A FerroFET based in-memory processor for solving distributed and iterative optimizations via least-squares method

Insik Yoon¹, Muya Chang¹, Kai Ni², Matthew Jerry², Samantak Gangopadhyay¹, Gus Henry Smith³, Tomer Hamam¹(Student Member, IEEE), Justin Romberg¹(Fellow, IEEE), Vijaykrishnan Narayanan³(Fellow, IEEE), Asif Khan¹(Member, IEEE), Suman Datta²(Fellow, IEEE), Arijit Raychowdhury¹(Senior Member, IEEE)

¹Georgia Institute of Technology, Atlanta, GA, 30332, USA
²University of Notre Dame, South Bend, IN, 46556, USA
³Penn State University, State College, PA, 16801, USA

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 $21\times$ in energy efficiency and $3\times$ in compute time compared to an SRAM based Processing-In-Memory (PIM) architecture.

Index Terms—In-memory processing, FerroFets, Emerging, Post-CMOS, Optimization, Distributed Computing, Least Square, Hardware,

I. INTRODUCTION

Modern computing systems based on the Von-Neumann architecture rely on a clear distinction between logic and memory, and processes information by executing a sequence of precise atomic instructions with periodic uploads to the memory. Such systems are the foundation of the digital revolution which began with the demonstration of the self-aligned planar-gate silicon MOSFET in the sixties and was accelerated by rapid advances in transistor technology. However, in the last one 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 requires 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 sub-array 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 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 a iterative dynamical system using a systolic 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 non-volatile memories (eNVM). Among all competing eNVM technologies, FerroFETs have emerged as promising candidates due to their compact size, multi-level storage, nano-second read-write and high energy-efficiency. We demonstrate that a systolic PIM architecture, using FerroFET pseudo-cross-point array can solve least squares minimization with $21\times$ improvement in energy-efficiency compared to 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 wide-spread applications in discrete signal processing. In particular, it is a common tool for signal reconstruction where the process of sampling is non-uniform [14], [15] such as in Computerized tomography (CT), magnetic resonance imaging (MRI) [16], radar signal processing, LIDAR systems etc. Consider (1) $u$ and $v$ are the horizontal and the vertical arguments of a continuous signal. (2) $x$ and $y$ are the discrete coordinate

![Fig. 1: (a) 2D continuous function $f(u,v)$ with non-uniform samples. (b) Spatial location of the non-uniform samples.](image-url)
indexes. (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 non-uniformly sampled and are stored in vector $b$, which are referred to as $f(x, y)$. The objective is to use the non-uniform samples to obtain complete reconstruction of $f(u, v)$ in $N_x \times N_y$ dimensional subspace. Fig. 1 shows an example of $f(u, v)$ and the results of non-uniform sampling. In this algorithm, we assume that $f(u, v)$ lies in an $N_x \times N_y$ dimensional subspace. To reconstruct the signal accurately we have used 2D 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 and $[k_x, k_y]$ represent a specific frame, $\omega_x$ and $\omega_y$ indicate the harmonic in horizontal and vertical directions.

$$\psi_{k_x, \omega_x, k_y, \omega_y}(u, v) = \sqrt{2} \cdot g(u - k_x, v - k_y) \cdot \cos((\omega_x + \frac{1}{2})\pi(u - k_x)) \cdot \cos((\omega_y + \frac{1}{2})\pi(v - k_y))$$  \hspace{1cm} (1)

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

$$f(u, v) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k_x=1}^{K_x} \sum_{k_y=1}^{K_y} \alpha(k_x, \omega_x, k_y, \omega_y) \psi_{k_x, \omega_x, k_y, \omega_y}(u, v)$$  \hspace{1cm} (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

$$Az = b$$  \hspace{1cm} (3)

Here $b$ is the sample vector, $z$ is the coefficient vector obtained by stacking the coefficients $\alpha(k_x, \omega_x, k_y, \omega_y)$, and $A$ is referred to as the Grammian (Gram) matrix of the basis. When the size of $A$ 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 $B = A^T A$ and $c = A^T b$.

$$z_j^k = B^{-1}_{jj}(c_j - \sum_{i \neq j} B_{ij}z_i^{k-1})$$  \hspace{1cm} (4)

Some observations are worth emphasizing: (1) To update $z_j^k$, only values from previous iterations are needed. (2) Columns of $A$ are coupled only with neighboring frames, which leads to simpler computation of $B_{ij}$. 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 sub-array – as will be described in the following sections.

**Fig. 2:** (a) – (d) show different FerroFET states, corresponding to different portions of ferroelectric domain switching. The yellow arrows indicate the polarization direction. The blue/red circles represent electron/hole, respectively. (e) shows the 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) shows the $G_{DS}$-$V_{GS}$ characteristics after each pulse. (g) shows the measured drain to source conductance as a function of applied pulse number. Here ideal case is presented, which shows linear and symmetrical potentiation and depression. [20]

**III. An 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 Ferroelectric FETs (FerroFETs) have recently received great interest for their application in nonvolatile memory (NVM) [17]. It is CMOS compatible and retains ferroelectricity for thin films with thickness around 10nm. By tuning the portion of switched ferroelectric domain, a FerroFET can exhibit multiple intermediate states, which has been used in neuromorphic computing [18], [19].

**Fig. 3:** (a) simulated FerroFET channel conductance (b) Measured FerroFET channel conductance ($G_{DS}$) as a function of pulse number.

The operation of FerroFET as an 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 a desired state [18], [19]. Various pulse schemes...
are proposed to tune the state, including identical pulse schemes [21], pulse-width modulation schemes [22], and pulse-amplitude modulation schemes [19] [23]. For illustration, Fig. 2 illustrates the operation with pulse-amplitude modulation scheme, which is used in this paper. Fig. 2(c) shows the applied pulse waveform. After each pulse, the percentage of switched ferroelectric domains is modified. The device states are shown in Fig. 2 (a)-(d). The device $I_{DS}$-$V_{GS}$ corresponding to different states are shown in Fig. 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, $G_D$, 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 atomistic simulation of domain dynamics with a drift-diffusion based FET model. The simulation results closely match the experimental data and is 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, 8 rows, with each row containing 8 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 the accuracy of signal reconstruction. From our analysis we identified 8x8 subspace dimensions is 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 analog-to-digital converters (ADCs) with different resolutions and design constraints.
4) For the current design the B-coefficients ($B_{ji}^{-1}$) and z-coefficients ($z_{ji}^{\pm}$) 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.

A. FerroFET cell structure

Fig. 5: FerroFET cell schematic (a) Conceptual (b) Transistor level implementation [20]

Fig 5 shows the schematic diagram 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] [26] [27], the proposed bit-cell allows both positive and negative values for stored values as well as the inputs. During a read operation the 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 SL. The accumulated current corresponding to $\Delta G$ and $\Delta V$ is given by:

$$I_1 = -\Delta V \left( G - \Delta G \right)$$
$$I_2 = \Delta V \left( G + \Delta G \right)$$
$$I = \Delta V \left( -G + \Delta G + G + \Delta G \right) = \Delta V \left( 2\Delta G \right)$$

The weights of B-coefficients are encoded as multiples of $\Delta G$ 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 detail 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 8 compute units corresponding to each neighbor. The details of the architecture and the sub-blocks 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 the Fig. 6. In each iteration, the memory array performs matrix-vector product of B and z using a pseudo-crossbar architecture.

2) Peripheral blocks: The current summing FerroFET subarrays have per-column analog-to-digital converters (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 (DAC) produce voltage signals corresponding to 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 4-phase handshake protocol has been used because of reduced logical complexity and competitive power and area efficiency when compared with respect to a 2-phase protocol. The details of the protocol has been discussed in the Supplementary material.

C. System Architecture

The proposed architecture comprises of 8 rows with 8 cores in each. The entire design is synthesized in the 28nm 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 I 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 single read and write ports and peripheral adders and multipliers to design a compute unit. The structure of cores in the SRAM PIM are 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 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 fixed point (6 bits for integer and 6 bits for the decimal). For example, the range corresponding to 2 bits with sign bit, i.e., [-4,3] is represented by 3bits/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 16b 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 bits, 14 bits 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 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 non-linearities where the conductance does not change linearly with the number of pulses. We analyze the effect of this non-linearity in conductance on the average normalized error of Z in Fig. 9. The non-linearity in conductance of FerroFET is modeled as a normalized sigmoid function.

\[ G(x) = \frac{\beta e^{\alpha x}}{1 + e^{\alpha x}} + G_{\text{min}}, \beta = G_{\text{max}} - G_{\text{min}} \]  

(7)

where \( G_{\text{max}} \) and \( G_{\text{min}} \) 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] [31] to model non-linearity. 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. 9(a) shows the nonlinear conductance of FerroFET as a function of the number of write pulses and (b) shows how non-linearity 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 cross-point 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 non-linearity when PIM architectures are used for solving linear-algebraic problems.

We study the effect of 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 64x256 memory cells (8 subarrays of 64x32 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 requires a total memory size of 64x384 cells (12 subarrays of 64x32 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 (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 6 LSB bits of all the z-coefficients and in the next cycle we evaluate the sum with the 6 MSB bits for all z-coefficients and eventually add them with appropriate scales using shift+add 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 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].

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 is 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.
Fig. 10 and Fig. 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 approach, respectively. It can be clearly seen from the two figures that in case of a sequential approach the computation time is 2-3X higher when compared to the parallel-computation approach. For parallel-computation (Fig. 10a-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 for the read current a higher resolution ADC is required and ADC latency increases super-linearly as the resolution increases. In Fig. 10a and Fig. 10b, 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. 10c and Fig. 10d 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 increase, we have to switch to a higher resolution ADC that adds to the energy consumed and off-sets the improvement due to 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 sub-arrays are fixed. Due to the energy trade-off between peripheral units and the ADC (discussed above), the trend for energy dissipation is similar to Fig. 10. Also it can be noted that 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 majority of design cases, as has been shown.

Fig.12 and Fig.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 non-uniform 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 ≥ 14 bits and number of bits per FerroFET cell is ≥ 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 and energy breakdown.
Fig. 12: Latency breakdown of the compute unit and communication channel of FerroFET based PIM and SRAM+ALU PIM

Fig. 13: Energy breakdown of the compute unit and communication channel of FerroFET based PIM and SRAM+ALU PIM

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

of the FerroFET based compute unit and communication as shown in Fig.12 (a) and Fig.13(a). Fig.12(b) and Fig.13(b) show the latency and energy breakdown of SRAM+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 6KB. ’SRAM’ in Fig.12 and Fig.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 percent of the total latency [32]. In case of SRAM + ALU PIM, the computation in ALU takes the 63 percent of the total latency. In Fig.13(a), communication between the neighboring cores dissipate 62 percent of the total energy since we use a 4-phase hand shaking mechanism with Muller-C elements(details in supplement material) which clock frequency is 1Ghz. In Fig.13(b), SRAM and its peripherals dissipates the most amount of power because the SRAM size expanded 3 times compared to the size of FerroFET cells to store all elements of B coefficients from Eq.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 Fig.14(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 sub-arrays and peripheral logic. As the number of bits/cell or the DAC resolution increase, we observe a reduction in number of shift+add array stages and memory subarrays, and this reduction causes an overall reduction in power. Further when Fig. 14(a) and Fig. 14(b) are compared to each other the parallel computation approach consumes higher power because of 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 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, shift+add and multi-stage 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 II presents the architectural results of compute time and energy for the baseline, SRAM PIM and FerroFET PIM architectures of 64 cores. FerroFET PIM shows 3x improvement in compute time and 21x 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 1D EEG Signals and (2) Recovery of CT Images used in medical imaging.

Typical examples have been shown in Fig. 16(a) and (b). Both the Peak signal-to-noise ratio (PSNR) & Structural similarity (SSIM) are shown in Fig. 17. We note that increasing the sub-space dimension increases the fidelity of the reconstruction process. This justifies the use of a
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2019.2930222, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Collective Electronics (EXCEL), an SRC-NRI Nanoelectronics Research Initiative under Research Task ID 2698.002 and JUMP ASCENT task 2776.044

References


