FEATURED ARTICLES

Emerging Memory Devices Beyond Conventional Data Storage: Paving the Path for Energy-Efficient Brain-Inspired Computing

Emerging Memory Devices Beyond Conventional Data Storage: Paving the Path for Energy-Efficient Brain-Inspired Computing

by Rashmi Jha

The current state of neuromorphic computing broadly encompasses domain-specific computing architectures designed to accelerate machine learning (ML) and artificial intelligence (AI) algorithms.[1] As is well known, AI/ML algorithms are limited by memory bandwidth.[2] Novel computing architectures are necessary to overcome this limitation. There are several options that are currently under investigation using both mature and emerging memory technologies. For example, mature memory technologies such as high-bandwidth memories (HBMs) are integrated with logic units on the same die to bring memory closer to the computing units.[3] There are also research efforts where in-memory computing architectures have been implemented using DRAMs or flash memory technologies.[4,5] However, DRAMs suffer from scaling limitations, while flash memory devices suffer from endurance issues.[6,7] Additionally, in spite of this significant progress, the massive energy consumption needed in neuromorphic processors while meeting the required training and inferencing performance for AI/ML algorithms for future applications needs to be addressed.[8] On the AI/ML algorithm side, there are several pending issues such as life-long learning, explainability, context-based decision making, multimodal association of data, adaptation to address personalized responses, and resiliency. These unresolved challenges in AI/ML have led researchers to explore brain-inspired computing architectures and paradigms. It is noteworthy that a biological brain naturally addresses these issues while consuming just a fraction of the amount of energy required by a conventional computer.

When it comes to brain-inspired paradigms of computing, memory devices used for storing weights in neuromorphic computers are compared to biological synapses. A biological process engine (PE) can be considered as an aggregate of neurons connected via synapses. A fundamental difference between neuromorphic PE (shown in Fig.1(a)) and biological PE (shown in Fig.1(b)) is that a biological synapse changes conductance based on learning rules, which reconfigures the signal transmission pathways between neuronal populations.[9] This seemingly simplistic approach serves as a basis for biological computing.

But then one ponders why it has been so difficul to replicate the computing paradigms of the brain? Biological synapses are diverse in morphology and functionality. They also demonstrate dynamic behavior on multiple time scales, such as short-term plasticity (STP), which forms the basis of working memory and sensory information filtering.[10] Dendritic architectures and distribution of synapses on dendrites also play a critical role in biological computing by modulating signal delays.[11] Several reports indicate that data is stored in the form of spatiotemporal clusters of synapses in the brain.[12] Additionally, beyond Hebbian learning based on pre- and post-neuronal spiking times, a third factor such as neurotransmitters or rewards that convey information about success can play an important role in learning which can be accommodated by biological synapses.[13] Conventional memory elements (such as DRAM, SRAM, flash) lack the versatility of biological synapses. This limitation is where the true benefit of emerging memory technologies can be leveraged, as many of the emerging memory devices can be engineered to manifest the "dynamic behavior."

There are several emerging memory devices that are currently under investigation to replace or complement the conventional memory technologies in neuromorphic architectures.[14] This article will discuss resistive random access memory (RRAM) devices as

(continued on next page)

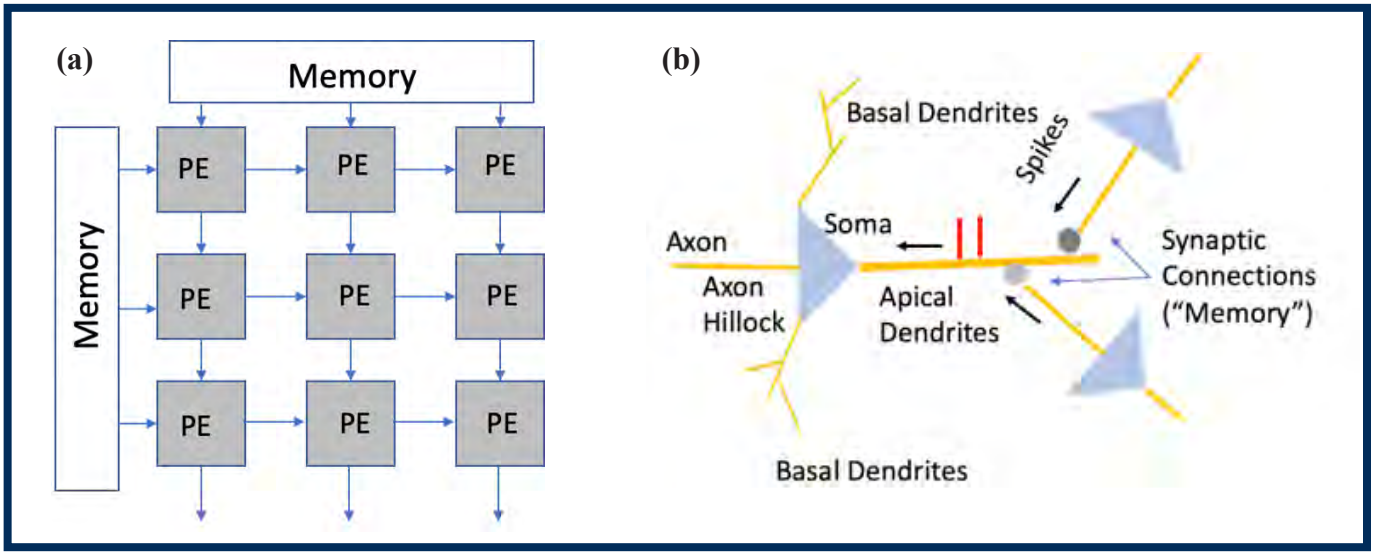

Fig. 1. *(a) Systolic array-based machine learning (ML) processing unit, (b) Neuro-synaptic processing unit in biological brain showing pyramidal neuron with complex dendritic architecture.*

(continued from previous page)

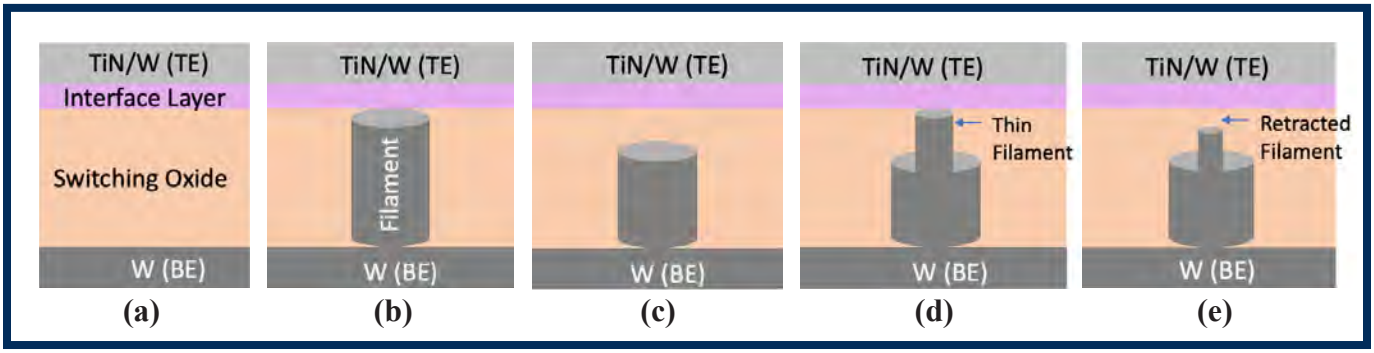

Fig. 2. *RRAM devices: (a) Fabricated state, (b) After electroforming showing thick filament in switching oxide, (c) First eset showing retracted filament, (d) Thin filament g owth during set process causing a low-resistance state (LRS), (e) Retracted filament during eset causing a high-resistance state (HRS).*

these are one of the more promising candidates and they have been widely studied in emerging neuromorphic architectures. RRAMs are two-terminal devices in a metal-insulator-metal (MIM) configuration, shown in Fig.2(a). The insulator is usually a metal oxide[15]. An interfacial layer can be designed to modulate the properties of metal oxide by serving as an oxygen exchange layer. Various dopants in metal oxides have also been widely investigated to achieve the desired switching characteristics.[16] These devices can be easily integrated on complementary metal oxide semiconductor (CMOS) platforms in back-end-of-line (BEOL) processing, adding computing value to the passive interconnects. There are two broad categories of RRAMs—filamentary and non-filamentar . In filamentary-RRAMs, the first step involves electroforming by applying positive electroforming voltage with compliance current (I_{cc}) control that leads to the formation of a defect-assisted filament, shown in Fig.2(b). These defects could be oxygen vacancies or metal ions. Then, the first reset is performed by applying negative voltage to retract the filament via a possible redox reaction, shown in Fig. 2(c). Finally, set operation is performed by applying positive voltage to reform the filament with relatively smaller I_{cc} to define the low-resistance state (LRS) (Fig. 2(d)). A subsequent reset operation leads to a high-resistance state (HRS) (Fig. 2(e)). The device can be switched between LRS and HRS with a write endurance of $>10^6$ cycles. Multiple resistive states can be achieved by modulating I_{cc} or reset voltages, which enables multi-bit weight storage in a single device, resulting in the densification of memory.[17] The resistive states in non-filamentary RRAMs are driven by the modulation of defect states at the oxide/metal interface or in oxide that alters the transport properties of electrons between top and bottom electrodes. Multiple analog resistance states can be achieved in these devices by using different programming conditions [18]

In a neuromorphic hardware, matrix multiplication is one of the most computationally intensive tasks limited by memory bandwidth. RRAM devices have been studied to enable in-memory computing in neuromorphic architectures, which has the potential to accelerate matrix multiplication.[19] RRAM devices in a 1 Diode-1 RRAM (1D1R) crossbar configuration are shown in Fig. 3(a). An access diode is necessary to mitigate the sneak current in the crossbars.[20] Though 1D1R is highly scalable, the desired specifications for access diodes have been difficul to meet and further research is needed in this area. Therefore, 1 Transistor 1 RRAM (1T1R), where the transistor serves as an access device, is a more practical implementation of RRAM in crossbar arrays currently (Fig. 3(b)). With these RRAM arrays, matrix multiplication is performed in analog fashion where the input voltage is intrinsically multiplied by the conductance state of an RRAM in a cell to result in an output current. The current through each cell is summed on the wire in column, resulting in matrix multiplication. Additional circuitry is needed to sense this current and transform it to the digital domain using analog-to-digital converters, or it is possible to continue processing in the analog domain. These architectures have been used to implement deep neural networks (DNNs).

Just like the brain, a neuromorphic hardware capable of real-time learning and inferencing is highly desirable. However, the training

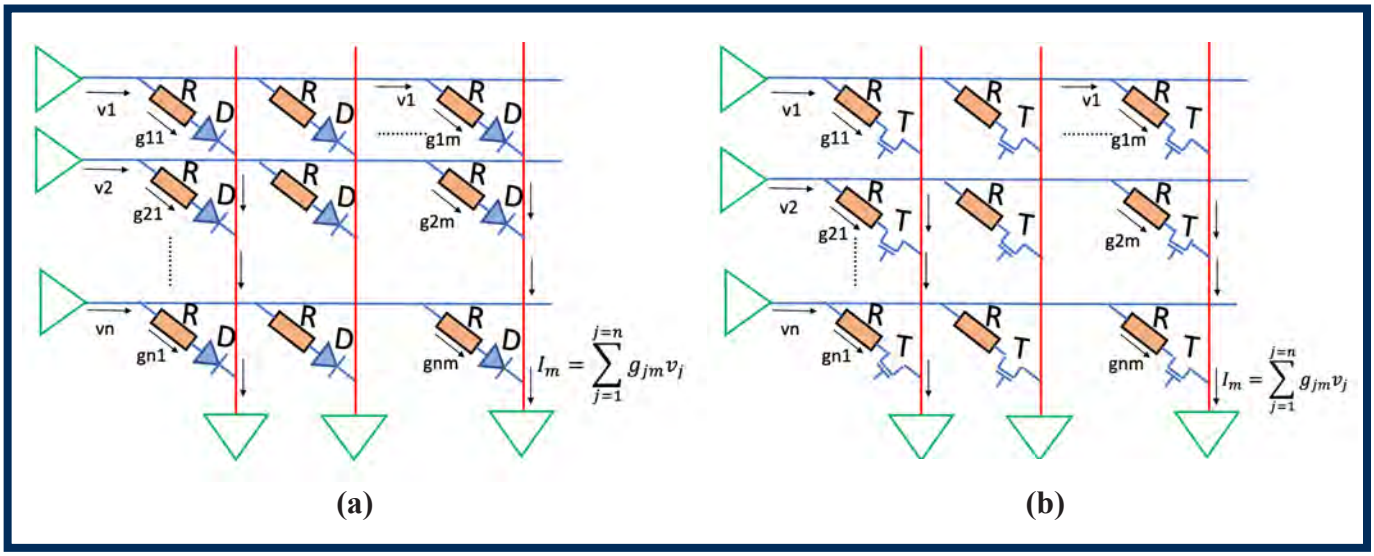

Fig. 3. *RRAM in crossbar array configuration in (a) 1D1R, (b) 1T1R. Input data ($_1$ to v_n) are applied into the array that gets transformed into current (I_1 to I_m) by multiplication with conductance value of corresponding RRAMs and summation in array.*

process is usually very complex, requiring additional hardware. Therefore, training and inferencing engines are designed separately to meet the optimum power, performance, area, and cost specifications. Inferencing engines based on RRAM neuromorphic architectures are of much interest for low-power edge-AI applications.[21]

One of the major drawbacks of RRAMs in neuromorphic architectures for inferencing is the drift in resistance states over time. The LRS and HRS retention over time has been reported to be a function of temperature, I_{cc}, or programming pulse-width.[22] The retention of resistive states over time can also be modulated by programming voltages—devices programmed with higher voltages or a higher number of pulses tend to have longer retention compared to devices programmed with lower voltages or a lower number of pulses.

Interestingly, while the time-dependent retention (or dynamic states) of these emerging memory devices is undesirable in a neuromorphic inferencing engine in its current implementation, this characteristic can be considered similar to the STP observed in biological synapses. Additionally, the ability to forget information has been shown to have a positive impact on learning.[23] A notable difference between a biological brain and RRAM-based neuromorphic inferencing engines is that a biological brain continues to learn from data even while inferencing. Therefore, time-dependent retention is useful because the system is dynamic. On the other hand, current inferencing engines based on RRAMs are static where states are expected to stay constant over time. A major challenge lies in understanding how we can use the dynamic nature of emerging memory devices to the advantage of neuromorphic systems. Indeed, these STP states of RRAM devices have been leveraged in spiking neural network architectures to demonstrate filtering of noise in sensory data and modified Hebbian learning.[24–26] While these preliminary reports are encouraging steps, further work is needed in this area to leverage these unique characteristics. Additionally, currently dynamic states in RRAMs are uncontrolled in nature. Once their applications are established, then they can be engineered to result in the desired performance.

In conclusion, RRAM devices hold promise for applications in neuromorphic computing, though there are some pending challenges that need to be addressed. Beyond their established applications for matrix multiplication in crossbar arrays, it is important to study time-dependent states and to develop techniques for controllably modulating the dynamic states. The reliability of these states needs to be studied as well. Complex dendritic architectures with RRAMs beyond crossbar arrays need to be investigated. A detailed understanding of these dynamic states can help implement cortical circuitries that utilize dynamic synaptic states in diverse distributions using these devices—which can have significant impact on advancing novel paradigms of computing.

Acknowledgment

This work is supported by National Science Foundation under award number ECCS 1926465.

 DOI: 10.1149/2.F10231IF

About the Author

Rashmi Jha, **Professor of Electrical Engineering and Computer Science, University of Cincinna ti**
Education: BTech (Indian Institute of Technology), MS and PhD (North Carolina State University) in Electrical Engineering.
Research Interests: Artificial intelligence (AI); Low-power neuromorphic systems; CMOS and other emerging logic and memory devices (e.g., RRAM, spintronics, and other memristive devices); On-die sensors; Cross-technology heterogenous integration and modeling; Cybersecurity with emphasis on hardware security; Additive, flexible, and wearable electronics; Nanoelectronics; Neuroscience and neuroelectronics; Bio-inspired computing and systems.
Work Experience: >18 years of experience in solid state electronics and nanoelectronic device design, modeling, fabrication, process integration, electrical characterization, data analysis, circuit design and simulation. Before NCSU, she was Assistant Professor and then Associate Professor in Electrical Engineering and Computer Science at the University of Toledo. Before that, she worked as a Process Integration Engineer at IBM's Semiconductor Research and Development Center.
Pubs + Patents: >106 peer-reviewed publications; 13 US patents.
Honors & Awards: AFOSR Summer Faculty Fellowship Award (2017); NSF CAREER Award (2013)
Website: https://researchdirectory.uc.edu/p/jhari
https://orcid.org/0000-0002-2656-5945

References

1. V. Sze, Y. -H. Chen, T. -J. Yang, and J. S. Emer, *Proceedings of the IEEE*, **105**, 2295 (2017).
2. M. Davies, *et al.*, *IEEE Micro*, **38**, 82 (2018).
3. Y.-C. Kwon, et al., *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 350 (2021).
4. F. Gao, G. Tziantzioulis, and D. Wentzlaff, *MICRO '52: Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 100 (2019).
5. H.-W. Hu, *et al.*, *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 138 (2022).
6. S. Shiratake, *2020 IEEE International Memory Workshop (IMW)*, Dresden, Germany, 1 (2020).
7. C. Zambelli, R. Micheloni, and P. Olivo, *2019 IEEE 11th International Memory Workshop (IMW)*, Monterey, CA, USA, 1 (2019).
8. J. Zhang, K. Rangineni, Z. Ghodsi, and S. Garg, *DAC '18: Proceedings of the 55th Annual Design Automation Conference,* 19 (2018).
9. N. Spruston, *Nat. Rev. Neurosci.*, **9**, 206 (2008).
10. L. Abbott and S. Nelson, *Nat. Neurosci.*, **3** (Suppl 11), 1178 (2000).
11. P. Poirazi and B. W. Mel BW, *Neuron.*, **29** (3), 779 (2001).
12. A. Govindarajan, R. Kelleher, and S. A. Tonegawa, *Nat. Rev. Neurosci.* **7**, 575 (2006).
13. See, for example: Ł. Kuśmierz, T. Isomura, and T. Toyoizumi, *Curr. Opin. Neurobiol.*, **46**, 170 (2017).
14. A. Sebastian, *et al. Nat., Nanotechnol.,* **15**, 529 (2020).
15. B. Govoreanu, *et al.*, *2011 International Electron Devices Meeting,* Washington, DC, USA, 31.6.1 (2011).
16. B. Long, *et al., ECS Trans.*, **53**, 115 (2013).
17. O. Golonzka, *et al.*, *2019 Symposium on VLSI Technology*, Kyoto, Japan, T230 (2019).
18. B. Long, Y. Li, and R. Jha, *IEEE Electron. Device Lett.*, **33** (5), 706 (2012).
19. W. Wan, *et al.*, *Nature*, **608**, 504 (2022).
20. H. Tsai, *et al.*, *J. Phys. D: App. Phys.*, **51** (28), 283001 (2018).
21. Z. Li, *et al.*, *IEEE J. Solid-State Circuits*, **56** (4), 1105 (2021).
22. Y. Y. Chen, *et al.*, *2013 IEEE International Electron Devices Meeting*, Washington, DC, USA, 10.1.1 (2013).
23. T. J. Ryan and P. W. Frankland, P.W., *Nat. Rev. Neurosci.*, **23**, 173 (2022).
24. T. J. Bailey, A. J. Ford, S. Barve, J. Wells, and R. Jha, *IEEE Trans. Very Large Scale Integ. (VLSI) Syst.*, **28** (11), 2410 (2020).
25. T. Chang, S.-H. Jo, and W. Lu, *ACS Nano*, **5** (9), 7669 (2011).
26. Z. Shen, *et al.*, *Nanomaterials (Basel)*, **10** (8), 1437 (2020).