Towards Reliable In-Memory Computing

From Emerging Devices to Post-von-Neumann Architectures

Amrouch, Hussam ; Du, Nan ; Gebregiorgis, Anteneh; Hamdioui, Said; Polian, Ilia

DOI
10.1109/VLSI-SoC53125.2021.9606966

Publication date
2021

Document Version
Final published version

Published in
Proceedings of the 2021 IFIP/IEEE International Conference on Very Large Scale Integration, VLSI-SoC 2021

2021 IFIP/IEEE 29th International Conference on Very Large Scale Integration (VLSI-SoC) | 978-1-6654-2614-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/VLSI-SoC53125.2021.9606966

Towards Reliable In-Memory Computing:From Emerging Devices to Post-von-Neumann Architectures

Hussam Amrouch[1] Nan Du[2,3,4,5] Anteneh Gebregiorgis[6] Said Hamdioui[6] Ilia Polian[1]

[1]University of Stuttgart
Institute of Computer Engineering
and Computer Architecture
Stuttgart, Germany
amrouch@iti.uni-stuttgart.de
ilia.polian@iti.uni-stuttgart.de

[2]Fraunhofer Institute for
Electronic Nano Systems (ENAS)
Department Nano Device Technology
Chemnitz, Germany
nan.du@enas.fraunhofer.de

[3]Chemnitz University of Technology
Faculty of Electrical Engineering
and Information Technology
Chemnitz, Germany

[4]Leibniz Institute of
Photonic Technology
Department of Quantum Detection
Jena, Germany

[5]Friedrich Schiller University Jena
Institute for Solid State Physics
Jena, Germany

[6]Delft University of Technology
Department of Quantum
and Computer Engineering
Delft, The Netherlands
{A.B.Gebregiorgis,S.Hamdioui}@tudelft.nl

***Abstract*—Breakthroughs in Deep neural networks (DNNs) steadily bring new innovations that substantially improve our daily life. However, DNNs overwhelm our existing computer architectures because the latter is largely bottlenecked by the data movement between memory and processing units. As a matter of fact, in the current von-Neumann architecture, which has remained unchanged since the beginning, data repeatedly moves back and forth between the physically-separated processing units (e.g., CPU, accelerator, etc.) and memory. This, in turn, inevitably leads to large latency and efficiency losses. In DNNs such a bottleneck becomes more and more prominent due to the massive amount of data that must be frequently transferred.**

This paper provides a cross-layer overview on how post-von-Neumann in-memory computing (IMC) architectures can be realized using three different emerging technologies: Charge-based ferroelectric transistors for logic-in-memory computations; memristive devices for unconventional brain-inspired computing; and ultra-low-power memristors especially suitable for Edge AI. Various levels of abstraction will be covered starting from semiconductor device physics to circuit and microarchitecture levels all the way up to the system level, but special attention will be put on reliability aspects.

***Index Terms*—In-memory computing, Ferroelectric FETs, Memristors, Brain-Inspired Compuing**

I. Introduction

In recent years, the accuracy of Deep Neural Networks (DNNs) has continuously improved. This is often associated with making the NN models deeper and more sophisticated, which, in turn, increases the already large demand for computing power and memory requirements. As a matter of fact, overwhelming data-centric workloads driven by DNNs impose a serious challenge for conventional von-Neumann architectures. In particular, data transfer between memory and processing elements largely contributes to the total energy consumption [1] and rapidly form a fundamental bottleneck.

Traditional Neural Processing Units (NPUs) to accelerate deep learning accelerators, such as Google TPU, employ huge systolic arrays of multiply-and-accumulate (MAC) units ($256 \times 256 = 64$K MACs) [2]. In such hardware accelerators, the data is transferred from external off-chip memory to a large on-chip SRAM-based memories and then repeatedly fed to the MAC array. Even through such an implementation minimizes the need for off-chip communications, which indeed helps in reducing the total energy, on-chip SRAM-based memories are power hungry and their access time is significantly larger than the processing time that MAC units requires to perform computations. Furthermore, the massive number of multiplication operations simultaneously performed within the systolic MAC array leads to excessive on-chip power densities due to the significant amount of power that is consumed within a small confined area. The latter quickly leads to elevated temperatures that form a thermal bottleneck for the entire NPU chip as it has been recently demonstrated [3].

Because the underlying core principle of von-Neumann architectures separates processing units from memory storage, data must be frequently moved back and forth, resulting in the so-called "*memory wall*". To overcome this challenge and significantly improve the efficiency, *beyond von-Neumann architectures*, in which computations are executed inside the memory itself, are being heavily researched by both academia and industry. The demand for such novel architectures be-

This work was supported by by the DFG (German Research Foundation) Priority Program Nano Security, Project MemCrypto (DU 1896/2–1, PO 1220/15–1). The work of H. Amrouch and I. Polian was partially supported by Advantest as part of the Graduate School "Intelligent Methods for Test and Reliability" (GS-IMTR) at the University of Stuttgart. We thank Simonn Thomann for his help in Section III.

comes even more prominent when it comes to data-centric workloads like those in DNNs.

To realize beyond von-Neumann architectures, recent works demonstrated how a Boolean logic function (e.g., XNOR, NAND, etc.) could be implemented using both conventional SRAM memories [4] as well as emerging Non-volatile memory (NVMs) such as resistive random access memory (ReRAM), phase change memory (PCM) and spin transfer torque magnetoresistive random access memory (STT-MRAM) [5], and recently FeFET [6]. In addition, NVM-based crossbar arrays are one excellent candidate for *in-memory computing* (IMC) due to its profound energy efficiency, stemming from analog computing [7], [8] when performing matrix multiplications, which are the core of any DNN accelerator.

The remainder of the paper is organized as follows. The next section provides the necessary background on IMC and its enabling technologies. Section III focuses on the ferroelectric FET (FeFET) technology. Section IV highlights the usage of emerging devices for brain-inspired computing. Section V shows how IMC can lead to ultra-low-power processing. Section VI concludes the paper.

II. Architecture Classification and Basics of In-Memory Computing

A. In-Memory computing architecture classification

In-memory computing (IMC), also known as "computing-in-memory" (CIM), is a paradigm in which the computation is performed within the memory core where the data resides [9]. This capability enables in-memory computing to achieve a higher energy-efficiency over the conventional von-Neumann paradigm by avoiding the costly data movements between processing and storage units in von-Neumann systems. In-memory computing can be realized using different emerging memristive technologies such as Resistive Random Access Memory (RRAM), Phase Changing Memory (PCM) and Magnetic RAM (MRAM) as well as conventional memory technologies such as SRAM, DRAM and Ferroelectric FETs [10], [11]. In-memory computing using emerging memrisitve devices benefits from their non-volatile nature and their practically zero leakage compared to their conventional memory technology counterparts.

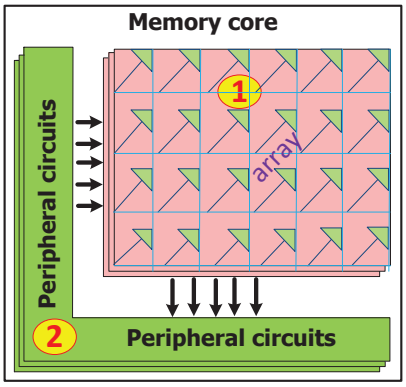

Fig. 1. Memory core for IMC architectures and its classification

Irrespective of the memory technology used, a memory core consists of *memory array* and *peripheral circuits*. Any in-memory computing block that implements a logic or arithmetic operation in memory core produces the computing result either within the memory array itself or within the periphery. Thus, based on the location where the result of the computation is generated, CIM architectures can be classified into two main categories, namely IMC Array (CIM-A) and CIM Periphery (CIM-P). Figure 1 shows the CIM classes in a memory core where labels 1 and 2 indicate CIM-A and CIM-P classes, respectively.

1) CIM-A (Array-oriented architecture): In this architecture, the computing result is produced within the array [12] (noted as position 1 in Figure 1). Typical examples of such architectures that use memristive logic designs include MAGIC and imply [13], [14]. These architectures can be further subdivided into two groups: (1) basic, where only changes inside the memory array are required to do the computation, and (2) hybrid, where, in addition to major changes in the memory array, minimal to medium changes are required in the peripheral circuit.

2) CIM-P (Periphery-oriented architecture): In this architecture, the computing result is produced within the peripheral circuitry [12] (noted as position 2 in Figure 1). Similar to array-oriented architectures, periphery-oriented architectures can be further classified into (1) basic, where only changes inside the peripheral is required and (2) hybrid architectures, where the majority of the changes take place in the peripheral circuit and minimal to medium changes in the memory array. Typical examples of such architectures involve logical operations and vector-matrix multiplications [15]–[17]

B. IMC technologies

A resistance-based computational memory device can be modulated between a low resistance state (LRS) and a high resistance state (HRS) (or among multiple resistance states) by an appropriate electric stimulus. The programmed resistance states are non-volatile. For implementing the IMC applications, the representative computational memory devices, such as ReRAM, PCM and STT-MRAM, are typically constructed in a crossbar array and require a selection transistor device in series (as demonstrated in Figure 2) for the sake of the elimination of sneak path currents during writing and reading operations [18].

A **resistive random access memory (ReRAM)** consists of one or multiple metal-oxide layers sandwiched between top electrode (TE) and bottom electrode (BE). The resistive switching process typically involves either the construction/disruption of conductive filaments, or the modulation of carrier transport barrier at the electrode/switching layer interface induced by ion migration. These two switching mechanisms, called filamentary or interfacial switching, respectively, are shown in insets of Figure 2(a). In filamentary switching devices [19], [20], a one-time application of stronger electric field strength upon device operation is required for the initial formation process of the conducting filament, i.e.

electroforming process. A compliance current is necessary to confine the current flows through the local path in LRS. For this reason, there is a substantial interest in the usage of interfacial switching devices [21] [22] for avoiding the electroforming step altogether. Their further advantage is their self-rectifying behavior, which is key for developing selector-free memristive crossbar arrays (MCAs).

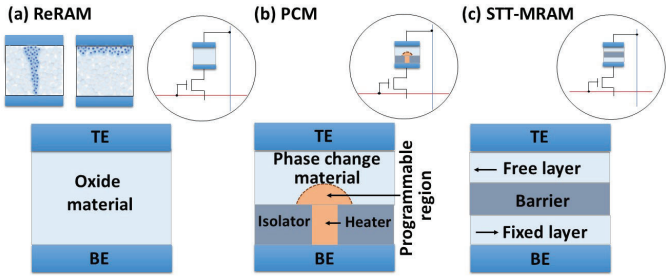

Fig. 2. Schematic illustrations of resistance-based computational memory units for (a) ReRAM, (b) PCM, (c) STT-RAM and their corresponding 1T1R structures applied in crossbar array. The insets in (a) demonstrate the distribution of mobil ions (blue dots) with filamentary and interfacial switching mechanisms in the ReRAM devices.

A **phase change memory (PCM)** is relying on the reversible transition between a highly resistive amorphous structure and a highly conductive crystalline structure in a phase change material within sub-nanosecond switching time [23]. The memory effect of PCM devices are driven by the thermal excitation. As shown in Figure 2(b), an application of a large current pulse with short duration leads to a near-hemispherical shape of amorphous phase region, thus exhibiting a HRS state. Upon an application of a current pulse for a relatively longer duration, the amorphous region is turned into crystalline, thus decreasing the resistance to LRS. The phase change materials $Ge_2Sb_2Te_4$ (GST) [23] and $TiTe_2/Sb_2Te_3$ [24] can be used for constructing PCM devices.

A relatively new alternative promising computational memory technology is **spin transfer torque magnetoresistive random access memory (STT-MRAM)**, where a magnetic tunnel junction structure is composed with a free and a pinned ferromagnetic metal layers, such as CoFeB alloys [25]. A thin insulating tunnel oxide barrier, such as MgO, separates these two layers as shown in Figure 2(c). The magnetic polarization in the free layer is free to change during the writing operation. The LRS or HRS of STT-MRAM devices are obtained upon an application of a current, which can change the free layer to be parallel or antiparallel with the pinned layer, respectively.

The aforementioned resistance based RAMs can be served as elements of a computational memory unit for low-cost in-memory computing applications. The promising alternative IMC technologies offer prospective gains in programming rate, energy consumption, device lifetime, and storage capacity, as well as in-memory storage and computing capabilities.

C. Potentials of IMC

IMC architecture provides efficient computing capability for a wide range of applications and computation kernels. Some examples application kernels which can benefit from IMC include:

- Database query: database query applications can benefit from IMC by accelerating the bulky bitwise AND/OR operations with scouting logic [26].
- Deep learning: AI and deep learning application can use IMC to accelerate the resource intensive vector matrix multiplication kernels [10].
- Automata processor: in automata processors State Transition Element (STE) matrix is usually huge, but it can be easily mapped to an IMC array in order to accelerate the automata processor.
- Hyperdimensional computing: similar to database query and deep learning, hyperdimensional computing is full of bitwise AND operations and vector matrix multiplication operations which are suitable for IMC acceleration.

In general, the IMC implementations are shown to achieve significant benefits in energy and area with respect to alternate implementations that do not use IMC. Recent published work based on circuit simulation and small-scale prototypes has shown the promise of IMC. Simulation-based work reported that IMC architecture provides two to three orders of magnitude improvement in energy-delay product and energy spent per operation compared to conventional von-Neumann architecture [9], and around 10 fJ per arithmetic operation (1 MAC = 255 arithmetic operations) can be realized [8]. Small-scale prototype work considering database query applications demonstrated that IMC architecture can achieve 6 fJ per logic operation. All these examples highlight the tremendous potential of IMC over von-Neumann architecture.

III. Reliable Computing Beyond von-Neumann on Unreliable Ferroelectric Transistors

Emerging Non-Volatile Memory NVM technologies keep gaining a significant attraction akin to their promise in building ultra-efficient Logic-in-Memory (LiM) and In-Memory Computing (IMC). In this section, we focus on discussing Ferroelectric Field-Effect Transistor (FeFET) technology as an example of charge-based memory devices and in the next section, we discuss Resistive Random-Access Memory (ReRAM) as an example of memristive devices.

After the discovery of how hafnium-based oxide material can be converted into a ferroelectric material, FeFET-based NVM has become fully compatible with the existing fabrication process of conventional CMOS. Several leading semiconductor vendors are currently exploring FeFETs for both memory and neuromorphic applications. For instance, GlobalFoundries has shown the successful fabrication of FeFETs using their commercial 28nm CMOS through a dual mask patterning [27] and demonstrated 10MiB memory chips using FeFET-based NVM that feature 1ns read latency. Further, Intel has recently demonstrated, for the first time, FeFET devices with an endurance of 10^{12} cycles [28].

FeFET Basic Operation: Hafnium-based high-k dielectric is the conventional material to construct the transistor's gate in the current CMOS technology. When it is doped with

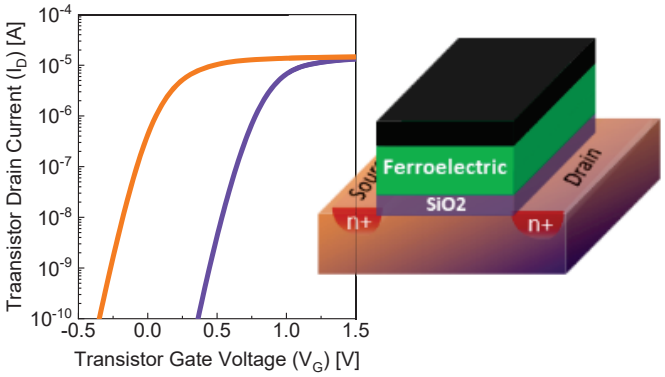

Fig. 3. FeFET-based non-volatile memory is realized by replacing the conventional high-k material with a ferroelectric material. Depending on the direction on the polarized dipoles, the underlying transistor exhibits either a low-V_{TH} or high-V_{TH} state.

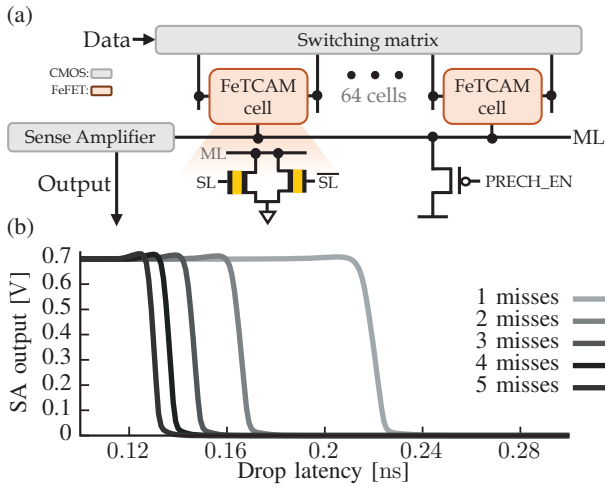

Fig. 4. FeFET-based CiM for Hamming distance calculation. Ternary Content-Addressable Memory (TCAM) is realized through connecting LiM-based FeFET-based XNOR. The intensity of the discharged current reflects the degree of mismatch between the stored vector and the searched vector [29].

zirconium, a ferroelectricity phenomenon is then realized, which turns an ordinary transistor into an NVM device. The basic concept to store the information is charge-based in which the polarized dipoles inside the ferroelectric (FE) layer interact with the electrical characteristics of the underlying transistor. When a vertical electric field across the transistor gate is applied, the FE dipoles switch towards a certain direction (up or down). As a result, the underlying transistor exhibits either a low threshold voltage (low V_T) or a high threshold voltage (high V_T) as illustrated in Fig. 3.

FeFET-based Logic-in-Memory: Through coupling two FeFET transistors together [30], as presented in Fig. 4, an XNOR logic function can be realized. When a value X is stored within the XNOR gate, it is always stored in a complementary manner. Hence, one of the two FeFETs is in a low-V_T state and the other FeFET is in a high-V_T state. When the FeFET-based XNOR receives an input (Y), depending whether a match (i.e., $X = Y$) or mismatch (i.e., $X \neq Y$), the output voltage either remains high or drops, leading to either '0' or '1', respectively. To achieve that, the match line is first charged (i.e., full V_{dd}) and when $X = Y$, both FeFET transistors will be OFF. Therefore, no conducting path is possible and hence the voltage remains high. In other words, the XNOR's output will be logic '1'. Only when $X \neq Y$, a path is formed and the current can go through the FeFET that is in low-V_T state. In other words, the voltage drops and the XNOR output becomes logic '0'. All in all, if and only if $X \neq Y$, the output is '0'. Otherwise, it is '1'. Hence, an LiM-based XNOR Boolean function is realized.

FeFET-based In-Memory Computing: In the majority of classification tasks, Hamming distance calculation is essential to compute the similarity between vectors. This holds even more when it comes to brain-inspired hyperdimensional computing. In order to accelerate classification, computing Hamming distance within the memory without accessing it is vital. To achieve that, several LiM-based XNOR gates (as explained above) can be connected in series. When a vector is compared against the stored vector, the output of every LiM can provide a discharging current or not based on whether a match or mismatch has occurred. Then, through sensing the overall output, one can estimate the number of mismatches and hence the corresponding Hamming distance [29].

IV. Emerging Devices for Unconventional Brain-Inspired Computing

Within the big data era, the conventional digital computers based on the von Neumann architecture are becoming ineffective while dealing with unstructured real-time big data flow. The human brain is an advanced information storage and computation platform, which is able to process massive real-time data in a parallel and adaptive manner with very low energy consumption of only approx. 10 W. Inspired by the biological human brain, the unconventional non-von Neumann computing architectures are attracting significant interest for handling vast amounts of data efficiently.

The aforementioned beyond-CMOS (complementary-metal-oxide-semiconductor) computational memory devices ReRAM, PCM and STT-MRAM are potential solutions for the implementation of power-efficient unconventional brain-inspired computing. Such computational memory devices possess direct interfaces with analog signals and offer an intrinsically electrically-tunable conductance. They can update their conductances (artificial synaptic weights) upon electrical stimuli (neuronal activities) and demonstrate stable resistive states within itheir dynamic range (analog behavior). Besides that, they provide a number of other beneficial functional properties [18], including low power consumption, reconfigurability, fast switching speed, high endurance/retention, and excellent scalability (e.g., 3D integration manufacturing techniques) [31].

For instance, memristive crossbar array with a 2 nm feature size and a single layer density up to 4.5 Tbit/in^2 [32] has been demonstrated, where the information density is comparable with the three-dimensional stacking in state-of-the-art 64-layer

and multilevel 3D-NAND flash memory [33]. Last but not least, the computational memory devices perform massive parallel computations supported by a dense array of millions of nanoscale compute units, which improves the time complexity and is the key to low-cost cognitive computing.

In the past decades, the vast amounts of data and huge cost of computational power are the main driving factors for the development of power-efficient brain-inspired computing, i.e. implementation of deep learning (DL) accelerators and spiking neural networks (SNN). The DL [32], inspired by biological neural networks, relies on the computational networks of connected computational units (plastic synapse) operating in parallel. By exploiting brain-inspired IMC, the inference (forward propagation) and training (backward propagation) of various layers of DNN are adapted to the computational units organized in a crossbar configuration. The propagation of data is performed in a single step by sourcing the data to the crossbar word lines and recording the feedback at each bit line. The synaptic weights are stored as the conductance state of computational memory units in crossbar. In contrast to DL networks, the distinct feature of SNN [34] is the incorporation of spike timing in the data processing according to the biologically inspired spike-timing dependent plasticity (STDP) rule [35]. For example, based on a simplified STDP model, the auto-associative pattern learning tasks are demonstrated by exploiting an integrated neuromorphic core with 256*256 PCM synapses fabricated along with Si CMOS neuron circuits with high learning efficiency [34]. Nevertheless the emerging technologies are still to realize their full potential in the promoted DL-based and spike-based learning and inferences.

V. Ultra-low Power Memristor based In-Memory Computing for Edge AI

Edge computing (aka edge-AI), is a promising solution to overcome the latency, data transfer bandwidth barriers of cloud-based systems by performing local computing (on the edge-devices) [36]–[38]. The main advantages of edge-AI over traditional AI applications are energy-efficiency, bandwidth minimization and real-time response. However, edge-AI has stringent requirements that must be dealt with in order to harness its full potential; edge-AI hardware must be fast, compact and extremely energy-efficient, as edge-devices have limited resource such as battery lifetime or harvested energy [36], [39], [40].

Memristor-based in-memory computing has the potential to break the aforementioned challenge (due to the nature of the architecture and the devices used to realize it) and deliver energy efficient implementations of hardware edge-AI [9]. Such architecture perform computation on the stored data and hence, circumventing the costly data movement of von-Neumann based systems [12].

A. In-memory computing architecture for Edge-AI

As shown in Figure 5(b), an in-memory computing core for edge applications has two main architectural units: Memory array commonly known as crossbar array unit, and periphery unit. The crossbar array stores the data and perform operation. Similarly, the periphery unit converts input/output data formats between analog and digital. Moreover, the periphery unit can also be used to perform basic logical and arithmetic operations.

Crossbar array: Neural networks for edge-AI applications use multiply and accumulate (MAC) extensively in order to perform matrix-matrix multiplication (MMM) with large operand sizes [8]. Such units can be easily mapped into a memristive-based crossbar array and perform their operation e.g., MMM in the crossbar unit. Figure 5(a) shows a subset of MMM operation *i.e.*, vector-matrix multiplication (VMM) using in-memory computing crossbar array. From Figure 5(a) it can be observed that the VMM is performed by applying a voltage vector $V = V_j$ (where $j \in \{1, m\}$) to a memristor-crossbar matrix of conductance values $G = G_{ij}$ (where $i \in \{1, n\}$, $j \in \{1, m\}$). At any instance, each column performs a vector-vector multiplication (VVM) or a MAC operation, with the output current vector I, in which each element is $I_i = \Sigma V_j \cdot G_{ij}$. Note that all n MAC operations are performed with O(1) time complexity.

Periphery: An in-memory computing core needs some major modifications to accommodate analog-based computing, as shown in Figure 5(b). The circuit blocks comprising the periphery that supports the bitcell array need to be modified to support in-memory operations. For example, the following is needed to perform VMM operation in a crossbar: 1) Row-decoder becomes complex as it involves enabling several rows in parallel. Also, *1-bit* row or word-line drivers are now replaced by digital-to-analog converters (DACs) that convert multi-bit VMM operands into an array of analog voltages. 2) Column periphery circuits performing read operations need to be replaced by analog-to-digital converters (ADCs). 3) Control block needs to deal with complex instructions such as handling intricacies of multi-operand VMM operations.

B. In-memory computing-based neuromorphic design for edge applications

Neuromoprhic computing is one of the application domains which can significantly benefit from in-memory computing architecture. The main reason for this is the fact that the main operation employed by neuromorphic systems involves intensive Matrix-Matrix Multiplication (MMM) or Vector-Matrix Multiplication (VMM). Since both MMM and VMM

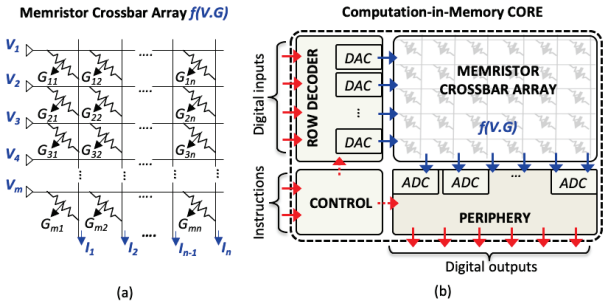

Fig. 5. Basic in-memory computing architecture for edge-AI (a) ReRAM based crossbar operation demo (b) In-memory computing core architecture i.e., Periphery + crossbar array

kernels can be easily accelerated using in-memory computing architecture, neuromorphic computing can achieve substantial improvement in energy-efficiency and alleviate data movement problems by employing in-memory computing architecture.

However, there are several open questions that need to be addressed in order to fully harness the potential of in-memory computing for edge-AI. At circuit level, issues such as device endurance, high resistance ratio between the off and on state of the devices, multi-level storage, precision of analog weight representation must be addressed. Similarly, at the circuit and architecture levels, various challenges have to be addressed; examples are high precision programming of memory elements, complexity of signal conversion circuits, scalability of the crossbars and their impact on computation accuracy etc. Moreover, maturity of system- and compiler-level tools e.g., profiling, simulation and design tools is of decisive importance.

VI. Conclusions

Post-von-Neumann in-memory computing architectures are an important foundation for emerging applications, and they can maximally benefit from novel devices. We reviewed several promising technologies, from ferroelectric transistors to ultra-low-power memristors, and architectures on their basis. Reliability turns out to be a central challenge that needs to be addressed by solutions coordinated across the layers.

References

[1] V. Sze *et al.*, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[2] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Int'l Symp. Computer Architecture*, 2017.

[3] H. Amrouch, G. Zervakis, S. Salamin, H. Kattan, I. Anagnostopoulos, and J. Henkel, "NPU thermal management," *IEEE Trans. CAD*, vol. 39, no. 11, pp. 3842–3855, 2020.

[4] A. Agrawal *et al.*, "X-SRAM: Enabling in-memory Boolean computations in CMOS static random access memories," *IEEE Trans. Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4219–4232, 2018.

[5] Editorial, "Beyond von Neumann," *Nature Nanotechnology*, vol. 15, no. 7, p. 507, 2020.

[6] E. T. Breyer *et al.*, "Compact FeFET circuit building blocks for fast and efficient nonvolatile logic-in-memory," *IEEE Jour. Electron Devices Society*, vol. 8, pp. 748–756, 2020.

[7] M.-L. Wei, H. Amrouch *et al.*, "Robust brain-inspired computing: On the reliability of spiking neural network using emerging non-volatile synapses," in *IEEE Int'l Reliability Physics Symp.*, 2021.

[8] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 14–26, 2016.

[9] S. Hamdioui *et al.*, "Memristor based computation-in-memory architecture for data-intensive applications," in *Design, Automation & Test in Europe Conf.*, 2015, pp. 1718–1725.

[10] A. Singh, S. Diware, A. Gebregiorgis *et al.*, "Low-power memristor-based computing for Edge-AI applications," in *IEEE Int'l Symp. Circuits and Systems*. IEEE, 2021, pp. 1–5.

[11] X. Si *et al.*, "A twin-8t sram computation-in-memory unit-macro for multibit cnn-based ai edge processors," *IEEE Jour. Solid-State Circuits*, vol. 55, no. 1, pp. 189–202, 2019.

[12] H. A. D. Nguyen *et al.*, "A classification of memory-centric computing," *ACM Jour. Emerging Technologies in Computing Systems*, vol. 16, no. 2, pp. 1–26, 2020.

[14] K. Kim, S. Shin, and S.-M. Kang, "Stateful logic pipeline architecture," in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*. IEEE, 2011, pp. 2497–2500.

[13] S. Kvatinsky *et al.*, "MAGIC——Memristor-aided logic," *IEEE Trans. Circuits and Systems II*, vol. 61, no. 11, pp. 895–899, 2014.

[15] P. Chi *et al.*, "Prime: A novel processing-in-memory architecture for neural network computation in ReRam-based main memory," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 27–39, 2016.

[16] L. Xie *et al.*, "Scouting logic: A novel memristor-based logic design for resistive computing," in *IEEE Computer Society Annual Symp. VLSI*, 2017, pp. 176–181.

[17] S. Li *et al.*, "Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories," in *DAC*, 2016.

[18] N. Du, H. Schmidt, and I. Polian, "Low-power emerging memristive designs towards secure hardware systems for applications in internet of things," *Nano Materials Science*, vol. 3(2), pp. 186–204, 2021.

[19] X. Xu *et al.*, "Superior retention of low-resistance state in conductive bridge random access memory with single filament formation," *IEEE Electron Device Lett.*, vol. 36, p. 129–131, 2015.

[20] A. Siemon *et al.*, "Realization of Boolean logic functionality using redox-based memristive devices," *Adv. Funct. Mater.*, vol. 25, p. 6414–6423, 2015.

[21] N. Du *et al.*, "Field-driven hopping transport of oxygen vacancies in memristive oxide switches with interface-mediated resistive switching," *Physical Review Applied*, vol. 10(5), p. 054025, 2018.

[22] M. Hansen *et al.*, "A double barrier memristive device," *Sci. Rep.*, vol. 5, p. 13753, 2015.

[23] H.-S. P. Wong *et al.*, "Phase change memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[24] J. Shen, S. Lv, X. Chen, T. Li, S. Zhang, Z. Song, and M. Zhu, "Thermal barrier phase change memory." *ACS Appl. Mater. Interfaces*, vol. 11(5), pp. 5336–5343, 2019.

[25] G. Jan *et al.*, "Achieving sub-ns switching of STT-MRAM for future embedded LLC applications through improvement of nucleation and propagation switching mechanisms," *IEEE Symp. VLSI Tech.*, 2016.

[26] I. Giannopoulos *et al.*, "In-memory database query," *Advanced Intelligent Systems*, vol. 2, no. 12, p. 2000141, 2020.

[27] S. Beyer *et al.*, "FeFET: A versatile CMOS compatible device with game-changing potential," in *IEEE Int'l Memory Workshop*, 2020.

[28] A. A. Sharma *et al.*, "High speed memory operation in channel-last, back-gated ferroelectric transistors," in *IEEE Int'l Electron Devices Meeting*, 2020, pp. 18.5.1–18.5.4.

[29] S. Thomann, C. Li, C. Zhuo, O. Prakash, X. Yin, X. S. Hu, and H. Amrouch, "On the reliability of in-memory computing: Impact of temperature on ferroelectric TCAM," in *IEEE VLSI Test Symp.*, 2021.

[30] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, no. 11, pp. 521–529, 2019.

[31] P. Lin *et al.*, "Three-dimensional memristor circuits as complex neural networks," *Nature Electronics*, vol. 3(4), pp. 225–232, 2020.

[32] S. Pi *et al.*, "Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension," *Nature Nanotechnology*, vol. 14, pp. 35–39, 2019.

[33] S. Lee *et al.*, "A 1 Tb 4b/cell 64-stacked-WL 3D NAND flash memory with 12 MB/s program throughput," in *IEEE Int'l Solid-State Circuits Conf.*, 2018, pp. 340–342.

[34] S. Kim *et al.*, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," in *IEEE Int'l Electron Devices Meeting*, 2015, p. 17.1.1– 17.1.4.

[35] N. Du *et al.*, "Synaptic plasticity in memristive artificial synapses and their robustness against noisy inputs," *Frontiers in Neuroscience*, vol. 15, p. 696, 2021.

[36] T. Rausch, W. Hummer, V. Muthusamy, A. Rashed, and S. Dustdar, "Towards a serverless platform for edge AI," in *2nd {USENIX} Workshop on Hot Topics in Edge Computing*, 2019.

[37] Y.-L. Lee, P.-K. Tsung, and M. Wu, "Techology trend of edge AI," in *Symposium on VLSI Design, Automation and Test*, 2018.

[38] P. G. López *et al.*, "Edge-centric computing: Vision and challenges," *Comput. Commun. Rev.*, vol. 45, no. 5, pp. 37–42, 2015.

[39] X. Xu *et al.*, "Scaling for edge inference of deep neural networks," *Nature Electronics*, 2018.

[40] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, 2016.