Memristors Development and In-Memory Computing Architecture

Wenjing Chen *

Department of Electronic Engineering, Fuzhou University, Fujian, China, 120000

* Corresponding Author Email: 832201115@fzu.edu.cn

Abstract. With the development of neural network, Von Neumann's bottleneck is hanging over heads. The emergence of memristor and computing in memory architecture has greatly improved data processing speed and energy efficiency. This review provides a brief literature review of memristors, focus on development, and provide applications, challenges, and prospects of memristors. Firstly, the process of memristor research is introduced. Secondly, this paper highlights memristors development, which is divided into three time periods: 2008 before, 2008 to 2013, and 2013 to 2018. In 1971, Cai Shaotang predicted the existence of memristors. The Stan William team of HP Laboratory first prepared a memristor through experiments in 2008. In 2015, UCSB used a 12 times 12 array to realize the recognition of three letters. Memristor optimization work of the Tsinghua University team is introduced. Thirdly, this paper discusses the applications, and challenges of memristors and gives useful advice on optimization work in the future. Application is divided into 3 parts. In addition to many other difficulties in circuit and system designs, memristive neural networks require much larger array sizes and peripheral circuit integration with computing kernels for maximum energy and throughput efficiency in order to be used for more complex computing tasks.

Keywords: Memristors, computing in memory, development.

1. Introduction

In 1971, Leon Chua introduced the memristor as the fourth basic circuit element [1]. In 2008, Strukov et al. declared that they had discovered the memristor after realizing that the behavior of their molecular electrical switches was caused by the titanium electrodes rather than the organic layer [2-3]. Ella Gale discussed memristor models, switching mechanisms, and material properties in 2014 [4]. According to P. Yao et al., they performed grey-scale face categorization utilizing an integrated 1024-cell array in 2017. Their experimental findings confirm the viability of an analog synaptic array and open the door to the development of a large-scale, energy-efficient neuromorphic system [5]. Utilizing physical laws directly, Crossbar arrays offer huge parallelism and quick in-memory computing. The computation may be completed very quickly and with vast parallelism utilizing only physical laws, which significantly increases computing throughput. Time and energy usage can be further decreased thanks to sensors' direct analog information processing capacity [6]. Mehonic, Adnan, et al. examined the memristor in 2020 as a possible means of implementing spiking neural networks, DL accelerators, and power-efficient in-memory computing. Two important topics are the need for specialized learning and inference algorithms and the usage of non-von-Neumann computing architectures [7]. Kumar S. et al. talk about how new material characteristics in memristor devices and systems define many orders of complexity and allow for complicated dynamics. New computer architectures that provide better energy efficiency and higher computational capacity, are made possible by these intrinsic complex dynamics at the device level [8]. In order to guide the hardware implementation of NCS based on large-scale CBAs, Li Y et al.studied the difficulties encountered by hardware ANNs in terms of device performance and array parameters, and also proposed effective solutions to deal with the difficulties [9]. A completely integrated memristor device with improved learning capability and reduced energy consumption was created in 2023 by Zhang W et al. In their research, the general methods of STELLAR has significant advantages, which enable on-chip learning through the use of a memristor crossbar array. In this investigation, speech recognition, picture categorization, and motion control were all performed [10]. The use of memristive devices to

achieve synaptic function makes it possible to modify the neuromorphic system through lower cost and physical size as well as energy savings [11]. Based on the existing research papers, this paper will analyze and discuss the prediction, development, application and future prospects of memristors.

2. Method

2.1. Prediction

As Figure 1 shows, in 1971, Cai Shaotang predicted the existence of memristors. He mentioned in the paper that there are four basic circuit elements. The current-voltage relationship corresponds to resistance, the relationship between voltage and charge corresponds to capacitance, and the current-magnetic flux relationship corresponds to inductance. He designated the memristor as the fourth element capable of reflecting the charge-magnetic flux relationship. The charge passing through a memristor determines its resistance. Memory charge is therefore caused by the memristor's resistance, which may be used to measure the amount of charge passing through it. To put it briefly, it is a part whose resistance is influenced by the charge of the current.

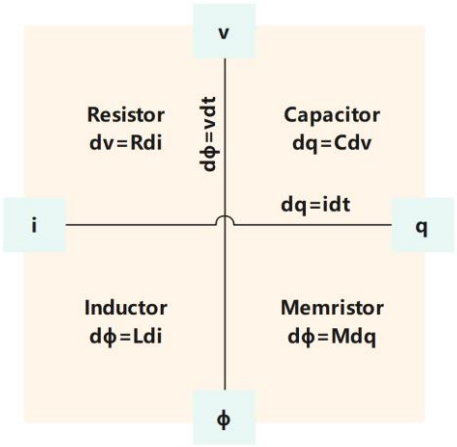


Figure 1. The existence of memristor

2.2. Development

In 2008, the Stan William team of HP Laboratory first prepared a memristor through experiments. In 2010, Professor Lu Wei of the University of Michigan first realized non-volatile continuous resistance switching and proposed the concept of electronic synapses. In 2013, Stanford University developed a high-speed and low power memristor.

2.3. Structure of the memristor

Figure 2 shows the structure of memristor. Firstly, semiconductor film is added between the two electrodes. The film consists of two parts, one is doped with cations with good conductivity, and the other is not doped poor conductivity. The cation will travel in the direction of undoping under the action of voltage if a positive voltage is provided at both ends of the electrode at this time, then the doping area will become larger, and the entire resistance becomes smaller. While when we apply reverse voltage, it will increase. Once the current stops, the resistance value will remain in the current state.

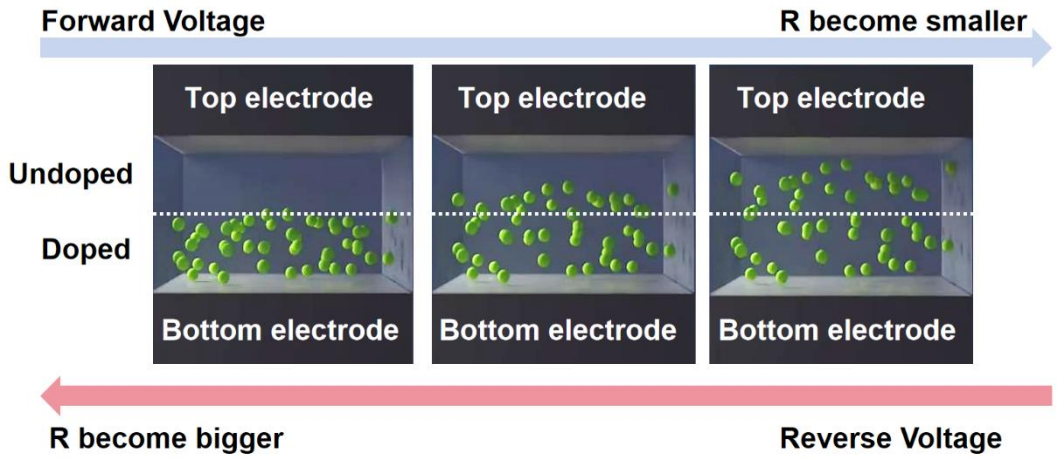


Figure 2. Structure of memristor

2.4. Computing in memory architecture

When we combine it with the operation in AI, a computing in memory architecture beyond the von Neumann architecture appears. As the Figure 3 shows, there are three horizontal wires and three vertical wires, nine intersections, and a memristor is connected at each intersection, a simple memristor array is formed. We apply three voltages on these three horizontal wires, and the current is generated on the wire. There are memristors at intersection of each wire. We correspond the conductance of the nine memristors to the nine values in the matrix. Current equals to conductance multiply voltage. We measure the total current on each vertical,which enables us to complete the convolution operation quickly. Compared with the traditional structure, memristor has the advantages of non-volatile and low power consumption, so many people try to use this array in machine learning calculation after the emergence of it.

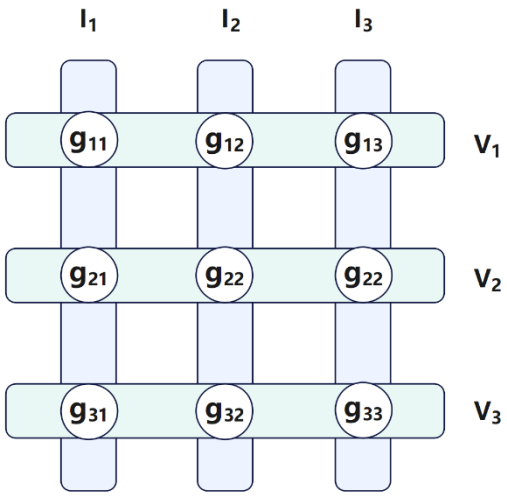


Figure 3. Cross bar array

2.5. Deeper research and Optimization Work

In 2015, UCSB used a 12 × 12 array to realize the recognition of three letters. In 2017, Tsinghua used a 128 × 8 array to realize face recognition in 2018, HP and Massachusetts University demonstrated MNIST data set recognition on 8K array.

With the deepening of research on memristors, many teams are not only expanding the scale of the cba, but also working on improving its accuracy and optimizing its structure. Tsinghua team is one of the teams.The Tsinghua team has mainly done optimization work at the following two levels, just as Figure 4 shows : The uncertainty of the ion diffusion causes the resistance change to be discontinuous and unstable.they alternately stack the conductive layer and the non-conductive layer,

and add a heat exchange layer with low thermal conductivity at the top.on the one hand, the ion diffusion is controlled by stratification, on the other hand, a more stable temperature is obtained.

The memristor unit is divided into two parts: positive weight and negative weight, which realizes the potential parallel function of the memristor. At the same time, a threshold is added to the error vector E to filter out some small errors in advance. The symbol vector given in this way has better training performance. They have made a complete integrated chip of storage and computing, and installed the chip on a car, so that the car can adapt to the new scene only through off-line on-chip learning, whether it is day or night can achieve smooth operation. During the character recognition test, the Tsinghua team also repeated the test for 48 days, and the results were stable. For the development of chips with significant learning capacity and excellent energy efficiency in the future, they have completed a significant step.

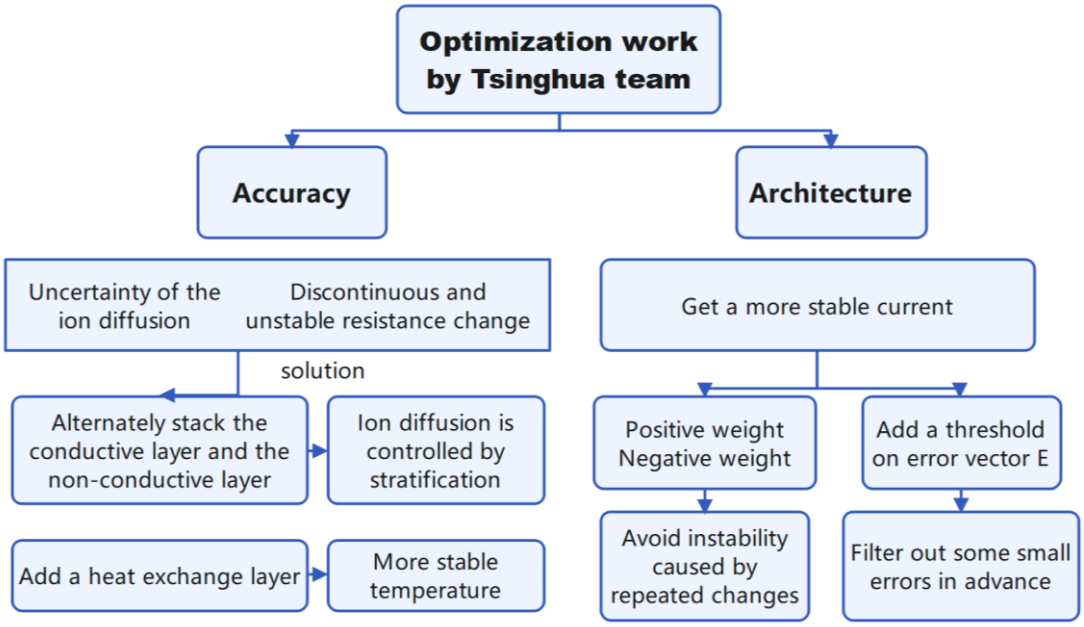


Figure 4. Optimization work by Tsinghua team [10]

3. Result

There are three applications of memristor: Memristors in Spiking Neural Networks (SNNs), Deep Learning Acceleration with Memristors, and In-Memory Computing with Memristors. To begin with, computing in memory, which store data by memory and conduct calculation at the same physical location, may greatly increase efficiency. Secondly, the use of memristors for DL acceleration has long been explored. Additionally, memristors can directly perform some of the activities of biological synapses and neurons. The three most significant ones are pulse firing, neuron-like integration, and synaptic plasticity.

The Von-Neumann design in 1940s separated the Memory and processor units. During carrying out various computational tasks, a significant quantity of data must be sent by many times between inner units. One of the main sources of performance bottlenecks for a variety of applications, especially common workloads connected to artificial intelligence, is the delay and energy incurred when retrieving data from memory units [12]. With in-memory computing, the memory wall of modern computing systems may be removed, and the traditional computer architecture subverted. Utilizing the device's mechanics, several approaches have been proposed to perform digital, analog, and stochastic computing inside resistive switching devices [13]. This is achieved by utilizing the physical properties, array-level design, peripheral circuitry, and control logic of the memory devices all at once. Not only does in-memory computing save the energy and latency costs of data transfer,

but it can also significantly lower the computational time complexity of certain computational jobs, which may pave the way for the development of non-von Neumann computing [14].

DNNs have shown to be one of the most effective techniques for achieving significant advancements in artificial intelligence, owing to advancements in CMOS scaling, algorithm development, and architectural design. Moreover, DNN may be built using memristive crossbars, representing synaptic weights in a massively parallel form. Different kinds of memristors might theoretically be used to build the neurons that link each layer, which are now generally constructed with amplifiers based on CMOS circuitry [15]. SNNs are expected to be more energy-efficient and more accurate in simulating the brain, particularly when processing temporal input like voice and video [16]. SNNs constructed with memristive devices, like DNNs, avoid the von Neumann bottleneck more effectively since they do not require a separate unit to hold synaptic weights. Memristors have been researched to mimic the brain integrate-and-fire and synaptic plasticity processes, as well as to incorporate intriguing learning principles seen in their biological counterparts [17–22]. To create extra crucial features, a great deal of effort has to be done on both materials and devices, despite the positive findings that have recently been obtained in showing Hebbian-like learning and engineering critical dynamics into standard memristors for SNNs [6].

Beyond the period of von Neumann and Moore's law, memristor-based designs have demonstrated considerable potential for constructing future computing systems [23]. Although memristor technology has great potential for many uses, in order to be commercially successful, a number of important issues must be resolved. Variability in device conductance has an impact on power usage, particularly in high-resistance regimes. Lack of an ideal selector, leading to issues such as sneak path current and half-select in dense crossbar arrays. In ultra-large networks, wire resistance influences synaptic weight programming accuracy during training. Intrinsic stochasticity of memristive switching behavior, which needs to be considered during synaptic learning. Nonlinear I-V relationship in lower conductance range, requiring research on how to use it efficiently.

Table 1 shows the development of experimentally memristor crossbar arrays. In the future, optimization work should be done in following parts: Material selection and engineering to improve stability and reduce drift of conductance states. Device and circuit engineering approaches to address variation in programming devices and reduce device variability. Exploiting the intrinsic stochasticity of memristive devices in emulating biological systems and optimizing computing. Exploring the use of the lower conductance range in memristors for efficient computing and reduced power consumption. Create methods for updating weights in memristive devices in a symmetric and linear manner to enable effective network training. Coherent collaboration across disciplines and levels to advance hardware, algorithms, architecture, system, circuit, and material research in memristor technologies.

Table 1. Development of experimentally memristor crossbar arrays

	Active arrays		Passive arrays		
Memristor size	4 × 4 μm2	NA	200 × 200 nm2	500 × 500 nm2	NA
Transistor node	2 μm	1.2 μm	No transistors	No transistors	No transistors
Max. array size	128 × 64	128 × 8	20 × 20	32 × 32	4 (wired) + 4 (software)
Conductance range	100–900 μS	10–60 μS	12–142 μS	0–5 μS	145–175 μS
Application demonstrations	Image compression/ filtering; MNIST classification; time-series regression; gait recognition. reinforcement learning	Face classification	3 × 3 or 4 × 4 bitmap classification	Sparse coding. reservoir computing	Unsupervised learning
Reference	24-28	5	29-30	31-32	33

4. Conclusion

Although the research of memristor is still in its infancy, especially in terms of architecture and algorithms, its development prospects are very broad. To advance the field and utilize the system's features for brain-like computation, an in-depth understanding of the nature and physical principles behind the memristor is indispensable [34]. At the same time, strategies for creating real-time, lower-power learning systems have always been greatly inspired by biology, and this will continue to be the case [35]. How to implement the in-depth research of memristors inspired by biology is essential for the future. It will need persistent and innovative research across several disciplines, including computer science, physics, chemistry, neurology, electrical and computer engineering, and others, to create cutting-edge computing systems of the future.

References

[1] Chua, L. "Memristor-the missing circuit element." IEEE Transactions on circuit theory 18.5: 507-519 (1971).

[2] Williams, R., "How we found the missing memristor," IEEE Spectr. (45) 28–35 (2008).

[3] Strukov, D. B., Snider, G. S., Stewart, D. R., & Williams, R. S. "The missing memristor found." nature 453.7191: 80-83 (2008).

[4] Gale, E. "TiO2-based memristors and ReRAM: materials, mechanisms and models (a review)." Semiconductor Science and Technology 29.10: 104004 (2014).

[5] Yao, P., Wu, H., Gao, B., Eryilmaz, S. B., Huang, X., Zhang, W., ... & Qian, H. "Face classification using electronic synapses." Nature communications 8.1: 15199(2017).

[6] Xia, Q., & Yang, J. J. "Memristive crossbar arrays for brain-inspired computing." Nature materials 18.4: 309-323 (2019).

[7] Mehonic, A., Sebastian, A., Rajendran, B., Simeone, O., Vasilaki, E., & Kenyon, A. J. "Memristors—From in-memory computing, deep learning acceleration, and spiking neural networks to the future of neuromorphic and bio-inspired computing." Advanced Intelligent Systems 2.11: 2000085(2020).

[8] Kumar, S., Wang, X., Strachan, J. P., Yang, Y., & Lu, W. D. "Dynamical memristors for higher-complexity neuromorphic computing." Nature Reviews Materials 7.7: 575-591 (2022).

[9] Li Y, Ang K W. "Hardware implementation of neuromorphic computing using large-scale memristor crossbar arrays." Advanced Intelligent Systems 3.1: 2000137(2021).

[10] Zhang, W., Yao, P., Gao, B., Liu, Q., Wu, D., Zhang, Q., ... & Wu, H. "Edge learning using a fully integrated neuro-inspired memristor chip." Science 381.6663: 1205-1211(2023).

[11] Nawrocki, Robert A., Richard M. Voyles, and Sean E. Shaheen. "A mini review of neuromorphic architectures and implementations." IEEE Transactions on Electron Devices 63.10: 3819-3829(2016).

[12] O. Mutlu, S. Ghose, J. Gómez-Luna, R. Ausavarungnirun, Microprocess. Microsyst. 67, 2 8 (2019)

[13] Ielmini, D., & Wong, H. S. P. "In-memory computing with resistive switching devices." Nature electronics 1.6: 333-343 (2018).

[14] Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. "Memory devices and applications for in-memory computing." Nature nanotechnology 15.7: 529-544(2020).

[15] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. "Mastering the game of Go with deep neural networks and tree search." nature 529.7587: 484-489(2016).

[16] Wang, W., Pedretti, G., Milo, V., Carboni, R., Calderoni, A., Ramaswamy, N., ... & Ielmini, D. Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses. Sci. Adv. 4, eaat4752 (2018).

[17] Jo, S. H., Chang, T., Ebong, I., Bhadviya, B. B., Mazumder, P., & Lu, W. Nanoscale memristor device as synapse in neuromorphic systems. Nano Lett. 10, 1297–1301 (2010).

[18] Ohno, T., Hasegawa, T., Tsuruoka, T., Terabe, K., Gimzewski, J. K., & Aono, M. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. Nat. Mater. 10, 591–595 (2011).

[19] Yu, S., Wu, Y., Jeyasingh, R., Kuzum, D. & Wong, H.-S. P . An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation. IEEE Trans. Elect. Dev. 58, 2729–2737 (2011).

[20] Wang, Z. Q., Xu, H. Y., Li, X. H., Yu, H., Liu, Y. C., & Zhu, X. J. Synaptic learning and memory functions achieved using oxygen ion migration/diffusion in an amorphous InGaZnO memristor. Adv. Funct. Mater. 22, 2759–2765 (2012).

[21] La Barbera, S., Vuillaume, D. & Alibart, F. Filamentary switching: synaptic plasticity through device volatility. ACS Nano 9, 941–949 (2015).

[22] Pickett, M. D., Medeiros-Ribeiro, G. & Williams, R. S. A scalable neuristor built with Mott memristors. Nat. Mater. 12, 114–117 (2013).

[23] Zidan, Mohammed A., John Paul Strachan, and Wei D. Lu. "The future of electronics based on memristive systems." Nature electronics 1.1: 22-29(2018).

[24] Li, C. et al. "Analogue signal and image processing with large memristor crossbars." Nat. Electron. 1, 52–59 (2018).

[25] Li, C. et al. "Efcient and self-adaptive in-situ learning in multilayer memristor neural networks." Nat. Commun. 9, 2385 (2018).

[26] Li, C. et al. "Long short-term memory networks in memristor crossbar arrays." Nat. Mach. Intell. 1, 49–57 (2019).

[27] Hu, M. et al. "Memristor-based analog computation and neural network classifcation with a dot product engine." Adv. Mater. 30, 1705914 (2018).

[28] Jiang, H. et al. "Sub-10 nm Ta channel responsible for superior performance of a HfO2 memristor." Sci. Rep. 6, 28525 (2016).

[29] Prezioso, M. et al. "Training and operation of an integrated neuromorphic network based on metal–oxide memristors." Nature 521, 61–64 (2015).

[30] Bayat, F. M. "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits." Nat. Commun. 9, 2331 (2018).

[31] Sheridan, P. M. et al. "Sparse coding with memristor networks." Nat. Nanotechnol. 12, 784–789 (2017).

[32] Du, C. et al. "Reservoir computing using dynamic memristors for temporal information processing." Nat. Commun. 8, 2204 (2017).

[33] Serb, A. et al. "Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses." Nat. Commun. 7, 12611 (2016).

[34] Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. "Memory devices and applications for in-memory computing." Nature nanotechnology 15.7: 529-544(2020).

[35] Lanza, M., Wong, H. S. P., Pop, E., Ielmini, D., Strukov, D., Regan, B. C., ... & Shi, Y. "Recommended Methods to Study Resistive Switching Devices http://researchonline. ljmu. ac. uk/id/eprint/9642."