

FerroElectronics for Edge Intelligence

Ali Keshavarzi

Stanford University

Kai Ni

Rochester Institute of Technology

Wilbert van den Hoek

Leading Edge Research

Suman Datta

University of Notre Dame

Arijit Raychowdhury

Georgia Institute of Technology

Abstract—The future data-centric world demands edge intelligence (EI)—the ability to analyze data locally and to decide on a course of action autonomously. Challenges with Moore’s Law scaling and limitations of von Neumann computing architectures are limiting the performance and energy efficiency of conventional electronics. Promising new discoveries of advanced CMOS-compatible HfO₂-based ferroelectric devices open the door for FerroElectronics; electronics based on ferroelectric building blocks integrated on advanced CMOS technology nodes. It will enable much needed improvement in computing capabilities making EI a reality. In-memory computing in data-flow architectures is at the core of FerroElectronics. This approach will enable building 1000X more compute-energy-efficient small-system AI engines needed for EI. Smart edge intelligent IoT devices enable new applications, for example, micro Drones (uDrones), that demand higher performance to support local embedded intelligence, real-time learning, and autonomy. They will drive the next phase of growth in the semiconductor industry.

■ **INTERNET OF THINGS** (IoT) in its “smart” form is becoming the next driver of the semiconductor industry. We live in a world where huge amounts of data from our physical world around

us are being sensed. These data need to be analyzed, reduced, and acted upon. Today these data are sent to a central location, the cloud, for analysis. In the cloud, using established computing architectures, data are processed to provide analytics and services for users based on known business models. This cloud-centric model is not sustainable and will not be capable of meeting the requirements of the smart IoT world as

Digital Object Identifier 10.1109/MM.2020.3026667

Date of publication 28 September 2020; date of current version 21 October 2020.

explained by Keshavarzi and van den Hoek.¹ It is interesting to note that in 1999 at its inception, IoT was defined by Proctor and Gamble's Kevin Ashton as "Sensor-technology enabled computers that observe, identify, and understand the world—without the limitations of human-entered data. These computers communicate with each other via the Internet." This describes a smart system, but due to a lack of capabilities of the available technologies, the implementation yielded passive (not smart) IoT devices. This turned the world of IoT into a communication-centric proposition where raw data collected locally by the IoT devices were transmitted by these IoT devices to a central location for processing and analysis.

The semiconductor industry for several decades has revolved around using versatile, pervasively available, programmable CPUs based on the von Neumann architecture with clear (physical and architectural) separation of memory blocks and logic/processing units. These architectures rely on a controller that moves data from cache to the compute element. Preserving the states and the control flow is critical in these architectures. While the von Neumann architecture for constructing microsystems has served us well and continues to be useful, it is proving to be insufficient to support today's new computing workloads, more focused on the flow of data and characterized by an overwhelming deluge of data.^{1,2} Today's and tomorrow's computing demand new capabilities driven by data centric applications to augment our legacy ecosystem. New architectures are needed to serve the demand of smart IoT.

One of the core challenges of many IoT applications is being able to operate in an environment where energy is scarce, and its sources are intermittent. To process massive amounts of data locally collected by these IoT devices with high energy efficiency while maintaining high throughput, the computing hardware will have to overcome energy waste associated with moving data back and forth between separately located memory and logic areas, i.e., addressing the memory wall and the von Neumann bottleneck (speed mismatch of memory and logic) challenges. This

points toward adopting near-memory and in-memory computing (IMC) architectures, i.e., moving toward a blurred boundary between logic and memory elements. Memory plays a critical role in these innovative data-flow architectures. The concept of a controller may get challenged because compute occurs by immediate access to the data, as it flows. For example, a mathematical function like a matrix multiplication may occur in the memory. The goal is for the computation to become analogous to a flow, where one cannot separate the data flow from the compute. These new architectures need to be considered while delivering continued performance gains at a rate exceeding the one provided by scaling microsystems utilizing the established von Neumann architecture.¹⁻³

At the forefront of the data centric computing paradigm is the vision that a trillion, connected, smart edge IoT device will be pervasively and seamlessly integrated into the fabric of life measuring physical world parameters. For this to become a reality, EI is required. EI is the ability to analyze data at the point of data collection and make decisions based on that data autonomously, locally at the

We live in a world where huge amounts of data from our physical world around us are being sensed. These data need to be analyzed, reduced, and acted upon.

edge in real time (see Figure 1). The EI ability will lead to unprecedented opportunities for contextually intelligent applications with far-reaching societal implications. EI will also ameliorate the communication bottleneck by allowing the communication of information bits rather than the raw data bits.¹ EI requires artificial intelligence (AI) to evolve from being performed in the cloud to being executed by "Small-System AI" engines in the smart IoT devices at the edge.

Small-system AI enabled autonomy in decision making requires a capable engine for these smart IoT devices. Energy autonomy is equally vital, especially when energy is scarce and its supply intermittent. This leads to a special class of smart IoT devices that will need to rely on intermittent computing.¹ Their small-system AI engines will operate on harvested energy and need to have a means to preserve their computational state when the energy source is depleted.¹

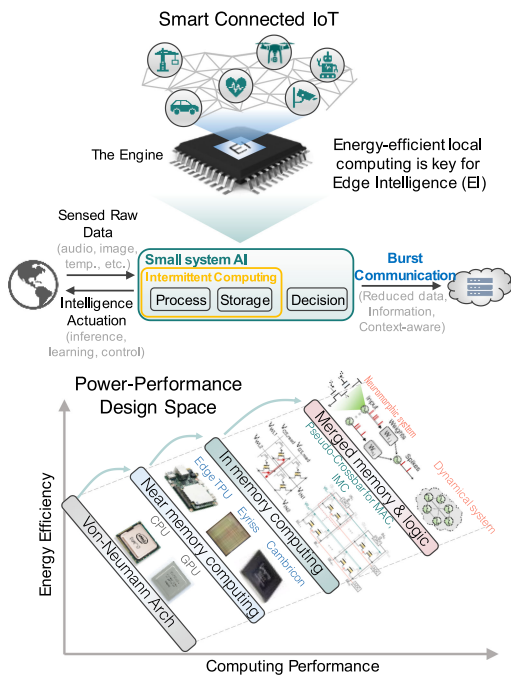


Figure 1. Energy-efficient local computing is the key for EI, a key to enabling the vision of a trillion smart connected IoT devices. The power-performance design space of computing hardware is depicted.

Altogether, if new semiconductor technology capabilities can support the upcoming computing paradigm shift effectively, the sheer number of required semiconductor devices will drive the next stage of exponential growth of the semiconductor industry. This motivates this article. Innovations at all levels of the computing hierarchy enabled by materials, devices, circuits, architecture, system, and algorithms will have to play in concert to deliver new functionalities beyond what is available today.^{1,3} FerroElectronics was introduced as a subfield of electronics based on HfO_2 ferroelectric thin films.^{1,4} Khan *et al.*⁴ discussed the wide range of devices from versatile embedded (non)volatile memory elements to compute elements made possible by CMOS compatible HfO_2 thin films. This technology shows promise to form the foundation of tomorrow's in-memory computing (IMC) paradigm.

This article addresses the small-system AI engine based on new architectures, IMC fabrics, and memory compute elements using ferroelectric key building blocks which are compatible with advanced CMOS technology platforms. It will

show how ferroelectrics can be leveraged at the circuit, microarchitecture, system, algorithm, and software level to deliver autonomy and energy efficiency and provide the performance gains resulting from logic-memory collocation. Here, we posit that FerroElectronics, its building blocks, in particular the ferroelectric field-effect transistor (FEFET) memory device with its extreme energy efficiency and functional diversity enables merged logic-memory functionalities and is the new paradigm of electronics, necessary for addressing the needs of these emerging data centric edge computing applications.^{4,5}

APPLICATION-DRIVEN HARDWARE REQUIREMENTS

Currently, GPUs, the main computing units used in a central location, the cloud, perform the computation needed for the neural networks (NN) used in image recognition applications. They (including TPUs) operate at a compute efficiency of approximately 1 TOPS/W (and are aspiring to reach 10 TOPS/W) while delivering 100 TOPS of performance (see Figure 2). TOPS stand for tera operations per second. This computing performance level is needed to achieve the required system level accuracy targets. It mainly relies on an array of processing cores with shared memory. In these GPU-based accelerators, weights and inputs/outputs move across graphics processing elements (PEs) accessed from a shared memory. The multiply-and-accumulate (MAC) function is performed digitally for the required linear algebra. The compute efficiency tops out at ~ 1 TOPS/W.

EI's requirements exceed today's capabilities. Take for example the case of a micro Drone (uDrone).^{3,6} A uDrone should be capable of doing local computing in order to move smoothly at speed of 10 m/s or higher and to control its movement without getting stalled in the air waiting to decide where to go next. It needs to process images at a data rate of 30 frames per second (fps) or more. To be able to do this, three computing vectors need to be addressed and accelerated: 1) computing for solving perception class of problems, such as inference (in machine learning) using neural networks; 2) computing for optimization class problems: particularly, solving large-dimensional

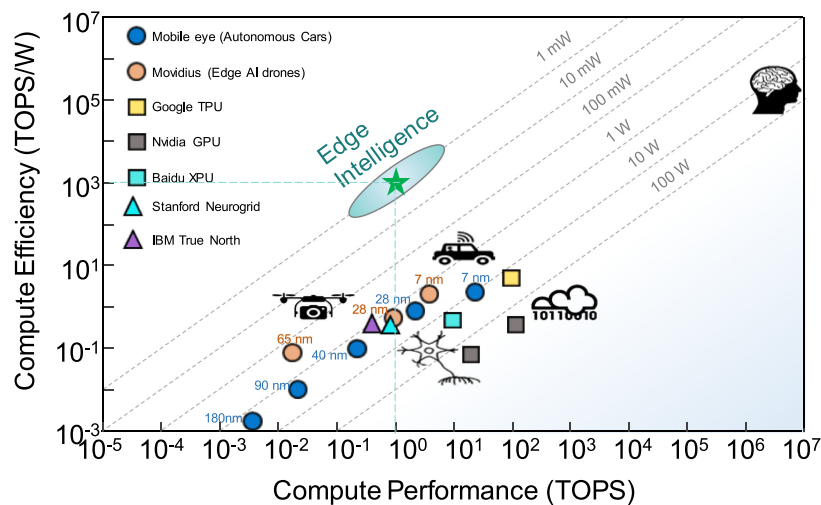


Figure 2. Compute performance versus compute efficiency with contours of constant power consumption. EI requires 1000 TOPS/W with 1-mW power consumption, delivering a performance of 1 TOPS.

optimization problems by the alternating direction method of multipliers (ADMM), by breaking the convex optimization problem into smaller pieces and solving with neural network; this method requires iteration capability and hence relies on local memory capability; and 3) computing to enable in-field learning, i.e., utilizing reinforcement learning (RL) as an efficient technique, particularly doing so by a combination of transfer learning (TL) and RL (referred to as RL+TL). Efficiently performing RL+TL, which requires a clever local memory hierarchy design and perhaps combining different memory computing elements, will be discussed in this article.

To quantify these computing needs, the following should be considered. As mentioned above, the uDrone relies on a vision-based navigation capability because it needs to move smoothly at speed of 10 m/s. It will process images at a data rate of 30 fps. Assuming it will use inference for navigation (which means training has happened somewhere else), this uDrone needs to deliver 1.8 TOPS of performance. This number is based on using the average of ResNet-50 and VGG-16 models (parameters and computing), which requires ~ 10 GMACs per inference in the required neural network. Each MAC corresponds to 2 OPS, so 10 GMACs translates to 20 GOPS. At 30 fps, we need 600 GOPS of performance. Since the uDrone relies on multispectral imaging, using 3-frequency imaging the total required computation is 1.8 TOPS for this inference-based image-based navigation.

Inference utilizes stationary weights in the neural network. If the uDrone needs the ability to learn in the field, RL for training on-the-fly is needed. RL can use neural network to learn a function approximator. We are assuming that our required RL needs 10 rounds (or passes) and 10 iterations, leading to 100 times more computation than in the case of inference only. In-the-field learning requirement increases the computation demand a hundred fold from 1.8 to 180 TOPS. With the uDrone being powered by a lightweight battery providing 100 mW over a period of 30 h of flight, this means that an engine with a performance of 100 TOPS requires a compute efficiency of 1000 TOPS/W (see Figure 2).

These calculations show that EI may demand delivering to a wide dynamic range of performance, spanning from 1 TOPS to over 100 TOPS. Considering the limited power budget at the edge, compute efficiency of exceeding 1000 TOPS/W would be necessary to deliver our required performance in the small systems to realize the vision of EI (see Figure 2) for low-latency and real-time decision-making ability.

WHY EMBEDDED NONVOLATILE MEMORY?

Various new elements, building blocks, and foundational technologies (bottom-up) will be necessary to march toward achieving these high compute efficiencies. IMC is one such foundation.

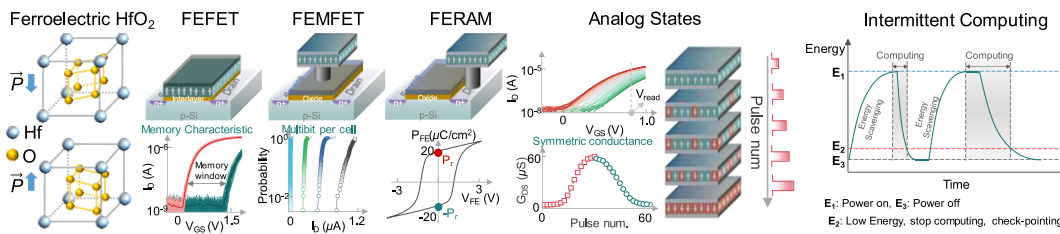


Figure 3. FerroElectronics building blocks based on orthorhombic phase ferroelectric HfO₂. Ferroelectric devices include: the 1T FEFET, 1T-1C FEMFET, and 1T-1C FERAM. A large memory window and multiple bits per cell are exhibited in FEFET and FEMFET. By harnessing the partial polarization switching in ferroelectric HfO₂, intermediate analog states can be created and utilized as synaptic weight cell. Symmetric weight tuning characteristics can be achieved in FEFET. Intermittent computing uses eNVM to store the computing state.

IMC needs a memory compute element made using an advanced technology node because both high performance and low power consumption are key requirements. Typically, embedded SRAM (eSRAM) is the only memory available to be integrated on the chip in these advanced technology nodes, but not at large densities nor cost effectively. Moreover, SRAM is a volatile embedded memory, at 120–150F² its bit cell is not dense and for IMC the situation is even worse because extra transistors (8 T to 10 T instead of typical 6 T SRAM) are needed to manage the write disturb. In addition, SRAM does not support multiple bits per cell, it is leaky and consumes significant amount of standby leakage power. Furthermore, embedded nonvolatile memory (eNVM) is required for our small-system AI engine to be capable of doing intermittent computing and burst communication. Being in the field where sources of energy are scarce and having a small-system AI engine capable of operating in the mW regime opens the door for energy scavenging as the main source of energy. Intermittent computing requires the capability to maintain the state of the compute engine when the energy source runs out. This requires the appropriate architecture and software in combination with eNVM to store the “state of the compute engine” so the system, at the time when energy becomes available again, can seamlessly progress forward from where it stalled rather than having to roll back the computation progress (resulting in wasting valuable energy) (see Figure 3).

The EI applications, of which the uDrone case is an example, impose additional requirements on the choice of embedded memory and the memory compute element. Considering

that the application should be capable of doing the three defined computing vectors, at least two classes of neural network topologies are considered in our discussions. These two neural network topologies are AlexNet neural network topology and ResNet-50 neural network topology. For these two neural network topologies, we will discuss the number of weight parameters and the required amount of MACs, assuming that each weight parameter needs approximately eight bits (8 b) or one-byte (1 B) of memory.

The AlexNet neural network model and its variants have been shown to serve EI applications that require solving a combination of perception and optimization problems because this neural network topology uses a combination of both convolutional and fully connected layers. It requires ~60 MB of memory and a computing performance of ~0.725 GMACs or ~1.5 GOPS. The AlexNet neural network topology has 5 convolutional and 3 fully connected layers representing a more balanced neural network approach. It requires a large amount of memory along with delivering a high computing performance. However, it is not just about computing. As a result, it cannot be serviced by a systolic array of processing elements alone. This topology applies to the uDrone example and for the in-field learning RL+TL using a mix of ~20% SRAM and ~80% eNVM to provide the right memory density, low write energy, and high endurance write capable implementation.^{6,7}

A dense and hierarchical embedded memory with a mix of volatile and nonvolatile capability is chosen to avoid the memory wall problem and the energy cost associated with an external

DRAM and FLASH memory combination used in computing solutions typically seen in the cloud. At the edge and in our discussed topology for the uDrone, we need to go beyond stationary weights for inference-only applications. First and foremost, we will be requiring a dense low write energy profile embedded memory solution.

The ResNet-50 topology can be more suitable for perception class problems (deployed in imaging and vision applications) that use a deeper 50-layer neural network with 49 convolutional layers and 1 fully connected layer. This neural network topology demands more computing and requires less memory. ResNet-50 asks for ~ 26 MB of memory and requires a computing performance of ~ 4 GMACs or ~ 8 GOPS. To support both topologies, a computing performance of >10 GOPS and a memory density of >60 MB at the neural network level are needed. For challenging EI applications such as the uDrone, >1 TOPS of performance with embedded memory density of >100 MB at low power are required.

FEFETs AND KEY FERROELECTRICS BUILDING BLOCKS OF FERROELECTRONICS

For the data centric smart IoT applications requiring EI, the key features of the memory compute element for the IMC and the memory compute fabric in the memory-centric computing approaches are discussed next. A dense, fast, and low energy embedded memory is essential. SRAM although readily available in advanced technology nodes (below 28-nm node) by itself is not the desired choice (low density, high cost in area, high leakage resulting in poor energy efficiency). It should be augmented with a denser and equally low energy profile embedded nonvolatile memory. Key EI application-driven requirements are as follows: 1) dense embedded memory capacity of >100 MB; achieving density metric improvement $>4\times$ bits/mm² based on a single bit per cell compared to an SRAM; 2) cell size of $20\text{--}30F^2$ corresponding to $>5\times$ smaller cell size than an SRAM; 3) multiple bits per cell capability, preferably 3 bits/cell; 4) density improvement of $>10\times$ per consumed area, for example, $\sim 5\times$ by cell size reduction and $\sim 3\times$ improvement by multiple bits per cell

for about $\sim 15\times$ improvement; 5) symmetric conductance for improved learning capability; 6) high endurance of $>10^{10}$ cycles as learning needs writing to the memory and not just reading and sensing it; 7) low write energy and in general a low energy profile in the class of fJ/bit; 8) speed in ns range; 9) transistor transconductance allowing for faster read and improved multibit per cell read; and 10) finally retention which may be traded for higher endurance for improving write performance as needed in an IMC operation. Such tradeoff in a computing-oriented memory has ramifications for speed and energy improvements. A table of eNVMS and embedded memories is captured in Figure 5.^{4,5} From an energy, speed, and density parameters point of view, a memory compute element based on century-old physics and a decade-old newly found ferroelectric material stands out for serving the requirements of EI. It also satisfies the critical attribute of being process compatible with advanced technology nodes, so it can be used in concert with eSRAM and scaled logic transistors.

Ferroelectricity has been around for 100 years. In fact, for decades there have been commercial products using PZT-based Perovskite Oxide Ferroelectric Capacitors (FeCAP) in a 2T-2C FERAM cell configuration (some applications use 1T-1C FERAM cell configuration) inside memory arrays that are used as nonvolatile embedded memories for microcontrollers and digital signal processors as well as standalone NOR memory solutions in niche applications like e-metering and RFIC.^{1,4} FERAM solutions based on Perovskite Oxides have not scaled beyond the 130-nm technology node. They suffer from a destructive read, are slow, and sensing for read requires an accurate reference, which adds complexity to the memory array design, typically requiring the larger 2T-2C cell configuration and it impacts array efficiency negatively. These commercial FERAM solutions have shown very good endurance of $>10^{14}$ cycles making them well suited for the DRAM type application (note that the FERAM operation is like a DRAM's with a destructive read, consuming some extra endurance cycles for rewriting, but with a longer retention, lowering the refresh burden at the cost of not being as dense as DRAM). The film thickness

for PZT-based FeCAP is large at 70 nm. This is a major reason why the technology has not scaled beyond 130 nm. The cell size is at best $3\times$ smaller than an SRAM's at that 130-nm node. Another reason for the niche status of this technology is the fragile nature of the PZT film, which makes it hard to integrate in a standard CMOS process flow. Higher spontaneous and remnant polarization for a 1T-1C FERAM architecture utilizing 1C FeCAP are required.

Ferroelectric-based devices have the best energy profile, primarily because of their physics. The polarization state changes by applying the small amount of energy associated with moving an atom by a distance less than an angstrom. This consumes less energy than needed to program charge-based devices and much less than needed to program magnetic, spin, phase change, or filament forming RRAM devices.⁴ Our quest for the next eNVM technology that meets all the requirements of our EI applications leads us to a novel thin film ferroelectric material used to create a device called FEFET. The FEFET device has been the subject of research lately because of the recent breakthrough observation of ferroelectricity in doped hafnium dioxide (HfO_2),^{8,9} an oxide that is compatible with leading edge high- κ -metal-gate (HKMG) CMOS process flows. HfO_2 -based ferroelectricity and related materials (such as HZO, i.e., Zr-doped HfO_2 also shown as $\text{Hf}_{1-x}\text{Zr}_x\text{O}_2$ where $x < 1$, but typically x is ~ 0.5) have opened the path to FEFETs becoming the preferred memory compute element to be used for IMC.^{1,4,5} hafnium dioxide (HfO_2) has been widely used in HKMG logic transistors since the mid-2000.^{2,4} Therefore, this compatibility and the availability of processing tools can unleash the promise of FEFETs in high volume semiconductor manufacturing.^{1,4} Provided the ferroelectric gate stack thickness is scaled concurrently, the FEFETs have similar scalability trends to state-of-the-art logic HKMG/FinFET transistors, which are scalable down to sub-10-nm nodes. FEFETs can be integrated in FEOL with a greatly reduced mask count of ~ 2 . FEFETs have already been integrated in 28-nm planar bulk CMOS and 22-nm fully depleted planar silicon-on-insulator CMOS platforms as an embedded memory technology.^{4,9}

FEFETs will be critical in addressing the needs of the data centric computing paradigm based on their extreme energy efficiency, high density, and diverse merged logic-memory functionalities. For example, FEFETs may prove to possess the ideal analog weight cell characteristics (critical for IMC). Ferroelectric devices operate based on polarization switching dynamics. In FEFETs, the intrinsic ferroelectric polarization dynamics are strongly coupled to the conductance state of the underlying transistor channel (see Figure 3).^{4,5} This device relies on voltage (electric field driven) switching and is not based on current switching like many of the proposed emerging eNVM solutions.

The FEFET is a three-terminal device having a high transconductance gain that allows for a wide range of circuit topologies and designs that can leverage its unique ferroelectric physics, serving the needs of both traditional and emerging computing applications. One aspect of ferroelectric physics in FEFETs is its plasticity based on the stable, partial switched states in the ferroelectric film of the FEFET, programmed by subcoercive voltages (V_c). In the FEFET, plasticity leads to multistates, nonvolatile operation, allowing multiple bits per cell capability as shown in Figure 3. Transconductance gain eases the operation of reading the multiple bits per cell. Another interesting characteristic, useful for learning, is the symmetric conductance behavior that can be achieved in the FEFET (see Figure 3).

The energy profile of the FEFET is the best-in-class among all nonvolatile memory technologies (see Figures 3 and 5) and approaches the realm of the volatile eSRAM. The transistor action in FEFETs, which is not available in other two terminal emerging (resistive) eNVM memories, allows not only for a fast, nondestructive read but also enables unique, efficient, and creative cell, array and circuit designs with a small cell size ($10\text{--}30F^2$ depending on the application). The write operation in ferroelectric devices can be extremely fast, taking less than 1 ns. It is the FEFET's transconductance that makes also the read fast.

FEFETs are a work-in-progress and are being heavily researched to address some of their challenges: variability (a problem toward achieving

high density arrays), endurance improvement requiring engineering of the Interfacial oxide Layer (IL), lowering of the HZO film thickness,¹⁴ and scaling of the FEFET device size. To date, endurance cycling performance of state-of-the-art FEFETs have been limited to the range of 10^5 – 10^9 cycles.⁴ Although this endurance is better than the endurance of most emerging memory alternatives, it is poor compared to the near unlimited endurance of eSRAM. Since IMC needs $>10^{10}$ endurance cycles, improving endurance is the subject of research with alternative device structures being considered (for example, the FEMFET shown in Figure 3, in which the IL layer is eliminated). Improving the design of the gate stack and the IL layer between the FE layer and transistor channel of the FEFET are also topics of intense research. The IL layer is the main cause of the limited endurance performance of the FEFET. The FE layer by itself if sandwiched between two metallic layers will have good endurance similar to the 1T-1C FERAM.⁴ Retention in FEFETs is good and meets the typical 10 years duration specification. It is important to note that retention may be traded for improving endurance for performing IMC and in-field learning in smart IoT devices at the edge.

FEFETs possess a set of key characteristics that are particularly important for creating either dense embedded memories for standard embedded memory applications or more importantly toward the memory compute element for IMC to accelerate computation, for example, for neural networks (providing multistate weight cells or so-called analog synapses).⁴ Multibit operation with 2–8 bits per cell (4–256 levels), in the order of 100-fold conductance modulation, fast nanosecond write time, as well as linear and symmetric conductance (potentiation-depression) leading to higher accuracy computation were discussed by Khan *et al.*⁴ An FEFET can also act both as the selector and the nonvolatile memory element in a ternary content addressable memory (TCAM) leading to the smallest footprint TCAM cell with just two transistors. Content addressable memory cells can be efficiently used for pattern matching applications, for fast and parallel database searches and in finding match locations. Cypress semiconductor commercially deployed nonvolatile SRAM

(combining SRAM and SONOS eNVM) to make TCAM solutions for network packet routing, but that TCAM cell comprised of over ten transistors resulting in a very large footprint.¹ Nonvolatile logic and fast data back-up and wake-up circuits for intermittent computing can also utilize combinations of these ferroelectric features and characteristics.⁴

The use of ferroelectric devices expands significantly beyond memory applications and today includes negative capacitance transistors for ultralow power, high-performance logic technology, artificial neurons for spiking neural networks (SNNs), and circuit primitives for stochastic computing.⁴ All are beyond the scope of this article. However, we will briefly elude to using ferroelectric-based coupled oscillatory networks for continuous time dynamical systems in our outlook section.^{4,10}

The FEFET is a foundational technology building block in FerroElectronics. However, FerroElectronics builds on additional ferroelectric-based devices and technologies for logic, analog, and RF transistors,¹ but these are not the subject of this article. The FEFETs provide desired features for the alternative computing paradigms. FEFET technology and the broader FerroElectronics are important elements for realizing EI, the bottom-up approach path.

IN-MEMORY COMPUTING BASED ON CIRCUITS WITH FEFETs

This section describes building the arrays, memory compute fabric, corresponding circuits, processing elements, and the cores for the compute engine for EI using the FEFET memory compute element.

The idea of IMC is not new. Back in the 1960s, even von Neumann himself was thinking about processing in memory. However, the question of what processing should be performed in the memory was not resolved then. Today, with data-centric computing demands and in particular with the need for vector multiply and add operations, it is worth revisiting IMC for small systems that need to process large amounts of data at the point of collection efficiently with low latency and high speed (with high throughput). IMC is essential in memory centric computing. IMC reduces the

movement of large amounts of data and hence addresses the memory bottleneck challenge. IMC brings the MAC operation into the memory. The way IMC works is that the vector-matrix multiplication (VMM) is conducted in a parallel manner: input vectors activate multiple rows in parallel in the memory. The input vectors are multiplied by the memory cell conductance (i.e., multiplication or dot-product) that contains the weights creating a partial sum on the bit line column, where the current of the bit line column represents the analog MAC value in VMM. IMC needs analog-to-digital converters (ADCs) at the periphery of the array in order to convert the analog MAC/VMM on the bit line to binary bits for digital processing in the small system. The parallel nature of conducting the math saves energy, but several tradeoff parameters such as the energy, array area efficiency, pitch-matching, types of ADCs used and the area and power consumed by the required ADCs (and extra complexity of mixed signal design used) need to be considered for IMC.

Next the design of the memory computing fabric delivering a more efficient IMC needs to be addressed. What memory compute element should be used?

SRAM is available and has been used by itself for IMC with mixed results. The throughput is high, the read is fast, and the write energy of this charge-based embedded memory solution is good. However, this memory is not dense enough (it consumes a large area in very expensive silicon on advanced technology nodes) for the workloads of interest in the data-centric systems utilizing IMC, for example, storing the weights for doing inference. SRAM leakage is high favoring the use of SRAM for systems that are computing with a high activity factor in order to amortize the leakage cost/penalty of the memory. SRAM does not support multiple bits per cell. It is based on 1 bit/cell and to achieve higher precision it needs to add them, using thermometer coding. Capacitive banks (based on the SAR ADC topology) add to its complexity (capacitive matching and offset cancelation) and add to the cost. SRAM IMC is based on the mixed signal design and it is more efficient than digital-only implementations. It targets lower precision applications and the periphery circuits are

complex. These issues lower the array efficiency for SRAM IMC. Particularly, it will be difficult to encode inputs in voltage or time. This limits SRAM to toy and smaller problems with a low number of weight parameters. SRAM IMC is mostly targeted for binary neural networks (BNNs) based on XNOR/XOR with low precision. Since multiple word lines are activated simultaneously, read disturb is a fundamental issue. A bit line discharge in the SRAM array, which is required for high dynamic-range readout, threatens to write-back into the cell causing destructive read. SRAM IMC happens locally on the bit line as we explained earlier (and it requires ADCs). A better approach is to leverage the SRAM's strengths and augment it with a dense eNVM.^{1,4,6,7} Based on our applications' requirements, >100 MB memory capacity is required, suggesting utilizing a mix of 20% SRAM (20 MB) and 80% dense eNVM (80 MB).^{6,7}

Referring to the three vectors of computing described earlier, more than computing for accelerating neural networks (e.g., CNNs/DNNs) is needed. Other statistical machine learning algorithms for supporting other linear algebra kernels (beside the covered vector matrix multiplication) may not be as complex as deep neural networks (DNNs). Thus, near-memory computing using embedded memory (however, not in 2.5D HBM style systems deployed in the cloud) may be used. For example, solving the optimization problems encountered in the case of autonomous uDrones relies on iterative and distributed architectures. One such standard algorithmic approach is ADMM, which is based on local computing and iterative communication (computing a result, communicate, and iterate on it) in these distributed array architectures. Studies addressing where the energy goes in systems designed for ADMM tasks show that the consumed energy is almost equally distributed between computing, communication, and memory operations.¹² This shows that pure systolic arrays of processing elements using a predetermined dataflow may not be suitable for solving optimization class problems. Rather, processing units used for optimization should be connected with their immediate and farther away neighbors for consensus for determining global optimization in ADMM as shown by Chang *et al.*¹² using a Network-on-a-Chip (NoC)

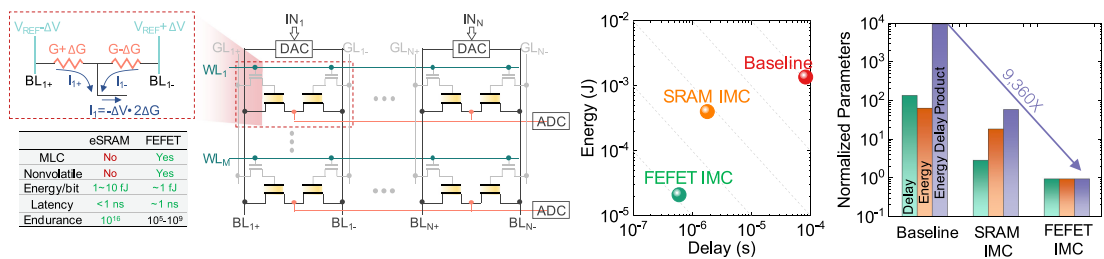


Figure 4. In memory computing with FEFETs. By using the FEFET conductance as the neural network weight, matrix vector multiplication can be accelerated in the analog domain. IMC computing based on FEFETs provides a 9360× improvement in energy-delay product compared to the von Neumann “Baseline.”











with an 8-neighbor hierarchical multicast network based on asynchronous communication and circuits with a 4-phase handshake protocol. Furthermore, optimization class problems will require higher bit precision in the compute stages and should support a programmable bit precision and data movement for the iterative algorithms.

Most AI solutions based on machine learning and accelerated computing for matrix multiplications for neural networks can be modified to use positive numbers only, which simplifies the math and implementation. To address the need of our three computing vectors and solving optimization class problems requires that both positive and negative operands are handled. This capability of doing the math with both positive and negative numbers will allow serving a wider range of useful algorithms. However, it will require designing a special array¹¹ with devices to emulate positive and negative operands, as well as appropriate peripheral circuits as shown in Figure 4. This figure shows using a pseudocrossbar with a 2T cell. What memory element would be most suited in this approach?

Most emerging eNVM solutions are based on resistive elements typically using a two-terminal 1R resistor element forming a 1T-1R memory cell that is deployed in crossbar resistive arrays. However, the FEFET can be a memory compute element in the proposed 2T cell configuration. Building a large array from FEFETs for IMC, doing computing and using them for EI applications will require working on improving device-to-device variability and increasing the endurance of the FEFETs. We described earlier FEFET’s potential capability for being an eNVM operating at fJ/bit energy consumption with nano second speed (latency) for IMC. Moreover, the FEFET for our

proposed 2T-cell configuration has a small cell size of $\sim 30F^2$ which is $\sim 4\text{-}5\times$ smaller than an SRAM cell size. Enabling multiple bits per cell, let us say 3 bits/cell, where the FEFET transistor gain plays a key role in distinguishing between the states also contributes to a higher overall density of $\sim 15\times$ compared to the same technology node eSRAM. FEFETs can operate at low voltage and are compatible with advanced technology nodes and hence can be integrated next to the embedded SRAM and advanced logic transistors with minimal extra masks. These listed characteristics and what we discussed earlier make FEFETs an ideal solution toward fulfilling our three asked for computing vectors for EI. The FEFET enables a memory compute element, leading to an IMC fabric, building a unit to go inside a core, and then configuring many cores with local memories as per a more capable data flow architecture while allowing for communication among neighboring cores to solve a broader scope of computing problems. Beyond fixed flow neural networks for vision and perception (in today’s systolic array data-flow architectures), this provides a path toward solving optimization with ADMM.

Many factors need to be considered in designing circuits and IMC arrays: cell size, cell configuration (more flexible positive and negative operands with a differential 2T FEFET cell configuration without impacting peripheral circuits), number of bits/cell, symmetric potentiation/depression for higher accuracy computation, BL capacitance, number of cells per BL, number of rows and columns, number of activated rows for IMC, DAC, and ADC resolutions, topology, type of ADC, and circuit style (requiring mixed signal circuit design) for IMC readout and sensing. A careful analysis of the circuit simulation of the full

	eSRAM	eDRAM	FG Flash	SONOS	ReRAM	PCM	STT-MRAM	FERAM	FEFET	FEMFET
Structure										
Cell structure	6T	1T-1C	1.5T	2T	1T-1R	1T-1R	1T-1R	1T-1C	1T	1T-1C
Mechanism	Cross-coupled inverter+charge	Charge on capacitor	Charge on FG	Charge in nitride	Filament formation	Phase change	Spin transfer torque, magnetic	Polarization switching	Polarization switching	Polarization switching
MLC	No	No	Yes	Yes	Yes	Yes	No	Potentially	Yes	Yes
R_{ON}/R_{OFF} ratio	N/A	N/A	$>10^4$	$>10^4$	10-100	10-100	<10	N/A	$>10^4$	$>10^4$
Status ^a	Available ^b	Development ^c	Available ^b	Development ^c	Development ^c	Development ^c	Development ^c	Available ^b	Research ^d	Research ^d
Integration node	7nm FinFET	22nm FinFET	40nm	28nm HKMG	22nm FinFET	40nm	22nm FinFET	130nm	22nm FDSOI	N/A
Cell size	120-150 F ²	40 F ²	50 F ²	60 F ²	60 F ²	60 F ²	50 F ²	50 F ²	20-30 F ²	30-40 F ²
Additional masks	0	5+	13+	5+	3+	3+	3+	2-3	1	3+
Energy/bit	~ 1 fJ	~ 1 pJ	100 pJ	10 pJ	>10 pJ	100 pJ	>10 pJ	~ 1 pJ	~ 1 fJ	~ 10 fJ
Latency	<1 ns	>10 ns	0.1-1 ms	10-100 ns	<100 ns	<100 ns	>10 ns	>10 ns	~ 1 ns	10 ns
Endurance	10^{16}	10^{16}	10^4 - 10^5	10^4 - 10^6	10^5 - 10^7	10^5 - 10^7	10^6 - 10^7	$>10^{14}$	10^5 - 10^9 ^e	10^{10}
Retention	Volatile	Refresh	10 yrs	10 yrs	10 yrs	10 yrs	10 yrs	10 yrs	10 yrs	10 yrs

^a "Status" row is added to distinguish results for the solutions that are in manufacturing phase and production compared to development or research phase.

^b Available - All the parameters are based on the reported data on memory macro in production and in manufacturing

^c Development - The parameters are based on memory macro data that is still under development, not in high volume manufacturing

^d Research - The parameters are based on memory device data that is still being researched. Also, there is a difference between macro and array results versus device level results.

^e The endurance of FEFETs are under research to increase performance $>10^{10}$ cycles [4].

Figure 5. Comparison table of ferroelectric devices with other embedded (nonvolatile) memory devices. Ferroelectric devices are advantageous in terms of energy-efficiency and overall balanced performance.^{4,5} Note that more references for this table are captured by Khan *et al.*⁴

integrated circuit chip using the measured FEFET device parameters based on 28-nm HKMG technology by Yoon *et al.*¹¹ compared three chip implementations of systems for solving iterative convex optimization problem by ADMM via least-squares-minimization: 1) a digital von Neumann architecture implementation based on ALUs and embedded SRAM (Baseline); 2) an SRAM-based 6T-cell IMC implementation (SRAM-based IMC); and 3) an FEFET-based 2T-Cell, in pseudo-cross-point array for IMC implementation (FEFET-based IMC). The results are shown in Figure 4. All three parameters: energy, delay, and energy-delay product (EDP) were better for implementation number (3), i.e., the FEFET-based IMC. In fact, the FEFET-based IMC consumes 65 \times less energy than the digital von Neumann baseline and 19 \times less energy than the SRAM-based IMC. The EDP for the FEFET-based IMC is $\sim 9400\times$ lower than the EDP for the digital von Neumann baseline and 60 \times lower than the EDP for SRAM-based IMC. It should be emphasized that improved EDP resulting from using the FEFET technology makes FEFET technology based IMC a preferred solution for inference applications, and not necessarily limited to just low bit precision inference. This technology is also positioned for in-field learning. Note that the SRAM-based IMC occupies $>3\times$ larger silicon area than

the FEFET-based IMC implementation. Exploring the design space shows that for parallel throughput computation using 12-bit DAC and 14-bit ADC with 3 bits/cell, the FEFET-based IMC implementation results in the lowest energy and compute time. Using higher than 3 bits/cell increases the overhead and when the DAC resolution increases, a higher ADC resolution is needed, increasing the energy consumption. Device variability was accounted for in these simulations.

The improved efficiency in computing achieved by the FEFET-based IMC was utilized for solving iterative convex optimization problems by ADMM via least-squares-minimization in two applications: 1) constructing signal from 1-D EEG (ElectroEncephaloGram) and 2) recovering of CT (computerized tomography) scan images which is relevant in medical imaging applications. The fidelity of reconstruction process increases as the subspace dimension increases because the FEFET-based IMC has more computing capability. This enables real-time reconstruction of the data in the field at low-power consumption. Furthermore, it is worth mentioning that solving an iterative algorithm like ADMM in the dataflow architecture resembles a hardware emulation of a distributed and discrete-time dynamical system that will be discussed later.

For these imaging applications, accuracy is a system-level metric. A similar exercise for the uDrone application would use a different metric like collision avoidance as a key system-level metric. A uDrone uses computing for path finding, mapping, depth, localization/SLAM, and control in addition to vision-based navigation. The kinds of circuits and arrays we described for the FEFET-based IMC can help doing more computation for these functions and algorithms.

SMALL-SYSTEM AI AND THE ENGINE FOR EDGE INTELLIGENCE

The engine for EI should be more than a simple inference engine. Perception and optimization class problems (model-free and even model-based), and ultimately in-field learning by RL for autonomy (for uDrone applications) all need to be handled by such an engine.

A small-system AI engine should be capable of handling the real-time low-latency learning in the field through RL+TL using a combination of eSRAM and FEFET memory. Learning is critical in the case of uDrones if there is no GPS coverage. When the uDrones need to be autonomous, providing a means to learn by interacting with their environment is critical and challenging, since the uDrone needs to move seamlessly at a reasonable speed of 10 m/s. In that case, model weights need to be updated frequently (for learning) and in a short amount of time with a latency of <10 ms. RL can be viewed as a form of learning by trial and error based on reward mechanisms. Learning places a demanding burden on write time, write energy, and endurance cycling performance of the embedded memory solution used in IMC. This explains why dense FEFET's potential for fast nanosecond write time at low fJ/bit write energy with an improved endurance of at least 10^{10} cycles is a game changer for these learning applications. eSRAM and magnetic/spin-based STT-MRAM were used in a memory hierarchy for RL in systems for uDrones by mapping the algorithm carefully into this memory hierarchy.⁶ The algorithm utilizes both convolutional and fully connected layers. The fast and changing fully connected layers use eSRAM while convolutional and slow changing fully connected layers are placed in a denser eNVM. In the case of Yoon

et al.,⁶ the choice of eNVM was STT-MRAM. The write time (latency) and write energy of this choice of eNVM determine the performance of the system and is limited by the magnetic/spin technology capability; although this hierarchical approach allowed for an overall improvement in speed of the uDrone by allowing processing performance at a higher data rate measured by fps, but it did not meet what is needed for 30 fps at 10 m/s and without any loss in system-level accuracy metric. Using FEFET-based eNVM and moving to more advanced technology nodes will significantly improve the performance of such small systems for these applications.

If more computing capability is needed for the uDrone application for path planning, mapping, depth, localization/SLAM, and control in addition to the vision-based navigation, then an FEFET-based IMC provides the more computing required. The uDrone example is an interesting platform for exploring various compute demands such as 1) model-free, learning-based statistical solutions by neural network, etc.; and 2) model-based solutions by the potential fields approach utilizing various linear and nonlinear processing units.

Therefore, capabilities enabled by new ferroelectric-based materials and devices such as the FEFET-based IMC circuits will be key to realizing small-system AI engine's computation demands. This goes beyond current research in curating data, pruning, compression, condensing the weights or techniques for tweaking the precision based on today's technology features that have been widely discussed and continue to be explored in the literature.^{2,3}

Going back to today's solutions (e.g., GPU engine) operating at an energy efficiency of 1 to 10 TOPS/W and considering the gains in delay reduction by a factor of $\sim 140\times$ and improvement in energy efficiency by a factor of $\sim 65\times$ for the FEFET-based IMC as shown in Figure 4, the system performance compared to these digital von Neumann architecture (Baseline) solutions can be enhanced by implementing an IMC architecture. The results of these improvements make it likely that the target of >1000 TOPS/W compute efficiency with a performance dynamic range of 1–100 TOPS for a corresponding power range of 1–100 mW as shown in Figure 2 is feasible for an FEFET-based IMC solution that can be

exploited by smart IoT devices. In comparison, central systems in the cloud deliver >100 TOPS of performance at much higher power consumption of >100 W. However, it will require pushing research vectors in all scales from materials, devices, circuits, architecture, and systems, incorporating both bottom-up and top-down approaches to achieve the efficiency.

ARCHITECTURAL FEATURES OF THE PROCESSING ENGINE

Until this point, we discussed how FerroElectronics, a bottom-up approach viewpoint, can enable much needed efficiency and performance improvements for EI. Let us also look at architecture and software from a top-down perspective.

Architecture plays a key role in achieving the goals of the small-system AI engine solutions deployed for the EI applications. So far, the architecture has evolved from many cores to domain specific architectures (DSA) with accelerators and ASIC SoC implementations in 2.5D heterogeneous integration for near memory computing. Furthermore, data-flow architectures are deployed based on a systolic array of processing elements where predetermined data flow paths are being implemented with shared memory, inching their way toward near-memory computing implementations. The next phase in architecture enhancement will be revisiting and implementing IMC with a dense and low-energy profile embedded memory to eliminate the “memory wall” problem, providing more energy efficient higher compute performance needed to move toward in-field learning. Such architectures are shown in Figure 6 and are discussed in greater details in the work of Raychowdhury.³ These architectures will be able to support the three vectors of computing described earlier in this article.

Data-flow architectures with arrays of systolic processing elements have been deployed and one could conceive moving toward IMC using other resistive memory elements, but we described its challenges earlier. However, implementing these modified data-flow architectures with our memory compute fabric in IMC, adopting new Ferroelectric-based eNVM in the form of a FEFET memory compute element, in conjunction with eSRAM to be used in the architecture shown in Figures 4 and 6 can deliver the performance and efficiency we

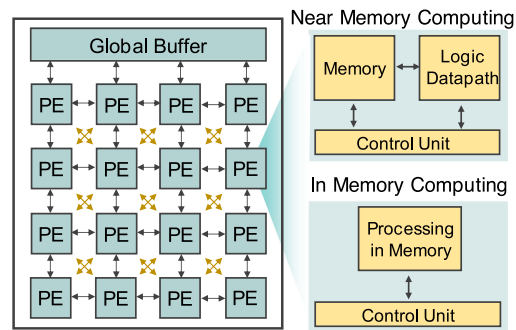


Figure 6. Architecture of computing hardware supporting near memory and in-memory computing.

discussed in this article. Small systems designed based on these technology features can solve interesting EI problems. Distributed optimization shows an interesting class of algorithms where computation, communication, and memory storage are almost equally important in terms of power consumption.¹² In-field learning’s computational demand will be served by IMC, employing TL+RL by using an eSRAM and FEFETs memory hierarchy.^{6,11}

The path toward merged logic and memory and the computational memory concepts will lead to a future when we create MANNs, RNNs, dynamical systems based on FerroElectronics computing foundational capabilities.

OUTLOOK AND A PATH TOWARD THE FUTURE

A small-system AI engine is at the heart of the much-needed efficient electronic hardware for EI that enables the explosion of data centric computing applications. This will in effect fuel the next wave of exponential growth of the semiconductor industry. We have explained in this article why FEFET technology and the broader FerroElectronics are important elements for realizing EI and how they fit in a data-flow architectural scheme with IMC.

Achieving a computing performance at the very high compute efficiency of 1000 TOPS/W based on FerroElectronics allows minimizing the system-level energy consumption by utilizing enough local computing to minimize the communication needs in these smart small systems.¹ Local computing lowers the burden of communication. Because transmission energy per bit has

not been successfully reduced below 1 nJ/bit and because of the bandwidth problems, communication needs to be kept to a minimum and reserved for valuable information bits. The solution path points to using IMC consuming 1 fJ/bit to convert raw data to valuable information bits, aggregate these information bits and communicate them in bursts while being aware of the quality of the channel, i.e., communicate when channel is good, further lowering the communication energy (see Figure 1). Note that by going from data to information using local computing reduces the number of bits by a factor of $10\,000\times$.¹

A recent paper¹³ shows how computing and communication can be managed by a small-system AI engine doing IMC to autonomously optimize the system energy consumption. The engine is implemented by a SoC that includes an image processor and a digitally adaptive radio for communication along with an IMC-based controller implemented in 65-nm technology. It emphasizes the point that AI can be applied even for the management task of determining how much computing versus communication should occur. The problem is a multivariable system level power optimization challenge.¹³ This paper is an interesting proof

of concept, but the compute performance efficiency is only at about 1 TOPS/W mainly constrained by the technological features and the limited capabilities supported by the 65-nm node. To improve the efficiency by greater than two orders of magnitude while delivering much improved performance that can be allocated for computing, for in-field learning, for implementing in-device in-hardware security solutions, FerroElectronics computing is a promising answer.

In dynamical systems, much like the human brain, logic and memory are not physically separated. In a dynamical system, computation evolves as a flow of physical variables like voltage, current, phase, etc., that interact with each other in a coupled system. In future machines deploying a dynamical-system-based compute engine, the states and the compute are not physically separable. States appear as analog variables (current, voltage, phase, etc.), which evolve based on

physical rules and computation is analogous to a flow. This can be thought of as the ultimate form of merged logic and memory, where states evolve autonomously based on the computation that we are performing (as per flow of the data). For example, in a dynamical system, we can set the phase of an oscillator to follow the function that we want to differentiate, and then observe the corresponding frequency [$\text{frequency} = d/dt(\text{phase})$]. Phase and frequency are not separable here—one is the result of another. However, if we think of phase as a state variable, then frequency is the output of the computation. Ferroelectric-based oscillators (by FEFETs and as part of FerroElectronics) for dynamical systems can create such future machines.^{4,5,10}

In closing, this article discussed how combining a bottom-up ferroelectric based material and device approach (see Figure 3) with a top-down architecture choice (see Figure 6) enabled an IMC capability to achieve the results shown in Figure 4 and the future small-system AI engine to meet the compute performance and efficiency requirements of EI (see Figure 2). The application-driven discussions in this article based on looking at the requirements from both bottom-up and

In both near-memory and IMC approaches, logic and memory are architecturally close but physically need to be designed explicitly. Thinking beyond, a dynamical system approach should be explored.

top-down identified several technical challenges and research vectors covering all scales—materials, devices, circuits, design, architectures, and implementation of an engine to realize the vision of EI. The archetype of AI in small efficient systems is a key enabler for a wide range of applications that require the devices to operate autonomously and sustainably in challenging and energy-constrained environments, projected to reach a trillion IoT devices. EI enables smart devices to sense, analyze, decide based on and act on locally collected data, and send information to the cloud, rather than relying on the cloud to analyze and decide based on the transmission of raw locally collected data that is sent to the cloud. Turning sensed data into information for actionable intelligence locally requires a careful balance of energy-efficient computing and communication demands in a small system. The system is operating at the intersection of Moore's Law and the Shannon–Hartley Theorem. The

careful tradeoff results in minimized system energy consumption. Ferroelectric building blocks enable a new capability and ushers in the era of FerroElectronics, paving the way for doing IMC in data-flow architectures, improving compute efficiency by 1000 \times , satisfying the requirements of EI. This will allow deployment of many smart IoT devices based on the small-system AI engines for a range of smart applications that will drive the next phase of the semiconductor industry growth.

REFERENCES

1. A. Keshavarzi and W. Van Den Hoek, "Edge intelligence-on the challenging road to a trillion smart connected IoT devices," *IEEE Des. Test*, vol. 36, no. 2, pp. 41–64, Apr. 2019.
2. S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nature Electron.*, vol. 1, no. 8, pp. 442–450, Aug. 2018.
3. A. Raychowdhury, "Towards memory centric autonomous systems: A technology and device perspective," in *Proc. Int. Electron Device Meeting Short Course*, 2019.
4. A. I. Khan, A. Keshavarzi, and S. Datta, "The future of ferroelectric field-effect transistor technology," *Nature Electron.*, vol. 2, pp. 580–586, 2020.
5. K. Ni, S. Dutta, and S. Datta, "Ferroelectrics: From memory to computing," in *Proc. Asia South Pacific Des. Autom. Conf.*, 2020, pp. 401–406.
6. I. Yoon, M. A. Anwar, R. V. Joshi, T. Rakshit, and A. Raychowdhury, "Hierarchical memory system with STT-MRAM and SRAM to support transfer and real-time reinforcement learning in autonomous drones," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 3, pp. 485–497, Sep. 2019.
7. M. Donato, L. Pentecost, D. Brooks, and G. Y. Wei, "MEMTI: Optimizing on-chip nonvolatile storage for visual multitask inference at the edge," *IEEE Micro*, vol. 39, no. 6, pp. 73–81, Nov. 2019.
8. J. Müller *et al.*, "Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories," in *Proc. IEEE Int. Electron Devices Meeting*, 2013, pp. 10.8.1–10.8.4.
9. T. Mikolajick *et al.*, "The Past, the present, and the future of ferroelectric memories," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1434–1443, Apr. 2020.
10. Y. Fang, Z. Wang, J. Gomez, S. Datta, A. I. Khan, and A. Raychowdhury, "A swarm optimization solver based on ferroelectric spiking neural networks," *Front. Neurosci.*, vol. 13, Aug. 2019, Art. no. 855.
11. I. Yoon, *et al.*, "A FerroFET-based in-memory processor for solving distributed and iterative optimizations via least-squares method," *IEEE J. Exploratory Solid-State Comput. Devices Circuits*, vol. 5, no. 2, pp. 132–141, Dec. 2019. doi: [10.1109/JXCDC.2019.2930222](https://doi.org/10.1109/JXCDC.2019.2930222).
12. M. Chang, L. H. Lin, J. Romberg, and A. Raychowdhury, "OPTIMO: A 65-nm 279-GOPS/W 16-b programmable spatial-array processor with on-chip network for solving distributed optimizations via the alternating direction method of multipliers," *IEEE J. Solid-State Circuits*, vol. 55, no. 3, pp. 629–638, Mar. 2020.
13. N. Cao, B. Chatterjee, M. Gong, M. Chang, S. Sen, and A. Raychowdhury, "A 65-nm image processing SoC supporting multiple DNN models and real-time computation-communication trade-off via actor-critical neuro-controller," in *Proc. IEEE Symp. VLSI CIRCUITS*, 2020, pp. 1–2. doi: [10.1109/VLSICircuits18222.2020.9162878](https://doi.org/10.1109/VLSICircuits18222.2020.9162878).
14. S. Cheema and S. Salahuddin, "Enhanced ferroelectricity in ultrathin films grown directly on silicon," *Nature*, vol. 580, pp. 478–482, 2020. [Online]. Available: <https://www.nature.com/articles/s41586-020-2208-x>

Ali Keshavarzi is an Adjunct Professor with Stanford University. He is an advisor to DARPA. He is also with Leading Edge Research. He was the Vice President of R&D and a Fellow at Cypress and had leading R&D roles at Intel, TSMC, and GLOBALFOUNDRIES. Keshavarzi received the Ph.D. degree in electrical engineering from Purdue University. He is a senior member of the IEEE. Contact him at akesh@stanford.edu.

Kai Ni is an Assistant Professor with Rochester Institute of Technology. Ni received the Ph.D. degree in electrical engineering from Vanderbilt University. He is a member of the IEEE. Contact him at kai.ni@rit.edu.

Wilbert van den Hoek is an independent adviser, a board member, and a Principal with Leading Edge Research. He was CTO and EVP of Novellus Systems, Inc. Van den Hoek has a Doctorandus in chemistry (*cum laude*) from Rijks Universiteit Utrecht. Contact him at wgm.vdh@gmail.com.

Suman Datta is the Stinson Professor of Nanotechnology with the University of Notre Dame. He is also the Director of the Multi-University Advanced Microelectronics Research Center, the ASCENT, funded by the Semiconductor Research Corporation and DARPA. He was with the Advanced Transistor Group, Intel Corporation, where he developed several generations of high-performance logic transistor technologies, including high-k/metal gate, tri-gate, and nonsilicon channel CMOS transistors. He has published more than 350 journals and refereed conference papers and holds 185 patents related to advanced semiconductors. He is a Fellow of the National Academy of Inventors and the IEEE. Contact him at sdatta@nd.edu.

Arijit Raychowdhury is a Professor with Georgia Institute of Technology. He is also a Co-Director of Georgia Tech's Quantum Alliance. He was a research scientist with Circuit Research Lab of Intel Corporation. He has won 11 best paper awards in his career. Raychowdhury received the Ph.D. degree in electrical engineering from Purdue University. He is a senior member of the IEEE. Contact him at arijit.raychowdhury@ece.gatech.edu.



Call for Articles

IEEE Pervasive Computing

seeks accessible, useful papers on the latest peer-reviewed developments in pervasive, mobile, and ubiquitous computing. Topics include hardware technology, software infrastructure, real-world sensing and interaction, human-computer interaction, and systems considerations, including deployment, scalability, security, and privacy.

Author guidelines:
www.computer.org/mcl/pervasive/author.htm

Further details:
pervasive@computer.org
www.computer.org/pervasive

IEEE pervasive COMPUTING
 MOBILE AND UBIQUITOUS SYSTEMS