khan[1], Saibal Mukhopadhyay[1]

[1] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332–0250, USA
[a] Address all correspondence to this author. e-mail: nathan.miller@gatech.edu

Received: 7 July 2021; accepted: 13 September 2021; published online: 28 September 2021

In this study, we analyze the impact of drain current (I_{DS}) variation in 28 nm high-K metal-gate and 22 nm fully-depleted silicon-on-insulator Ferroelectric FET devices on processing-in-memory (PIM) deep neural network (DNN) accelerators. When performing repeated read operations on several devices at various read frequencies and under various biasing and programming conditions, non-Normal variation in I_{DS} is observed. Device-circuit co-analysis is used to emulate PIM performance subject to noise when classifying images. Marginal degradation is observed in Fashion-MNIST classification accuracy using LeNet-5, and more significant degradation is observed in CIFAR-10 classification accuracy using MobileNetV2. Variation-aware training is shown to fully recover minor drops in LeNet-5 accuracy but becomes difficult for large workloads like MobileNetV2. We demonstrate that I_{DS} variation in individual FeFETs over many read cycles is not prohibitive to designing DNN accelerators with small workloads, but advanced design techniques are required to mitigate error for larger workloads.

Introduction

In various non von Neumann computing applications, ferroelectric FETs (FeFETs) have been demonstrated as useful building blocks for functionality which surpasses the capabilities of MOSFETs alone [1–11]. These devices include a ferroelectric layer in the gate stack composed of silicon doped hafnium oxide (Si:HfO_2), which can be electrically polarized to a high or low threshold voltage (V_{th}) state to effectively store a memory bit within the transistor itself, thus making FeFETs an excellent candidate to fulfill both logic and memory functionalities in a wide variety of applications [1, 2]. Logic-in-memory [3], content-addressable memory [4], coupled oscillators [5], and reconfigurable computing [6, 7] are a few of many applications of this technology. For machine learning (ML) accelerators based on processing-in-memory (PIM) architectures, FeFETs have shown exceptional promise [8–11]. PIM accelerators are primarily used to perform vector matrix multiplication (VMM), in which the FeFET crossbar array (used as a memory) stores the synapse matrix (DNN weights), the input vector is applied via the rows to the gate voltages of the FeFETs, and the output is obtained from the columns each performing analog summation of I_{DS} currents from each FeFET in that column. In some designs, the FeFETs are used as analog synapses where the channel transconductance behaves as an analog weight. As prior works have shown, this weight can be tuned to achieve symmetric potentiation and depression characteristics [8, 9].

Alternatively, some designs consider a digital approach where the weights are quantized to multiple bits and each FeFET in the crossbar is used to represent a single weight bit [10]. The ferroelectric oxide layer of Si:HfO_2 in the gate stack of FeFET devices can be electrically polarized via a gate voltage pulse to distinct high and low threshold voltage (V_{th}) states, as shown in Fig. 1. Thus, these FeFET devices store logic states of '0' or '1' depending on how V_{th} is programmed. Combining this stored bit determined by V_{th} with the input bit provided by a high or low V_{GS}, an AND function is produced with I_{DS} as the output [10] using a system architecture such as Fig. 2. The input bit V_{GS} and the stored weight bit V_{th} thus create single-bit multiplication to produce the output bit I_{DS}. An analog summation of the I_{DS} of each FeFET in each column enables multiply-and-accumulate operation for each bit as shown in Fig. 2. This analog current is then digitized using an analog-to-digital converter, and each of the columns are combined to produce the final multi-bit VMM result using a hierarchical shift-and-add logic. This full VMM

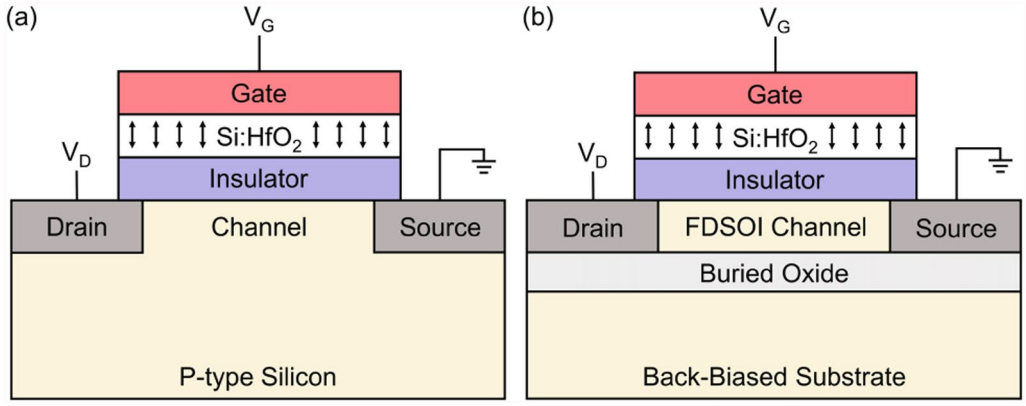

Figure 1: (a) HKMG and (b) FDSOI FeFET device structures including an electrically polarizable ferroelectric layer in the gate stack.

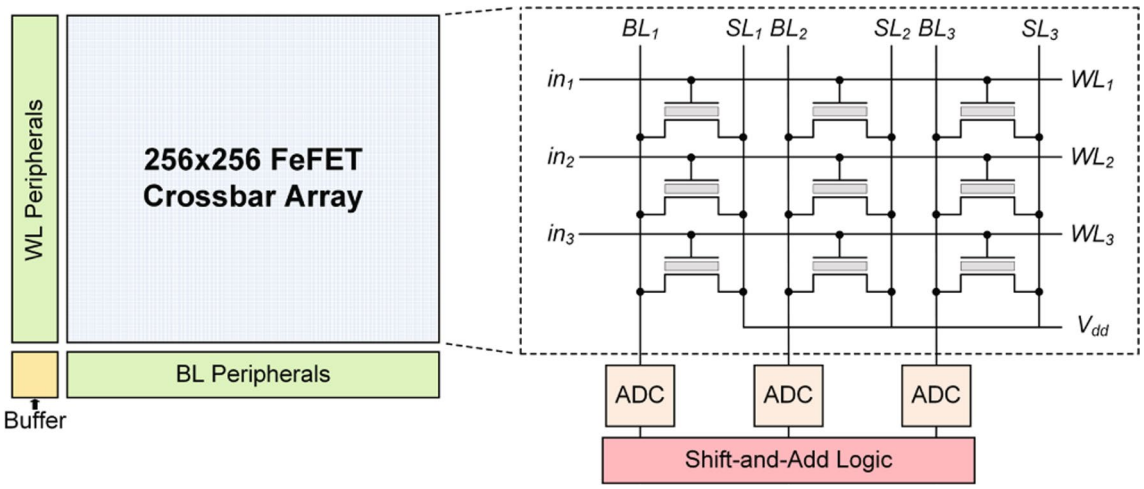

Figure 2: An FeFET crossbar-based PIM architecture for DNN acceleration, similar to [10].

engine design can thus accelerate deep neural networks (DNNs) in hardware [10].

I_{DS} variation within each individual FeFET device within the crossbar is a key challenge which can lead to inaccuracies in PIM computation [10]. We continue our prior work [12] in this study in order to characterize the impact of I_{DS} variation in FeFETs on the system accuracy of PIM-based DNN accelerators using hardware measurements of various individual FeFET devices. We differentiate from our original work by comparing measured I_{DS} distributions from both 28 nm HKMG and 22 nm FDSOI FeFET devices, by analyzing the effects of lowered V_{GS} read voltage and partial V_{th} programming and by introducing more complex DNN acceleration tasks. Unlike other works which have acknowledged the issue of I_{DS} variation in FeFETs and characterize variation from sources including V_{DS} and V_{GS} variation and retention loss over time [3, 13, 14], we utilize device-circuit co-analysis to determine the impact of this variation in practical FeFET system applications such as PIM-based DNN accelerators. We measure three 28 nm HKMG FeFETs and two 22 nm FDSOI FeFETs, each with different channel dimensions, and use the measured I_{DS} distributions to emulate the performance of a PIM-based DNN accelerator architecture considering FeFET I_{DS} variation under a variety of conditions (including various biasing and programming conditions and various read frequencies). We task our FeFET PIM architecture with classifying images in the presence of these sources of variation, including classifying the Fashion-MNIST dataset [15] using the LeNet-5 convolutional neural network architecture [16] and the CIFAR-10 dataset [17] using the MobileNetV2 architecture [18].

Results

First, we characterize three 28 nm FeFET devices [19] with channel dimensions 500 nm × 80 nm (Devices 1a and 1b, or D1a and D1b) and 80 nm × 34 nm (D2). Note that D1a and D2

are shown in our prior work as 'D1' and 'D2' [12]. The measured I_{DS}–V_{GS} hysteresis curves of D1a, D1b and D2 are shown in Fig. 3. The hysteresis is measured by applying a voltage sweep to the gate, 50 mV to the drain and ground to the source and body. The voltage sweep from −3 to 3 V on the gate implements a full erase and program cycle on the device. Based on the measured hysteresis, a logic 0 V_{GS} of − 1 V and a logic 1 V_{GS} of 0 V are chosen to represent the input bits. The only combination of V_{th} and V_{GS} which produces an output current greater than 1 μA (logic 1) occurs where V_{th} is programmed low and V_{GS} is 0 V (thus, V_{th} and V_{GS} are both logic 1). All other cases produce a current less than 0.1 nA for D1a, approximately 10 nA for D1b, and less than 0.01 nA for D2, yielding an I_{on}/I_{off} ratio of greater than 10^4 in D1a, approximately 10^2 in D1b, and greater than 10^5 in D2. A high I_{on}/I_{off} ratio leads to robustness in the full system to noise in $I_{off}I_{off}$, so maximizing this parameter is desirable.

With this configuration, the logic 1 I_{DS}, also called I_{on}, occurs only where V_{GS} and V_{th} are logic 1. We perform repeated reads of I_{on} of the FeFET by first programming the V_{th} to its logic high state, then applying ground to the source and body terminals, 50 mV to the drain, and a voltage pulse to the gate. The gate pulse has a base voltage of − 1 V (logic 0 V_{GS}), a peak voltage of 0 V (logic 1 V_{GS}), and a pulse width of 10 μs. V_{th} of the device is not reprogrammed between read cycles, and thus over repeated read cycles we measure both read endurance (long term change in average I_{DS}) and cycle-to-cycle variation.

Measurements are performed with read frequencies of 15, 30, 60, and 120 Hz. In a VMM engine as in Fig. 2, each weight in the FeFET crossbar is read once while processing an image. Therefore, the frequency at which each FeFET is read represents the frame rate of the VMM engine. In application spaces where the full DNN weight matrix can be loaded into the FeFET crossbars at once, the layers are accessed once per image, the outputs of that layer are fed to the next layer, and so on. Each layer is not accessed again until the next image is passed. Since the devices in that layer are accessed once per image, a 15 Hz read frequency of the individual FeFETs represents a 15 frames per second (FPS) image processing rate. Note that one could also consider a pipelined design, wherein one image is processed through a given layer, then passed to the next layer, at which point the next image is passed into the first layer, and so on. This allows for parallel processing of images in different layers, thus leading to a higher frame rate.

Figure 4 shows 30,000 I_{DS} read measurements for the three 28 nm HKMG devices at various read frequencies. The measurement distributions are fit to normal distributions with mean 1 as shown in Fig. 4b. In all three devices and for all tested frequencies, there is a significant trend in the first 2000 cycles where the average I_{DS} increases which we call the ramp up period. We hypothesize that a parasitic capacitance present in the FeFET or the measurement system could be the cause of this ramp up period. Emulation of the PIM architecture's performance subject to variation in both the ramp up period and in the long term case is performed by bootstrap sampling the 2000 cycle ramp up period and the full I_{DS} dataset, respectively. In certain instances such as in the 30 Hz measurement of D1b and the 60 Hz measurement of D2, we also note the presence of abrupt drops in measured I_{DS}. We attribute these sudden drops to ferroelectric breakdown causing V_{th} retention loss due to repeated measurement. The device characterization from GLOBALFOUNDRIES shows that ferroelectric breakdown of the low V_{th} case (which results in I_{on} as we measure here) can occur around 10^5 bipolar stress cycles [19]. While we do not perform full bipolar stress from −3 to 3 V, the voltage swing of the read pulse from −1 to 0 V still appears to cause some breakdown, particularly in D2. The measurements are taken successively, i.e., 30,000 read cycles are measured at 15 Hz, then the device is reprogrammed and 30,000 measurements are taken at 30 Hz, and so on. We note that in D2, the average value of I_{DS} drops with each successive test, and the sudden drops occur near where 10^5 total read cycles have occurred between all tests, which is where we expect ferroelectric breakdown from bipolar stress. It is also possible that these abrupt I_{DS} drops occur purely from PVT variations in some cases, such as in D1b where the drops occur in the 30 Hz case and the following tests restore their average I_{DS} after reprogramming the device.

While the original characterization shows the V_{th} retention loss from bipolar stress leading to a drop in average I_{DS} over time, it does not analyze variation due to repeated read operations on individual devices. All of the measured I_{DS} distributions show some degree of non-Normality, shown by the 120 Hz distributions in Fig. 4b, for example. We expect that by measuring many devices of the same dimension, these distributions would be drawn closer to Gaussian. Since we use distributions from single devices in our emulation of the PIM architecture, our study represents a possible worst case for PIM classification accuracy wherein each device demonstrates similar non-Normal variation.

The low output current I_{off} can be measured by applying $V_{GS} = -1$ V. When V_{GS} is low in any of the three devices, the measured I_{off} is below 0.1 nA or even as low as the pA range, which is the noise floor of the measurement system. The more problematic case of I_{off} occurs in D1b where the V_{th} is high (logic 0) and V_{GS} is high, since I_{off} in this case is in the 10 nA range due to D1b's low memory window for an I_{on}/I_{off} ratio near 10^2. However, the I_{on}/I_{off} ratio remains larger than 10^4 in all other cases. In general use cases, we can assume noise in I_{off} is negligible, but D1b shows a case where noise in I_{off} may become noticeable due to its smaller I_{on}/I_{off} ratio compared to other devices. In this study, we primarily analyze how noise in I_{on} affects PIM accuracy, as this is representative of PIM

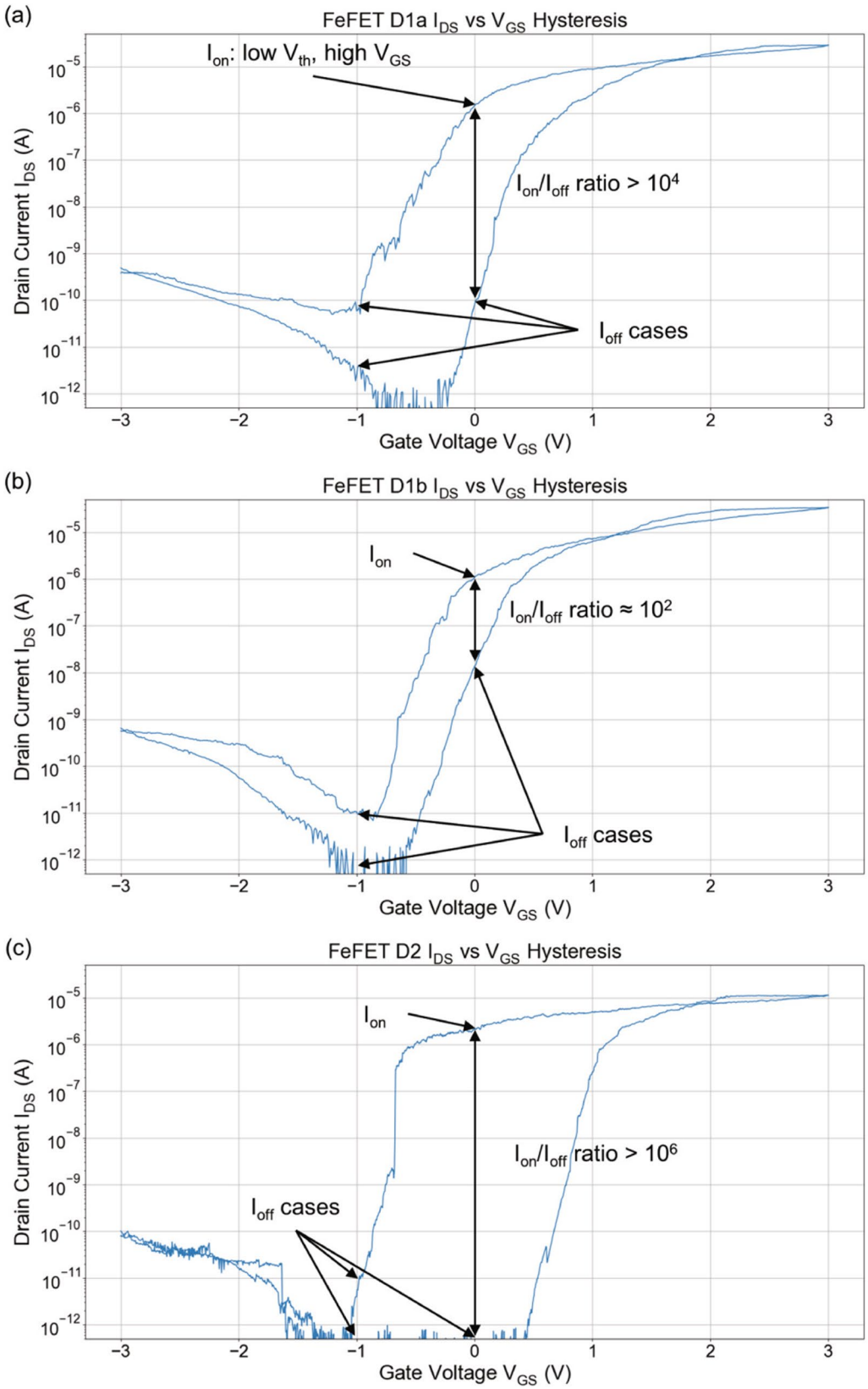

Figure 3: Measured I_{DS}–V_{GS} hysteresis for 28 nm HKMG FeFET devices of various size: (a) Device 1a (or D1a) with dimensions 500 nm × 80 nm, (b) Device 1b (D1b) with the same dimensions as D1a, and (c) Device 2 (D2) with dimensions 80 nm × 34 nm. V_{DS} of 50 mV and a V_{GS} sweep are applied to produce the hysteresis curves. 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.

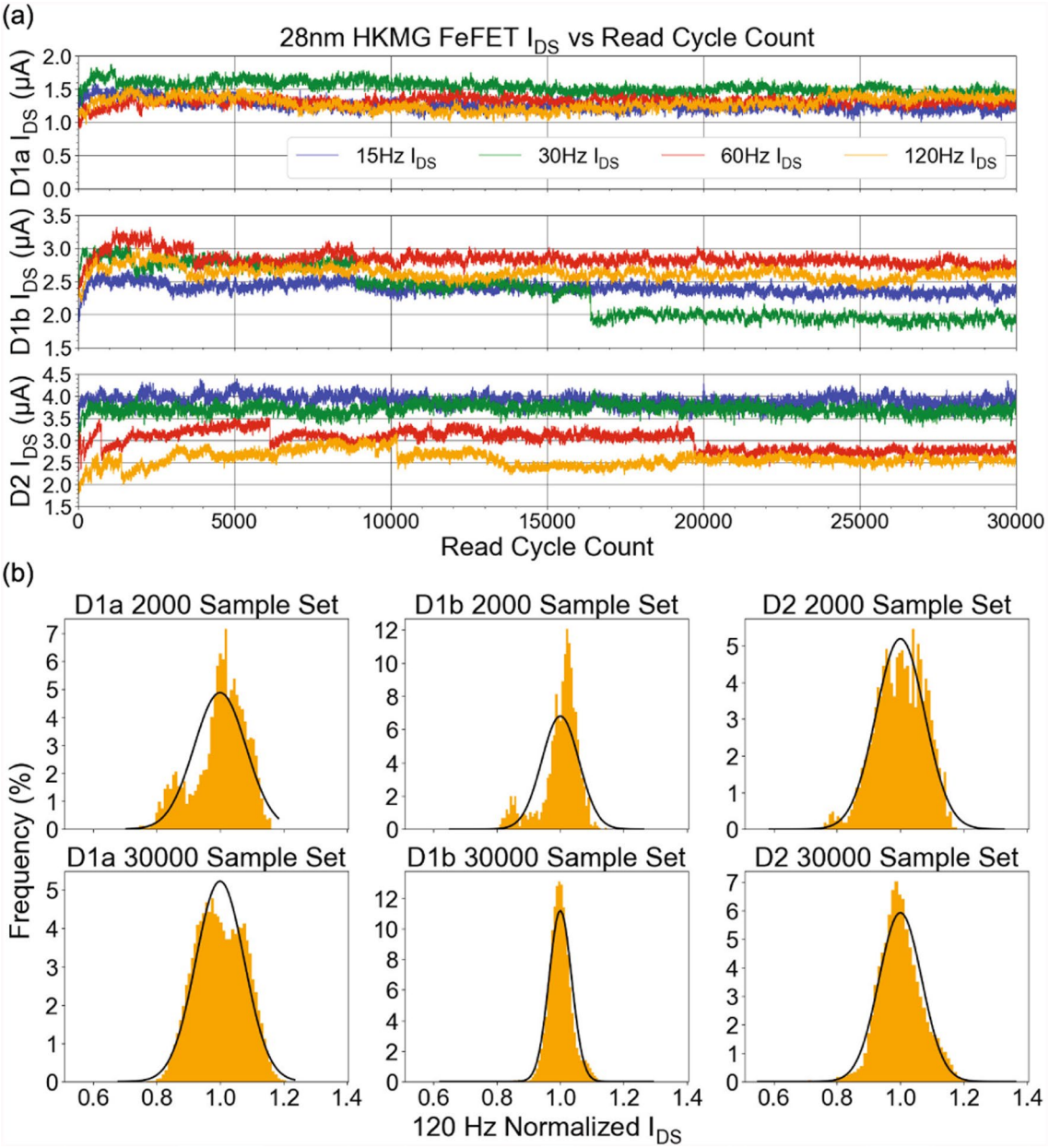

Figure 4: (a) Measured I_{DS} for D1a, D1b and D2. The devices are programmed only once before the first cycle. The distributions are normalized to Gaussian distributions of mean 1. (b) 120 Hz distributions for D1a, D1b and D2 are bootstrap sampled from the full 30,000 sample set and the 2000 sample ramp up period.

architectures composed of FeFET devices designed with sufficiently large memory windows.

According to the hysteresis curve in Fig. 3b, D1b is expected to produce a higher I_{on}/I_{off} ratio closer to 10^3 if the logic high V_{GS} is lowered to −0.5 V instead of 0 V. Therefore, we measure I_{DS} variation in this case as well, as shown in Fig. 5. With the lowered V_{GS}, we generally observe slightly less variation in the measured I_{DS} distributions than when V_{GS} of 0 V is used. The compiled standard deviations and skew of every measured dataset for all three of the 28 nm HKMG devices is shown in Table 1a.

We also measure I_{DS} variation in two 22 nm FDSOI devices [20]. Device 3 (or D3) has channel dimensions of 1 μm × 70 nm, and Device 4 (D4) consists of ten parallel FeFETs, each with channel dimensions of 170 nm × 24 nm. The hysteresis curves for these devices are shown in Fig. 6. We follow a similar approach for logic parameter assignment for these devices as for the 28 nm devices, wherein the only state which produces a

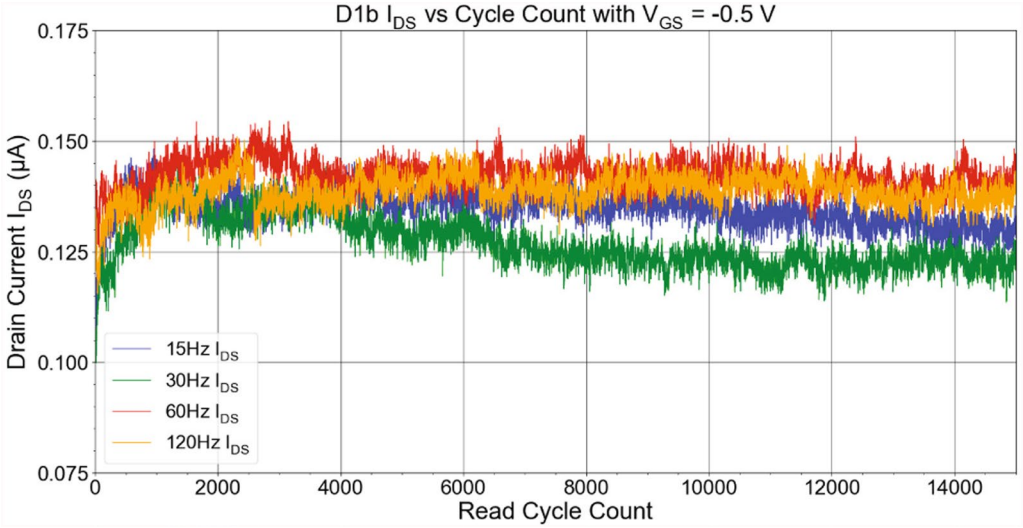

Figure 5: I_{DS} for D1b is measured with a lower V_{GS} of -0.5 V to achieve a higher I_{on}/I_{off} ratio with a measurement setup similar to that of Fig. 4.

high I_{DS} occurs where V_{th} is programmed low and V_{GS} is high. We note that because of the large memory window of D4, a logic 0 V_{GS} near −1.5 V would be needed for successful PIM operation as opposed to the −1 V we have assigned for other devices. We also observe that the I_{on}/I_{off} ratio can be increased for both devices by lowering the high V_{GS} to −0.5 V similarly to D1b, but we perform read endurance tests with a V_{GS} waveform with a base of −1 V and a peak of 0 V at a pulse width of 10 μs to maintain consistency with the tests performed on the 28 nm devices.

All tests performed on the 22 nm devices are performed at 30 Hz, and we study the effects of partially programming V_{th} on the I_{DS} distributions of D4. This process is performed by first erasing the device, then sweeping the V_{GS} to a maximum program voltage value between 1.5 V and 3 V, where 3 V represents a full program.

The 22 nm devices show significant variations in I_{DS} in the first 2000 cycles, similarly to the 28 nm devices. D3 shows a ramp up followed by a drop to a value which stays relatively steady throughout the remainder of the cycles, whereas D4 shows a steady drop in its early cycles. We hypothesize that these trends are also caused by parasitic capacitance in the devices, and the difference in the directions of these trends in D3 and D4 may be explained by D4 being composed of ten parallel FeFETs rather than one single device. To capture these trends, we perform bootstrap sampling of both the first 2000 measurements and the full dataset, similarly to the 28 nm devices. The measurement results for both D3 and D4 are shown in Fig. 7, and the standard deviation and skewness of each distribution when fit to mean 1 are shown in Table 1b and c.

By varying the program voltage of D4 as shown in Fig. 7b, we notice that the average I_{DS} value depends greatly on the program voltage. Increasing the program voltage increases the average I_{DS} value until a maximum point is reached near 2.25 V. The 2.25 V and 2.5 V results show very similar average I_{DS} results, while increasing the program voltage to 3 V actually results in a slight decrease in the average I_{DS}. For maximum I_{on}/I_{off}, it may be beneficial therefore to choose a program voltage closer to 2.25 V. However, we also notice that the partial program at 1.5 V shows the least variation in I_{DS} from cycle to cycle, as shown by its very low standard deviation and skew in Table 1c. Therefore, to minimize cycle to cycle variation, a lower program voltage may be beneficial for certain applications at the expense of the I_{on}/I_{off} ratio.

Using all of these results from both the 28 and 22 nm FeFET devices, we perform device-circuit co-simulation using PyTorch [21] to study the effects I_{DS} variation on our PIM architecture's accuracy when classifying the Fashion-MNIST dataset [15] using the LeNet-5 convolutional neural network model [16], as well as classifying the CIFAR-10 dataset [17] using MobileNetV2 [18] (Fig. 8a). We specifically emulate the PIM architecture designed by Yun Long et al. [10], which is composed of many coupled FeFET crossbars to form the VMM engine, each of which are similar to Fig. 2. We consider two PIM operating modes, ASIC Mode and Accelerator Mode (shown in Fig. 8b), which introduce FeFET noise which is bootstrap sampled from the full measured distributions and the 2000 cycle ramp up period distributions, respectively.

We define ASIC mode as the mode of PIM operation where the system of coupled FeFET crossbar arrays is large enough to store the full DNN weight matrix at once. In this operation mode, the weights are written to the FeFET arrays once and streaming inputs (images) are used for inference. The FeFET V_{th} are thus programmed once at device startup and the I_{DS} is measured many times during inference without rewriting V_{th} between

TABLE 1: Measured I_{DS} distribution standard deviation and skewness when normalized to mean 1.

(a) 28 nm HKMG devices

	D1a				D1b, $V_{GS} = 0V$				D1b, $V_{GS} = -0.5V$				D2			
	Full set		Ramp up		Full set		Ramp up		Full set		Ramp up		Full set		Ramp up	
Frequency	σ	Skew	σ	Skew	σ	Skew	σ	Skew	σ	Skew	σ	Skew	σ	Skew	σ	Skew
15	0.063	0.709	0.052	− 1.322	0.034	− 0.182	0.055	− 2.489	0.028	− 0.728	0.042	− 1.670	0.031	− 0.021	0.027	− 0.900
30	0.057	0.209	0.062	− 0.557	0.159	0.204	0.040	− 1.257	0.047	0.429	0.059	− 1.161	0.030	− 0.431	0.042	− 1.932
60	0.048	− 0.893	0.079	− 0.427	0.037	0.885	0.073	− 0.971	0.023	0.001	0.031	− 0.160	0.072	− 0.134	0.063	− 0.073
120	0.076	0.005	0.082	− 0.669	0.036	0.291	0.059	− 1.265	0.025	− 0.759	0.033	− 1.039	0.067	0.088	0.077	− 0.313

(b) 22 nm FDSOI D3: 30 Hz, PGM 3 V

	Full set		Ramp up	
PGM (V)	σ	Skew	σ	Skew
3	0.055	1.535	0.106	1.247

(c) 22 nm FDSOI D4: 30 Hz, Varied PGM

	Full set		Ramp up	
PGM (V)	σ	Skew	σ	Skew
1.5	0.015	− 0.244	0.015	− 0.243
1.75	0.043	2.183	0.078	0.079
2.0	0.048	2.023	0.042	0.489
2.25	0.032	4.746	0.061	1.934
2.5	0.037	3.003	0.053	1.548
3.0	0.059	4.415	0.075	3.897

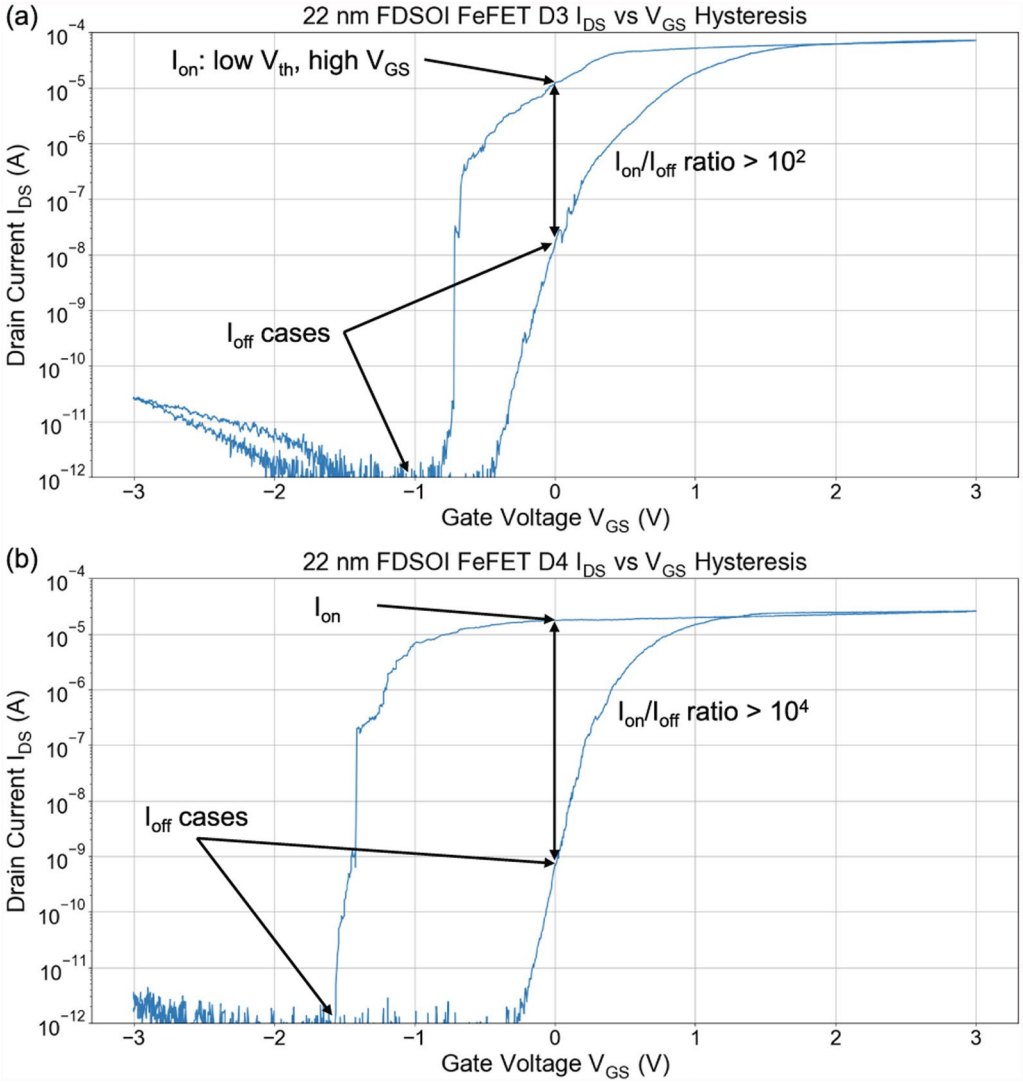

Figure 6: I_{DS}–V_{GS} hysteresis for two 22 nm FDSOI FeFET devices: (a) Device 3 (D3) with channel dimensions of 1 μm × 70 nm and (b) Device 4 (D4) which consists of ten parallel devices with dimensions 170 nm × 24 nm. The hysteresis curves are measured with a measurement setup similar to that which is used for the data in Fig. 3.

image passes. Therefore, this operation mode is impacted most by long term variation in I_{DS}, which we emulate by boostrap sampling the full 30,000 measurement set. As each FeFET in the weight matrix is read once during the passing of a single image, the read measurement frequency represents the frame rate of the full system. For example, a 15 Hz FeFET read frequency corresponds to a frame rate of 15 FPS.

We define accelerator mode as the mode of PIM operation where the on-chip storage of the VMM engine is not sufficient to store the entire weight matrix for the DNN. In this case, the weights are time-multiplexed to compute a large network (similar to the process used by Yun Long et al. [10]). In this mode of operation, the FeFET weights are programmed to contain the first neural network layer, each image is processed consecutively, the weights are rewritten with the next DNN layer, and so on. As the cells are frequently rewritten as the weights are time-multiplexed, the system is most impacted by I_{DS} variation shortly after reprogramming, and we therefore bootstrap sample I_{DS} values from the 2000 cycle ramp up period distributions. The frequency of read measurement in this operation mode represents the ratio of the size of the weight matrix to the total capacity of the crossbar. Processing a batch of 20 images with a 20 FPS throughput with a 5 layer neural network 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. We note that since write operation occurs on the order of a few ns [19, 20] for these devices, write time is negligible by comparison to the read time when calculating this frequency. In cases where the batch size is equal to 1 (processing a single image as in a digital camera), the read frequency is irrelevant.

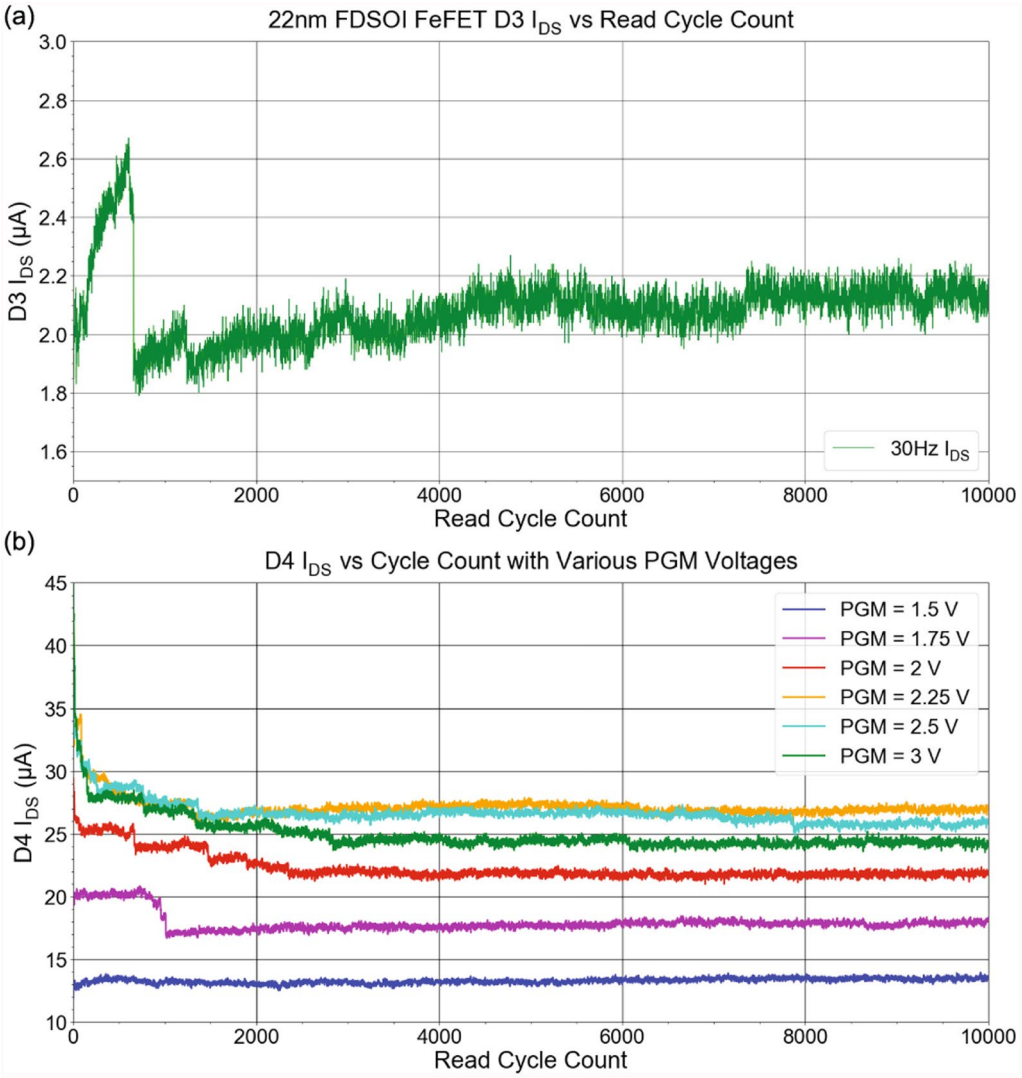

Figure 7: (a) Measured I_{DS} variation for D3 and D4 at 30 Hz read frequency. (b) Measured I_{DS} variation in D4 due to various program voltages in the initial ERS-PGM cycle. All measurements are acquired utilizing a similar measurement setup to that which is used for the data in Fig. 4.

We emulate our PIM architecture's performance in the presence of I_{DS} variation using device-circuit co-analysis. First, each model parameter is quantized to 8 bits, then the I_{DS} variation is introduced to each logic 1 bit coming from the FeFET outputs to represent variation in I_{on} during PIM operation. The I_{DS} variation in each of the individual devices accumulates in the columns of the crossbar and leads to degradation of the classification accuracy of the system. Two emulation tests are studied in this work. The Fashion-MNIST dataset [15] is classified using the LeNet-5 convolutional neural network architecture [16], and the CIFAR-10 dataset [17] is classified using MobileNetV2 [18]. The baseline, noiseless classification accuracy of the 8-bit quantized networks are shown to be 85.86% for Fashion-MNIST classification using LeNet-5, and 90.06% for CIFAR-10 classification using MobileNetV2. Table 2 shows the full results of the emulation using each of the measured distributions, including the mean and standard deviation in the PIM architecture's classification accuracy due to I_{DS} variation.

In most cases of the Fashion-MNIST classification on LeNet-5, we see only a marginal drop in classification accuracy of about 1 to 3% depending on the standard deviation and skewness of the measurement distributions. The primary outlier is the 30 Hz case of D1b with $V_{GS} = 0$ V, in which the measurements show a nearly trimodal distribution caused by the sudden drops in average I_{DS} at regular intervals in the test due to V_{th} retention loss. In most devices, there is not a clear dependence on measurement frequency, although we do observe a generally decreasing accuracy trend as frequency increases specifically in accelerator mode for D1a and D2. In the program voltage tests for D4, we observe high accuracy with low variance in the cases

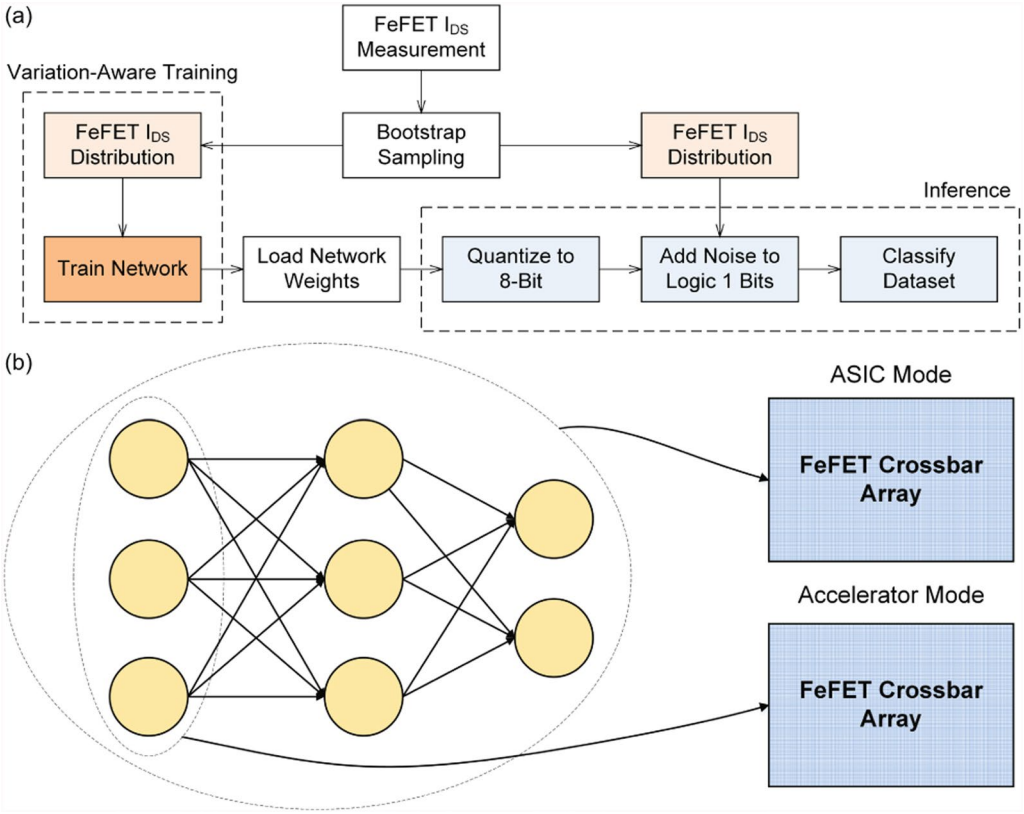

Figure 8: (a) Analysis algorithm for device-circuit co-analysis using I_{DS} variation in PIM architecture emulation and (b) the two PIM architecture operating modes. The two operating modes depend on the size of the DNN weight matrix. In ASIC mode, the DNN is small enough to be fully stored in the FeFET crossbars at once. Alternatively in accelerator mode, the DNN is too large to be loaded into the array at once, so a single layer is written into the array at a time and then overwritten by the next layer.

where I_{DS} remains steadiest, showing further evidence that tuning V_{GS} to certain values with stable I_{DS} (typically lower than the maximum allowable program voltage) can be beneficial for system accuracy.

When classifying CIFAR-10 with MobileNetV2, the degradation of the classification accuracy due to I_{DS} variation is much more pronounced. In general, the trends in accuracy reduction with respect to frequency and program voltage tend to follow the trends shown in the Fashion-MNIST results, wherein test cases with degraded accuracy on Fashion-MNIST show significant degradation under this larger workload. In the worst case of 30 Hz measurement in D1b as mentioned previously, where a nearly trimodal I_{DS} distribution occurs, the CIFAR-10 accuracy drops to a miniscule 12.14%. Another case in which the CIFAR-10 classification shows significant degradation as compared to Fashion-MNIST classification is in D3, where the large I_{DS} spike in the first 1000 cycles drops the classification accuracy as low as 26.61% in accelerator mode. However, there are also many test cases which show accuracy over 85%, showing that our PIM design still has very reasonable accuracy even for highly complex workloads when the I_{DS} variation is minor. This is especially true in the case of D1b with lowered V_{GS} read voltage, which shows some of the highest accuracy results, demonstrating that lowering the read voltage can improve system accuracy if the I_{on}/I_{off} ratio is kept sufficiently high.

To recover the accuracy loss caused by I_{DS} variation, we can sample the I_{DS} measurements to add noise while training the DNN. This process, called variation-aware training, is documented by Yun Long et al. in a similar DNN accelerator design using ReRAM [22]. We observe full accuracy recovery to the 85.86% baseline by using variation aware training in the LeNet-5 emulation for Fashion-MNIST classification. This accuracy recovery becomes much more difficult in the MobileNetV2 classification of CIFAR-10 due to the significantly larger network and the large accuracy reductions caused by noise [23]. Therefore, the drop in classification accuracy caused by the FeFET I_{DS} variation measured in this study is shown to not be a limiting factor to our PIM accelerator design in the case of small workloads, but larger workloads present challenges which require more advanced design techniques to remedy.

TABLE 2: Emulated PIM classification accuracy using measured I_{DS} distributions.

(a) 28 nm HKMG devices

		D1a				D1b: $V_{GS} = 0V$				D1b: $V_{GS} = -0.5V$				D2			
		ASIC mode		Accel. mode		ASIC mode		Accel. mode		ASIC mode		Accel. mode		ASIC mode		Accel. mode	
Classification	Freq. (Hz)	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Fashion-MNIST with LeNet-5	15	84.22	0.23	84.78	0.20	85.42	0.18	85.36	0.19	85.48	0.17	85.45	0.18	85.48	0.18	85.55	0.15
	30	84.59	0.25	84.33	0.23	78.94	0.37	85.43	0.18	85.43	0.15	85.29	0.18	85.49	0.17	85.18	0.18
	60	84.99	0.19	83.27	0.30	85.54	0.21	85.05	0.25	85.42	0.17	85.47	0.17	83.67	0.27	84.31	0.24
	120	83.40	0.31	83.00	0.29	85.43	0.20	85.28	0.23	85.41	0.16	85.51	0.17	84.05	0.25	83.37	0.24
CIFAR-10 with MobileNetV2	15	64.99	0.59	74.21	0.60	84.62	0.26	72.86	0.51	86.47	0.18	80.92	0.39	85.55	0.15	86.65	0.23
	30	70.42	0.65	66.23	1.19	12.14	0.42	82.06	0.37	77.94	0.45	69.79	0.61	85.87	0.15	80.98	0.56
	60	77.72	0.84	49.88	1.08	83.43	0.23	55.54	1.22	87.87	0.25	85.56	0.30	56.79	0.90	65.75	0.62
	120	52.06	0.87	46.73	0.80	84.07	0.34	69.24	0.74	87.21	0.23	84.85	0.18	61.29	0.60	51.99	1.16

(b) 22 nm FDSOI D3: 30 Hz, PGM 3 V

	ASIC mode		Accel. mode	
Classification	Mean	σ	Mean	σ
Fashion-MNIST with LeNet-5	85.21	0.18	84.90	0.19
CIFAR-10 with MobileNetV2	72.34	0.56	26.61	0.63

TABLE 2: (continued)

(c) 22 nm FDSOI D4: 30 Hz, Varied PGM

Classification	PGM (V)	ASIC mode Mean	Accel. mode σ	Mean	σ
Fashion-MNIST with LeNet-5	1.5	85.45	0.13	85.42	0.17
	1.75	85.46	0.21	84.90	0.25
	2.0	85.40	0.13	85.44	0.18
	2.25	85.42	0.16	85.32	0.15
	2.5	85.38	0.13	85.42	0.25
	3.0	85.22	0.27	84.85	0.21
CIFAR-10 with MobileNetV2	1.5	89.16	0.08	89.15	0.15
	1.75	80.29	0.34	50.46	0.79
	2.0	76.89	0.62	80.87	0.60
	2.25	85.43	0.30	66.75	0.64
	2.5	83.06	0.20	73.11	0.80
	3.0	67.54	0.58	51.51	0.93

Note that the quantized, noiseless classification accuracy of Fashion-MNIST with LeNet-5 is 85.86%, and of CIFAR-10 with MobileNetV2 is 90.06%.

Discussion

In this study, measured I_{DS} variation in 28 nm HKMG and 22 nm FDSOI FeFET devices in both the short term and the long term is applied to the emulation of a PIM-based DNN accelerator. The accuracy of the accelerator is tested when classifying the Fashion-MNIST dataset using the LeNet-5 convolutional neural network and when classifying the CIFAR-10 dataset with MobileNetV2. In each measured I_{DS} dataset, we tend to see the largest variation in the initial read cycles, which we use to emulate a mode of PIM operation we call accelerator mode. In the 28 nm HKMG devices, this initial variation manifests as a ramp up period before the average I_{DS} settles to a steady average, which we attribute to a parasitic capacitance present in the device or measurement system. In the 22 nm FDSOI devices, we notice a ramp up period in D3 followed by a sudden drop, and a ramp down period in D4. The differences in these responses can likely be explained by the device structure, as D4 is composed of ten parallel FeFETs instead of a single device as in D3. Additionally, we study long term variation in I_{DS} to emulate PIM operation in ASIC mode. In most of the test cases demonstrated in this study, ASIC mode outperforms accelerator mode in the metric of classification accuracy, with the few exceptions occurring where sudden drops in average I_{DS} create skewed or multimodal distributions.

In the 28 nm HKMG FeFETs, we analyze the effects of different read frequencies on the I_{DS} variation to determine how a higher frame rate of the system impacts device variation. In D1a and D2 we observe a decreasing PIM classification accuracy for both of the tested classification tasks as frequency increases in accelerator mode, but this trend does not hold in ASIC mode, nor does it hold for D1b. In D1a and D2, the decreasing trend is caused by higher frequencies showing more significant variation in the ramp up period. However, since this trend is not consistent in D1b, we cannot say with certainty that this trend will hold for all HKMG FeFET devices across varying frequencies.

Although they are theoretically identical devices with the same channel dimensions, we see significant differences between the memory windows and measured I_{DS} variations of D1a and D1b. D1a and D1b show a similar $I_{on}I_{on}$ on the order of 1 μA, but D1b has a lower V_{th} in the logic 0 state, leading to a significantly higher I_{off}. This is likely caused by manufacturing variability in the fabrication process, particularly in the ferroelectric layer, leading to a lower degree of electrical polarization in the gate stack differentiating the programmed and erased states and thus a narrower memory window. As clearly evidenced by just these two devices, device-to-device variation within the FeFET crossbar could be a very critical component leading to inaccuracies in overall computation, and is an avenue we do not explore in depth in this study. With more devices to characterize, this could be a very interesting avenue for future work.

To address some of these differences, we test D1b with a lower V_{GS} of −0.5 V which helps to raise its I_{on}/I_{off} ratio. Through this test, we observe that the lower V_{GS} actually leads to less cycle-to-cycle variation in the I_{DS} measurements, thereby leading to improved PIM performance with a higher classification accuracy. When assigning and tuning V_{GS} parameters, designers can choose to lower V_{GS} in order to lower cycle-to-cycle I_{DS} variation, while also considering the tradeoff between V_{GS} and I_{on}/I_{off} ratio in specific devices.

The results shown by D2 clearly show that V_{th} retention loss from ferroelectric breakdown can occur due to repeated read operation. As each successive test is performed, the average I_{DS} for this device decreases, including abrupt drops in the 60 Hz test. These drops occur near the range of 10^5 total read cycles, which is the same range shown in the original device characterization [19] where V_{th} reduction can occur due to bipolar stress. Although our test does not perform true bipolar stress since the voltage swing between −1 and 0 V is smaller than the full bipolar stress voltage swing from the device characterization [19], it is still evident that this voltage swing, when repeated over many cycles, can lead to V_{th} loss. This is a critical source of degradation to consider when designing PIM systems and could severely limit system accuracy over the life of the device.

We expect this effect to be compounded by write endurance effects in accelerator mode. In accelerator mode, we assume that the FeFET crossbar will be frequently rewritten as each layer of weights in the network overwrite the previous layer. Performing repeated program and erase operations to overwrite the weights stored in the FeFETs produces bipolar stress effects which degrade the low-V_{th} state by steadily increasing V_{th} (reducing I_{on}). As shown by the original characterizations of both devices [19, 20], bipolar stress effects become significant in the range of 10^4 to 10^5 cycles. In another work studying SiON and SiO_2 FeFET devices, the effects of repeated program and erase cycles in isolation are also observed to cause significant breakdown in I_{DS} due to charge trapping effects [24]. Therefore, we expect that in accelerator mode, I_{DS} would show additional decay as read and write cycles both increase.

When measuring 22 nm FDSOI devices, we study the impact of partial V_{th} programming on I_{DS} variation. We observe the highest accuracy in the measurement cases where I_{DS} remains the steadiest, typically for V_{GS} programming pulses lower than the maximum program voltage. In fact, the lowest PIM classification accuracy for both ASIC and accelerator mode tends to occur where the highest program voltage of $V_{GS} = 3$ V is used. For the MobileNetV2 classification of CIFAR-10 in particular, a 15 to 18% improvement is observed when using a program voltage of 2.5 or 2.25 V rather than the full 3 V in ASIC mode, and an improvement of over 21% is observed when using 2.5 V rather than 3 V in accelerator mode. This is a crucial result showing that partial programming of the V_{th} of the FeFETs can be beneficial to reduce I_{DS} variation in the system and thereby improve overall classification accuracy for the PIM system. However, lowering the program voltage can also lead to reduced I_{on}/I_{off}, so this decision must be optimized for maximum performance.

As shown by the 28 nm devices and as confirmed by D3 and D4, the memory windows of each device can vary significantly based on channel dimension and device-to-device variation. Devices used to create a full PIM architecture in hardware must therefore be carefully designed and tuned to maintain a high I_{on}/I_{off} ratio and optimal memory window for operation based on chosen V_{GS} parameters. For example, the memory window of D4 was too large to keep the same V_{GS} parameters as the other devices, and would thus create different design constraints if a full PIM system were to be created from this device structure.

In spite of non-Normal variation in I_{DS} in all test cases, the accuracy of the PIM architecture remains very close to the baseline quantization accuracy with only a 1 to 3% accuracy drop in most cases for the classification of Fashion-MNIST with LeNet-5. An approximately 7% drop in D1b at 30 Hz with $V_{GS} = 0$ V occurs due to abrupt I_{DS} drops, which we call an outlier since these drops are recovered when D1b is reprogrammed for the 60 and 120 Hz cases. In larger workloads such as classifying CIFAR-10 with MobileNetV2, the accuracy drops become significantly worse when modeling with I_{DS} datasets which have significant noise. However, the PIM architecture still shows promise with accuracy well over 80% in many cases. Noise aware training can fully recover the accuracy drops observed in PIM emulation of LeNet-5, although this becomes more challenging for the larger and less accurate MobileNetV2 classification of CIFAR-10 [23]. We determine that these sources of noise are not preventative of PIM architecture design for small workloads, and that larger workloads may require further advanced design techniques to ensure high accuracy.

A significant area of future work includes exploring these advanced design techniques to mitigate PIM errors due to I_{DS} variation. In our emulation, we assume that the crossbar array is large enough to hold the entire weight matrix of each layer in accelerator mode, and the entire weight matrix for the full network in ASIC mode. Reducing the array size would reduce the sum of the noise in each column, leading to higher accuracy at the cost of computation time and more write endurance-induced drift due to the need to cycle through the array more frequently per network pass. This need for rewriting the arrays can be avoided by using many networked FeFET arrays as demonstrated by Yun Long et al [10]. With this networked design, one could also explore tuning the size of each array, as too large of an array leads to higher summed error and slower read times, while using too small of an array leads to increased errors from inter-array communication. One could also explore using refresh operations to reduce average I_{DS} drift in the case

of ASIC mode, with the clearest area of need for this being demonstrated by the 30 Hz measurements from D1b which drift far below their starting value and are refreshed for the 60 Hz and 120 Hz measurements.

Future work in the area of device measurement includes deeper study of device-to-device variation and parameter tuning which would be crucial to developing an FeFET PIM-based DNN accelerator in hardware. Variation observed between devices that should be theoretically identical, as in the case of D1a and D1b, is especially crucial to study. Should a hardware DNN accelerator be developed from many devices of these same dimensions, it is possible that there will be large variations in the memory windows and I_{on}/I_{off} ratios of each individual device due to material and process variations, leading to inaccuracies in overall system accuracy. Testing many more devices would improve understanding of this issue. Additionally, this work has not explored FeFETs built on TFTs or FeFETs with different ferroelectric materials than Si:HfO_2. Performing I_{DS} measurement for these other types of FeFETs would provide further insight into noise behavior in devices outside of the small sample tested in this study.

Conclusion

In conclusion, measurements of the I_{DS} of individual 28 nm HKMG and 22 nm FDSOI FeFET devices over many read cycles show non-Gaussian variation which can lead to errors in the overall accuracy of PIM-based DNN accelerators. Based on measurements of three 28 nm and two 22 nm FeFETs in this work, the I_{DS} variations do not show conclusive dependence on read frequency, nor is there a stark difference between the two process technology nodes. Device-circuit co-analysis demonstrates that the PIM classification accuracy reductions caused by these measured I_{DS} variations are only marginal (between 1 to 3%) when classifying the Fashion-MNIST dataset with LeNet-5, and can be worse in more difficult workloads such as classifying CIFAR-10 with MobileNetV2. Some crucial elements of system design shown in this study include that partially programming the V_{th} of the FeFETs or using a lowered read voltage can both lead to improved accuracy in certain cases as long as the I_{on}/I_{off} ratio of the device is preserved, and that short term variation (shown in accelerator mode) tends to lead to worse PIM accuracy degradation than long term variation (shown in ASIC mode). Using variation-aware DNN training, wherein the measured I_{DS} variations are used to introduce noise during the DNN training phase, we observe that accuracy loss caused by I_{DS} variation can be fully recovered back to its noiseless baseline in the case of Fashion-MNIST classification with LeNet-5, though this noise aware training becomes less effective for larger networks like MobileNetV2. Therefore, individual FeFET device current variation due to many read cycles is shown to not be preventative of our FeFET-based DNN accelerator design in either 28 nm HKMG or 22 nm FDSOI technology nodes when accelerating small PIM workloads, but advanced design techniques are required to mitigate error in the case larger workloads.

Materials and Methods

The FeFET devices measured in this study are provided by GLOBALFOUNDRIES for both 28 nm HKMG and 22 nm FDSOI process technology nodes [19, 20]. Both devices include a ferroelectric layer of doped HfO_2 which is studied as the main source of material and process variation. Further study of both devices can be found in their original technology characterizations [19, 20], while we perform measurements which are designed with the PIM-based DNN accelerator application space in mind. All device measurements are performed using a Keysight Technologies B1530A Waveform Generator/Fast Measurement Unit.

To create a more accurate method for sampling small I_{DS} measurement datasets for PIM emulation, we perform bootstrap sampling of each of the measured datasets by sampling with replacement until each bootstrapped set contains 15,000,000 samples. These datasets are then sampled by our emulation system in PyTorch, in which we add the noise from I_{DS} to the 1 bits in the outputs to emulate noisy FeFETs within the crossbar which composes the VMM engine. During variation-aware training, this I_{DS} variation is also added during the training process in order to train the model to recover inaccuracies produced by noisy weights.

Acknowledgments

A.I.K. thanks GLOBALFOUNDRIES for providing FeFET technology wafers.

Author contributions

FeFET measurement data was collected by Nathan Eli Miller and Zheng Wang under the advisement of Asif Islam Khan. Device-circuit co-simulation was performed by Saurabh Dash and Nathan Eli Miller. Data analysis was performed primarily by Nathan Eli Miller. Saibal Mukhopadhyay advised the study and provided direction. The original manuscript was written by Nathan Eli Miller and read, reviewed and approved by all authors.

Funding

This material is based on work supported by National Science Foundation (1810005).

Data availability

The datasets generated and analyzed in this study, as well as the code which produced them, are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest.

References

1. T. Mikolajick, U. Schroeder, S. Slesazeck, IEEE Trans. Electron Dev. **67**(4), 1434 (2020). https://doi.org/10.1109/TED.2020.2976148
2. A.I. Khan, A. Keshavarzi, S. Datta, Nat. Electron. **3**(10), 588 (2020). https://doi.org/10.1038/s41928-020-00492-7
3. E.T. Breyer, H. Mulaosmanovic, J. Trommer, T. Melde, S. Dünkel, M. Trentzsch, S. Beyer, S. Slesazeck, T. Mikolajick, IEEE J. Electron Dev. Soc. **8**, 748 (2020). https://doi.org/10.1109/JEDS.2020.2987084
4. K. Ni, X. Yin, A.F. Laguna, S. Joshi, S. Dünkel, M. Trentzsch, J. Müller, S. Beyer, M. Niemier, X.S. Hu, S. Datta, Nat. Electron. **2**(11), 521 (2019). https://doi.org/10.1038/s41928-019-0321-3
5. Z. Wang, S. Khandelwal, A.I. Khan, IEEE Electron Dev. Lett. **38**(11), 1614 (2017). https://doi.org/10.1109/LED.2017.2754138
6. S.K. Thirumala, S.K. Gupta, IEEE Trans. Electron Dev. **66**(6), 2771 (2019). https://doi.org/10.1109/TED.2019.2897960
7. N. Tasneem, A.I. Khan, in *2018 76th Device Research Conference (DRC)* (2018), pp. 1–2. https://doi.org/10.1109/DRC.2018.8442250
8. M. Jerry, P. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, S. Datta, in *2017 IEEE International Electron Devices Meeting, IEDM 2017* (Institute of Electrical and Electronics Engineers Inc., 2018), Technical Digest—International Electron Devices Meeting, IEDM, pp. 6.2.1–6.2.4. https://doi.org/10.1109/IEDM.2017.8268338
9. H. Mulaosmanovic, J. Ocker, S. Müller, M. Noack, J. Müller, P. Polakowski, T. Mikolajick, S. Slesazeck, in *2017 Symposium on VLSI Technology* (2017), pp. T176–T177. https://doi.org/10.23919/VLSIT.2017.7998165
10. Y. Long, D. Kim, E. Lee, P. Saha, B.A. Mudassar, X. She, A.I. Khan, S. Mukhopadhyay, IEEE J. Explor. Solid-State Comput. Dev. Circ. **5**(2), 113 (2019). https://doi.org/10.1109/JXCDC.2019.2923745
11. I. Yoon, M. Jerry, S. Datta, A. Raychowdhury. Design space exploration of ferroelectric fet based processing-in-memory DNN accelerator (2019). https://arxiv.org/abs/1908.07942
12. N.E. Miller, Z. Wang, S. Dash, A.I. Khan, S. Mukhopadhyay, in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (2021), pp. 1–4. https://doi.org/10.1109/AICAS51828.2021.9458437
13. C. Besleaga, R. Radu, L.M. Balescu, V. Stancu, A. Costas, V. Dumitru, G. Stan, L. Pintilie, IEEE J. Electron Dev. Soc. **7**, 268 (2019). https://doi.org/10.1109/JEDS.2019.2895367
14. I. Katsouras, D. Zhao, M.J. Spijkman, M. Li, P.W.M. Blom, D.M.D. Leeuw, K. Asadi, Sci. Rep. **5**(1), 12094 (2015). https://doi.org/10.1038/srep12094
15. H. Xiao, K. Rasul, R. Vollgraf, CoRR (2017). http://arxiv.org/abs/1708.07747
16. Y. LeCun. Lenet-5, convolutional neural networks (2015). http://yann.lecun.com/exdb/lenet
17. A. Krizhevsky, Learning multiple layers of features from tiny images. Tech. rep. (2009)
18. M. Sandler, A.G. Howard, M. Zhu, A. Zhmoginov, L. Chen, CoRR **abs/1801.04381** (2018). http://arxiv.org/abs/1801.04381
19. M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck, J. Ocker, M. Noack, J. Müller, P. Polakowski, J. Schreiter, S. Beyer, T. Mikolajick, B. Rice, in *2016 IEEE International Electron Devices Meeting (IEDM)* (2016), pp. 11.5.1–11.5.4. https://doi.org/10.1109/IEDM.2016.7838397
20. S. Dünkel, M. Trentzsch, R. Richter, P. Moll, C. Fuchs, O. Gehring, M. Majer, S. Wittek, B. Müller, T. Melde, H. Mulaosmanovic, S. Slesazeck, S. Müller, J. Ocker, M. Noack, D. Löhr, P. Polakowski, J. Müller, T. Mikolajick, J. Höntschel, B. Rice, J. Pellerin, S. Beyer, 2017 IEEE International Electron Devices Meeting (IEDM) pp. 19.7.1–19.7.4 (2017). https://doi.org/10.1109/IEDM.2017.8268425
21. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, CoRR (2019). http://arxiv.org/abs/1912.01703
22. Y. Long, X. She, S. Mukhopadhyay, in 2019 Design. Automation Test in Europe Conference Exhibition (DATE) , 1769–1774 (2019)
23. S. Dash, S. Mukhopadhyay, in *Proceedings of the 39th International Conference on Computer-Aided Design* (Association for Computing Machinery, New York, NY, USA, 2020), ICCAD '20. https://doi.org/10.1145/3400302.3415679
24. T. Ali, P. Polakowski, S. Riedel, T. Büttner, T. Kämpfe, M. Rudolph, B. Pätzold, K. Seidel, D. Löhr, R. Hoffmann, M. Czernohorsky, K. Kühnel, P. Steinke, J. Calvo, K. Zimmermann, J. Müller, IEEE Trans. Electron Dev. **65**(9), 3769 (2018). https://doi.org/10.1109/TED.2018.2856818