A CIM-Digital Heterogeneous Accelerating System with Analog Interconnection for Neural Networks

Rui Xu, Zikun Xiang, Linfeng Tao and Xi Jin

Key Laboratory of Strongly-Coupled Quantum Matter Physics, Chinese Academy of Sciences, School of Physical Sciences, University of Science and Technology of China, Hefei, Anhui, China
Email: xray@mail.ustc.edu.cn, jinxi@ustc.edu.cn

Abstract. In recent years, computing in memory (CIM) has been regarded as a promising competitor for low-power accelerators for neural networks. Different implementations based on various memories like ReRAM and flash memory have been proposed. However, a ‘cloud’ is hovering on the horizon of CIM. The limited frequency and high power consumption of DACs and ADCs have become major obstacles of further improvement of the power efficiency of CIM accelerators. The method of utilizing a more advanced process to increase the frequency and reduce the power consumption of DACs and ADCs can hardly be used because of the tight coupling of memory process and CMOS logic process in CIM accelerators. To solve this problem, this paper proposes a CIM-digital heterogeneous neural network accelerating system with analog interconnection. It decouples the CMOS logic process and the memory process in CIM accelerators by using analog interconnection, and implements ADCs and other units in an independent CMOS process. According to experimental results, compared to prior CIM accelerators, the proposed architecture can achieve an increase in the sampling rate of 26.7 times, to 100 MHz, and an increase in the performance of 26.7 times accordingly. The power efficiency can be improved 5.8 times.

1. Introduction
The rising of edge computing and on-device computing breads a large number of applications of neural networks and provide a great possibility for putting neural networks to work in the real world. However, embedded systems at edge and end side pose high demands on power consumption of neural network accelerators and accelerating systems. Applications such as face verification on mobile and wearable devices which are generally powered by a battery and require durability, pose a stringent requirement on power consumption. Compared to traditional digital accelerators, emerging computing-in-memory (CIM) technology based on resistive random-access memory (ReRAM) [1-3], NOR flash [4-6], and other non-volatile memories (NVMs) [7-9] saves the entire neural network model on chip, and avoids high bandwidth requirements and additional power consumption incurred by repeatedly moving data between the processing unit and the memory unit. Moreover, CIM technology utilizes the high density of memory cells of NVMs and makes computation more power-efficient by using analog characteristics of devices instead of digital methods. A power efficiency of over 100 TOPS/W, which is way ahead traditional digital accelerators, has been achieved by CIM-based accelerators. Therefore, CIM technology has great potential in promoting further development of neural networks at edge and end side.

However, there are still challenges in the application of CIM technology in neural networks. First, as shown in Figure 1, in addition to a memory array for computing and some digital logic circuits, a CIM system includes many digital-to-analog converters (DACs) and analog-to-digital converters.
(ADCs), which are used to generate input signals and measure output signals respectively. In fact, in a common CIM architecture, usually the DACs and ADCs take up a majority of the area and power consumption of the entire system and even low precision converters dominate in the area and power consumption of the system [10]. According to [11], ADCs and DACs can take up 85% to 98% of the total area/power of a neuromorphic computing system. Second, memory process and CMOS process in a CIM accelerator are tightly coupled because they are designed in the same chip. In terms of requirements of CIM itself, it may not be a good choice to make a fabricating process for CIM as advanced as possible for reasons of write endurance, date retention, analog characteristics, etc. Thus, CMOS-compatible memory processes have significant limitations. Embedded memory processes also have their limitations like lower density, the process node lagged behind CMOS process and limited choices. Finally, CIM is usually used for computing convolutional layers and fully connected layers that are transformed to general matrix multiply (GEMM) or directly decomposed to multiply and accumulate (MAC) operation, but for operations such as pooling and local response normalization (LRN) which cannot be simply implemented in CIM, digital processing unit are required. As the frequency of the digital processing unit is limited by the CMOS process, the performance of these operations is also limited. To sum up, the lagged CMOS process has limited the further improvement of performance and power efficiency of CIM accelerators. If the memory processes and the CMOS process in CIM accelerators can be decoupled and a more advanced CMOS process is used, the performance and energy efficiency of the CIM accelerator can be greatly improved.

![CIM Macro](image)

**Figure 1.** A Typical CIM System

![MCM Diagrams](image)

**Figure 2.** (a) A 2D MCM Connected with Bond Wires; (b) A 2D MCM Connected with Flip-chip; (c) A 3D MCM Connected with Bond Wires and Flip-chip; (d) A 3D MCM Connected with Flip-chip and TSVs
Figure 3. (a) A CIM system with Two Modules Connected with Digital Interfaces; (b) A CIM System with Two Modules Connected with Analog Interfaces

The development of semiconductor packaging technology makes it possible to package multiple chips into a package, i.e., multi-chip package or multi-chip module (MCM). Figure 2 shows 2D and 3D MCMs with different configurations. An obvious solution for process decoupling in CIM accelerators is using two modules and packaging them into one package. But as shown in Figure 3a, with common digital interconnection, DACs and ADCs will have to be in the same module with the memory array and cannot take advantages of the advanced process of Module 2. According to [12], with stacked ICs and packages (SIP) technologies like 2.5D ICs and 3D ICs, the length of connection in a package can be further shortened and the impedance and noise can be reduced to an extent that satisfies the requirements for transferring high precision analog signals in a CIM system. Based on this, this paper proposes a CIM-digital heterogeneous neural network accelerating system architecture with analog interconnection, as shown in Figure 3b. By using analog interconnection, the memory process in CIM accelerators is decoupled from the CMOS process of other circuits, and DACs, ADCs, digital processing units and the like are implemented on an independent chip using a more advanced CMOS process, thereby greatly reducing power consumption of CIM accelerators and improving their performance. The main contributions of this paper are: (1) proposing a CIM-digital heterogeneous architecture with analog interconnection for neural network accelerating systems; (2) designing an analog interface for CIM accelerators; and (3) using a CIM chip and an FPGA to implement and verify this architecture.

2. Related Work

CIM accelerators based on NVMs for neural networks has been widely studied these years. Different NVMs, such as NOR flash, ReRAM, and MRAM, are used for CIM because of their different characteristics. Wherein, ReRAM and NOR flash are the most commonly employed NVMs because of their good characteristics or mature manufacturing process. CIM accelerators already have achieved excellent performance based on these two NVMs.

ReRAM has excellent characteristics of high speed, and high density, making it a good choice for CIM. RAND chip [1] is a binarized ReRAM-based CIM processor which employs a current-controlled writing scheme and a flexible network architecture to obtain high accuracy and to support any neural networks. RAND chip achieves 90.8% on the MNIST dataset [13] with a multi-layer perceptron (MLP) and a power efficiency of 66.5 TOPS/W. A hardware-driven binary-input ternary-weighted network using positive and negative pseudo-binary nvCIM macros is proposed in [2]. With a 65 nm CMOS logic process, it achieves a path-delay of 15.6 ns and MNIST numerical recognition rate. A multi-bit CIM macro of [2] is proposed in [3]. Instead of using two macros, two columns in a macro are used as positive and negative columns. The chip fabricated in a 55 nm CMOS process achieves an 88.52% inference accuracy on the CIFAR-10 dataset [14] and a peak power efficiency of 21.9 TOPS/W. Although it is CMOS-compatible, ReRAM is immature and suffers from retention and stability problem [15].

NOR flash has been used in not only researches on CIM but also many commercial companies interested in CIM because it is mature and stable, low-cost, and good in signal characteristics. However, a NOR flash memory is not CMOS-compatible and NOR flash process is far behind CMOS process at present. An energy-efficient analog mode CIM macro with a digital input/output interface
and configurable precision are proposed in [6]. Circuit-level optimizations are performed for sensing circuitry and ADCs to achieve high performance and power efficiency. According to post-layout simulation based on a 55 nm embedded NOR flash memory process, up to 400 MHz operation and 1.68 POPS/J power efficiency can be achieved. A NOR flash-based CIM chip has been proposed in [4] and achieves 94.7% classification fidelity on the MNIST dataset.

Other memories [7–9] are also used for CIM based on a similar principle to their NOR flash and ReRAM counterparts.

3. CIM-Digital Heterogeneous Accelerating System with Analog Interconnection

3.1. System Architecture

Figure 4 shows the architecture of a CIM-Digital heterogeneous system with analog interconnection. The system includes two modules, a CIM Module in memory process and a Digital Module in an advanced CMOS process. The CIM Module can be implemented based on various NVMs and the Digital Module can be implemented as an ASIC or based on an FPGA. The two modules are packaged into one package through the SIP technology to reduce the interference of noise and impedance. Considering characteristics of CIM and digital processing units, the convolutional layers and fully connected layers of the neural network are performed on the memory array in CIM Module, and operations such as accumulation, activation function, pooling, and normalization and the like are performed by digital processing units. The two modules are connected by two digital interfaces for transferring the input data and control instructions for the CIM Module, and one analog interface for transferring analog signals representing results.

All key circuits other than measurement circuits of CIM are implemented in the CIM module. A CIM Module includes one or more CIM macros used for computing and circuits used for writing weights to a memory array. Since the binarized neural network like BNN [16] and XNOR-NET [17] are used, the input signals of the memory array can only be “0” or “1”. Therefore, instead of DACs, wordline drivers (WLDRs) can be used to generate a “high” or “low” signal. Then, output signals of the memory array that represent the results are generated according to the stored weights and signals input by WLDRs. The output signals generated by the memory array are output after being adjusted by an analog signal processing unit and then transferred to the ADCs in the Digital Module through the analog interface. The Digital Module is mainly responsible for storing data, scheduling, sampling and converting analog signals from the analog interconnection, and some computation. First, the microcontroller obtains data to be processed from a sensor or other data sources and stores it in an on-chip memory. Then, the digital preprocessing module obtains a slice of the data in the on-chip memory. The data are reshaped, normalized, binarized, and then sent to WLDRs for computation through a digital interface. After computation in the memory array, the analog signals from the analog interconnect are sampled and converted into digital signals and are then transferred to the digital postprocessing module for accumulation, activation, pooling, and other operations. Finally, the results are stored back to the on-chip memory and will be used by the digital preprocessing module again. After repeating the procedures many times, all the computation of a neural network can be finished and the final results can be obtained.

![Figure 4](image_url)

Figure 4. Architecture of a CIM-Digital Heterogeneous System with Analog Interconnection
Two key factors, i.e., performance and power efficiency, need to be considered in designing a CIM accelerator. The performance is determined based on the sampling rate and number of ADCs. The power efficiency is determined based on performance and the power consumption of ADCs. With an advance CMOS process, the proposed architecture which allows more ADCs with lower power consumption can achieve a higher performance and power efficiency. What's more, compared to digital signals, an analog signal can transfer more information, so that the width of the interface can be largely reduced to one eighth or one quarter.

3.2. CIM Module

The detailed implementation of the CIM Module is shown in Figure 5. The figure takes NOR flash memory as an example, but other memories can also be used with a similar method. Each wordline (WL) has two states: “on” and “off”, determined based on the input signal generated by WLDRs. Cells in the memory array work in the saturation region and are turned “on” or “off” for different weights of “0”, “1”, or “-1”. To represent a binarized weight, a pair of cells are used, one on a positive bitline (BLP), the other on a negative bitline (BLN). The two cells are located on the same WL to share the input. When the value of a weight is “1”, the corresponding cell on BLP will be turned “on” and the corresponding cell on BLN will be programmed and turned “off”. When the value of a weight is “-1”, the corresponding cell on BLP will be programmed and turned “off” and the corresponding cell on BLN will be turned “on”. When the value of a weight is “0”, both cells are programmed and turned “off”. The same constant driving voltage \( V_{DD} \) is applied to all enabled bitlines (BLs) so as to satisfy the condition for the cells to operate in the saturation region, and all SLs are set to ground. Thus, there will be a current in a cell only when the corresponding WL and the cell itself are turned “on” at meantime and the current \( I_0 \) of a cell is mapped to the result of a binary multiplication.

![Figure 5. The CIM Module](image)

Except for computing cells used for multiplication, bleeding cells are used to reduce clock feedthrough and charge injection effect caused by the flipping of input signals, and the effect caused by impedance variation in large cell arrays. All cells on a BLP and a BLN constitute a process element (PE) and each PE can compute a dot product of two vectors at a time. The total current \( I_{BL} \) on the two BLs represents the result of the dot product. With multiple PEs sharing the same WLs, a vector-matrix multiplication (VMM) can be performed.

Each \( I_{BL} \) on BLP and BLN is caught by a current mirror, converted to voltage and sent to an analog multiplexer. The current mirror ratio \( m \) is set based on \( I_{BL} \) to ensure that the maximum voltage value on the analog interconnection will not exceed the range of ADCs. Finally, the analog multiplexer selects one or more analog signals and sends them to the Digital Module according to the number of ADCs in the Digital Module. These analog signals are sampled and converted to digital signals by the ADCs and then processed by a digital postprocessing unit and stored back to the on-chip memory.
3.3. Digital Module
The detailed implementation of the Digital Module is shown in Figure 6. Under the control of a microcontroller, a slice of data in on-chip memory are sent to a digital preprocessing unit. In the digital preprocessing unit, data are normalized by the batch normalization (BN) unit. The normalized data are then binarized by the binarization (BIN) module. According to different binarization algorithm, data may be binarized to \{-1, 0, 1\}, \{-1, 1\}, or \{0, 1\}. The binarized data are reshaped by a reshape module to a sequence which can be used by the CIM module. After the analog signals on the analog interconnection are sampled and converted to digital signals, the accumulation (ACC) unit in the digital postprocessing unit will accumulate it with data from the on-chip memory to get weighted sums. Then, the activation function (ReLU in this case) will be applied to the weighted sum and the activation will be stored back to on-chip memory or sent to a pooling unit according to the neural network model. The control logic unit provides an interface for configuring the clock frequency, the driving voltage value on WLs and BLs and other trim values of the CIM module, and get the values of status registers to monitor the status of the CIM Module.

![Figure 6. The Digital Module](image)

3.4. Analog Interconnection
The Digital Module can be implemented as an ASIC or based on an FPGA. As most units of the Digital Module are digital logic circuits, the key problem is to implement an ADC on the Digital Module. Various implementations based on CMOS process work for the proposed system architecture as long as the requirement of high performance, high integration, and low power consumption are met. In this paper, an FPGA is used to implement the Digital Module. But for FPGAs, existing ADCs [18, 19] do not have enough performance and integration to be used in CIM. An ADC implementation

![Figure 7. (a) An FPGA-based Implementation of Analog Interconnection for Proposed Heterogeneous System; (b) Waveforms of a TDC-based ADC](image)
based on time-digital converters (TDC) on FPGA for a quantum system in a liquid Helium environment is proposed in [20]. However, the original implementation is not fit for CIM which uses SIP technology instead of PCB, needs to be more accurate and meets tight timing constraints. In this paper, an TDC-based implementation of analog interconnection which can reduce the inference of temperature and voltage variation and improve the integration are proposed based on a similar principle to [20], as shown in Figure 7a.

The voltage signal $V_{result}$ from above-mentioned analog multiplexer and a reference voltage signal $V_{ref}$ are input to a differential input buffer with complementary outputs. $V_{ref}$ is generated by a triangle wave generator in the CIM Module which is driven by a slow clock $CLK\_SLOW$. The differential input buffer works as a comparator and compares the two voltage signals to generate a digital signal CMP. As shown in Figure 7b, the width $t$ of a CMP pulse represents the value of sampled $V_{result}$:

$$V_{result} = \left(1 - \frac{t}{\text{period}(CLK\_SLOW)}\right) \cdot \text{amplitude}(V_{ref}) \quad (1)$$

The complementary outputs of the differential buffer, i.e., CMP and $\overline{CMP}$, are sent to an input serializer/deserializer (ISERDES) respectively. The two ISERDESes act as a TDC based on 8 equidistant phase-shifted clock and all clocks for the two ISERDESes and $CLK\_SLOW$ are generated by the same PLL. The sequence of sampling result of two ISERDESes are converted to the measured value of $V_{result}$ by the Seq2Val module. According to the resolution of the TDC, the cycle of $CLK\_SLOW$ is an integral multiple of the cycle of CLK. Except for $CLK\_SLOW$, $V_{ref}$, and $V_{result}$, a valid signal is needed for synchronization and indicating the start and end of the data transfer. As $V_{result}$ is constant in a cycle of $CLK\_SLOW$, a small difference between the delay time of $V_{ref}$ and $V_{result}$ will not influence the result. A sequence of training voltage values can be used to synchronize $V_{result}$ and valid and after the synchronization, valid can be used to eliminate the influence of ringing.

![Figure 8](image)

**Figure 8.** (a) Layout of a Memory Array, Including Peripheral Circuits; (b) Layout of an ADC

4. Experimental Results

4.1. Test Systems Setup

To verify the efficiency of the proposed architecture, two accelerating systems are built. The first one is a common CIM system based on a 65 nm NOR flash memory process and 0.22 $\mu$m CMOS process in a single chip. The layouts of a memory array and an ADC are shown in Figure 8. This system includes WLDRs, a memory array, SAR ADCs and other circuits and can perform computation of convolutional layers and fully connected layers with a precision of 1b-input, 1b-weight and 6b-output. The experimental results of this system are based on a post-layout simulation by Synopsys HSIMPlus. The second one is a CIM-FPGA heterogeneous accelerating system based on the proposed architecture, with a CIM module in a 65 nm NOR flash memory process and 0.22 $\mu$m CMOS process and a Digital Module implemented on Xilinx Artix-7 XC7A100T. Circuit and layout of units in the CIM Module
are the same as the first system. The frequency of the fast clock for TDC units is 800 MHz and the frequency of \textit{CLK\_SLOW} driving the result outputting logic of TDC units, the digital preprocessing unit, the digital postprocessing unit and other units is 100 MHz. With this configuration, the resolution of the implemented ADCs is also 6 bits. The experimental results of the CIM Module is also based on a post-layout simulation by Synopsys HSIMplus. The synthesis and simulation of the Digital Module are based on Vivado 2018.3. Besides, MATLAB R2016b is used to simulate the analog interconnection in the system-level co-simulation of the heterogeneous system to obtain accuracies of networks.

4.2. Resource Utilization and Area

For the Digital Module, the resource cost of each unit is evaluated by resource utilization. With 24 TDCs implemented on the FPGA, the resource utilization of the TDC unit, the digital preprocessing unit and the digital postprocessing unit are shown in Table 1. The resource cost of the all units is very small compared to the available resources on Artix-7 XC7A100T. The percentage of used LUTs, LUTs as memory, registers, Slices and DSPs are no greater than 3.33\% for all units, which means these resources are totally excess. Clock resources like PLLs and BUFGCTRLs do not increase as the number of TDCs increases. The resource bottleneck of the Digital Module is the available number of ILOGICs, i.e., the available number of ISERDESes. According to [21], 24 TDCs can be implemented in each high range I/O bank or high performance I/O bank. As there are many HR or HP I/O bank in FPGAs (e.g., 6 banks in XC7A100T), a lot of TDCs can be integrated with an FPGA and provide enough performance for the heterogeneous system.

| Resource    | Available | TDC | TDC (%) | Pre | Pre (%) | Post | Post (%) |
|-------------|-----------|-----|---------|-----|---------|------|---------|
| LUT         | 63400     | 409 | 0.65\%  | 23  | 0.04\%  | 110  | 0.17\%  |
| LUT RAM     | 19000     | 192 | 1.01\%  | 8   | 0.04\%  | 24   | 0.13\%  |
| FF          | 126800    | 637 | 0.50\%  | 371 | 0.29\%  | 1267 | 1.00\%  |
| Slice       | 15850     | 169 | 1.07\%  | 117 | 0.74\%  | 366  | 2.31\%  |
| DSP         | 240       | 0   | 0.00\%  | 4   | 1.67\%  | 8    | 3.33\%  |
| ILOGIC      | 210       | 48  | 22.86\% | 0   | 0.00\%  | 0    | 0.00\%  |
| PLL         | 6         | 1   | 16.67\% | 0   | 0.00\%  | 0    | 0.00\%  |
| BUFGCTRL    | 32        | 7   | 21.88\% | 0   | 0.00\%  | 0    | 0.00\%  |

TDC: TDC unit; Pre: Digital Preprocessing Unit; Post: Digital Postprocessing Unit

As for the CIM system, ADCs takes up about 49.4\% of the total area of a CIM macro, while the memory array only takes 0.4\%. Thus, the area of the CIM Module in the heterogeneous system can be greatly reduced to a half that of the CIM system. Because of the high area cost of ADCs in the CIM system, it is nearly impossible for the CIM system to have as many ADCs as the Digital Module implemented on an FPGA.

4.3. Performance and Power Breakdown

According to post-layout simulation, the CIM system can achieve a computing precision of 6bit while the operational clock of ADCs is 60 MHz, i.e., the sampling rate is 3.75 MSPS. The power breakdown of the CIM system is shown in Figure 9a. As stated above, the ADCs take up a large proportion of the power consumption. The reason is that to improve the resolution and the performance, the operational clock of ADCs is much faster than the clock for WLDR and other drivers, and the DAC in the ADCs generating reference voltages and the comparator are power consuming to achieve a high performance. The power consumption of WLDRs is small because they only drive the gate terminals of memory cells. The power breakdown of the heterogeneous system is shown in Figure 9b. The power consumption of digital preprocessing unit, digital postprocessing unit and some other units are not
considered as they have no counterparts in the CIM system. For the heterogeneous system, the power consumption of ADCs accounts for a higher percentage and the total power consumption is 4.6 times that of the CIM system. But the sampling rate of ADCs in the FPGA is 100 MHz, which is 26.7 times the sampling rate of ADCs in the CIM system. With the same number of ADCs, the performance of computing will also increase 26.7 times compared to the CIM system. As a result, there will be an increase of 5.8 times in the power efficiency. Considering that much more ADCs can be implemented in an FPGA, the improvement of performance can be even higher.

![Figure 9](a) Power Breakdown of the CIM System; (b) Power Breakdown of the Heterogeneous System

4.4. Accuracies of Networks
Based on MATLAB and Vivado, using a random continuous function as an input, the signal noise distortion ratio (SNDR) and effective number of bits (ENOB) of the TDC-based ADC can be obtained. According to the results in the frequency domain, the SNDR and ENOB of the ADC is 32.5 dB and 5.11 bits respectively. According to this, an evaluation on the influence of using the ADC on the accuracies of LeNet-5 on MNIST and ResNet-18 on CIFAR-10 are performed by injecting noise to the intermediate results. The results show that for MNIST, an accuracy of 99.00% is achieved and for CIFAR-10, an accuracy of 88.98% is achieved. Compared to accuracies of 99.02% and 89.05% of original networks, the losses of accuracy are 0.02% and 0.07% respectively.

5. Conclusions
To decouple the memory and CMOS process and achieve higher performance and power efficiency for CIM technology, this paper proposes a CIM-digital heterogeneous architecture with analog interconnection for neural network accelerating systems. By implementing ADCs and other digital units in a more advanced CMOS process, the proposed architecture can achieve much higher performance, higher power efficiency, and lower area cost with no loss of accuracy in neural networks. Based on the proposed heterogeneous architecture, we designed a CIM-FPGA heterogeneous accelerating system, including a CIM module in a 65 nm NOR flash memory process and 0.22 µm CMOS process and a Digital Module implemented on Xilinx Artix-7 XC7A100T. According to experimental results, it is proved that the heterogeneous system can improve the performance and power efficiency greatly. Experimental results show that it achieves an increase in the performance of 26.7 times and in the power efficiency of 5.8 times.

6. References
[1] R. Mochida, K. Kouno, Y. Hayata et al., "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," 2018 IEEE Symposium on VLSI Technology, 2018, pp. 175-176
[2] W. H. Chen, K. X. Li, W. Y. Lin et al., "A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), 2018, pp. 494-496

[3] C.-X. Xue, W.-H. Chen, J.-S. Liu et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6 ns Parallel MAC Computing Time for CNN Based AI Edge Processors," 2019 IEEE International Solid-State Circuits Conference-(ISSCC), 2019, pp. 388-390

[4] F. Merrikh-Bayat, X. Guo, M. Klachko et al., "High-Performance Mixed-Signal Neurocomputing With Nanoscale Floating-Gate Memory Cell Arrays," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 4782-4790, 2018

[5] R. Z. Han, P. Huang, Y. C. Xiang et al., "A Novel Convolution Computing Paradigm Based on NOR Flash Array with High Computing Speed and Energy Efficient," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1-4

[6] M. R. Mahmoodi and D. Strukov, "An Ultra-Low Energy Internally Analog, Externally Digital Vector-Matrix Multiplier Based on NOR Flash Memory Technology," 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC), 2018, pp. 1-6

[7] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in Memory With Spin-Transfer Torque Magnetic RAM," Ieee T Vlsi Syst, vol. 26, no. 3, pp. 470-483, 2018

[8] D. Fan and S. Angizi, "Energy efficient in-memory binary deep neural network accelerator with dual-mode sot-mram," 2017 IEEE International Conference on Computer Design (ICCD), 2017, pp. 609-612

[9] A. Aziz, E. T. Breyer, A. Chen et al., "Computing with ferroelectric FETs: Devices, models, systems, and applications," 2018 Design, Automation \& Test in Europe Conference \& Exhibition (DATE), 2018, pp. 1989-1998

[10] N. Verma, H. Jia, H. Valavi et al., "In-Memory Computing: Advances and Prospects," IEEE Solid-State Circuits Magazine, vol. 11, no. 3, pp. 43-55, 2019

[11] S. Mittal, "A Survey of ReRAM-Based Architectures for Processing-In-Memory and Neural Networks," Machine Learning and Knowledge Extraction, vol. 1, no. 1, pp. 75-114, 2018

[12] M. Božanić and S. Sinha, Systems-Level Packaging for Millimeter-Wave Transceivers. Springer, 2019.

[13] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits, 1998," online: http://yann.lecun.com/exdb/mnist, 1998

[14] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," online: http://www.cs.toronto.edu/kriz/cifar. html, 2014

[15] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," Solid-State Electronics, vol. 125, pp. 25-38, 2016

[16] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," arXiv preprint arXiv:1602.02830, 2016

[17] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," European conference on computer vision, 2016, pp. 525-542

[18] Xilinx, "LogiCORE IP XPS SYSMON ADC," Datasheet - DS620 (v3.00.b). Xilinx Inc., 2011

[19] Xilinx, "7 Series FPGAs and Zynq-7000 SoC XADC Dual 12-Bit 1 MSPS Analog-to-Digital Converter," User Guide - UG480 (v1.10.1). Xilinx Inc., 2018

[20] Z. Xiang, T. Wang, T. Geng et al., "Soft-Core, Multiple-Lane, FPGA-based ADCs for a Liquid Helium Environment," 2018 IEEE High Performance extreme computing Conference (HPEC), 2018, pp. 1-6

[21] Xilinx, " I/O and Memory Interfacing Features and Benefits in 7 Series Architecture," White Paper: 7 Series FPGAs - WP393 (v1.1). Xilinx Inc., 2012