Reliability Analysis of Memristor Crossbar Routers: Collisions and On/off Ratio Requirement

Junren Chen, Chenxi Wu, Giacomo Indiveri, Melika Payvand
Institute of Neuroinformatics, *University of Zurich and ETH Zurich*, Switzerland
junren, chenxi, giacomo, melika@ini.uzh.ch

2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS) | 978-1-6654-8823-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECS202256217.2022.9970901

***Abstract*—Memristors are commonly used in crossbar arrays as "in-memory computing" elements to solve the von-Neumann bottleneck problem. However, they can also be used as "in-memory routing" elements to configure on-chip interconnection schemes and route signals among computing elements in configurable multi-core neuromorphic processors. While there has been a significant focus on the use of memristive devices as in-memory computing elements, to date, studies on the fundamental reliability properties of memristors as routing elements are still missing. In this paper, we analyze the reliability issues of using these devices in routing crossbar arrays, caused by sharing routing resources (collisions), and undesired pulses due to the leakage paths (on/off ratio requirement). We show that there is a trade-off between routing collision probability and the degree of connectivity (i.e., fan-in) of the receivers sharing routing channels. We provide specifications and guidelines based on a theoretical analysis for engineering the properties of memristive devices, and for designing routing systems based on memristor crossbars.**

***Index Terms*—memristor crossbar, router, reliability, collision, on/off ratio**

I. Introduction

Unlike static random-access memory (SRAM) and dynamic random access memory (DRAM), memristors can operate in a non-volatile fashion, keeping their stored values for several years, without burning static power. They are reported to have low energy consumption, low switching latency, and small area [1], which make them a promising candidate for the next generation of memory devices. Furthermore, thanks to their non-volatile, nano-scale and multi-bit properties, memristors have been extensively used to model synaptic properties in hardware architectures for spiking neural networks (SNN) [2] and artificial neural networks (ANN) [3]. In addition to emulating synaptic functions, their property as memory devices can also be used to configure and re-program the connectivity between digital logic elements [4]. In digital re-configurable hardware, e.g., field programmable gate arrays (FPGA), where programmable interconnections are crucial for chip area and power consumption, several studies have demonstrated the advantage of memristor-based architectures. These memristor-based FPGAs show significant gain in routing resources and configuration storage improvements in system performance, area and power consumption [5]–[8]. Recently, we have proposed to use this memristor-based routing approach also in multi-core SNN neuromorphic processors, in which neurons communicate with each other via spike events [9].

Currently, neuromorphic systems use the Address-Event Representation (AER) protocol to transmit spike events among neurons, by using the source address, destination address, or a combination of source-destination address encoding in digital data packets [10], [11]. These addresses are typically stored in on-chip SRAM or/and Content Addressable Memory (CAM) cells. These digital memory and routing logic circuits use a significant area overhead, and can take up to more than half of the chip area in mixed-signal neuromorphic processors [11].

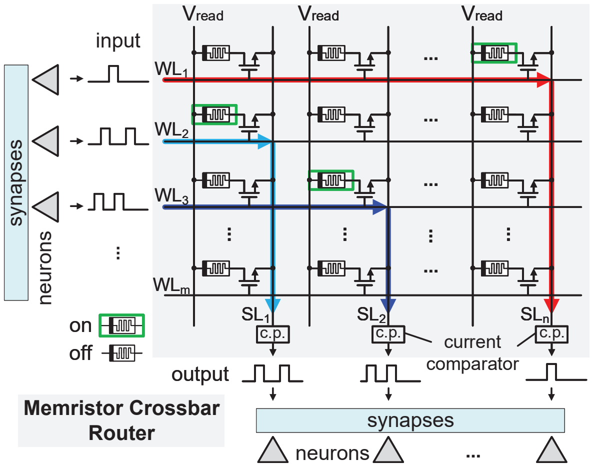


Fig. 1. An approach of using memristor crossbars for routing signals in neuromorphic systems. Identical rectangular voltage pulses depict spike events of spiking neurons. "On" and "Off" states of the memristors determine whether the signals are transmitted. Each column in the crossbar array is one routing channel. This idea can be generalized for systems that transmit information using voltage pulses.

Motivated by the enhancement of memristor-based routing schemes, we propose to replace the CMOS-based digital routing logic and memory cells of mixed-signal neuromorphic chips, used for neuron connectivity configurations, with memristor crossbars. Fig. 1 illustrates this approach. A 1T1R (1-transistor-1-memristor) crossbar array is used as a crossbar router. Input signals are applied as voltage pulses on the Word Lines (WLs), connected to the gate terminal of transistors. In the routing mode, the read voltages (V_{read}, e.g., 0.2 V) are constantly applied onto the top electrode of memristors. On its arrival, the spike-event pulse, enables the N-type transistors, resulting in a current flow through memristors. All these currents are summed on the Source Lines (SLs), and the result is sensed by current comparators ("c.p."). These elements compare the total current on each SL to a reference current. If the current on one SL is greater than the reference, then the comparator generates an output voltage pulse. Otherwise, the output remains at zero voltage. Therefore, by configuring the state of the memristors, the inputs are transmitted to the desired destinations or are blocked: “On” state, i.e., Low Resistance State (LRS) propagates the inputs, while “Off” state, i.e., High Resistance State (HRS), blocks the inputs.

While this scheme has many potential advantages over the pure CMOS AER scheme, it has some reliability issues.

1) Since this routing scheme does not include arbiters, spike-events being routed might collide through shared paths. We call this the “collision” problem, shown in red in Fig. 2.
2) Since “Off” cells can leak currents, there could be undesired output signals generated by the comparators. We call this the “undesired pulse” issue, shown in blue in Fig. 2.

To the best of our knowledge, these reliability issues have not been analyzed and characterized quantitatively so far. Here, we perform a theoretical study on the trade-off between the degree of network connectivity (fan-in of receivers) and routing collision probability, along with the required memristor’s on/off ratio for the application of routing information.

In the next section, we describe these reliability issues in detail; in Section III we present our methodology, and in Section IV we report the results of our analysis.

II. Reliability Issues

A. Collisions: derived from “On” Cells

The “collision” problem happens when multiple senders try to use the same routing channel at the same time to send information to a receiver. Each column in the crossbar array is an individual routing channel. The fan-in of the receiver can be increased by programming multiple cells in one column to “On” state to share the channel. This strategy increases the utilization ratio of the hardware resources, but input events may collide. As shown in Fig. 2, in the left column, when the input pulses of the two “On” cells (marked with red boxes) overlap during the pulse width T_{pw}, the two inputs merge together into one pulse on the same output wire. Thus, the receiver misses some information from the senders when collision happens in the signal transmission process. This collision issue leads to the study of collision probability in Section IV-A.

B. Undesired Output Pulse: derived from “Off” Cells

The other reliability issue originates from the physical characteristics of the memristor crossbar. The high resistive “Off” cells produce leakage currents to the output wire at the arrival of the input pulses, which enable the select transistors (shown in blue arrows in Fig. 2). When multiple inputs arrive at the same time, the leakage currents can accumulate. If the net accumulated leakage current from the “Off” cells is equal to or greater than the current of an “On” cell ($I_{leak} \geq I_{on}$) in one column, then an undesired output voltage pulse is generated without any input pulse applied to the “On” cell, or without any “On” cells present in the column.

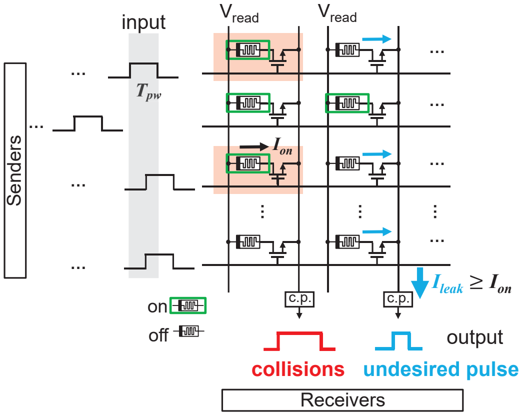


Fig. 2. Illustration of the two reliability issues in memristor crossbar routers when the input pulses overlap: collisions caused by multiple “On” cells in the same column (in red), and the undesired output voltage pulse caused by the leakage current from “Off” cells in the same column (in blue) if the accumulated leakage current is higher than the current of an “On” cell.

This “undesired pulse” issue indicates that there is a lower bound on the conductance on/off ratio of memristors for the receivers to avoid receiving extra “error” information without any valid inputs, which leads to the study of on/off ratio requirement of memristors in Section IV-B.

III. Methods

Reliability issues appear when multiple input pulses arrive at the same time, i.e., when they overlap. It is important therefore to study the number of incoming overlapping pulses.

We perform our analysis under the assumption that the inputs to the crossbar router are independent identically distributed (iid) Poisson processes. This assumption is made due to the fact that both the spike trains of spiking neurons in neuromorphic systems [10], and data frames in computer networks [12] are typically modeled as Poisson processes. As visualized in Fig. 3, the probability of k pulses occurring in one column (routing channel) during T_{pw} is given by

$$Pr\left\{X=k\right\} = \frac{(\lambda T_{pw})^k}{k!}e^{-\lambda T_{pw}} \tag{1}$$

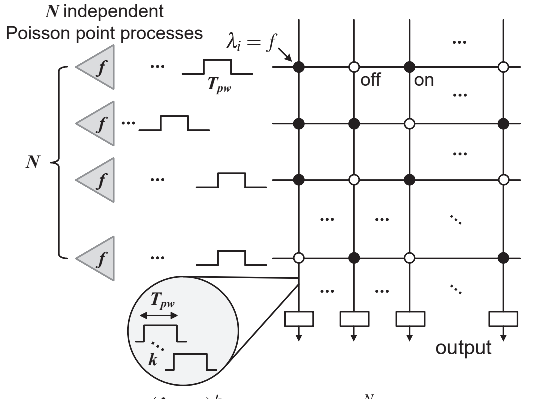
$$\lambda = \sum_{i=1}^{N} \lambda_i = Nf \tag{2}$$

where X is a random variable that depicts the number of incoming input pulses, f is the mean rate of each input pulse train, T_{pw} is the routing pulse width, λ is the mean rate of summed N inputs, λT_{pw} is the expected number of input pulses in T_{pw}.

IV. Results

A. Collision Probability

To transmit a pulse without collisions, the previous pulse must occur at least T_{pw} seconds earlier, and the next pulse must occur at least T_{pw} seconds later. Pulses are forbidden in $2T_{pw}$, centered around the time that the transmission starts [10]. Thus, the probability of a pulse without collision



$$Pr\{X = k\} = \frac{(\lambda T_{pw})^k}{k!} e^{-\lambda T_{pw}} \qquad \lambda = \sum_{i=1}^{N} \lambda_i = Nf$$

Fig. 3. Method for analyzing the two reliability issues. Modeling the inputs as independent Poisson pulse trains. $Pr\{X = k\}$ is the probability of k input pulses occurring in T_{pw}, which is the key to study the reliability of the router.

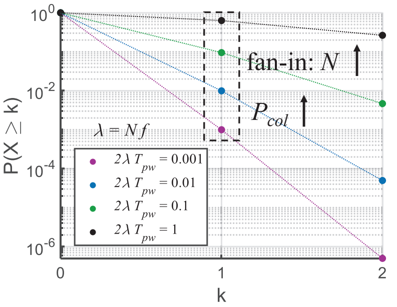


Fig. 4. The collision probability per pulse is the cumulative probability $P(X \geq 1)$ in $2T_{pw}$ (equation 3). Lower rate of inputs (Nf) and pulse width (T_{pw}) reduces signal transmission collision probability. There is a trade-off between the increase of fan-in (N) of receivers and the collision probability.

is $P(0, 2T_{pw}) = e^{-2\lambda T_{pw}}$, and collision probability (P_{col}) per pulse is expressed as:

$$P_{col} = 1 - e^{-2\lambda T_{pw}} \tag{3}$$

If $2\lambda T_{pw}$ is small (e.g. $2\lambda T_{pw} < 0.1$), we can approximate P_{col} as:

$$\begin{aligned} P_{col} &= 1 - e^{-2\lambda T_{pw}} \\ &= 1 - \left[1 - 2\lambda T_{pw} + \frac{(-2\lambda T_{pw})^2}{2!} + ... + \frac{(-2\lambda T_{pw})^n}{n!}\right] \\ &\approx 2\lambda T_{pw} = 2NfT_{pw} \end{aligned} \tag{4}$$

Therefore, when λT_{pw} is small, the collision probability is proportional to the fan-in (N) of the receiver, and T_{pw}. Lower rate of inputs (λ) and pulse width values (T_{pw}) reduce the signal transmission collision probability. As it is shown in Fig. 4, when $2\lambda T_{pw} < 0.1$, one order of magnitude increase in λT_{pw} results in one order of magnitude of increase in collision probability as well. To make the best of the resources of dense crossbars, this trade-off between the fan-in and collision probability should be taken into account when mapping interconnections between computing elements.

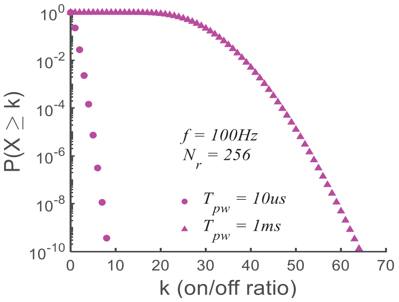


Fig. 5. Probability of an undesired output pulse from the crossbar router with independent Poisson input pulse trains. N_r is the size of the crossbar.

B. On/off Ratio Requirement of Memristors

As is discussed in Section II-B, the on/off ratio of memristors should be high enough to ensure that the accumulated leakage current in one column is lower than the current of one "On" cell. Otherwise, the receiver would receive an undesired signal, without the senders having sent any signals. Therefore, the number of incoming inputs occurring in T_{pw}, i.e., k, is the required minimum on/off ratio of memristors. We study and find the bounds on k for four cases with different correlations within the input.

1) Average case: All inputs are independent Poisson processes. Based on equation 1, we have calculated the cumulative probability $P(X \geq k)$ in T_{pw}. This is shown in Fig. 5 plotting the probability of an undesired output pulse from the router with respect to the memristive on/off ratio, k. We assume the size of the crossbar router $N_r = 256$, and the mean rate of each input $f = 100Hz$ (e.g. the firing rate of spiking neurons in neuromorphic systems usually is lower than hundreds of Hertz). If we route signals with $T_{pw} = 10\mu s$, the on/off ratio, k, of 10 can reduce the probability of an undesired output to 10^{-10}. If $T_{pw} = 1ms$, the required on/off ratio for this amount of probability is 65.

2) α fraction of synchrony in inputs: α fraction of the inputs are synchronized, where $0 < \alpha < 1$, and the remaining 1-α fraction of the inputs are independent Poisson processes. The illustration is shown in Fig. 6. $\hat{k}$ is the number of inputs during T_{pw} with Poisson distribution. Considering the worst condition that the synchronized αN_r inputs and the $\hat{k}$ inputs overlap, the total number of inputs that could arrive in T_{pw} is given by:

$$k = \hat{k} + \alpha N_r \tag{5}$$

with $\hat{k}$ inputs following the distribution:

$$Pr\left\{X = \hat{k}\right\} = \frac{(\hat{\lambda} T_{pw})^{\hat{k}}}{\hat{k}!} e^{-\hat{\lambda} T_{pw}} \tag{6}$$

where $\hat{\lambda}$ is:

$$\hat{\lambda} = (1 - \alpha) N_r f \tag{7}$$

According to the equations 5, 6, 7, the probability of an undesired output pulse generated by the router is shown in Fig. 7. It plots cumulative probability of equation 6 and the curves are shifted to the right by αN_r. In other words, the

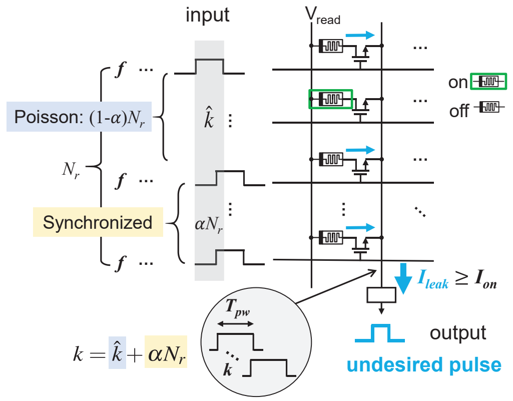


Fig. 6. Illustration of α percent of the inputs are synchronized and the others are independent Poisson processes. N_r is the size of the crossbar.

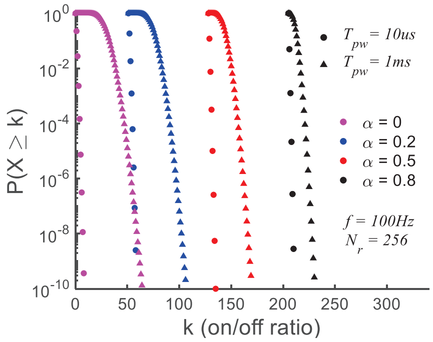


Fig. 7. Probability of an undesired output pulse with α fraction of synchronized inputs and $1-\alpha$ fraction of independent Poisson inputs. Synchronization in inputs leads to higher on/off ratio requirement.

higher synchrony of the inputs leads to a higher on/off ratio requirement. The variables f, N_r, and α all influence $\hat{\lambda}$, which in turn affects the slope of the curves $P(X \geq k)$ vs. k. Smaller $\hat{\lambda}$ and T_{pw} has steeper slope, which implies the higher efficiency of reducing the probability of undesired output pulses by increasing the on/off ratio of memristors.

3) Extreme case: $\alpha = 1$, all inputs are exactly synchronized. The requirement is on/off ratio $> N_r$ (size of the router). But in practice, this rarely happens in a system.

4) Correlations in inputs: Some inputs of the router are correlated, i.e. part of the input space always follow another part of it, after a certain time duration. In terms of on/off ratio requirement, it is similar to α fraction of synchrony. This case will be a case between the synchrony and the extreme cases, which we already calculated.

According to the literature, values of on/off ratio $> 10^3$ are reachable [13], [14]. Hence, building a memristor crossbar router that can support $> 10^3$ inputs is desirable.

Conclusion

We analyzed both system and device requirements for using memristive devices in crossbar arrays as routing elements. Based on this analysis, we can conclude that:

(i) From a circuits and system perspective, the fan-in of the receiving nodes can be increased by sharing routing channels in the memristor crossbar routers with low collision probability (e.g., $< 1\%$) depending on the routing pulse width and frequency of inputs, and by using circuit design techniques for reliably transmitting and receiving short pulses. (ii) From a device perspective, using devices with an on/off ratio > 10 helps in reducing the probability of possible undesired "error" output signals to very low probability (e.g., 10^{-10} under reasonable use case assumptions). Synchrony and correlations in inputs however lead to a higher on/off ratio requirements.

Although we carried out our analysis for event-based neuromorphic systems, our results are in principle applicable to any electronic systems which routes signals using voltage pulses through crossbar routers.

Acknowledgment

J. Chen would like to thank his colleague Zhe Su for discussions. This work is supported by China Scholarship Council, ETH Zurich Research Partnership Grant (RPG 072021_03) and in part by the innovation programme for ECSEL grants ANDANTE (grant agreement No. 876925).

References

[1] M. Zangeneh and A. Joshi, "Design and optimization of nonvolatile multibit 1t1r resistive ram," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 8, pp. 1815–1828, 2014.

[2] M. Payvand, M. V. Nair, L. K. Müller, and G. Indiveri, "A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: from mitigation to exploitation," *Faraday Discuss.*, vol. 213, pp. 487–510, 2019.

[3] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature*, 2020.

[4] Q. Xia, W. Robinett, M. W. Cumbie, N. Banerjee, T. J. Cardinali, J. J. Yang, W. Wu, X. Li, W. M. Tong, D. B. Strukov *et al.*, "Memristor-cmos hybrid integrated circuits for reconfigurable logic," *Nano letters*, vol. 9, no. 10, pp. 3640–3645, 2009.

[5] Y. Chen, J. Zhao, and Y. Xie, "3d-nonfar: Three-dimensional non-volatile fpga architecture using phase change memory," in *Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design*, 2010, pp. 55–60.

[6] Y. Y. Liauw, Z. Zhang, W. Kim, A. El Gamal, and S. S. Wong, "Non-volatile 3d-fpga with monolithically stacked rram-based configuration memory," in *2012 IEEE International Solid-State Circuits Conference*. IEEE, 2012, pp. 406–408.

[7] J. Cong and B. Xiao, "Fpga-rpi: A novel fpga architecture with rram-based programmable interconnects," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 4, pp. 864–877, 2013.

[8] X. Tang, P.-E. Gaillardon, and G. De Micheli, "A high-performance low-power near-vt rram-based fpga," in *2014 International Conference on Field-Programmable Technology (FPT)*. IEEE, 2014, pp. 207–214.

[9] T. Dalgaty, F. Moro, Y. Demirag, A. De Pra, G. Indiveri, E. Vianello, and M. Payvand, "The neuromorphic mosaic: re-configurable in-memory small-world graphs," 2021.

[10] K. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 5, pp. 416–434, 2000.

[11] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (dynaps)," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 1, pp. 106–122, 2018.

[12] A. S. Tanenbaum and D. Wetherall, *Computer networks*, 5th ed., 2011.

[13] H. Lee, P. Chen, T. Wu, Y. Chen, C. Wang, P. Tzeng, C. Lin, F. Chen, C. Lien, and M.-J. Tsai, "Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust hfo2 based rram," in *2008 IEEE International Electron Devices Meeting*. IEEE, 2008, pp. 1–4.

[14] W. Guan, S. Long, Q. Liu, M. Liu, and W. Wang, "Nonpolar nonvolatile resistive switching in cu doped ZrO_2," *IEEE Electron Device Letters*, vol. 29, no. 5, pp. 434–437, 2008.