SCIENCE CHINA
Information Sciences

• INSIGHT •

June 2018, Vol. 61 060425:1–060425:3
https://doi.org/10.1007/s11432-017-9378-3

Special Focus on Memristive Devices and Neuromorphic Computing

Challenges of memristor based neuromorphic computing system

Bonan YAN*, Yiran CHEN & Hai LI

Electrical and Computer Engineering, Duke University, Durham NC 27707, USA

Received 6 October 2017/Revised 17 December 2017/Accepted 23 January 2018/Published online 15 May 2018

Citation Yan B N, Chen Y R, Li H. Challenges of memristor based neuromorphic computing system. Sci China Inf Sci, 2018, 61(6): 060425, https://doi.org/10.1007/s11432-017-9378-3

The high efficiency of human brain illuminates the development of neuromorphic computing system (NCS). In human brain, membrane potential that goes beyond certain threshold voltage triggers the propagation of spike signals that carries the information to proceeding neurons. This process is usually abstracted as the integration and fire model, inspiring circuit design and architecture solutions for neuromorphic computing. Different from conventional von Neumann architecture, neuromorphic systems tend to closely integrate the computing units and memories, efficiently reducing the distance between computing units and memory and consequently breaking the so-called “memory wall” [1].

Confronted with the fast-growing computational demands of neural networks, many neuromorphic ASICs on conventional CMOS technology have been developed and demonstrated. Examples include tensor processing units (TPU) by Google and TrueNorth by IBM [2]. In these designs, the fundamental operations are fulfilled by high performance multiplication-and-accumulation (MAC) units. The synapse design is usually implemented by SRAM or capacitor, making high-density application very challenging.

As one prospective solution to this difficulty, memristor (a.k.a. resistive memory or ReRAM) is a two-terminal nonvolatile resistive device. When external programming circuit injects a current flow through a memristor, its resistivity changes accordingly within a certain range — as in an empirical model, the resistance (memristance) varies between the highest and the lowest resistances. The pulse programming scheme of a memristor hence demonstrates the plasticity as an electrical synapse and emulates its biological counterpart. The neuromorphic design integrating such memristor-based synapses also exhibits high power efficiency.

Memristor with up to 7-bit precision [3] was reported, which is sufficient to provide weight accuracy for implementing many neural networks. Spiking-timing-dependent plasticity (STDP) was demonstrated with second-order memristor device [4], greatly extending the possible network structures. Process-in-memory architecture employs memristor crossbar to assist general-purpose CPU to boost the performance and efficiency of machine learning applications [5]. Image reconstruction [6] and convolutional neural network (CNN) [7] are also paradigms to use memristors to merge memory and computation.

As memristor technology still requires much effort on development, the obtained yield is much lower than the commercialization criteria. Tech-

* Corresponding author (email: bonan.yan@duke.edu)

niques to mitigate process variations, dynamic disturbance and integration of emerging devices upon CMOS process are vital to enable wide use of memristor devices. In this stage, the enhancement from circuit and system perspectives is an effective approach to enable neuromorphic computing systems as practical use.

Architecture of memristor based NCS. The memristor based neuromorphic design can be realized in spiking-based or level-based forms. The explanations of these two different designs were explained in [8,9], respectively, by using the handwritten digit recognition as examples.

Figure 1 illustrates an array of memristors, where each cell on the cross point is associated with a selecting device. The selecting device can be either resistive selector, diode or transistor. In a neuromorphic operation, a set of input voltages $\boldsymbol{V} = \{V_1, V_2, \ldots, V_m\}$ are fed to source lines (SLs), together forming an input vector $\boldsymbol{V}$. According to Kirchhoff's law, the output currents at bit lines (BLs) $\boldsymbol{I} = \{I_1, I_2, \ldots, I_n\}$, can be represented as $\boldsymbol{I} = \boldsymbol{VG}$, where $\boldsymbol{G}$ is the conductance matrix composed of the conductance of memristor array. For those weights that are negative, the single weight is represented by a pair of conductance, on which the same voltage applies [10]. And the difference between the two currents is the effective negative weight.

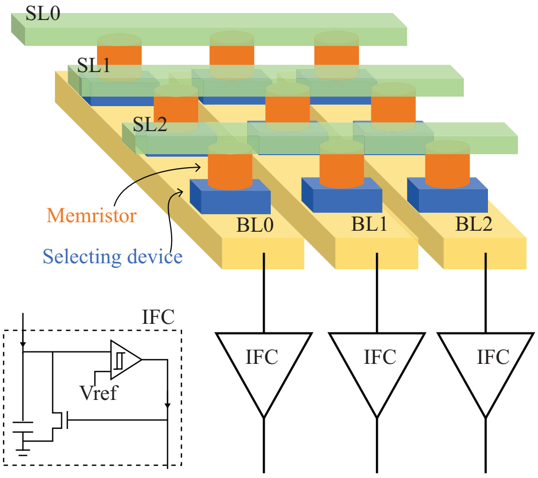

Figure 1 (Color online) Memristor based NCS diagram.

In the spiking-based NCS design, integrate and fire circuit (IFC) is commonly adopted to realize neuron functionality [11]. IFC, compared to analog/digital converter (ADC), features higher speed and lower power consumption [8, 11]. As illustrated in Figure 1, an IFC drains the current from BL and generates output spikes. The working mechanism of IFC can be summarized as a feedback loop that controls the charging and discharging of a capacitor within it [8].

The capacitor is used to track the duration and mimic the behavior of membrane potential firing output spikes. When such a design operates recall (testing/inference) function, the current from a BL charges the capacitor in its corresponding IFC. Once the voltage of the capacitor reaches a preset reference value Vref, the following comparator module switches on the loop and discharges the capacitor. By restoring the voltage on capacitor to binary digital signals, the output spikes are created. Within this scheme, higher current amplitude logically results in faster charging/discharging of capacitor, i.e., more output spikes in a given duration, and consequently lower output digitalization error.

The array usually is constructed with one-transistor-one-memristor (1T1R) or one-selector-one-memristor (1S1R) cells. Note that the effective cell conductance is sensitive to the voltage difference between SL and BL. For the 1T1R array, the body effect of the selective transistor affects the effective drain to source resistance. For the 1S1R based design, the non-steep nonlinearity yields large difference under different voltages [12].

Challenges of reliable memristor based NCS.

- Device level challenges. In memrisor-based NCS design, uniformity is the major challenge in device fabrication. It originates from randomness of filaments formation [13]. Moreover, this process variation is vulnerable to be magnified by the nonlinear dynamics of memrisor switching [14].

In contrast, ideally, a neural network trained at software level can be mapped to memristor arrays for recall operation. Unfortunately, real devices are not perfect, resulting in significant degradation in system reliability and robustness. Aside from device engineering efforts, techniques from circuit and system perspectives light the way of reliable design of memristor based NCS.

- Circuit level challenges. The integration of memristor crossbar array and CMOS based neuron circuits needs further calibration for better cooperation. In view of these problems, general solutions, facing different types of NCSs, are reviewed in this part.

For example, the interconnect IR-drop in memristor arrays could severely limit the scale of memristor crossbar based NCS and hence hinders the design scalability. The distortion of weight matrix can be ameliorated with IR-drop compensation scheme [15]. Another approach is to partition a large network [16] into many small ones in lower dimension. The impact of interconnect IR-drop is thereby lowered.

As the fabrication process of memristor devices is still under development, the yield at array level is not satisfying yet. Stuck-on and stuck-off defects can always be observed. At the algorithm level, we note that each individual weight has different impact on the system performance. Therefore, a possible solution is to remap the weights with less impact to defective cells while keeping those significant weights on qualified devices in an array [17].

Due to the intrinsic mechanism of memristor, analog memristance values are vulnerable to read disturbance in many times of use, that is, the memristor could drift from its originally programmed value after a number of accesses. An efficient realtime feedback controller was presented in [17] to lower the cost of disturbance compensation.

Acknowledgements This work was supported by National Science Foundation (NSF) (Grant No. CSR-1717885), and Air Force Research Laboratory (AFRL) (Grant No. FA8750-15-2-0048). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of AFRL or its contractors.

References

1 Wulf W A, McKee S A. Hitting the memory wall: implications of the obvious. ACM SIGARCH Comput Archit News, 1995, 23: 20–24

2 Schneider D. Deeper and cheaper machine learning. IEEE Spectr, 2017, 54: 42–43

3 Alibart F, Gao L, Hoskins B D, et al. High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. Nanotechnology, 2012, 23: 075201

4 Kim S, Du C, Sheridan P, et al. Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity. Nano Lett, 2015, 15: 2203–2211

5 Chi P, Li S C, Xu C, et al. Prime: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. In: Proceedings of the 43rd Annual International Symposium on Computer Architecture, Seoul, 2016. 27–39

6 Ma W, Cai F, Du C, et al. Device nonideality effects on image reconstruction using memristor arrays. In: Proceedings of International Electron Devices Meeting, San Francisco, 2016

7 Shafiee A, Nag A, Muralimanohar N, et al. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In: Proceedings of the 43rd Annual International Symposium on Computer Architecture, Seoul, 2016. 14–26

8 Liu C C, Yan B N, Yang C F, et al. A spiking neuromorphic design with resistive crossbar. In: Proceedings of the 52nd ACM/EDAC/IEEE Design Automation Conference, San Francisco, 2015

9 Indiveri G. A low-power adaptive integrate-and-fire neuron circuit. In: Proceedings of International Symposium on Circuits and Systems, Bangkok, 2003

10 Yan B N, Yang J H, Wu Q, et al. A Closed-loop design to enhance weight stability of memristor based neural network chips. In: Proceedings of International Conference on Computer-Aided Design, Irvine, 2017. 541–548

11 Liu C C, Yang Q, Yan B N, et al. A memristor crossbar based computing engine optimized for high speed and accuracy. In: Proceedings of IEEE Computer Society Annual Symposium on VLSI, Pittsburgh, 2016. 110–115

12 Yan B, Monmouth A M, Yang J, et al. A neuromorphic ASIC design using one-selector-one-memristor crossbar. In: Proceedings of International Symposium on Circuits and Systems, Montreal, 2016. 1390–1393

13 Liu Q, Long S B, Lv H B, et al. Controllable growth of nanoscale conductive filaments in solid-electrolyte-based ReRAM by using a metal nanocrystal covered bottom electrode. ACS Nano, 2010, 4: 6162–6168

14 Chua L O. Local activity is the origin of complexity. Int J Bifurcat Chaos, 2005, 15: 3435–3456

15 Liu B Y, Li H, Chen Y R, et al. Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems. In: Proceedings of International Conference on Computer-Aided Design, San Jose, 2014. 63–70

16 Wang Y D, Wen W, Liu B Y, et al. Group scissor: scaling neuromorphic computing design to big neural networks. In: Proceedings of the 54th Annual Design Automation Conference, Austin, 2017

17 Liu C, Hu M, Strachan J P, et al. Rescuing memristor-based neuromorphic design with high defects. In: Proceedings of the 54th Annual Design Automation Conference, Austin, 2017