Computation-in-Memory Architectures

Muath Abu Lebdeh*, Uljana Reinsalu†, Hoang Anh Du Nguyen*, Stephan Wong* and Said Hamdioui*

*Computer Engineering Lab, Delft University of Technology, Delft, The Netherlands
†Department of Computer Engineering, Tallinn University of Technology, Tallinn, Estonia
Email: {m.f.m.abulebdeh,h.a.dunguyen,j.s.s.m.wong,s.hamdioui}@tudelft.nl uljana@ati.ttu.ee

Abstract—Emerging computing applications (such as big-data and Internet-of-things) are extremely demanding in terms of storage, energy and computational efficiency, while todays architectures and device technologies are facing major challenges making them incapable to meet these demands. Computation-in-Memory (CIM) architecture based on memristive devices is one of the alternative computing architectures being explored to address these limitations. Enabling such architectures relies on the development of efficient memristive circuits being able to perform logic and arithmetic operations within the non-volatile memory core. This paper addresses memristive circuit designs for CIM architectures. It gives a complete overview of all designs, both for logic as well as arithmetic operations, and presents the most popular designs in details. In addition, it analyzes and classifies them, shows how they result in different CIM flavours and how these architectures distinguish themselves from traditional ones. The paper also presents different potential applications that could significantly benefit from CIM architectures, based on their kernel that could be accelerated.

I. INTRODUCTION

The modern-day data explosion is triggered by many factors and applications; e.g., content creation by consumers, genomics, Internet-of-things, (video) surveillance, and autonomous driving cars. This results in a certain pervasiveness of (large amounts of) data surrounding us in our daily lives, which needs processing to extract meaningful information to enhance our lives. The data ubiquity directly contradicts the data ubiquity directly contradicts the von Neumann architecture as it would require the continuous transportation of data towards and from central processing units. In particular, high energy consumption, limited bandwidth, guaranteed response times are obstacles that prove to be increasingly difficult to overcome [1], [2]. On a smaller scale (between CPUs and memory), we are talking about the memory, ILP, and power walls [1]. Consequently, researchers have been looking at computation-in-memory (CIM) architectures based on memristive devices [3]–[6] as a possible solution to the aforementioned problems. The CIM approach keeps data at the location of creation or consumption, and avoids any unnecessary transfer of data. Furthermore, all needed processing is "brought" to and performed at the data location, i.e., the memory or storage. Enabling such architecture requires the design of memristive-based circuits being able to perform both logic as well as arithmetic operations in an efficient manner within the memory core.

Many memristive based circuits have been proposed in the last decade to enable the implementation of some primitive functions. Most of this work addressed logic bitwise operations such as NAND, AND, OR, NOR and XOR [7]–[12]. Recent work has focused more on some (limited) arithmetic operations such as vector-matrix multiplication [13]–[15], addition [16]–[19] and multiplication [16], [20], [21]. In addition, there is some limited work aiming at reviewing and analyzing such circuits from different perspectives. Maan et al. [22] focused on the memristive threshold logic circuits and their different implementations. Vourkas et al. [23] presented a general overview of the memristive logic circuits. Reuben et al. [24] discussed a framework for the stateful logic circuits, which are suitable for enabling CIM architectures. Zidan et al. [25] overviewed the potential of memristive devices in embedded memory design, biologically inspired computing, and CIM circuits. Ielmini et al. [26] discussed the emerging resistive devices to implement digital and analog CIM circuits. Although all this work tries to analyze the memristive circuit designs from different perspectives, none of them is able to extract the impact of these designs on CIM architecture and the potential applications they could enable. Inspecting these circuit designs reveals that they are not "generic" designs; i.e., they impose constrains on the kind of the CIM architecture they support. For example, Scouting logic [12] requires both inputs (operands) to be stored and aligned in the crossbar memory, whereas the vector-matrix multiplication [13] requires the matrix to be stored in the memory array and the vector to be provided via the memory port. Understanding such constrains and their impact on the CIM architecture is the key towards defining both the details of the architecture and the applications that could be targeted.

This paper presents an overview of the memristive circuit designs (for executing primitive logic and arithmetic operations), classifies them to present the different types of CIM architectures they enable, and discusses some potential applications that could benefit (depending on the available kernels and their designs).

The rest of the paper is organized as follows. Section II classifies the CIM architectures based on the location where the memristive-based circuit design of a kernel (primitive
function) produces its results; either in the memory Array (CIM-A) or in the Peripheral circuit of the memory (CIM-P). Section III discusses CIM-A class and reveals the details of two popular designs. Section IV presents the same, but then for CIM-P class. Section V shows the available kernels, their details in two classes as shown in Figure 1.

Finally, section VI concludes the paper.

### II. CLASSIFICATION

Computation-in-memory (CIM) aims at integrating processing within the memory itself; i.e., the computation takes place within the memory core. In addition, it aims at using the non-volatile memristive device technology (e.g., Resistive RAM), as it has practically no leakage and its nature enables both storage and computing capabilities.

As it is known, any memory core (including memristive memory cores) consists of a memory array and its peripheral circuits. Each memristive circuit design (aiming at implementing any logic or arithmetic operation in memory core) produces the computing result either within the array or within the periphery. Hence, depending on where the result of the computation is produced, the CIM architecture can be divided into two classes as shown in Figure 1.

- **CIM-Array (CIM-A):** the computing result is produced within the memory array. Hence, the output should be stored in a memristive device in the array in form of a resistance state.

- **CIM-Periphery (CIM-P):** the computing result is produced within the peripheral circuitry. Given the fact that memory periphery is based on CMOS technology, the nature of the produced output is voltage.

The memristive circuit designs that enable CIM architecture (by implementing logic and arithmetic operations) are confined to hold at least one of its inputs (operands) in the array. In other words, the operator being executed within the memory needs to have all operands stored in the array (hence their logic values are resistive) or only part of the operands is stored in the array and the other part is received via the memory port(s) (hence their logic values are hybrid, i.e., resistive and voltage). This results into four sub-classes as shown in Figure 1: CIM-A, CIM-Ah, CIM-Pr and CIM-Ph; the additional letters ‘r’ and ‘h’ denote the nature of the inputs (operands), namely resistive and hybrid, respectively. The figure maps the existing memristive circuit designs into the classification. In the next section, some of these designs will be discussed as examples.

### III. MEMRISTIVE CIRCUIT DESIGN FOR CIM-A ARCHITECTURE

CIM-A architecture uses the memristive CIM-Ar and CIM-Ah circuit designs to perform primitive logic (e.g., NOR) and arithmetic (e.g., addition) operations. The existing CIM-Ar circuit designs include Snider logic [7], implication logic [8], memristor aided logic (MAGIC) [9], and fast boolean logic (FBL) [10]. These designs implement primitively bitwise logic operations such as material implication, NOR, and NOT functions. On the other hand, the existing CIM-Ah circuit designs include majority logic [27] and resistive accumulator (nanoscale abacus) [18]. These two designs implement primitively logic operations such as majority function, and arithmetic operations such as addition.

The following subsections discuss two popular design examples in details: implication logic [8] as an example of CIM-Ar design, and majority logic [27] as an example of CIM-Ah design. Snider, MAGIC and FBL are quite similar to implication logic, while the resistive accumulator is similar to majority logic.

#### A. Implication Logic (CIM-Ar design example)

The implication logic (IMPLY) design executes the material implication as a universal logic function. An example of IMPLY gate is shown in Fig. 2, where the memristive devices $P$ and $Q$ present the operands. The design requires $R_{on}<<R_{G}<<R_{off}$ ($R_{on}$ is the low Ohmic state and presents logic 1, and $R_{off}$ is the high Ohmic state of the memristive device and presents logic 0). In this design, the device $Q$ acts also as an output. Assuming that the devices $P$ and $Q$ are already programmed, the execution of IMPLY is done by applying two control voltages $V_{wh}$ and $V_{w}$ as shown in Fig. 2; note that the design requires $V_{w}>V_{th}>V_{wh}$, where $V_{th}$ is the threshold voltage switching of the device. The magnitude of $V_{w}$ across a device is sufficient to switch the state of the device from $R_{off}$ to $R_{on}$, whereas $V_{w}-V_{wh}$ is not sufficient enough to switch the device state. Hence, after execution, and depending on the values of $P$ and $Q$, the state of $Q$...
may or may not switch. E.g., if \( P=Q=0 \), the state of \( Q \) will switch from \( R_{\text{off}} \) to \( R_{\text{on}} \). Therefore, this logic design is input destructive. More complex logic functions can be implemented by cascading IMPLY sequentially.

### B. Majority Logic (CIM-Ah design example)

The majority logic design executes the majority logic function. An example of such gate is shown in Fig. 3; the circuit has two voltage inputs (\( P \) and \( Q \)), and one resistive input \( Z \) that acts also as an output. Assuming that the device \( Z \) is already programmed, the execution of the function is done by applying the voltages \( V_{wP} \) and \( V_{wQ} \) to the top and bottom electrodes of the memristive device, respectively, as shown in Fig. 3. Depending on the state of the device \( Z \) and the applied voltages, the device may or may not switch. E.g., if initially \( Z=0 \) (\( R_{\text{off}} \)), and we apply \( V_{wP}>V_{th} \) and \( V_{wQ}=0 \), then \( Z \) will switch to \( R_{\text{on}} \). Therefore, this operation is destructive for input \( Z \). This majority gate can be used to implement other CIM-Ah logic functions, such as OR and AND, by fixing one of the voltage inputs (\( P \) or \( Q \)) to a specific logic value. Other than the majority logic design, the resistive accumulator is another CIM-Ah circuit that utilizes multiple resistance storage of some memristive devices to implement addition (accumulation) operation primitively [4], [18].

### C. Common Aspects

In summary, the existing memristive circuit designs that enable CIM-A architecture share the following aspects:

- Provide maximum level of parallelism within the memory, as execution is independent from sense amplifiers.
- Allow cascading of operations without being fed back to the array.
- Affect the memory endurance, as execution requires changing the state of the memristive devices.
- Require large drivers for computing, as execution requires high voltages to be applied to the memory array (high power consumption).
- Require redesigning the memory array to support computing, as the conventional optimized structure of memory array (including bitlines, wordlines and memory cells) may not allow for correct computation-in-memory operations.

### IV. MEMRISTIVE CIRCUIT DESIGN FOR CIM-P ARCHITECTURE

The CIM-P architecture uses the memristive CIM-Pr and CIM-Ph circuit designs to perform primitive logic (e.g., OR) and arithmetic (e.g., vector-matrix multiplication) operations. The existing CIM-Pr circuit designs include Pinatubo [11], MPIM [28], Scouting [12], STT-CiM [29], and Hielm [30]. These designs implement primitive logic operations such as OR, AND and XOR functions. On the other hand, the existing CIM-Ph circuit designs execute logic and arithmetic operations; including vector-matrix multiplication [13]–[15], [31] and vector-matrix-matrix multiplication [32].

The following subsections discuss two design examples in details: scouting logic [12] as an example of CIM-Pr design, and vector-matrix multiplication [15] as an example of CIM-Ph design. Pinatubo, MPIM, STT-CiM and Hielm are similar to scouting logic. Similarly, the other existing CIM-Ph designs (including designs described in [13], [31], [32]) are quite similar to the discussed example.

### A. Scouting Logic (CIM-Pr design example)

Scouting logic design executes bit-wise OR, AND and XOR logic functions. The concept of scouting logic is shown in Fig. 4, where devices \( M_1 \) and \( M_2 \) represent the resistive operands. The output is produced as voltage by the sense amplifier. The execution of scouting logic operations is based on reading multiple rows simultaneously, and activating the appropriate reference current of the desired bitwise operation. The value of the reference current depends on the executed operation as shown in Fig. 4. The reference current of OR function must be selected between \( \frac{V_{r}}{R_{GN}} \) and \( \frac{V_{r}}{R_{OFF}} \) (where \( V_{r} \) is the read voltage), whereas the reference current of the AND function must be selected between \( \frac{V_{r}}{R_{ON}} \) and \( \frac{V_{r}}{R_{GN}} \) [12]. The XOR function requires two reference currents. The design of scouting logic is based on read operations; hence, it does not destruct input data (states) neither impact the endurance.

### B. Vector-Matrix Multiplication (CIM-Ph design)

The vector-matrix multiplication design described in [15] is implemented using 1T1R (1 transistor and 1 resistive device) array as shown in Fig. 5; the inputs consist of the binary vector \( a=[a_1, a_2, a_3] \) provided via the wordlines of the array, and the binary matrix \( B=[b_{ij}] \) stored in the array. Each bit line will provide one element of the output vector. In this design, the...
multiplication is performed as AND function, and the addition as OR function. The execution starts by precharging the bitlines of the array, and then applying the vector elements to the wordlines. A precharged bitline will discharge (resulting in logic 1) only if at least one low ohmic path is created as a result of applying the vector elements; see Fig. 5. E.g., $BL_1$ will result in output 1 if $a_1 = 1$ and $b_{11} = 1$ (low ohmic state). As it is based on the read operations, this design does not destruct the state of the matrix elements, neither reduce their endurance. More complex designs of vector-matrix multiplication (e.g., Dot Product Engine [13], and ISAAC [14]) utilizes analog circuits in the periphery (e.g., ADC and DAC) to perform arithmetic multiplications and additions using ohm’s law and kirchhoff’s current law, respectively.

### C. Common Aspects

The existing memristive circuit designs that enable CIM-P architecture share the following aspects:

- Do not affect the memory endurance, as execution does not change the states of the memristive devices.
- Require small drivers, as execution requires low voltages to be applied to the memory array (low power consumption).
- Provide limited parallelism within the memory, as execution is dependent on the sense amplifiers. In the extreme case, each column may have a sense amplifier, resulting in a maximum parallelism at the cost area of overhead.
- Limit cascading of operations without feeding (or writing) the intermediate results back to the memory array, as the output is produced in the periphery.
- Entail less impact on the structure of memory array, as their designs focus on modifying the peripheral circuits to realize logic or arithmetic operations.

## V. POTENTIAL APPLICATIONS

The CIM architecture has a great potential to improve the overall performance and power consumption of at least some of emerging applications. Table I shows the different kernels (primitive operations) that can be implemented using memristive devices, the class of CIM architecture they enable, and some potential applications which make use of such kernels; hence, they could be accelerated. The kernels include bitwise operations (e.g., OR, XOR, and implication material functions) and arithmetic operations (e.g., addition, multiplication and vector-matrix multiplication operations).

<table>
<thead>
<tr>
<th>Kernels (operations)</th>
<th>CIM</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR</td>
<td>Ar</td>
<td>Ah database (bitmap indices, bitWeaving) [33]</td>
</tr>
<tr>
<td>AND</td>
<td>Ar</td>
<td>Ah hyper-dimentional computing, language recognition, biosignal processing [34]</td>
</tr>
<tr>
<td>XOR</td>
<td>Pr</td>
<td>Ph database (bitmap indices, bitWeaving), encryption, hyper-dimentional computing: language recognition, biosignal processing, k-mean clustering [35]</td>
</tr>
<tr>
<td>IMPLY-Majority</td>
<td>Pr</td>
<td>CAM [36]</td>
</tr>
<tr>
<td>Addition</td>
<td>Ar</td>
<td>Ah temporal correlation, factorization [4]</td>
</tr>
<tr>
<td>Multiplication</td>
<td>Ar</td>
<td>Ph</td>
</tr>
</tbody>
</table>

Note that a kernel may be implemented with different designs resulting in different CIM architectures. For example, OR logic function can be implemented using IMPLY, Majority logic, or Scouting logic, resulting in CIM-Ar, CIM-Ah and CIM-Pr, respectively. Note also that this is not applicable to all kernels, at least as of today. The third column of the table shows different applications that could make use of the corresponding kernel/architecture. For example, the temporal correlation utilizes the addition kernel implemented in CIM-Ah architecture [4]; the one-time-pad cryptography encryption is accelerated using the XOR kernel implemented in CIM-Pr architecture [33]; the convolutional neural networks can be implemented more efficiently using vector-matrix multiplication in CIM-Ph architecture [14].

It is worth noting that most of the applications found in the literature investigate the potential of vector-matrix multiplication kernel in CIM-Ph architecture to boost the efficiency of computation. In addition, and as Table I clearly shows, it is not clear yet which applications could make use of some kernels and its associated architectures, as it is the case for OR kernel in CIM-Ar and CIM-Ah architectures (first row in the table).

## VI. CONCLUSION

This paper has shown the importance of understanding and analyzing the nature of each memristive device based circuit design for logic or arithmetic operations. It does not allow only for selection of the right computation-in-memory architectures, but also for definition of the right applications that could significantly benefit from them.