IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 29 April 2019; revised 18 June 2019 and 3 July 2019; accepted 17 July 2019.
Date of publication 22 July 2019; date of current version 3 February 2020.

*Digital Object Identifier 10.1109/JXCDC.2019.2930222*

# A FerroFET-Based In-Memory Processor for Solving Distributed and Iterative Optimizations via Least-Squares Method

**INSIK YOON[1] (Student Member, IEEE), MUYA CHANG[1], KAI NI[2] (Member, IEEE), MATTHEW JERRY[2], SAMANTAK GANGOPADHYAY[1] (Student Member, IEEE), GUS HENRY SMITH[3], TOMER HAMAM[1] (Member, IEEE), JUSTIN ROMBERG[1] (Fellow, IEEE), VIJAYKRISHNAN NARAYANAN[3] (Fellow, IEEE), ASIF KHAN[1] (Member, IEEE), SUMAN DATTA[2] (Fellow, IEEE), and ARIJIT RAYCHOWDHURY[1] (Senior Member, IEEE)**

[1]Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA
[2]Department of Electrical Engineering, University of Notre Dame, South Bend, IN 46556 USA
[3]Department of Computer Science and Engineering and Electrical Engineering, Penn State University, State College, PA 16801 USA

CORRESPONDING AUTHOR: A. RAYCHOWDHURY (arijit.raychowdhury@ece.gatech.edu)

This work was supported in part by the Semiconductor Research Corporation (SRC) through Extremely Energy Efficient Collective Electronics (EXCEL), an SRC-Nanoelectronics Research Initiative (NRI) under Grant ID 2698.002 and in part by the Joint University Microelectronics Program (JUMP) Applications and Systems driven Center for Energy-Efficient Integrated NanoTechnologies (ASCENT) under Grant 2776.044.

**ABSTRACT** In recent years, several designs that use in-memory processing to accelerate machine-learning inference problems have been proposed. Such designs are also a perfect fit for discrete, dynamic, and distributed systems that can solve large-dimensional optimization problems using iterative algorithms. For in-memory computations, ferroelectric field-effect transistors (FerroFETs) owing to their compact area and distinguishable multiple states offer promising possibilities. We present a distributed architecture that uses FerroFET memory and implements in-memory processing to solve a template problem of least squares minimization. Through this architecture, we demonstrate an improvement of 21× in energy efficiency and 3× in compute time compared to a static random access memory (SRAM)-based *processing-in-memory* (PIM) architecture.

**INDEX TERMS** Distributed computing, emerging, ferroelectric field-effect transistors (FerroFETs), hardware, in-memory processing, least square, optimization, post-CMOS.

## I. INTRODUCTION

MODERN computing systems based on the Von-Neumann architecture rely on a clear distinction between logic and memory and process information by executing a sequence of precise atomic instructions with periodic uploads to the memory. Such systems are the foundation of the digital revolution that began with the demonstration of the self-aligned planar-gate silicon MOSFET in the 1960s and was accelerated by rapid advances in transistor technology. However, in the last decade, the volume of data collected by distributed sensors and networks has grown exponentially. Ingesting, processing, and extracting actionable intelligence out of this abundant data require a large amount of data traffic between logic and memory blocks, leading to the problem of memory bottleneck. This requires novel ways of architecting the compute platform. For example, by embedding processing elements in the memory subarray itself in so-called processing-in-memory (PIM) architectures [1]–[5], the traditional Von-Neumann bottleneck can be addressed and significant acceleration and improved power efficiency can be achieved. In order to solve the memory bottleneck problem, current research focuses on architectures and memory arrays that can accelerate memory-based processing for machine learning applications. Designs explore the use of static random access memory (SRAM) arrays [6], crossbar arrays with ReRAMs [7]–[9], memristors [10]–[12], and spintronic MRAMs [13].

Apart from inference, one ubiquitous algorithm in signal processing and autonomous systems is optimization—in particular, convex optimization. Least squares minimization is such a template problem and is the focus of this paper. We demonstrate that distributed convex optimization via least squares method can be efficiently implemented in an iterative dynamical system using a systolic

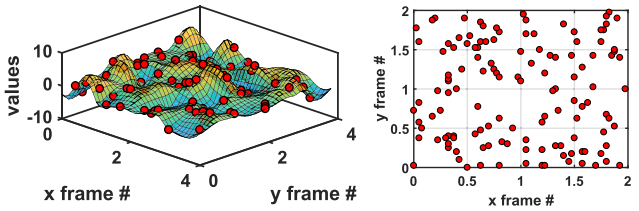


**FIGURE 1. (a) 2-D continuous function *f(u, v)* with nonuniform samples. (b) Spatial location of the nonuniform samples.**

PIM architecture, with breakthrough energy efficiency and performance. In particular, the iterative and parallel nature of memory-read makes the systolic PIM a good candidate for the proposed algorithm. This is further made possible by a parallel development in device technologies, namely, the advent of multiple embedded nonvolatile memories (eNVM). Among all competing eNVM technologies, ferroelectric field-effect transistors (FerroFETs) have emerged as promising candidates due to their compact size, multi-level storage, nanosecond read–write, and high energy efficiency. We demonstrate that a systolic PIM architecture, using FerroFET pseudo-crosspoint array can solve least squares minimization with 21× improvement in energy efficiency compared with an SRAM PIM architecture.

## II. CONVEX LEAST SQUARE MINIMIZATION

Before discussing the systolic PIM architecture, we present a brief overview of distributed least squares minimization as a template problem, with widespread applications in discrete signal processing. In particular, it is a common tool for signal reconstruction where the process of sampling is nonuniform [14], [15] such as in computerized tomography (CT), magnetic resonance imaging (MRI) [16], radar signal processing, light detection and ranging (LIDAR) systems, and so on. Consider (1) $u$ and $v$ are the horizontal and the vertical arguments of a continuous signal; (2) $x$ and $y$ are the discrete coordinate indexes; and (3) $\omega_x$ and $\omega_y$ are horizontal and vertical spatial frequencies. Let $f(u, v)$ be a band-limited signal in $\mathbb{R}^2$. The signal is nonuniformly sampled and are stored in vector $\boldsymbol{b}$, which are referred to as $f(x, y)$. The objective is to use the nonuniform samples to obtain complete reconstruction of $f(u, v)$ in $N_x \cdot N_y$ dimensional subspace. Fig. 1 shows an example of $f(u, v)$ and the results of nonuniform sampling. In this algorithm, we assume that $f(u, v)$ lies in an $N_x \cdot N_y$ dimensional subspace. To reconstruct the signal accurately we have used 2-D lapped orthogonal transform (LOT) cosine-IV harmonics as the basis functions. A smoothing function $g(u, v)$ is applied to all the basis functions to avoid distortions. Equation (1) shows a general LOT cosine-IV basis function. Here, $f(u, v)$ is split into $K_x$ by $K_y$ frames, $[k_x, k_y]$ represents a specific frame, $\omega_x$ and $\omega_y$ indicate the harmonic in horizontal and vertical directions

$$\begin{aligned}&\psi_{k_x,\omega_x,k_y,\omega_y}(u, v)\\&= \sqrt{2} \cdot g(u - k_x, v - k_y)\\&\quad \cdot \cos\left(\left(\omega_x + \frac{1}{2}\right)\pi(u - k_x)\right) \cos\left(\left(\omega_y + \frac{1}{2}\right)\pi(v - k_y)\right). \quad (1)\end{aligned}$$

Since $f(u, v)$ lies in a $N_x \cdot N_y$ dimensional subspace, it can be expressed as

$$f(u, v) = \sum_{\omega_x=1}^{N_x} \sum_{\omega_y=1}^{N_y} \sum_{k_x=1}^{K_x} \sum_{k_y=1}^{K_y} \alpha(k_x, \omega_x, k_y, \omega_y)\psi_{k_x,\omega_x,k_y,\omega_y}(u, v). \quad (2)$$

The key point to note here would be that LOT cosine-IV has compact support and the different frames are loosely coupled to each other. In fact, for samples in each frame, the nontrivial dependence would extend only to the adjacent frames apart from itself. According to (2), we can write an equation for each sample and collect them into matrix–vector product form and the coefficients can be found by solving the inverse-linear problem of

$$\boldsymbol{A}\boldsymbol{z} = \boldsymbol{b}. \quad (3)$$

Here $\boldsymbol{b}$ is the sample vector, $z$ is the coefficient vector obtained by stacking the coefficients $\alpha(k_x, \omega_x, k_y, \omega_y)$, and $\boldsymbol{A}$ is referred to as the Grammian (Gram) matrix of the basis.

When the size of $\boldsymbol{A}$ matrix is large (as in most applications), a direct solution is not possible. Therefore, alternatively we follow an iterative approach, the Jacobi method. A general update of $z$ in $j$th component at the $k$th iteration is given as (4), where $\boldsymbol{B} = \boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{c} = \boldsymbol{A}^T\boldsymbol{b}$

$$z_j^k = \boldsymbol{B}_{jj}^{-1}\left(\boldsymbol{c}_j - \sum_{i \neq j} \boldsymbol{B}_{ji} z_i^{k-1}\right). \quad (4)$$

Some observations are worth emphasizing: 1) to update $z_j^k$, only values from previous iterations are needed and 2) columns of $\boldsymbol{A}$ are coupled only with neighboring frames, which leads to simpler computation of $\boldsymbol{B}_{ji}$. Such a system maps naturally to a systolic PIM architecture with: 1) near neighbor connections and 2) embedded linear algebraic operators on the periphery of the subarray—as will be described in Sections III and IV.

## III. OVERVIEW OF FerroFETs-BASED PIM: MODELING AND EXPERIMENTAL VERIFICATION

In this paper, we explore FerroFETs as the technology of choice for implementing resistive cross-bar architectures that can accelerate linear algebraic operations. In particular, $HfO_2$-based FerroFETs have recently received great interest for its application in nonvolatile memory (NVM) [17]. It is CMOS-compatible and retains ferroelectricity for thin films with thickness around 10 nm. By tuning the portion of the switched ferroelectric domain, a FerroFET can exhibit multiple intermediate states, which has been used in neuromorphic computing [18], [19].

The operation of FerroFET as a multi-valued eNVM storage is different from a traditional binary memory [17] in that a series of weak pulses are applied to set the device in the desired state [18], [19]. Various pulse schemes are proposed to tune the state, including identical pulse schemes [21], pulsewidth modulation schemes [22], and pulse-amplitude

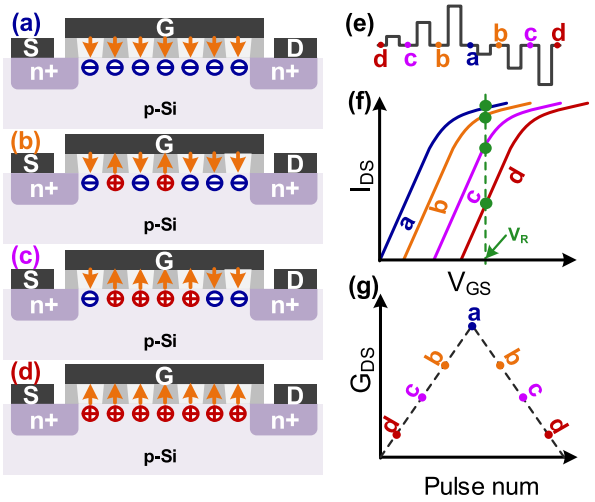


**FIGURE 2.** **(a)–(d) Different FerroFET states, corresponding to different portions of ferroelectric domain switching. Yellow arrow: polarization direction. Blue and red circles: electrons and holes, respectively. (e) Applied pulse amplitude modulation scheme. The states after each pulse are also illustrated. The initial state is assumed to be all polarizations are pointing toward the gate. (f) $I_{DS}$–$V_{GS}$ characteristics after each pulse. (g) Measured drain to source conductance as a function of applied pulse number. Here the ideal case is presented, which shows linear and symmetrical potentiation and depression [20].**

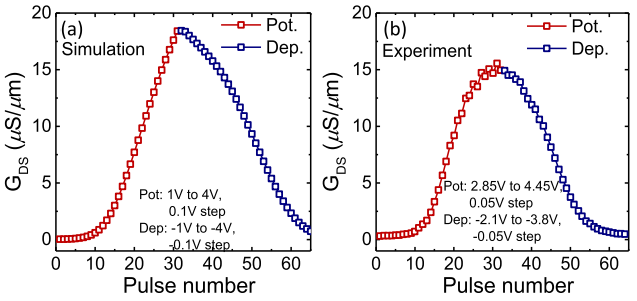


**FIGURE 3.** **(a) Simulated FerroFET channel conductance and (b) measured FerroFET channel conductance ($G_{DS}$) as a function of pulse number.**

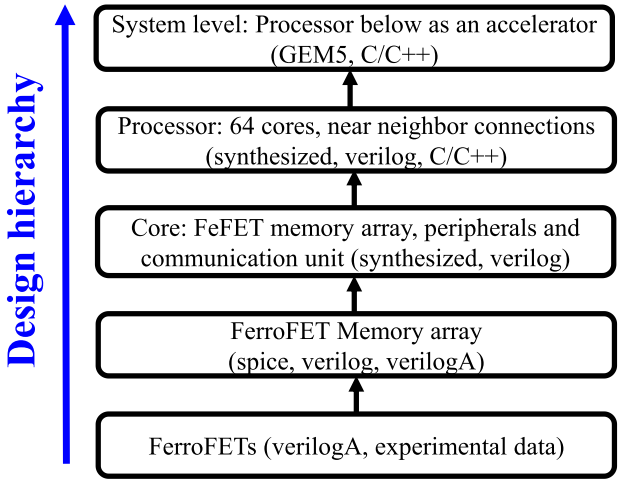


**FIGURE 4.** **Flowchart of design hierarchy from device to system.**

modulation schemes [19], [23]. For illustration, Fig. 2 illustrates the operation with a pulse-amplitude modulation scheme, which is used in this paper. Fig. 2(e) shows the applied pulse waveform. After each pulse, the percentage of switched ferroelectric domains is modified. The device states are shown in Fig. 2(a)–(d). The device $I_{DS}$–$V_{GS}$ values corresponding to different states are shown in Fig. 2(f), which shows the intermediate states. The different states could be sensed by applying a read pulse, $V_R$, the corresponding drain-to-source conductance, $G_{DS}$, can be sensed. Fig. 2(g) shows the ideal $G_{DS}$ as a function of applied pulse numbers. $G_{DS}$ increases/decreases linearly with pulse number during potentiation/depression, respectively. A symmetrical potentiation/depression is necessary for high accuracy computation. The experimental procedure is outside the scope of this paper and is described in [24]. The FerroFET model includes the atomistic simulation of domain dynamics with a drift-diffusion-based FET model. The simulation results closely match the experimental data and are shown in Fig. 3, where the different conductance levels are shown as a function of the number of programming pulses.

## IV. FerroFet PIM ARCHITECTURE AND END-TO-END TOOL CHAIN DEVELOPMENT

In this paper, we explore the FerroFET memory-based processing in memory (PIM) architecture in a hierarchical manner. A short description of each layer of the design abstraction is provided here. Fig. 4 provides the flowchart of the entire design cycle from devices to the PIM architecture. The salient features are as follows.

1) There are 64 cores, eight rows, with each row containing eight cores. With respect to Section II, this implies $N_x = N_y = 8$.
2) Each core is capable of performing Jacobi iterations with subspace dimensions $K_x$ and $K_y$ (horizontal and vertical dimensions) equal to 8. The subspace dimensions determine the core complexity and accuracy of signal reconstruction. From our analysis, we identified that 8 × 8 subspace dimensions are sufficient for signal-processing applications in hand.
3) Analog-to-digital converters (ADCs) are critical in terms of determining the latency and power consumption. In order to explore the design space properly, we have used ADCs with different resolutions and design constraints.
4) For the current design, the $B$-coefficients ($B_{jj}^{-1}B_{ji}$) and $z$-coefficients ($z_j^k$) are represented in 12-bit fixed-point representations where the MSB 6 bits represent the integer part and last 6 bits represent the fractional part.
5) To model the system, we have used Spice for simulating bit cells, Verilog and VerilogA models for array-level circuit architecture simulations and gem5 for architectural simulations.

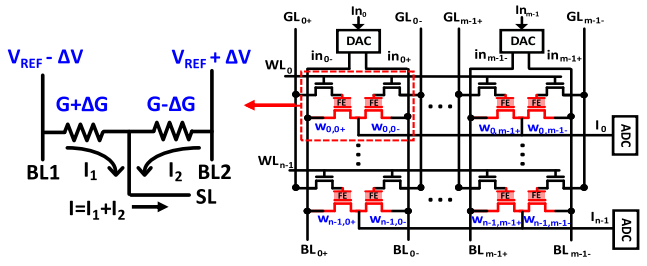


**FIGURE 5.** **FerroFET cell schematic. (a) Conceptual and (b) transistor-level implementation [20].**

### A. FerroFET CELL STRUCTURE

Fig. 5 shows the schematic for a differential FerroFET memory cell. The cell, apart from storage, provides the facility to compute 12-bit by 3-bit in-memory multiplications. Unlike previous work [25]–[27], the proposed bit-cell allows both positive and negative values for stored values as well as the inputs. During a read operation, the word line (WL) is fully turned on, appropriate $V_{GS}$ values are provided through GL1 and GL2. The entire row is read simultaneously through the current that is accumulated on source line (SL). The accumulated current corresponding to $\Delta G$ and $\Delta V$ is given by

$$I_1 = -\Delta V.(G - \Delta G), I_2 = \Delta V.(G + \Delta G) \quad (5)$$

$$I = \Delta V.(-G + \Delta G + G + \Delta G) = \Delta V.(2\Delta G). \quad (6)$$

The weights of $B$-coefficients are encoded as multiples of $2\Delta G$, and the inputs or $z$-coefficients are coded as multiples of $\Delta V$. Here, both the $\Delta G$ ($B$-coefficients) and $\Delta V$ ($z$-coefficients) can be positive or negative; or in other words, no additional peripheral structure is required that is determined by the sign of the number being multiplied. The FerroFET-based product evaluation has been done by implementing the full design through spice simulation.

This cell structure allows *in situ* analog computation of multiply and accumulate (with both positive and negative operands) in the memory array itself.

### B. CORE ARCHITECTURE

Fig. 6 shows the block diagram for the entire core and provides the detailed structure of the FerroFET memory array. Cores can be divided into three major blocks: 1) the FerroFET memory array that computes vector dot product (sum of products); 2) peripheral blocks; and 3) the communication block. The memory array and the peripheral blocks together form the compute unit. Each core has a maximum of eight compute units corresponding to each neighbor. The details of the architecture and the subblocks are shown as a part of the supplementary material. Here, we discuss the salient features only.

#### 1) FerroFET MEMORY ARRAY STRUCTURE

The hierarchy of the FerroFET memory array has been shown in detail in Fig. 6. In each iteration, the memory array performs matrix–vector product of $B$ and $z$ using a pseudo-crossbar architecture.

**TABLE 1.** **Specifications of baseline Von-Neumann architecture in 28-nm CMOS process.**

| Parameter | Value |
|---|---|
| Simulation Mode | Syscall Emulation |
| CPU Type | DerivO3CPU |
| CPU Width | 3 |
| L1 Inst. Cache Size | 64kB |
| L1 Data Cache Size | 64kB |
| L2 Cache Size | 2MB |
| Main Memory | 32GB DDR4 |

#### 2) PERIPHERAL BLOCKS

The current summing FerroFET subarrays have per-column ADCs to digitize the summation of the inner products. The peripheral blocks include, shift plus add (S + A) arrays, adders to collect the output of each compute unit, followed by a subtraction block. Once these blocks finish their operation the $z$-coefficients are computed and sent to the communication blocks. Each core receives inputs from the neighboring cores. Digital to analog converters (DACs) produce voltage signals corresponding to a digital value of $z$-coefficients and these voltages are asserted on bit-lines (BL1, BL2) of the memory array.

#### 3) COMMUNICATION UNIT

Communication between cores is done through an asynchronous mechanism. In this design, a four-phase handshake protocol has been used because of reduced logical complexity and competitive power and area efficiency when compared with respect to a two-phase protocol. The details of the protocol have been discussed in the Supplementary material.

### C. SYSTEM ARCHITECTURE

The proposed architecture comprises of eight rows with eight cores in each. The entire design is synthesized in the 28-nm CMOS process. To simulate and obtain latency and power estimations for the baseline Von-Neumann architecture, we used the gem5 simulator [28] and McPAT [29]. Table 1 shows the system specifications for the gem5 simulator. For each iteration of the baseline Von-Neumann architecture, we collect a set of workload statistics. The system configuration and the data for a single iteration are then run through McPAT to obtain power estimations.

Simultaneously, we construct an SRAM PIM to compare its performance with the proposed FerroFET-based PIM architecture. In this design, we use a single read and write ports and peripheral adders and multipliers to design a compute unit. The structure of cores in the SRAM PIM is identical to that of the FerroFET PIM. The SRAM PIM prototype also consists of 64 cores.

## V. DESIGN SPACE EXPLORATION

Fig. 7(a) illustrates how the average normalized error changes with respect to the number of iterations for a varying number of bits per FerroFET cell. The average normalized error

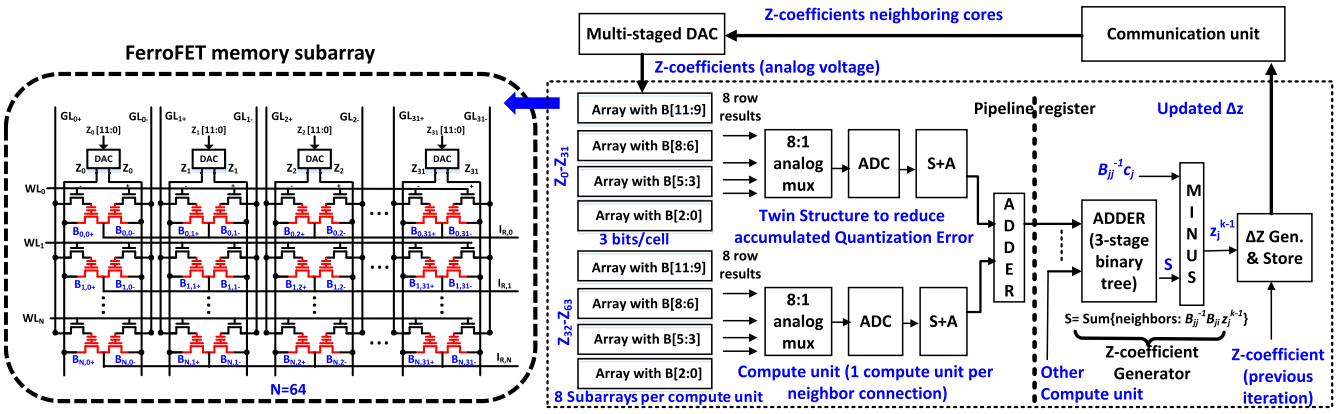


**FIGURE 6.** Schematic of a typical core.

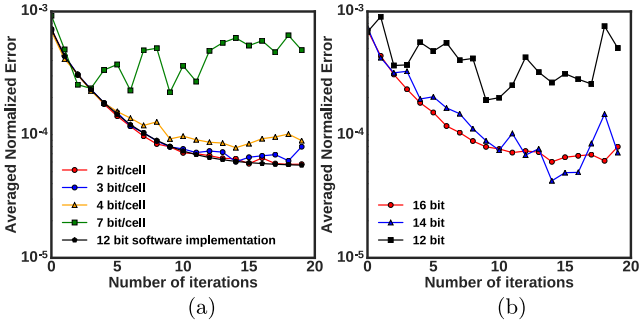


**FIGURE 7.** **(a) Average normalized error in signal reconstruction via distributed least-squares method as a function of the number of bits/cell of FerroFET. The ADC bit resolution is fixed to 16. (b) Average normalized error of $Z$ in nonuniform sampling algorithm with respect to different ADC bit resolution.**

is defined as the L2 norm of the difference of $Z$ between the proposed architecture and a corresponding floating-point architecture. In our design, we use 2/3/4/7 bits/cell to store 12 bits (excluding sign bit) of the fixed point (6 bits for integer and 6 bits for the decimal). For example, the range corresponding to 2 bits with the sign bit, that is, $[-4, 3]$ is represented by 3 bits/cell (due to the cell architecture). In our design the default ADC resolution is 16 bits; and we also study the effect of 16-bit data converters on the design. We use the linear part of the FerroFET's conductance, as discussed above.

We observe that the average normalized error increases as the number of bits/cell increases as shown in Fig. 7(a). This is attributed to the fact that the use of a larger number of bits/cell requires higher ADC and DAC bit resolution to maintain precision. average normalized error from 7 bits/cell FerroFET array is much larger than 2, 3, 4 bit/cell FerroFET array mainly due to the loss of precision during data conversion. A higher resolution from the data converters beyond 16 bits requires noise-shaping and advanced architectures that are not amenable for low-power designs.

In order to quantify the effect of the finite resolution of the ADC/DAC on the fidelity of the final results, we plot the average normalized error of $Z$ in Fig. 7(b). Three cases corresponding to the ADC/DAC resolution of 12, 14, and 16 bits are studied. Here the number of bits per FerroFET cell is considered to be 3. We observe that an ADC/DAC of 14-bit resolution results in convergence, whereas the quantization offered resulting for a 12-bit ADC/DAC is unacceptable. This leads to the design point where 14-bit ADC/DACs are used in the peripherals.

So far, we have studied the effect of the peripheral circuits and storage architecture on the convergence of the optimization algorithm. FerroFETs, in spite of their multi-state storage capability, suffer from inherent nonlinearities where the conductance does not change linearly with the number of pulses. We analyze the effect of this nonlinearity in conductance on the average normalized error of $Z$ in Fig. 9. The nonlinearity in conductance of FerroFET is modeled as a normalized sigmoid function

$$G(x) = \frac{\beta e^{\alpha x}}{1 + e^{\alpha x}} + G_{\min}, \beta = G_{\max} - G_{\min} \tag{7}$$

where $G_{\max}$ and $G_{\min}$ are the maximum and minimum conductance values, $\alpha$ is an empirically derived parameter. This is in contrast to the convex/concave functions that have been used in [30], [25], and [31] to model nonlinearity. We note that in the case of FerroFETs, the sigmoidal function is 1) a better fit and 2) physically meaningful. The sigmoidal conductance response manifests from the approximately Gaussian distribution of coercive fields among individual domains within the ferroelectric. Therefore, an amplitude-modulated pulse scheme, which in essence integrates across the domain distribution is expected to produce sigmoidal characteristics.

Fig. 8(a) presents the experimental data of device-to-device $V_{th}$ variation among 40 FerroFETs and the maximum variation in $V_{th}$ is 30%. More detailed experimental data are shown in [33]. Like other algorithms, we note that increasing variations will increase the error in computation. Fig. 8(b) shows the averaged normalized error of the algorithm with respect to % of random variation on $V_{th}$ after 20 iterations of the algorithm. As expected, the error increases as the number of bits of storage per cell increases.

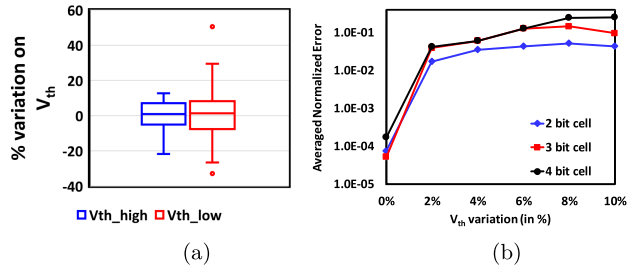


**FIGURE 8.** **(a) Experimental data of device-to-device $V_{th}$ variation over 40 FerroFET devices. $V_{th}$ high/low means the variation on the maximum/minimum $V_{th}$. (b) Averaged normalized error of the algorithm with respect to % random variation on $V_{th}$ of FerroFET.**

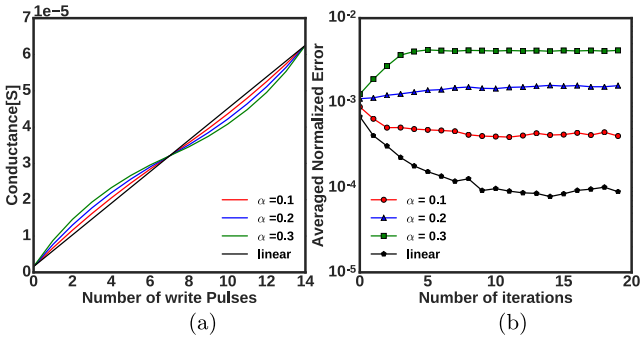


**FIGURE 9.** **(a) Nonlinear conductance of 4 bit/cell FerroFET. (b) Average normalized error of as a function of the nonlinear conductance of FerroFETs (4 bits/cell FerroFET and 16-bit ADCs are considered).**

Fig. 9(a) shows the nonlinear conductance of FerroFET as a function of the number of write pulses and (b) shows how nonlinearity in conductance affects the average normalized error. In this design, the number of bits per FerroFET cell is assumed to be 3. It is shown that if $\alpha$ is greater than 0.1, the average normalized error increases as the number of iterations progresses. This illustrates that the use of FerroFETs in optimizations for PIM architectures require linear changes in conductance during potentiation and depression. In [30], the authors have shown that when resistive processing units (RPUs) are used in crosspoint architectures for solving inference in deep neural network architectures, the resistive units need high degrees of linearity. We arrive at a similar conclusion when such resistive elements are used in solving optimization problems. This motivates further research in the device community to address the issue of nonlinearity when PIM architectures are used for solving linear-algebraic problems.

We study the effect of the design space on critical system parameters such as compute time, energy, power, and area. The number of bits that can be stored in a FerroFET decides the FerroFET array size. Our baseline design uses a cell with 4 bits/cell. We also consider the case of 5 bits/cell where we need 64 $\times$ 256 memory cells (eight subarrays of 64 $\times$ 32 dimension) to store all the $B$-coefficients. As we decrease the number of bits/cell, the total number of memory cells required increases. For example, a design with 3 bits/cell

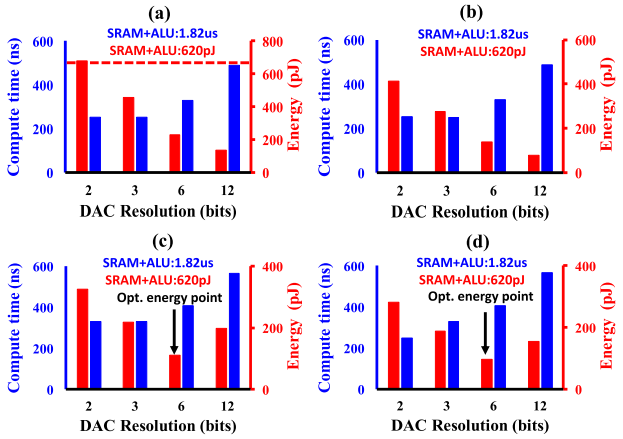


**FIGURE 10.** **Compute time and energy behavior of the compute unit versus DAC resolution for the parallel-computation approach and storage per FerroFET memory cell is (a) 2 bit/cell, (b) 3 bits/cell, (c) 4 bits/cell, and (d) 5 bits/cell.**

requires a total memory size of 64 $\times$ 384 cells (12 subarrays of 64 $\times$ 32 cells per subarray), and so on.

Similarly, the DAC resolution also affects the compute unit area and other critical metrics. In this architecture, the multi-stage DAC resolution can be configured to 2, 3, 6, and 12 bits. The main role of the DAC is to provide analog values of the $z$-coefficients, which are represented in a 12-bit fixed-point format. As we reduce the DAC resolution, there are two options that can be pursued in the design: 1) duplicate the subarrays to compute in parallel and maintain the compute time at the expense of area overhead and 2) perform the computations sequentially. The sequential computation can be explained by the following simple example. For a 6-bit DAC, we first evaluate the sum with six LSB bits of all the $z$-coefficients, and in the next cycle, we evaluate the sum with the six MSB bits for all $z$-coefficients and eventually add them with appropriate scales using S + A blocks. We define the first approach as parallel computation which results in higher throughput but lower area efficiency and the second approach as sequential computation, which consumes the lower area at the cost of lower throughput. Another important fact to note is that decreasing the number of bits/cell or the DAC resolution reduces the dynamic range of the read current out of SL lines resulting in simpler peripheral design. In our case studies, we have optimized the read peripheral circuits and ADCs based on the DAC configuration [32].

Figs. 10 and 11 illustrate the compute time and energy as the DAC resolution and number of bits/cell are varied for the parallel-computation and sequential-computation approaches, respectively. It can be clearly seen from the two figures that in case of a sequential approach, the computation time is 2–3$\times$ higher than in the parallel-computation approach. For parallel computation [Fig. 10(a)–(d)], we observe a trend that the compute time goes up as the DAC resolution increases. This is because the ADC starts to dominate the system latency. As we increase the DAC resolution, to maintain the same quantization error

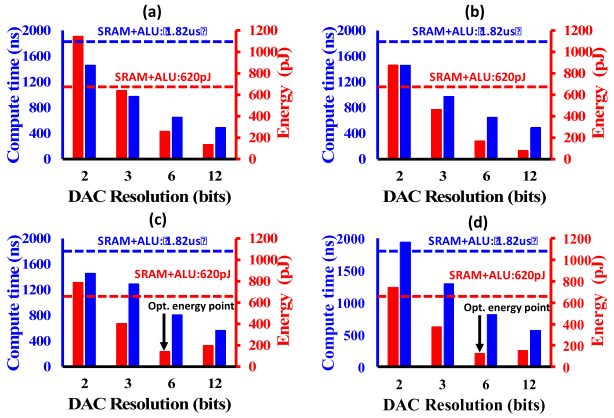


**FIGURE 11.** Compute time and energy behavior of the compute unit versus DAC resolution for the sequential-computation approach and storage per FerroFET memory cell is (a) 2 bit/cell, (b) 3 bits/cell, (c) 4 bits/cell, and (d) 5 bits/cell.

for the read current a higher resolution ADC is required and ADC latency increases super-linearly as the resolution increases. In Fig. 10(a) and (b), a monotonic decrease in energy is noted as the DAC resolution increases. This is because for both cases, the parallel memory array and associated peripheral hardware overhead is the dominant factor, which decreases as the DAC resolution increases and eventually causes a reduction in the overall energy consumed. However, for Fig. 10(c) and (d) that have higher bits/cell (4 and 5 bits respectively) the ADC overhead starts to be significant. As mentioned before, as the DAC resolution for these two cases increases, we have to switch to a higher resolution ADC that adds to the energy consumed and off-sets the improvement due to the reduction of the parallel subarrays and adders.

Fig. 11 exhibits an increasing trend of compute time as the DAC resolution and bits/cell decrease. With less bits/cell and DAC resolution, it results in multiple iterations of compute cycle since the number of subarrays is fixed. Due to the energy tradeoff between peripheral units and the ADC (discussed above), the trend for energy dissipation is similar to Fig. 10. Also, it can be noted that the sequential approach consumes higher energy than the parallel approach due to the multiple iterations that are required. The comparison with an SRAM PIM structure has been shown using a dotted line in each of the histograms. The proposed design outperforms SRAM PIM structure in terms of compute time and energy for the majority of design cases, as has been shown.

Figs. 12 and 13 present the latency and energy breakdown of each block in the computation and communication units. In Fig. 7, we present the analysis of the averaged normalized error of the nonuniform sampling algorithm with respect to the ADC bit resolution and the number of bits that a single FerroFET cell can store. Based on this analysis, the normalized error is minimized when the ADC bit resolution is $\geq$14 bits and number of bits per FerroFET cell is $\geq$3. With the same system configuration as shown in Fig. 6, a 12-bit DAC, a 14-bit ADC and 3 bits/cell, we calculated the latency

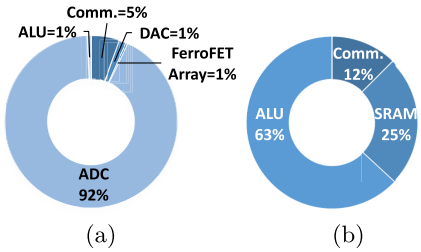


**FIGURE 12.** (a) Latency breakdown of the compute unit and communication channel (comm.) of FerroFET-based PIM. (b) Latency breakdown of the compute unit and communication channel (comm.) of SRAM + ALU PIM.

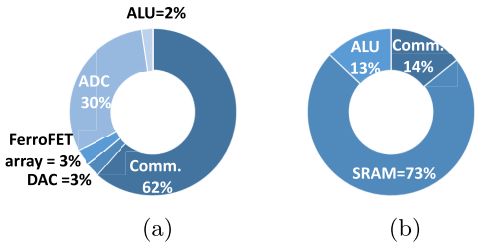


**FIGURE 13.** (a) Energy breakdown of the compute unit and communication channel (comm.) of FerroFET-based PIM. (b) Energy breakdown of the compute unit and communication channel (comm.) of SRAM + ALU PIM.

and energy breakdown of the FerroFET-based compute unit and communication as shown in Figs. 12(a) and 13(a). Figs. 12(b) and 13(b) show the latency and energy breakdown of SRAM + arithmetic logic unit (ALU) PIM, where SRAM is used as a storage and all computation is handled in multipliers and adders. Instead of DAC and ADC, SRAM+ALU PIM core has multipliers and adders and the memory size is 6 KB. "SRAM" in Figs. 12 and 13 note the SRAM with its peripheral. From Fig. 12(a), the block that takes the most latency is 14-bit ADC, which has 92% of the total latency [32]. In case of SRAM+ALU PIM, the computation in ALU takes the 63% of the total latency. In Fig. 13(a), communication between the neighboring cores dissipate 62% of the total energy since we use a four-phase handshaking mechanism with Muller-C elements (details in supplement material) whose clock frequency is 1 GHz. In Fig. 13(b), SRAM and its peripherals dissipate the most amount of power because the SRAM size expanded three times compared to the size of FerroFET cells to store all elements of $B$ coefficients from (4).

Fig. 14 shows the total power of the computation unit when the number of bits/cell and DAC resolution are varied for the parallel and sequential cases. From both Fig. 14(a) and (b) we observe that power consumption reduces as we increase either the number of bits/cell or the DAC resolution. From this, we conclude that the total power consumed is determined by both the memory subarrays and peripheral logic. As the number of bits/cell or the DAC resolution increase, we observe a reduction in number of S + A array stages and memory subarrays, and this reduction causes an overall reduction in

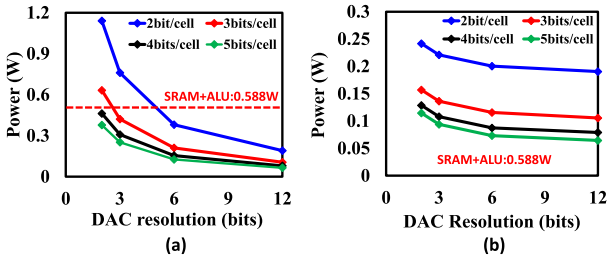


**FIGURE 14.** Power consumption of the compute unit when bits/memory cell and DAC resolution are varied for (a) parallel computation and (b) sequential computation.

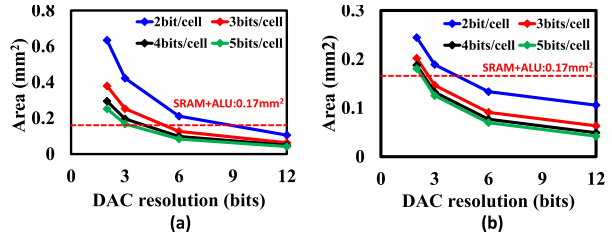


**FIGURE 15.** Estimated area of the compute unit when bits/memory cell and DAC resolution are varied for (a) parallel computation and (b) sequential computation.

**TABLE 2.** Compute time and energy comparison in different architectures.

| Architecture | Baseline | SRAM PIM | FerroFET PIM | Performance |
|---|---|---|---|---|
| Compute Time[s] | $83\mu$ | $1.83\mu$ | $0.57\mu$ | 3x wrt SRAM PIM |
| Energy[J] | $1.36m$ | $460\mu$ | $21\mu$ | 21x wrt SRAM PIM |

power. Further when Fig. 14(a) and (b) are compared to each other the parallel computation approach consumes higher power because of the additional memory array and associated peripheral hardware requirements.

Fig. 15 shows the total area of the computation unit when the number of bits/cell and the DAC resolution are varied for the parallel and sequential cases. For the parallel computation approach [Fig. 15(a)], the area is larger than in the sequential approach [Fig. 15(b)] since the computations are executed in parallel with a higher number of memory subarrays and peripheral blocks. As the DAC resolution and the number of bits/cell increase the total area increases because the memory subarray, S + A and multistage adders required are lesser in number, and they dominate any increase caused by the ADC area. For all the figures the dotted lines show the performance of a corresponding SRAM + ALU Von-Neumann architecture (baseline).

Table 2 presents the architectural results of compute time and energy for the baseline, SRAM PIM and FerroFET PIM architectures of 64 cores. FerroFET PIM shows 3× improvement in compute time and 21× improvements in energy efficiency compared to SRAM PIM.

## VI. APPLICATIONS

As examples of prototypical problems that can be solved using the proposed algorithm and architecture, we present two applications: 1) signal reconstruction from 1-D EEG signals and 2) recovery of CT Images used in medical imaging.

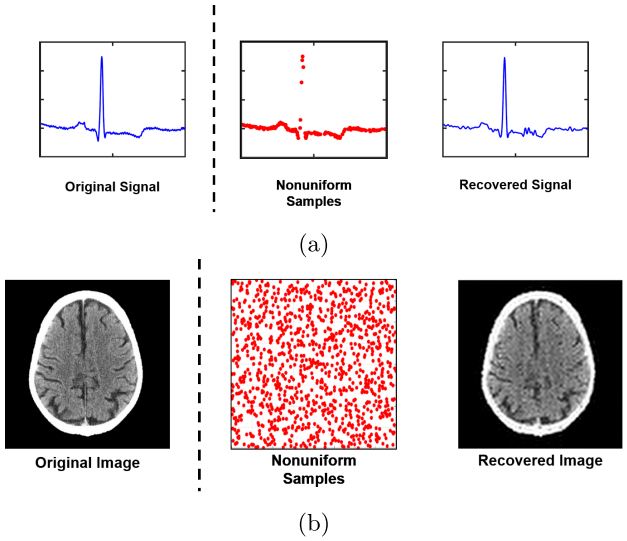


**FIGURE 16.** Reconstruction steps. (a) 1-D example: recovery of EEG signal profile. (b) 2-D example: brain computed topography recovery [20].

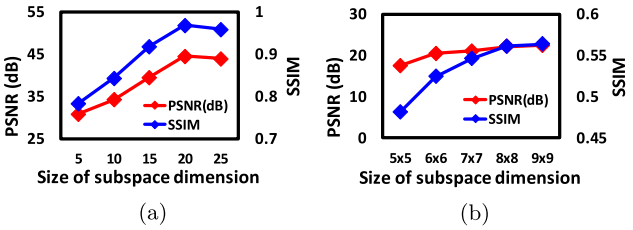


**FIGURE 17.** PSNR and SSIM. (a) 1-D example: recovery of a nonuniformly sampled 1-D signal from an EEG probe. (b) 2-D example: recovery of a sampled image from the CT scan of a brain.

Typical examples have been shown in Fig. 16(a) and (b). Both the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are shown in Fig. 17. We note that increasing the subspace dimension increases the fidelity of the reconstruction process. This justifies the use of a subspace dimension of 8 × 8 for the current applications in hand. It also shows the power of iterative algorithms in systolic PIM architectures for solving distributed convex optimization.

## VII. CONCLUSION

This paper presents a systolic PIM architecture based on analog FerroFet pseudo-crosspoint arrays with *in situ* computation to enable distributed convex optimization via least square minimization. Key contributions of the paper are as follows.

1) A FerroFET-based differential cell can compute matrix multiplication of both positive and negative numbers.
2) A FerroFET-based PIM architecture for solving a least squares minimization.
3) Development of a complete end-to-end tool chain and demonstration of 21× in energy efficiency and 3× in compute time compared to an SRAM-based PIM architecture.

We demonstrate that cross-bar resistive architectures are not only capable of accelerating machine-learning algorithms, but also distributed optimization in a systolic array.

## REFERENCES


[1] M. Gokhale, B. Holmes, and K. Iobst, "Processing in memory: The Terasys massively parallel PIM array," *Computer*, vol. 28, no. 4, pp. 23–31, Apr. 1995.

[2] J. Draper *et al.*, "The architecture of the DIVA processing-in-memory chip," in *Proc. 16th Int. Conf. Supercomput. (ICS)*, New York, NY, USA, 2002, pp. 14–25.

[3] J. Suh, E.-G. Kim, S. P. Crago, L. Srinivasan, and M. C. French, "A performance analysis of PIM, stream processing, and tiled processing on memory-intensive signal processing kernels," in *Proc. 30th Annu. Int. Symp. Comput. Archit.*, Jun. 2003, pp. 410–419.

[4] M. Hall *et al.*, "Mapping irregular applications to DIVA, a PIM-based data-intensive architecture," in *Proc. ACM/IEEE Conf. Supercomput.*, Nov. 1999, p. 57.

[5] X. Sun, P. Wang, K. Ni, S. Datta, and S. Yu, "Exploiting hybrid precision for training and inference: A 2T-1FeFET based analog synaptic weight cell," in *IEDM Tech. Dig.*, Dec. 2018, pp. 3.1.1–3.1.4.

[6] S. K. Gonugondla, M. Kang, and N. Shanbhag, "A 42pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training," in *IEEE Int. Solid-State Circuits Conf. ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 490–492.

[7] P. Chi *et al.*, "PRIME: A novel processing-in-memory architecture for neural network computation in reram-based main memory," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 27–39.

[8] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with *in-situ* analog arithmetic in crossbars," in *Proc. 43rd Int. Annu. Symp. Comput. Archit.*, Jun. 2016, pp. 14–26.

[9] Z. Chen *et al.*, "Optimized learning scheme for grayscale image recognition in a RRAM based analog neuromorphic system," in *IEDM Tech. Dig.*, Dec. 2015, pp. 17.7.1–17.7.4.

[10] Y. Kim, Y. Zhang, and P. Li, "A reconfigurable digital neuromorphic processor with memristive synaptic crossbar for cognitive computing," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 11, no. 4, p. 38, Apr. 2015.

[11] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, pp. 61–64, May 2015.

[12] G. W. Burr *et al.*, "Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)," in *IEDM Tech. Dig.*, Dec. 2015, pp. 4.4.1–4.4.4.

[13] D. Fan, Y. Shim, A. Raghunathan, and K. Roy, "STT-SNN: A spin-transfer-torque based soft-limiting non-linear neuron for low-power artificial neural networks," *IEEE Trans. Nanotechnol.*, vol. 14, no. 6, pp. 1013–1023, Nov. 2015.

[14] P. J. S. G. Ferreira, "The stability of a procedure for the recovery of lost samples in band-limited signals," *Signal Process.*, vol. 40, nos. 2–3, pp. 195–205, 1994.

[15] R. Marks, "Restoring lost samples from an oversampled band-limited signal," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 3, pp. 752–755, Jun. 1983.

[16] H. Stark, "Polar, spiral, and generalized sampling and interpolation," in *Advanced Topics in Shannon Sampling and Interpolation Theory*. New York, NY, USA: Springer, 1993, pp. 185–218.

[17] J. Müller *et al.*, "Ferroelectricity in $HfO_2$ enables nonvolatile data storage in 28 nm HKMG," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. 25–26.

[18] S. Müeller, S. Slesazeck, T. Mikolajick, J. Müller, P. Polakowski, and S. Flachowsky, "Next-generation ferroelectric memories based on FE-$HfO_2$," in *Proc. Joint IEEE Int. Symp. Appl. Ferroelectric (ISAF), Int. Symp. Integr. Functionalities (ISIF), Piezoelectr. Force Microsc. Workshop (PFM)*, May 2015, pp. 233–236.

[19] S. Oh *et al.*, "$Hfzro_x$-based ferroelectric synapse device with 32 levels of conductance states for neuromorphic applications," *IEEE Electron Device Lett.*, vol. 38, no. 6, pp. 732–735, Jun. 2017.

[20] I. Yoon *et al.*, "A FeFET based processing-in-memory architecture for solving distributed least-square optimizations," in *Proc. 76th Device Res. Conf. (DRC)*, Jun. 2018, pp. 1–2.

[21] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using $AlO_x/HfO_2$ bilayer RRAM array for neuromorphic systems," *Electron Device Lett.*, vol. 37, no. 8, pp. 994–997, Aug. 2016.

[22] S. Park *et al.*, "Neuromorphic speech systems using advanced ReRam-based synapse," in *IEDM Tech. Dig.*, Dec. 2013, pp. 25.6.1–25.6.4.

[23] H. Mulaosmanovic *et al.*, "Evidence of single domain switching in hafnium oxide based FeFETs: Enabler for multi-level FeFET memory cells," in *IEDM Tech. Dig.*, Dec. 2015, pp. 26.8.1–26.8.3.

[24] K. Ni *et al.*, "In-memory computing primitive for sensor data fusion in 28 nm HKMG FeFET technology," in *IEDM Tech. Dig.*, Dec. 2018, pp. 16.1.1–16.1.4.

[25] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *IEDM Tech. Dig.*, Dec. 2015, pp. 17.3.1–17.3.4.

[26] M. Hu, H. Li, Q. Wu, G. S. Rose, and Y. Chen, "Memristor crossbar based hardware realization of BSB recall function," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2012, pp. 1–7.

[27] B. Li, Y. Wang, Y. Wang, Y. Chen, and H. Yang, "Training itself: Mixed-signal training acceleration for memristor-based neural network," in *Proc. 19th Asia South Pacific Design Autom. Conf.*, Jan. 2014, pp. 361–366.

[28] N. Binkert *et al.*, "The gem5 simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, 2011.

[29] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, New York, NY, USA, Dec. 2009, pp. 469–480.

[30] T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations," *Frontiers Neurosci.*, vol. 10, p. 333, Jul. 2016.

[31] M. Jerry *et al.*, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *IEDM Tech. Dig.*, Dec. 2017, pp. 6.2.1–6.2.4.

[32] B. Murmann. *ADC Performance Survey 1997–2017*. Accessed: Dec. 2017. [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html

[33] K. Ni, W. Chakraborty, J. Smith, B. Grisafe, and S. Datta, "Fundamental understanding and control of device-to-device variation in deeply scaled ferroelectric FETs," in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, 2019, pp. T40–T41. doi: 10.23919/VLSIT.2019.8776497.



**INSIK YOON** (S'17) received the B.S. and M.S. degrees from Carnegie Mellon University, Pittsburgh, PA, USA, in 2009 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Georgia Institute of Technology, Atlanta, GA, USA.

From 2010 to 2015, he was with Memory and Display Interface Design, TLi and SK Hynix, Icheon, South Korea. His current research interests include emerging memory technologies and in-memory computing for machine learning acceleration.

**MUYA CHANG** is currently a dual-degree Graduate Student with Georgia Institute of Technology, Atlanta, GA, USA, where he is pursuing the Ph.D. degree in electrical and computer engineering (ECE) and the M.S. degree in computer science.

He is also a member of the Integrated Circuits and Systems Research Laboratory and is advised by ECE Associate Professor A. Raychowdhury. His current research interests include energy-efficient hardware design for distributed optimizations.

**KAI NI** (S'13–M'16) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2011, and the Ph.D. degree in electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2016, with a focus on characterization, modeling, and reliability of III–V MOSFETs.

Since then, he became a Postdoctoral Associate at the University of Notre Dame, Notre Dame, IN, USA, working on novel memory technologies and computing paradigms. He will be an Assistant Professor of microsystems engineering at the Rochester Institute of Technology, Rochester, NY, USA. His current interests include nanoelectronic devices empowering revolutionary transformation in artificial intelligence accelerator design, unconventional computing, security, and 3-D memory technology.

**MATTHEW JERRY** received the B.S. degree in physics from the Department of Physics and Astronomy, University of Delaware, Newark, DE, USA, in 2013, and the Ph.D. degree in electrical engineering from the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA, in 2018.

He was a Research Assistant with the Nanoelectronic Devices and Circuits Laboratory, University of Notre Dame. He is currently an Emerging Memory Engineer with Micron Technologies, Boise, ID, USA. His current research interests include the characterization and design of emerging solid-state circuits and devices based on transition metal oxide and ferroelectric materials.

**SAMANTAK GANGOPADHYAY** (S'13) received the B.Tech. and M.Tech. degrees in electronics and electrical communication engineering with a specialization in microelectronics and VLSI design from the IIT Kharagpur, Kharagpur, India, in May 2009, and the Ph.D. degree from the Integrated Circuits and Systems Research Laboratory, Georgia Institute of Technology, Atlanta, GA, USA, in 2017.

After graduation, he joined IBM India as a Physical Design R&D Engineer to work on the power series and Z-mainframe microprocessors. In this role, he was responsible for the implementation of synthesizable circuits while satisfying timing, power, noise, electromigration, design-for-test, and design-for-manufacturing specifications. His current research interests include the design of low-power digital circuits, low-dropout voltage regulators, and power management for wide dynamic range computation.

**GUS HENRY SMITH** received the M.S. degree in processing in memory from Pennsylvania State University, State College, PA, USA, under the supervision of Dr. V. Narayanan. He is currently pursuing the Ph.D. degree with the Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA, with a focus on the intersection of programming languages and computer architecture, under the supervision of Dr. L. Ceze.

**TOMER HAMAM** (S'12–M'13) received the B.Sc. degree in electrical engineering from Technion—Israel Institute of Technology, Haifa, Israel, in 2012. He is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

His current research interests include the intersection of optimization, signal processing, and machine learning. His current work focuses on the means to use structure in locally coupled problems to advance communication and computation performance in distributed optimization settings.

**JUSTIN ROMBERG** (S'96–M'03–SM'10–F'18) received the B.S.E.E., M.S., and Ph.D. degrees from Rice University, Houston, TX, USA, in 1997, 1999, and 2004, respectively.

From 2003 to 2006, he was a Postdoctoral Scholar in applied and computational mathematics with the California Institute of Technology, Pasadena, CA, USA. He is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, where he has been with the faculty since 2006.

**VIJAYKRISHNAN NARAYANAN** (S'97–A'98–M'03–SM'08–F'11) is currently the Robert Noll Chair Professor of Computer Science and Engineering and Electrical Engineering with Pennsylvania State University, State College, PA, USA. He also leads the National Science Foundation Expeditions in Computing Center. He is also a Thrust-Leader of the DARPA/SRC JUMP Center for Brain Inspired Computing. He has mentored more than 100 graduate students and published more than 400 papers. His current research interests include embedded systems, computer architecture, and design automation.

Dr. Narayanan is a fellow of the Association of Computing Machinery.

**ASIF KHAN** (M'15) received the B.S. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2007, and the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 2015.

He is currently an Assistant Professor with the Georgia Institute of Technology, Atlanta, GA, USA.

**SUMAN DATTA** (S'98–M'99–SM'06–F'13) was with the Advanced Transistor Group, Intel Corporation, Hillsboro, OR, USA, from 1999 to 2007, where he developed several generations of high-performance logic transistor technologies, including high-*k*/metal gate, tri-gate, and non-silicon channel CMOS transistors. His research group focuses on emerging device concepts that support and enable new computational models. He was a Professor of electrical engineering with the Pennsylvania State University, University Park, PA, USA, from 2007 to 2011. He is currently the Frank M. Freimann Chair Professor of engineering with the University of Notre Dame, Notre Dame, IN, USA. He is also the Director of a multi-university advanced microelectronics research center, the ASCENT, funded by the Semiconductor Research Corporation and the Defense Advanced Research Projects Agency. He has published over 300 journals and refereed conference papers and holds 175 patents related to advanced semiconductors.

Dr. Datta is a fellow of the National Academy of Inventors. He was a recipient of the Intel Logic Technology Quality Award in 2002, the Intel Achievement Award in 2003, the IEEE Device Research Conference Best Paper Awards in 2010 and 2011, the Penn State Engineering Alumni Association (PSEAS) Outstanding Research Award in 2012, the SEMI Award for North America in 2012, and the PSEAS Premier Research Award in 2015.

**ARIJIT RAYCHOWDHURY** (M'08–SM'13) received the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2007.

He joined Georgia Institute of Technology, Atlanta, GA, USA, in 2013. His industry experience includes five years as a Staff Scientist with the Circuits Research Laboratory, Intel Corporation, Hillsboro, OR, USA, and a year as an Analog Circuit Researcher with Texas Instruments, Inc. He is currently an Associate Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, where he also holds the ON Semiconductor Junior Professorship. He holds more than 25 U.S. and international patents and has published over 150 articles in journals and refereed conferences. His current research interests include low-power digital and mixed-signal circuit design, device circuit interactions, and novel computing models and hardware realizations.

Dr. Raychowdhury received the Best Thesis Award from the College of Engineering, Purdue University, in 2007, the Dimitris N. Chorafas Award for Outstanding Doctoral Research in 2007, the Intel Labs Technical Contribution Award in 2011, the NSF CISE Research Initiation Initiative Award in 2015, the Intel Early Faculty Award in 2015, the IEEE DAC Innovator under 40 Award in 2018, Georgia Tech's Outstanding Young Faculty Award in 2018, and multiple best paper awards and fellowships.