

aCortex: An Energy-Efficient Multipurpose Mixed-Signal Inference Accelerator

MOHAMMAD BAVANDPOUR^{ID},
MOHAMMAD R. MAHMOODI^{ID} (Graduate Student Member, IEEE),
and DMITRI B. STRUKOV^{ID} (Senior Member, IEEE)

Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93117 USA

CORRESPONDING AUTHOR: M. BAVANDPOUR (mbavandpour@ece.ucsb.edu)

This article has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors.

ABSTRACT We introduce “aCortex,” an extremely energy-efficient, fast, compact, and versatile neuromorphic processor architecture suitable for the acceleration of a wide range of neural network inference models. The most important feature of our processor is a configurable mixed-signal computing array of vector-by-matrix multiplier (VMM) blocks utilizing embedded nonvolatile memory arrays for storing weight matrices. Analog peripheral circuitry for data conversion and high-voltage programming are shared among a large array of VMM blocks to facilitate compact and energy-efficient analog-domain VMM operation of different types of neural network layers. Other unique features of aCortex include configurable chain of buffers and data buses, simple and efficient instruction set architecture and its corresponding multiagent controller, programmable quantization range, and a customized refresh-free embedded dynamic random access memory. The energy-optimal aCortex with 4-bit analog computing precision was designed in a 55-nm process with embedded NOR flash memory. Its physical performance was evaluated using experimental data from testing individual circuit elements and physical layout of key components for several common benchmarks, namely, Inception-v1 and ResNet-152, two state-of-the-art deep feedforward networks for image classification, and GNTM, Google’s deep recurrent network for language translation. The system-level simulation results for these benchmarks show the energy efficiency of 97, 106, and 336 TOP/J, respectively, combined with up to 15 TOP/s computing throughput and 0.27-MB/mm² storage efficiency. Such estimated performance results compare favorably with those of previously reported mixed-signal accelerators based on much less mature aggressively scaled resistive switching memories.

INDEX TERMS Artificial neural networks, floating-gate memory, machine learning, mixed-signal circuits, neuromorphic inference accelerator, nonvolatile memory (NVM).

I. INTRODUCTION

THE rapidly growing range of applications of machine learning for image classification, speech recognition, and natural language processing along with maturing of the neural network algorithms, especially for deep learning, led to an urgent need in specialized neuromorphic hardware [1]–[3]. At least, for the next several years, the demand for fast, low-precision inference accelerators will remain higher than for higher-precision systems for network training, as projected by Nvidia Corporation, a leading company in the machine learning hardware [4].

The vast majority of the proposed neuromorphic accelerators from industry and academia are digital [5]–[8]—see also an extensive review of various proposals in [2]. The most natural approaches, however, are based on analog and mixed-signal circuits [9]–[30]. Though the core principles of analog computing had been developed almost four decades

ago [9], [10], its efficient implementations were enabled only recently [14]–[30] due to the emergence of novel continuous-state, nonvolatile, memory devices [31], [32]. Such memories enable very dense implementation of weights and in-memory computing for vector-by-matrix multiplication, the most common operation in machine learning. Among different candidates, the resistive switching memories, including phase change and conductive bridge memories, and metal-oxide memristors (also known as ReRAM or RRAM [31]) are perhaps the most promising due to their excellent scaling prospects. Their technology, however, is still in need of improvement, which is less of a problem for another excellent candidate, floating gate (FG) memories, e.g., those based on redesigned commercial-grade-embedded NOR flash [22], [32], [33]. Though planar FG cells are less dense than passively integrated memristors, their main advantage is FG cell amplification, which simplifies and reduces the overhead

of peripheral circuitry. It is worth noting that the limited endurance of memristors and FG memories is less of an issue for inference applications since the weights are typically reprogrammed infrequently.

In this article, we present a multipurpose inference accelerator, dubbed “aCortex,” which is designed to capitalize on in-memory mixed-signal computing with nonvolatile memories. Though the idea of employing a mixed-signal vector-by-matrix multiplier (VMM) based on nonvolatile memories for multipurpose inference accelerators is not new [23]–[27], our work is novel in several aspects. Its key advantage is more extensive use of analog computing not only for VMM computation but also for data transfer. Such an approach minimizes the area/energy/delay overhead of the sensing and data conversion peripheries that are key factors limiting the efficiency of the mixed-signal neural accelerators [23]–[30]. A more compact design, in turn, reduces communication overhead due to shorter distances for data transfer. Moreover, data transfer overhead is further reduced by using a configurable chain of buffers exploiting the data reuse for convolution operation, programmable data buses that can be efficiently tailored on the fly to a particular utilization of mixed-signal array, and a custom-designed refresh-free embedded dynamic random access memory (eDRAM) tailored to meet the retention time requirement. We also propose a simple and efficient instruction set architecture (ISA) along with a multiagent controller, which takes advantage of the eligible time-overlap between consecutive micro-operations while minimizing the instruction memory (IM) requirement. Finally, we developed a system-level estimator that imports the target network’s computational graph along with experimental and circuit-level simulation results for different architecture components, including digital-to-analog converters (DACs), analog-to-digital converters (ADCs), sense amplifiers, memory cells, digital blocks, and buses, maps the weight kernels onto the 2-D array of nonvolatile memory (NVM) blocks, and finally produces a comprehensive performance report considering various non-idealities, such as leakages and line parasitics. Using such a simulator, we perform a detailed performance analysis based on the actual layout in the 55-nm process with embedded NOR flash memory. Note that unlike many proposals based on emerging memory technologies, core components have been previously taped out using commercial processes and successfully tested, and we used such experimental data in our analysis.

In Section S.I in the Supplementary Material, we introduce today’s major neural layer types and present their hardware-friendly rearrangement targeting a weight-stationary implementation. The overall aCortex architecture and operation scheme, as well as the internal structure of its main components, are presented in Section II. This section also introduces the proposed ISA along with the controller architecture. More details on the ISA are provided in Section S.II in the Supplementary Material. The general framework for mapping applications into aCortex and our case study for three representative neural network inference tasks are provided in Section III. In Section IV, we provide the circuit diagram and experimental/simulation results for the FG-based implementation of aCortex’s core computing units in a 55-nm technology node. We then perform a design space exploration for

architectural parameters and provide a detailed system-level report for a semi-optimal design point. Related prior works are discussed and compared with aCortex in Section S.III in the Supplementary Material.

II. aCortex ARCHITECTURE

A. TOP-LEVEL ARCHITECTURE

As shown in Fig. 1(a), the major processor’s components are auxiliary unit (AUX), microcontroller, main memory, and two mixed-signal processing units (MSPUs). Each MSPU includes a configurable chain of input digital buffer blocks (DBBs), a flexible 2-D array of VMM blocks, and an array of output neuron blocks. The architecture can be loosely characterized as Harvard weight-major type [2]. The instructions are stored in a microcontroller’s dedicated SRAM-based IM. All frequently changing data, i.e., input, output, and temporary data, are kept in eDRAM-based main memory, while fixed weights, which would be typically precomputed at *ex-situ* training, are stored in NVM arrays of MSPU’s VMM blocks.

The inference task is specified by a program code based on custom instructions and the corresponding set of neural layer weight matrices. Assuming that the code is loaded and all weights are set up accordingly, the inference is computed by loading input data to the main memory, executing code to perform the inference task, and storing the computed results back in the main memory. In particular, the stationary weight matrices corresponding to various network layers are packed in the 2-D array of VMM blocks [see Fig. 1(b)], and the inference is performed in a layer-by-layer manner by sequentially reading the layer input from the main memory into the DBBs, activating the appropriate VMM and neuron blocks to perform the target neural layer, and then temporarily storing the intermediate results in the main memory for computing the next layer. Note that some of the neural layers, such as CNV and LSTM, require multiple VMM operations with various input data on the same weight matrix to complete. In this case, the corresponding VMM/neuron blocks are activated multiple times during the execution of each neural network layer—more details on that are provided in Section III and Section S.I in the Supplementary Material.

Flexible activation of VMM/neuron blocks enables the compact implementation of a set of neural network layers with various VMM sizes while maximizing the energy efficiency (EE) by cutting off the active power consumption of unutilized VMM/neuron blocks. Moreover, aCortex minimizes the energy overhead of data transfer by cutting off the unutilized portion of data buses and effectively reducing their length via disabling further data propagation. For example, Fig. 1(c) shows the active blocks and buses when the processor is computing the first layer of the neural network shown in Fig. 1(b).

All inference data, i.e., inputs and outputs of the MSPU’s VMM blocks, as well as effective weight precision, are p bit. A set of p bits is defined as one data *word*. We specifically consider $p = 4$, which is typically sufficient for running state-of-the-art image classification inference without loss of functional performance [34], [35]. Blocks and buses are sized according to a global architecture parameter K , defined as the granularity of neural computation and data transfer on aCortex.

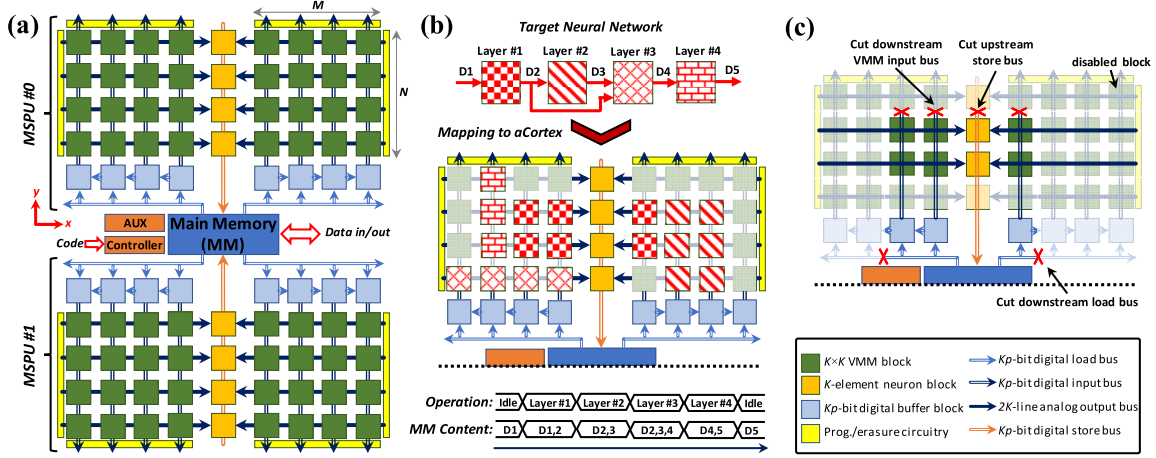


FIGURE 1. (a) aCortex top-level architecture. The location of the components crudely corresponds to the actual layout and is chosen to reduce data transfer overhead. For clarity, the architecture is shown for $N = M = 4$, and most of the control lines and the circuitry for testing/weight tuning are omitted. (b) Example of a weight kernel mapping on aCortex VMM blocks, layer-by-layer operation scheduling, and corresponding content of main memory over time. (c) aCortex active blocks/bus portions during the execution of layer #1 of the neural network shown in (b).

B. MAIN MEMORY

Main memory is implemented with eDRAM technology, whose retention time is tailored for a refresh-free operation. In particular, it is organized as an array of eDRAM blocks, each with a Kp -bit I/O data port. The multibanked structure allows us to read and write K data words simultaneously, i.e., to supply data to one DBB and receive data from one neuron block. The required memory capacity and its retention time are calculated by monitoring the memory content and the longest lifetime of intermediate data for the inference of the target neural networks [see Fig. 1(b)].

C. MIXED-SIGNAL PROCESSING UNIT: ARRAY OF VMM BLOCKS

Each MSPU is comprised of two N -by- M arrays (quadrants) of VMM circuit blocks located on each side of a column with N neuron blocks. Each VMM block features K -by- $2K$ array of NVM cells, which is suitable for implementing analog-mode differential K -by- K VMM operation; Kp -bit front-end DACs; and $2K$ back-end local sensing circuitry [see Fig. 2(a)]. The data to a single column of VMM block array are fed via Kp -bit wide digital programmable “VMM input” bus from the corresponding distributed memory buffer block. The VMM block outputs are connected via analog “VMM output” buses, which are $2K$ lines wide, to the corresponding neuron blocks. More circuit details on the data conversion and sensing/summation for the considered VMM design based on 2-D-NOR flash memory technology are provided in Section IV-A.

Column/row enable lines (denotes as CE/RE) span the MSPU quadrants in vertical/horizontal directions and are used to activate the desired VMM blocks at each processing step. Specifically, a given VMM block is activated only if its compute enable signal ($\text{CoE} = \text{CE} \wedge \text{RE}$) is equal to “1” [see Fig. 2(b)]. These control lines allow us to flexibly implement a wide range of VMM sizes (from $K \times K$ to $2MK \times NK$) while cutting off the active power consumption of unutilized VMM/neuron blocks. Moreover, a Kp -bit wide VMM input bus repeater with horizontally shared enable control line (RBE) is integrated into each VMM block [see Fig. 2(a)]

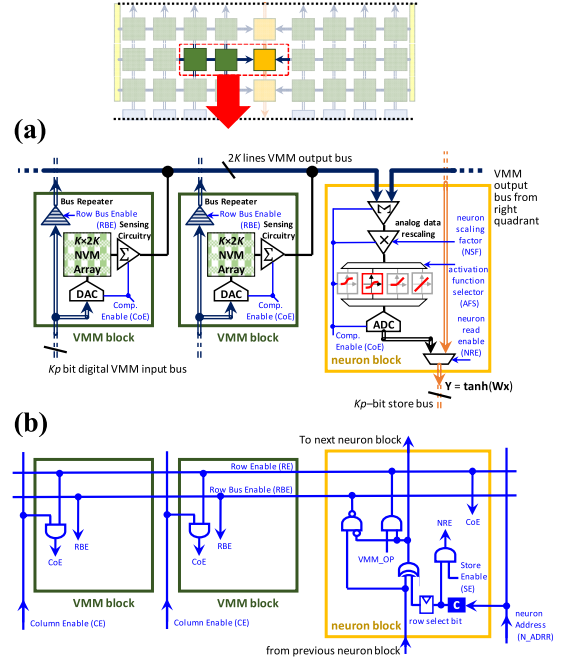


FIGURE 2. (a) Schematic of mixed-signal VMM and neuron blocks, including their connectivity and required control signals (blue). Neuron block includes four stages, namely, global sensing, rescaling unit, activation function, and ADC. (b) Control circuitry that facilitates flexible activation of the target VMM/neuron block and cuts unutilized portions of the VMM input and store buses. Square labeled with “c” denotes comparator for selecting a specific neuron. For clarity, some of the details, e.g., circuitry for setting up enable bits and some neuron control circuitry, are not shown.

to speed up data propagation as well as to minimize its energy overhead by cutting the unutilized portion of the VMM input bus. The logic circuitry and single-bit registers producing CE and RE/RBE control lines are integrated into the digital buffer and neuron blocks, respectively.

The programming/erasure circuitry consists of decoders and level-shifters that are shared by NVM arrays of VMM

blocks and are placed at the outer margins of the VMM block arrays [see Fig. 1(a)] and row/column access switches, placed at the periphery of each VMM block, controlling the applied signals.

D. MIXED-SIGNAL PROCESSING UNIT: NEURON BLOCK

A neuron block includes K identical neuron units. All units perform in parallel summation/integration of the analog data supplied from the corresponding VMM output lines, rescale the integrated data (if needed), apply the selected activation function, and finally convert the results to the digital domain using ADCs [see Fig. 2(a)]. The rescaling unit enables a wide range of quantization ranges, e.g., needed to operate with different VMM input sizes, via adjusting neuron's analog input amplitude to match the fixed operating input range of the activation functions and ADC units. The activation function can be selected from linear, rectified linear, sigmoid, or hyperbolic tangent types, to support a wide range of neural layers (see Section S.I in the Supplementary Material). Neuron block outputs are digitized with p -bit ADCs and temporarily latched in their embedded digital memory. The digital results are then passed via Kp -bit wide digital “store” bus to the main memory. Such data transfer, i.e., the “store” operation, is performed in one step per neuron block so that, e.g., a total of N steps are required to transfer data from all neurons in one MSPU. The specific number of processor cycles required for each step varies based on the location of the neuron block, i.e., its distance from the main memory. In particular, the store bus data are passed via a Kp -bit 2-to-1 multiplexer in each neuron block. These multiplexers act as a bus repeater for the utilized portion of the store bus and are also used to decouple and deactivate its unutilized portion.

The RE and RBE signals (which control the target neuron block and the corresponding VMM blocks it serves and cut the VMM input bus, respectively) are configured with “row-select” single-bit flip-flops of the neuron blocks [see Fig. 2(b)]. Specifically, the row-select flip-flops of neuron blocks are connected via xor gates, with the output of the xor gate directly controlling the RE line. To preselect RE signals for the contiguous set of rows (e.g., rows 10–15), row-select bits of neuron blocks in the first (10), closest to the main memory, and in the last+1 (16) row of the selected set of rows are set to “1,” while others are set to “0.” This implementation results in activating all selected rows (i.e., setting RE = 1 for rows 10–15). On the other hand, using a simple nand gate that detects a transition from selected to unselected row, RBE is deasserted only for the last+1 row of the selected group of rows [see Fig. 2(b)], which effectively cuts the downstream VMM input bus.

E. MIXED-SIGNAL PROCESSING UNIT: DIGITAL BUFFER BLOCK

The data to DBBs in each quadrant are supplied from the main memory via Kp -bit wide digital “load” bus [see Fig. 1(a)]. The DBB's internal logic circuitry is designed to flexibly support various data flow scenarios—from the simple load (Fig. 3(b): scenario #1) to “load and shift” configuration (Fig. 3(b): scenarios #2 and #3). The flexible “load and shift” configuration enables efficient computation of CNV layers with a wide range of specifications (filter size and

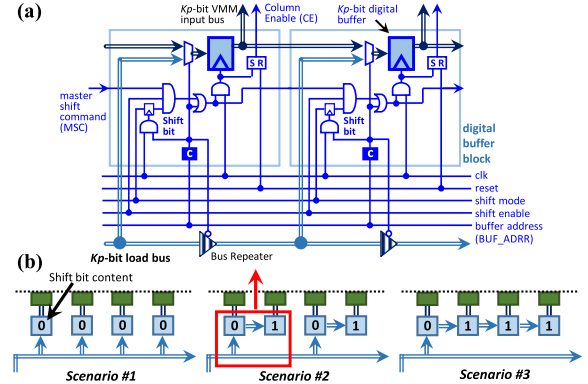


FIGURE 3. (a) Detailed circuit implementation of the DBB. The control circuitry is shown in blue color. Square labeled with “c” denotes comparator for selecting a specific DBB. For clarity, some of the details are not shown. (b) Examples of three DBB chain configurations scenarios. Specifically, scenario #1 shows loading (sequentially) data exclusively from the load bus. Scenarios #2 and #3 show loading data from load bus into DBBs and simultaneously shifting data between DBBs connected in a two-block shift register (#2) or a four-block shift register configuration (#3).

stride) while taking advantage of the rowwise data reuse [see Fig. S1b in the Supplementary Material).

Specifically, the input to each DBB is supplied by a Kp -bit-wide 2-to-1 multiplexer, which selects the input source between the load bus (i.e., main memory) and the previous DBB. For simplicity, we assume that only one DBB can be loaded from the load bus at one step, by setting the address of the target DBB on the buffer address bus. Similar to the data transfer from neurons, the number of processor cycles required for each step varies based on the location of the DBB.

The shift operations are supported by properly configuring a “shift-bit” flip-flop in each DBB and are masked by the “shift enable” line. In particular, the shift register configurations are programmed by setting shift bits to “1” for all the DBBs except for the first block of shift register [see Fig. 3(b)]. Setting shift bit to “1” of a particular DBB is performed concurrently with loading that DBB from load bus by asserting “shift mode” control line. (Note that rowwise execution of the CNV layer typically starts by loading all DBBs with row data from main memory, with all shift operations disabled by deasserting shift enable signal line so that all shift bits are typically configured at this time.) Once shift bits are properly configured, simultaneously with loading new data into DBB, the already loaded data in DBBs can be shifted between the remaining blocks of that shift register [e.g., shifted to the right in scenarios #2 and #3 of Fig. 3(b)]. This is performed with an internal “master shift command” signal that enables latching data from previous DBBs of the specific shift register whenever its shift bit and enable signals are set to “1.”

In addition, each DBB has a single-bit SR latch to specify the target VMM block columns, i.e., to set column enable (CE) signal, which is configured similar to shift-bit flip-flop. It also has a bus repeater on the load bus, which is disabled when the DBB is loaded, to stop downstream data propagation on the bus [see Fig. 3(a)].

F. MIXED-SIGNAL PROCESSING UNIT: VMM COMPUTATION

Assuming that the target VMM columns are preselected ($CE=1$) and the desired row-select bits are set, VMM operation is performed by applying a positive pulse to the “VMM_OP” command signal. This signal enables the neuron blocks associated with the selected rows as well as their input VMM blocks for which $CE = 1$ via propagating through the RE and then CoE lines. Naturally, the VMM_OP pulsewidth should be longer than the worst case end-to-end VMM operation time, i.e., the time it takes from the moment inputs are applied to DAC to that of latching ADC outputs.

G. AUXILIARY UNIT

Auxiliary unit (AUX) is provisioned to perform less frequent digital computations in neuromorphic inference. In particular, this block is used to perform in parallel Kp -bit vector-by-vector operations, such as additions, subtractions, and fixed-precision multiplications. It also performs comparison in max-pooling operation, which is typically used in CNNs. (An average pooling, another typical operation, can be implemented directly in a mixed-signal domain with properly adjusted weights.) AUX consists of an array of arithmetic logic units (ALUs), multipliers, and internal registers. During max-pooling operation, the register holds the current maximum value and feeds it to one of the ALU's inputs to compare with the next input value fed from the main memory through the load bus. The outputs of AUX blocks can be written back to the main memory via the store bus.

H. CONTROLLER AND INSTRUCTION SET ARCHITECTURE

The most typical operations on aCortex involve loading multiple DBBs with data from the main memory, performing the vector-by-matrix computation, and then moving the results from neuron blocks back into the main memory. Accordingly, aCortex controller includes three separate agents as loader, operator, and collector, each dedicated to performing one of these frequently used operations [see Fig. 4(a)]. These agents are configured and synchronized using the main controller. Such multiagent structure shrinks the code size, enables eligible time overlapping between these operations, and minimizes the engagement of the main controller, which, in turn, reduces the controller time overhead.

The main components of each agent are listed in Fig. 4(c). The loader is responsible for reading data from the main memory into the DBB/AUX blocks using the load bus. This agent uses two counters to support burst mode in which multiple data packs with a preconfigured main memory/DBB initial addresses and strides are loaded. Moreover, the loader uses a load bus counter to adjust the load time based on the physical distance between the main memory and the target DBB. Such distance is extracted from the most significant bits of the DBB address considering that the locations of neuron and DBBs are counted from the center of the chip outwards, starting from 1. For example, $\text{load_time} = \text{memory_read_time} + \text{ceiling}(\text{DBB_address}/32)$ for the case when passing data on the load bus through 32 DBBs takes one controller cycle. The collector that is responsible for storing back the neuron/AUX's output to the main memory has a similar

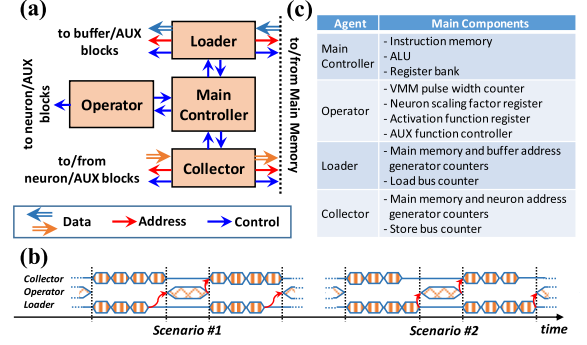


FIGURE 4. (a) Top-level representation of the multiagent controller and connectivity between different agents and the rest of the architecture. (b) Examples of agent operation timing diagram when the load operation is completed before (scenario #1) and after (scenario #2) the collect operation. (c) Main components of each controller's agent.

structure to the loader. The operator has an embedded counter to produce the appropriate pulsewidth of the VMM_OP control signal for the VMM operation. It also holds the neuron scaling factor, the type of activation function selector, and AUX function control bits. The main controller instructions are further detailed in Section S.II in the Supplementary Material.

III. APPLICATION MAPPING

Application mapping process of a given neural network on aCortex involves checking accelerator resource requirements (i.e., main memory and MSPU) and producing a weight mapping and machine code to set up the accelerator. A similar process is also used in the accelerator design to estimate the required main memory and MSPU resources and to optimize their specifications (M, N, K) for the target range of applications. Such a process starts with extracting the computational graph and finding the optimal topological partitioning and ordering of the computational steps [see Fig. 5(a)]. In such a graph, the vertices are computation kernels (i.e., neural layers), while the edges are input/output data. In general, finding the optimal sorting of a given graph is an NP-hard problem. However, the layer-by-layer operation scheme of aCortex, resulted from our energy-optimal design, reduces the complexity of the problem. After obtaining the graph and scheduling order, the next step is to check the memory requirement throughout the inference steps. In aCortex, 3-D data are stored in the main memory in a row first, column second, and channel/feature map third order [see Fig. 5(b)]. Following this order, multiple channels of one data pixel (and adjacent pixels in a row in the case of CNV stride > 1) are always grouped into K -word-long data packs (see Fig. S1b in the Supplementary Material). One pack of K -word data is then mapped into a word line of the main memory block and can be read/written simultaneously. Such data placement in the main memory enables burst mode read/write using the controller's loader/collector. Note that, due to quantization, such a scheme may result in underutilized memory, e.g., when the number of channels/feature maps is not divisible by K .

Considering such data arrangement, the memory usage after each inference step (i.e., neural layer) is calculated by

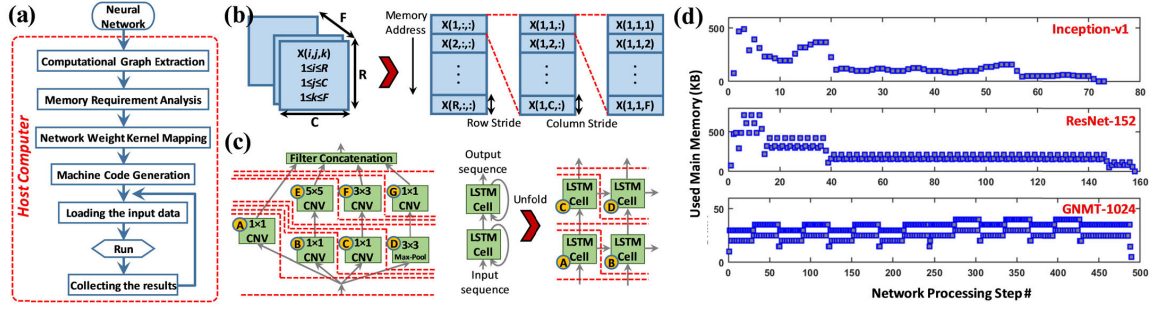


FIGURE 5. (a) Application mapping flowchart performed on the host computer. (b) Mapping scheme of a 3-D data structure into the aCortex's main memory. (c) Example showing computational graph cuts for evaluating the amount of main memory occupied during various steps of inference for a single Inception layer (left) and a multilayer LSTM network (right). (d) Utilized the main memory graph as a function of the network processing step for the studied benchmark networks.

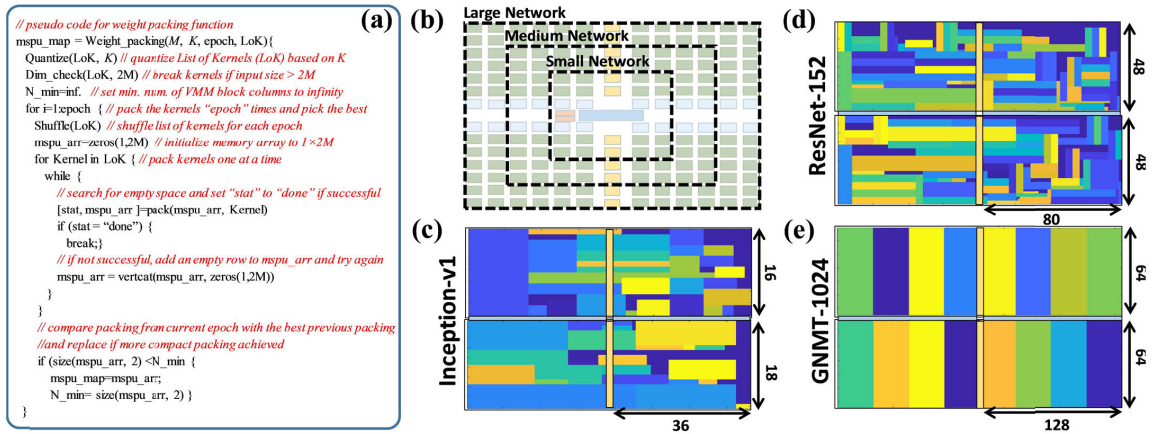


FIGURE 6. (a) Pseudocode for aCortex weight kernel packing algorithm. (b) Preferred mapping locations of various neural network sizes. (c)–(e) Weight kernel mapping results for Inception-v1, ResNet-152, and GNMT-1024 for $K = 64$. Each pixel shows one VMM block colored according to the neural layer occupying it.

drawing a cut in the computational graph that separates the already computed portions of the graph (network processing steps) from the upcoming ones and adding up all the edges that are crossed. Since, during the execution of each layer, both its inputs and outputs are present in the main memory, the upper bound for the total memory usage is calculated by adding up the memory usages for two consecutive cuts and subtracting their overlap edges.

We applied this algorithm to three studied networks, namely, Inception-v1 [36] and ResNet-152 [37] DNNs for image classification, and GNMT [38], Google's neural machine translation network featuring a 16-layer LSTM network with bidirectional encoder layers, with the vector length of 1024 and the sequence size of 10. For example, Fig. 5(c) shows such a process for an Inception neural layer and an unfolded two-layer LSTM network. As shown in Fig. 5(c), the memory requirement is limited by the initial layers of the DNNs (for 4-bit computation with the data pack quantization of $K = 64$).

In the next step, the weight matrices are mapped into MSPUs using a greedy search algorithm [see Fig. 6(a)], for which input parameters are the number of available MSPU's columns (M'), architecture granularity parameter (K), the number of tries (*epoch*), and the list of weight kernels (*LoK*). In one iteration of the algorithm, the kernels are first randomly ordered and then greedily mapped

in a row-first manner, in the given order, to the array of VMM blocks. Such procedure is repeated *epoch* times, and mapping configuration with the smallest number of occupied VMM blocks is selected. Furthermore, the mapping process is repeated for different $M' \leq M$ to search for a square-shaped mapping of occupied VMM blocks [see Fig. 6(b)] to minimize the average data transfer distance between MSPU's active VMM/neuron blocks and the main memory (and hence to increase energy efficiency)—see more discussion on that in Section IV-B. The results of this optimization process for our target networks are shown in Fig. 6(c)–(e).

The mapped locations of weight kernels are then used in the compilation process to generate the machine code for handling the data flow and operation of the network. The number of instructions for initializing and operating each layer type is estimated for the system-level analysis based on the proposed ISA (presented in Section II-H and Section S.II in the Supplementary Material).

IV. CIRCUIT DESIGN AND PERFORMANCE EVALUATION

A. 55-nm FG-BASED IMPLEMENTATION OF MIXED-SIGNAL BLOCKS

While aCortex can be realized with a variety of NVM technologies, the focus of this article is on a mature industrial-grade flash-memory technology that has already enabled

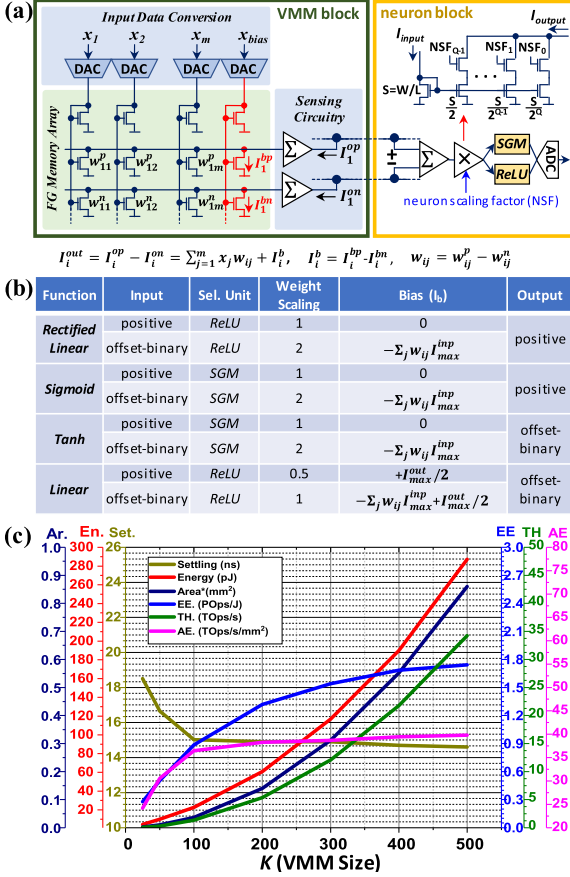


FIGURE 7. (a) VMM/neuron schematic for the FG-based mixed-signal universal neural computing scheme supporting both positive/negative input/output as well as various (non)linear activation functions. (b) Configuration of the computing elements for various scenarios (desired activation function and input type). (c) Performance results for a VMM connected to a neuron block as a function of K assuming 4-bit computing precision, maximum cell current of 16 nA, the maximum current of 1 μ A and unity gain for local sensing, maximum ADC input current of 5 μ A.

extremely compact and energy-efficient implementations of mixed-signal circuits [20]. The key advantage of such implementation, compared with, e.g., those based on resistive switching devices [31], is the FG cell's inherent signal amplification and low operation currents that greatly relax the requirement for sensing circuitry gain and enable very compact peripheral circuits [20]. More specifically, the proposed architecture was evaluated for the implementation based on ESF-3 (embedded split-gate flash) technology in which FG cells were redesigned for analog computing applications [17].

The proposed mixed-signal circuits allow for efficient implementation of different types of activation functions, kernel sizes, and quantization ranges, which are essential features of multipurpose inference accelerator [see Fig. 7(a)]. In particular, the VMM operation is implemented using a gate-coupled design [17]. In such an approach, inputs and outputs are both encoded as currents, and the weights are encoded as the memory state of the array FG devices.

The front-end input conversion is realized using current steering (CS) DAC architecture, a viable choice consid-

ering its low power consumption, compact footprint, and fast turn ON/OFF time at relatively low precision. A 4-bit PMOS-based CS DAC circuits source the current into peripheral FG devices.

The developed current-mode global sensing circuit of a neuron block has excellent wideband current following behavior and provides very low input impedance while limiting drain voltage distortion. To reduce process variation overhead associated with an offset error, two additional FG devices are provisioned in each channel (drain line) and are used either to source or sink the input-referred offset current.

The input current scaling of a neuron block is implemented using a binary-weighted current mirror structure controlled by the multibit digital input [see Fig. 7(a)]. As already mentioned in Section II.D, this feature is needed to adjust quantization range in accordance with the maximum VMM circuit output currents (which would vary, e.g., with the size of the weight kernels) and, hence, to minimize losses in functional performance due to quantization of activation function outputs.

The 4-bit current-mode ADC has a 1-bit per stage cyclic design that generates the 4-bit digital output in four cycles. FG transistors are also employed for offset/compensation in the high-speed comparator and to generate reference currents.

The sigmoid and ReLU activation functions are implemented directly in the ADC, without using any other additional circuitry, by appropriately choosing reference currents, i.e., quantization levels of the ADC. Linear and hyperbolic tangent functions are emulated via input biasing and appropriate weight scaling [see Fig. 7(b)]. DACs/ADCs are designed for unipolar data and can be reused, without any circuit modification, for bipolar data by utilizing offset-binary representation. More details on the circuit structure for the design of PVT-resilient CS DAC circuit and algorithmic ADC can be found in [20].

Fig. 7(c) provides the VMM performance results as a function of K based on the measurements of ESF3 memory devices and postlayout simulations of peripheral circuitry. These data are used to estimate system-level performance.

B. MAIN MEMORY AND BUSES IN 55-nm TECHNOLOGY NODE

The main memory is implemented using asymmetric 2T gain embedded DRAM cells with boosted power supply [39]. The retention time of eDRAM cells was changed to 100 μ s by reducing leakage and adjusting biases, with a 99.9% bit yield confirmed by block-level Monte Carlo simulations [39]. The memory performance was modeled using the CACTI tool [40]. We also developed a bus area/energy/delay model as a function of bus length and repeater size using postlayout simulations considering all device/interconnect parasitics.

C. SYSTEM-LEVEL RESULTS AND DESIGN SPACE EXPLORATION

We have developed a software framework that utilizes the postlayout energy/speed/area metrics for all the aCortex's building blocks (buffers, buses, DACs, ADCs, integrators, and digital circuits) to evaluate the system-level performance for any target DNN/RNN network. This framework uses the list of processing tasks for a given neural network to map the

VMM kernels on the NVM devices embedded in the VMM blocks and then generate a detailed performance report for the given set of architecture specifications. Using such a tool, we have performed a preliminary exploration of architectural parameters (i.e., K and MSPU aspect ratio, $AR = M/N$) to optimize the processor performance for the aforementioned target neural networks. A detailed study of these benchmark networks (see Figs. 5 and 6) has shown that a 1MB MM is sufficient to store all intermediate data, while the flow control program requires at most 4KB of IM. Moreover, the controller energy/delay is estimated in an instruction-by-instruction manner in which the required machine code for initializing and performing each layer type has been evaluated.

Fig. S2 in the Supplementary Material shows the system-level EE, throughput, and area with respect to K and AR for these networks. Larger K typically results in higher throughput due to wider bus widths and consequently higher data transfer rate. It also improves the EE by reducing the VMM block peripheral circuitry energy consumption (trend clearly seen in Fig. S2c in the Supplementary Material). Moreover, the increase in throughput results in lower leakage energy, which further increases EE. However, for the networks with medium and small weight kernels (i.e., ResNet-152 and Inception-v1), larger K results in underutilization of active blocks and buses as well as an increase in the number of required VMM blocks to map the network, which, in turn, increases the energy overhead of the analog peripheral circuitry and buses. As such negative effects outweigh the positive ones, the overall system-level EE and throughput are getting worse for larger K . For DNNs, the load bus delay/energy typically plays a more significant role in the system throughput/EE compared with the store bus due to a higher input-to-output data transfer ratio in convolution operations. Such property leads to higher throughput/EE for smaller AR (i.e., relatively shorter load bus) in these networks. The opposite of this trend is observed for GNMT in which the LSTM layers have a smaller input-to-output data transfer ratio. Note that such trends do not consistently hold due to network-specific weight packing efficiency with respect to AR .

The accelerator area decreases for larger values of K due to the larger sharing factor of analog peripheries, hence smaller area overhead of DACs and sensing circuitry (see Fig. S2 in the Supplementary Material). This trend does not hold for Inception-v1 in which smaller weight kernels result in the VMM block underutilization and lower weight packing efficiency which outweigh the gain in peripheral circuitry area efficiency.

The results for different networks indicate $K = 64$ and $AR \approx 2$ as a semi-optimal design point for which block utilization and load/store bus energy/delay are somewhat balanced. Detailed performance report and area/energy breakdowns for this design point are presented in Table 1. As these results show, the energy consumption is dominated by data transfer and intermediate data storage for DNNs with smaller size weight kernels, such as Inception-v1 and ResNet-152 networks. On the other hand, the energy consumption is dominated by sensing circuitry and DACs, even despite larger accelerator area (and, hence, larger data transfer energy consumption) for GNMT inference task. This is because

TABLE 1. aCortex system-level results and breakdowns *.

	Inc.-v1	ResNet-152	GNMT
Network Specification			
Number of Parameters	7.2×10^6	5.52×10^7	1.3×10^8
Number of Operations	5.2×10^9	2.0×10^{10}	2.6×10^9
Architectural Specification			
K	64	64	64
M	38	80	128
N (top/bottom)	16/18	48/48	64/64
Main Memory Capacity	1 MB		
Number of Memory R/W	3.3×10^5	8.1×10^5	1.7×10^4
Main Memory Utilizat. (%)	47.8	59.8	5.07
VMM block Utilization (%)	67.5	87.6	100
Area Breakdown (%)			
Main Memory	17.3	4.4	2.2
Sensing Circuitry	15.6	23.5	25.1
FG Arrays	24.5	36.8	39.3
DACs	4.5	6.8	7.3
Neuron Blocks	0.06	0.04	0.03
Programming/Erase**	26.5	14.2	11.3
Others	11.5	14.2	14.8
Energy Breakdown (%)			
Main Memory	36.3	23.2	10.9
Sensing Circuitry	16.4	12	33.6
FG Arrays	3.5	2.5	7.1
DACs	6.8	4.9	13.8
Neuron Blocks	0.8	1.1	0.9
Buses	30.1	41.4	16.2
Leakage	3.3	13.5	16.8
Others	2.8	1.4	0.7
Performance Summary			
Area (mm ²)	37	146	293
Power (mW)	20.6	26.5	44.5
Inference Time (ms)	2.37	6.72	0.16
EE (TOP/J)	97	106.2	335.8
Throughput (TOP/s)	2.00	2.9	14.94

* Because of more optimal overlapped execution and refined estimates for DAC circuits, the numbers are slightly adjusted compared to preliminary ones reported in Ref. [19]. ** The area overhead of programming/erasure circuitry is estimated as $(245.5M + 120N) \times K \mu\text{m}^2$.

GNMT inference involves larger size VMM operations (weight kernels), which ultimately leads to a larger compute-to-communication ratio and allows us to take better advantage of analog-domain computing. Also, note that area overhead of high-voltage programming/erasure and VMM peripheral circuits is quite low (compared with other NVM-based accelerators) due to their effective sharing, and, in fact, the area of aCortex is dominated by FG memory cells. Thus, the detailed results show that the integration density is one of the key properties of memory devices for the energy-efficient inference accelerators.

The performance comparison of aCortex against its major fully digital [5]–[7] and mixed-signal [23], [24] competitors shows that aCortex achieves significantly higher performance, especially for mobile/IoT applications, for which the storage efficiency (MB/mm²) and EE are the most important metrics (see Fig. S3 in the Supplementary Material). In order to make a fair comparison, we performed a highly optimistic projection of the performance metrics for the mixed-signal architectures to 55-nm, 4-bit design point. According to these estimations, aCortex achieves $\sim 28 \times / \sim 65 \times$ improvement in EE over ISAAC [23]/PUMA [24] while maintaining a comparable SE and enduring a relatively small drop in throughput ($\sim 0.3 \times / \sim 0.4 \times$). Note that these architectures do not consider the overhead of programming/erasure circuitry that could impact the performance results.

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [2] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.
- [3] N. P. Jouppi, C. Young, N. Patil, and D. Patterson, "A domain-specific architecture for deep neural networks," *Commun. ACM*, vol. 61, no. 9, pp. 50–59, Aug. 2018.
- [4] *Investor Day Presentation*, NVIDIA Corp., Santa Clara, CA, USA, 2017.
- [5] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. Int. Symp. Comput. Archit. (ISCA)*, Toronto, ON, Canada, Jun. 2017, pp. 1–12.
- [6] Y. Chen *et al.*, "DaDianNao: A machine-learning supercomputer," in *Proc. 47th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Cambridge, MA, USA, Dec. 2014, pp. 609–622.
- [7] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019.
- [8] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.
- [9] C. Mead, *Analog VLSI and Neural Systems*. Boston, MA, USA: Addison-Wesley, 1989.
- [10] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Comput.*, vol. 10, no. 7, pp. 1601–1638, Oct. 1998.
- [11] S. Chakrabarty and G. Cauwenberghs, "Sub-microwatt analog VLSI trainable pattern classifier," *IEEE J. Solid-State Circuits*, vol. 42, no. 5, pp. 1169–1179, May 2007.
- [12] G. Indiveri *et al.*, "Neuromorphic silicon neuron circuits," *Frontiers Neurosci.*, vol. 5, May 2011, Art. no. 73.
- [13] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Frontiers Neurosci.*, vol. 7, Sep. 2013, Art. no. 118.
- [14] F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, and D. Strukov, "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits," *Nature Commun.*, vol. 9, no. 1, Dec. 2018, Art. no. 2331.
- [15] M. J. Marinella *et al.*, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 86–101, Mar. 2018.
- [16] C. Li *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electron.*, vol. 1, no. 1, Jan. 2018, Art. no. 52.
- [17] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Austin, TX, USA, Apr./May 2017, pp. 1–4.
- [18] M. Bavandpour, M. R. Mahmoodi, and D. B. Strukov, "Energy-efficient time-domain Vector-by-Matrix multiplier for neurocomputing and beyond," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 9, pp. 1512–1516, Sep. 2019.
- [19] M. Bavandpour *et al.*, "Mixed-signal neuromorphic inference accelerators: Recent results and future prospects," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2018, pp. 20.4.1–20.4.4.
- [20] M. R. Mahmoodi and D. Strukov, "An ultra-low energy internally analog, externally digital vector-matrix multiplier based on NOR flash memory technology," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, San Francisco, CA, USA, Apr. 2018, Art. no. 22.
- [21] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Jul. 2015.
- [22] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2017, pp. 6.5.1–6.5.4.
- [23] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with *in-situ* analog arithmetic in crossbars," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Seoul, South Korea, Jun. 2016, pp. 14–26.
- [24] A. Ankit *et al.*, "PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Proc. 24th Int. Conf. Architectural Support Program. Lang. Operating Syst.*, Providence, RI, USA, Apr. 2019, pp. 715–731.
- [25] X. Liu *et al.*, "RENO: A high-efficient reconfigurable neuromorphic computing accelerator design," in *Proc. Design Autom. Conf. (DAC)*, San Francisco, CA, USA, Jun. 2015, pp. 1–6.
- [26] X. Liu *et al.*, "Harmonica: A framework of heterogeneous computing systems with memristor-based neuromorphic computing accelerators," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 5, pp. 617–628, May 2016.
- [27] L. Song, X. Qian, H. Li, and Y. Chen, "PipeLayer: A pipelined ReRAM-based accelerator for deep learning," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Austin, TX, USA, Feb. 2017, pp. 541–552.
- [28] P. Chi *et al.*, "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Seoul, South Korea, Jun. 2016, pp. 27–39.
- [29] M. Imani, M. Samragh, Y. Kim, S. Gupta, F. Koushanfar, and T. Rosing, "RAPIDNN: In-memory deep neural network acceleration framework," 2018, *arXiv:1806.05794*. [Online]. Available: <http://arxiv.org/abs/1806.05794>
- [30] H.-Y. Chang *et al.*, "AI hardware acceleration with analog memory: Microarchitectures for low energy at high speed," *IBM J. Res. Develop.*, vol. 63, no. 6, pp. 8:1–8:14, Nov. 2019.
- [31] D. B. Strukov and H. Kohlstedt, "Resistive switching phenomena in thin films: Materials, devices, and applications," *MRS Bull.*, vol. 37, no. 2, pp. 108–114, Feb. 2012.
- [32] F. M. Bayat, X. Guo, M. Klachko, N. Do, K. Likharev, and D. Strukov, "Model-based high-precision tuning of NOR flash memory cells for analog computing applications," in *Proc. 74th Annu. Device Res. Conf. (DRC)*, Newark, DE, USA, Jun. 2016, pp. 1–2.
- [33] F. M. Bayat, X. Guo, H. A. Om'mani, N. Do, K. K. Likharev, and D. B. Strukov, "Redesigning commercial floating-gate memory for analog computing applications," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Lisbon, Portugal, May 2015, pp. 1921–1924.
- [34] A. Mishra, E. Nurvitadhi, J. J. Cook, and D. Marr, "WRPN: Wide reduced-precision networks," 2017, *arXiv:1709.01134*. [Online]. Available: <http://arxiv.org/abs/1709.01134>
- [35] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," 2016, *arXiv:1609.07061*. [Online]. Available: <http://arxiv.org/abs/1609.07061>
- [36] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun./Jul. 2015, pp. 1–9.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun./Jul. 2016, pp. 770–778.
- [38] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: <http://arxiv.org/abs/1609.08144>
- [39] K. C. Chun, P. Jain, T.-H. Kim, and C. H. Kim, "A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches," *IEEE J. Solid-State Circuits*, vol. 47, no. 2, pp. 547–559, Feb. 2012.
- [40] N. Muralimanohar, R. Balasubramanian, and N. P. Jouppi, "CACTI 6.0: A tool to understand large caches," HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2009-85, 2009.