


COMMENT

DOI: 10.1038/s41467-018-07565-4

OPEN

Challenges hindering memristive neuromorphic hardware from going mainstream

Gina C. Adam¹, Ali Khat² & Themis Prodromakis ²

Memristive devices have elicited intense research in the past decade thanks to their inherent low voltage operation, multi-bit storage and cost-effective manufacturability. Nonetheless, several outstanding performance and manufacturability challenges have prevented the widespread industry adoption of redox-based memristive matrices. Here, we discuss these challenges in terms of key metrics and propose a roadmap towards realizing competitive memristive-based neuromorphic processing systems.

The promise of redox memristors

Heterogeneous hardware that combines traditional digital circuitry with two-terminal analog memory devices, promises to handle the Zettabyte storage and processing requirements of modern applications such as the Internet of Things (IoT) and Artificial Intelligence (AI). Several emerging device concepts, based on electrochemical metallization, phase change, and redox phenomena have been intensely explored. This work led to some successful commercial products, like Adesto's Moneta electrochemical metallization memory for low-energy applications and Intel-Micron's phase change memory Optane for storage class memory. Despite this fact, the most highly sought application nowadays, namely non-volatile neuromorphic processors, has yet to become industrially feasible.

We believe that redox memristive memory will be the technology to fuel the AI era in the upcoming decades by enabling competitive implementations of neuromorphic processors. These switches can facilitate the energy and space efficiency required for emulating synaptic weights—the programmable connections that equip a neuromorphic system with its learning and memory capabilities. The synaptic weights can be implemented with commercially available technologies, but they typically require tens of devices for emulating a single synapse, which renders large-scale systems impractical. For comparison, redox memristive cells can outshine by 2–3 orders of magnitude in density and lower energy consumption of the implementations featuring more mature technologies¹. To emulate the complexity and ultra-low power consumption of biological neural networks, neuromorphic hardware platforms have to deliver an ultra-high density (>1 Tb/cm²) and energy efficient (<10 fJ/operation) solution. If we want to implement large neural networks with billions of synaptic devices, resistive switches are particularly suited thanks to three disruptive attributes: low-voltage multi-bit programmability, an inherent non-volatility of their resistance state, and a scalable two-terminal structure appropriate for matrix integration.

¹Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA. ²Electronic Materials & Devices, Zepler Institute for Photonics and Nanoelectronics, University of Southampton, Southampton, SO17 1BJ, UK. Correspondence and requests for materials should be addressed to T.P. (email: t.prodromakis@soton.ac.uk)

The physics of resistive switching is on our side from an energy-consumption perspective, since in theory the state of the device can change through the movement of just a few ions under a very low voltage. Once the voltage is removed, the ions halt in place and the state is retained without any further use of energy. The fine synaptic programmability is a key element for neuromorphic algorithms and redox resistive devices have achieved the best analog capacity to date (>100 discernible states per single cell)². Redox resistive devices are bipolar so a desired state can be accessed either during set or reset, which decreases the latency to program the matrix. Redox memristors typically report the lowest energy consumption/switching among emerging analog memory solutions, $\sim 10\text{fJ}$ ³. Moreover, the switching time has been shown to be as low as 85 ps⁴ for nitride materials.

An ideal neuromorphic platform would take advantage of these properties in an integrated fashion. Such a system would have hundreds of layers of resistive switching matrices integrated over traditional digital circuitry to achieve high performance at a low manufacturing cost.

Performance vs manufacturability challenges

This bold dream has fueled intense research in the field. Significant progress has been made, but in all honesty, at a slower pace than anticipated. No miracle material stack that leads to the perfect device properties has been discovered yet. Several performance and manufacturability challenges prevent industry adoption. Yet we are optimistic that our community will

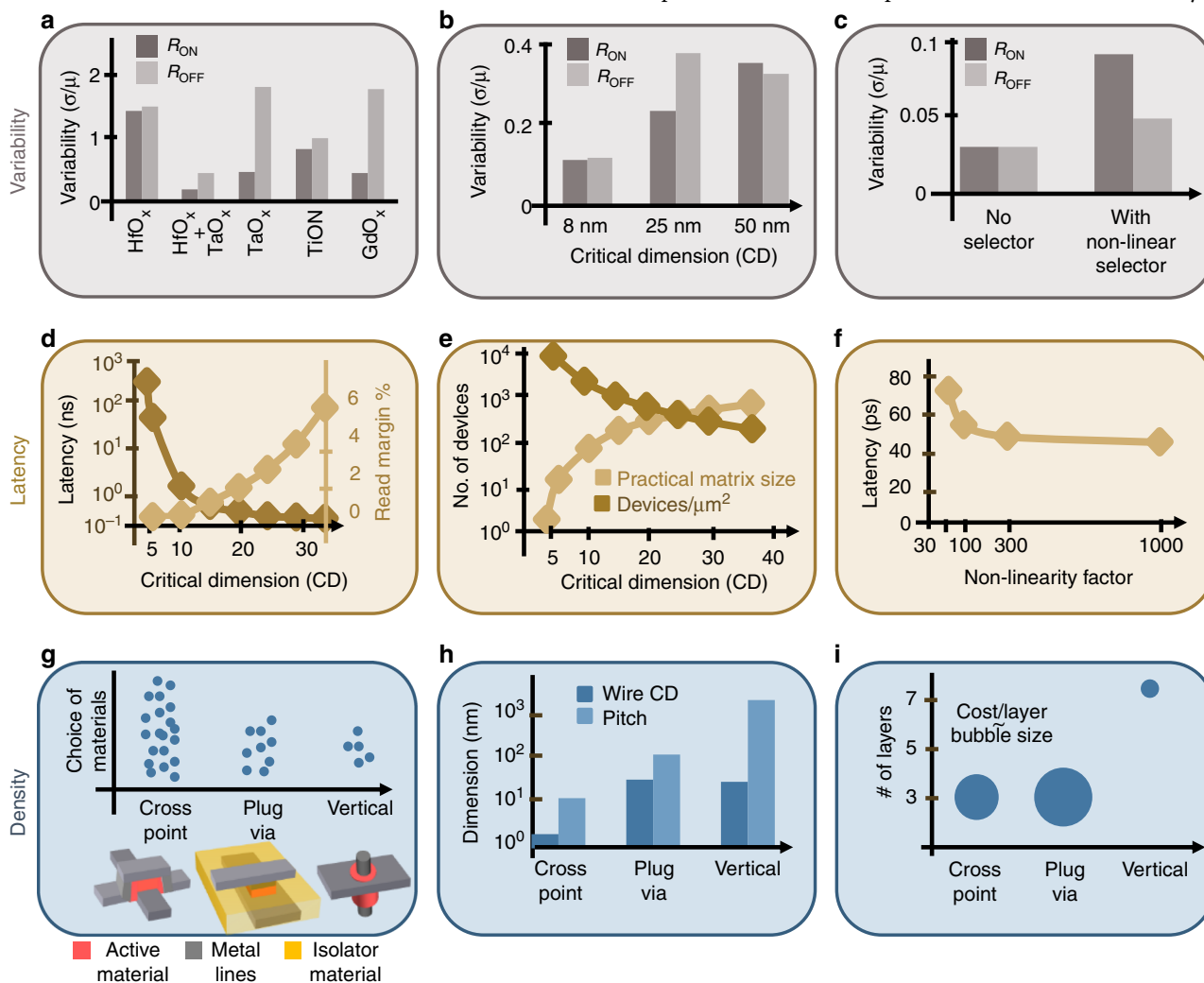


Fig. 1 Matrix-level metrics and manufacturing choices impacting them. **a–c** Variability metric. The variability is a measure of the spread of device performance (in this example, the two extreme resistance states R_{ON} and R_{OFF}) in a memristive matrix as defined based on the standard deviation and the means of the resistance distributions (σ/μ). The variability of the resistance states R_{ON} and R_{OFF} across a matrix is heavily influenced by **a** the choice of active material and of the material stack (e.g., single material HfO_x vs. bilayer $\text{HfO}_x + \text{TaO}_x$)^{5,12}; **b** the device scaling as determined by the smallest feature dimension (also known as critical dimension or CD);⁶ and **c** the presence of a series selector/cell which has its own variability profile⁸. The variability results presented in **a–c** are extracted from different studies so they have different orders of magnitude depending on the manufacturing process used. **d–f** Latency metric. The latency is a measure of the delay in accessing the desired device, delay caused by the charging and discharging of the wires. **d** Impact of the wire downscaling on latency and read margin, which is a measure of the capability to discriminate between the two extreme states (R_{ON} and R_{OFF}) of the memristive device¹³. **e** Practical matrix size limited by latency vs. the density (number of devices in a μm^2) allowed by the critical dimension of the manufacturing process. **f** The impact of the device / selector non-linearity on latency¹⁴. **g–i** Density metric discussed from the perspective of the most common device designs—crosspoint, plug-via and vertical. **g** The availability of materials suitable for each device design, given aspects such as uniformity, conformal deposition, etc. **h** The state-of-the-art scalability for each design (crosspoint: 2 nm CD/12 nm pitch⁹, plug-via $\sim 30\text{ nm}/100\text{ nm}$ ¹⁵, vertical structure has yet to be optimized for scalability¹²). **i** State-of-the-art stackability for each design and its approximate cost per matrix layer (represented by the relative size of the bubble)

overcome these challenges and develop a resistive switching technology of unparalleled performance for the next generation of neuromorphic hardware.

Variability. While neuromorphic computation is considered to be resilient to hardware defects, memristor variability is costly. If each device performs slightly different and its characteristics vary in time, programming to a desired state becomes a personalized endeavor. This approach is not feasible for training large matrices with billions of devices, as it consumes time, energy, and chip real-estate for supporting circuitry.

High-density integration and mass production will not be possible until the variability is fixed. And fixing it is challenging. This is a new technology that requires significant investment for refining the design and manufacturing process. More alarming is, however, the intrinsic stochastic nature of the switching. The resistive switching technology has been extensively shown in amorphous or polycrystalline materials. These materials have the advantage of low temperature deposition, so multiple matrix layers can be manufactured without disturbing the digital circuitry below. However, their uncontrolled high density of defects induces a high degree of variability. The choice of materials plays a critical role⁵ (Fig. 1a). Extreme scaling has also been shown to reduce variability, probably through confining the area where switching occurs⁶ (Fig. 1b). In the meantime, more complex cells, like the multi-memristor cell used to emulate a single synaptic unit⁷, can help alleviate some of these challenges, but at the cost of lower integration density.

Latency. While variability limits the size of the system that we can build, this is not our only challenge. The practical size of the matrix is limited also by the accessibility of individual devices in the matrix. The line resistance can determine a non-negligible voltage drop across the wires, increasing the latency (the time it takes to access a device) and the energy consumption and affecting the write/read margin (Fig. 1d). Sneak paths are another issue that aggravates with increased matrix size. A highly non-linear selecting device (called selector) in series with each memristor offers increased accessibility, as higher nonlinearity is desirable for reduced latency (Fig. 1f). Nevertheless, selectors have their own variability that further adds to the deterioration of performance⁸ (Fig. 1c). These issues become more acute with drastic technology scaling and limit the realistic matrix size (Fig. 1e).

Density. Despite the abovementioned limitations, the promise for an extremely small footprint provides a clear advantage by comparison with more mature technologies like flash memory. Various designs can be used, with the crosspoint, plug-via and vertical topologies being the most explored. Each has its merits and challenges, requiring trade-offs in scalability, stackability, selector integration capabilities and cost effectiveness. The crosspoint is the most common, due to easy manufacturing with a wide range of materials (Fig. 1g) and its extreme scalability, down to ~ 2 nm for an estimated density of >0.7 Tb/cm²⁹ (Fig. 1h). However, it has the major disadvantage of the active material stepping over the bottom line which can cause uncontrolled film thinning, increased device variability, or even electrodes shorting. The plug-via design has no step, but needs the etching of the via which damages the active film, increases the variability and requires additional masks. The vertical design is, by comparison, highly cost effective (Fig. 1i). The number of masks is independent of the number of layers, similar to the three-dimensional flash technology¹⁰. However, the requirement for conformal vertical deposition limits the choice of materials and of selector integration.

While the quest for the densest matrix design is admirable, a memristor-based neuromorphic processor is more than memristor matrices. Additional circuitry is typically required for selection, reading and programming of cells. Ideally, this circuitry would be implemented entirely below the memristor matrix stack for attaining highest chip space occupancy. However, high speed programming requirements can increase the circuitry footprint, thus straying away from the ideal density¹¹.

Reaching technological feasibility

Driven by its potential for extreme density, resistive switching matrices will benefit from the latest advances in nanofabrication, like the extreme ultraviolet lithography (EUV) which has already shown <10 nm half pitch lines. However, the industry can benefit from its technological potential only when the issues of variability and latency are solved, so that should be the short-term focus in our opinion (Fig. 2).

Tackling them requires a data-driven approach to accelerate the understanding and gaining control over the physics of switching, the materials and the manufacturing process. The necessity of having low access resistance and selector devices introduces extra complexity, requiring designs with higher

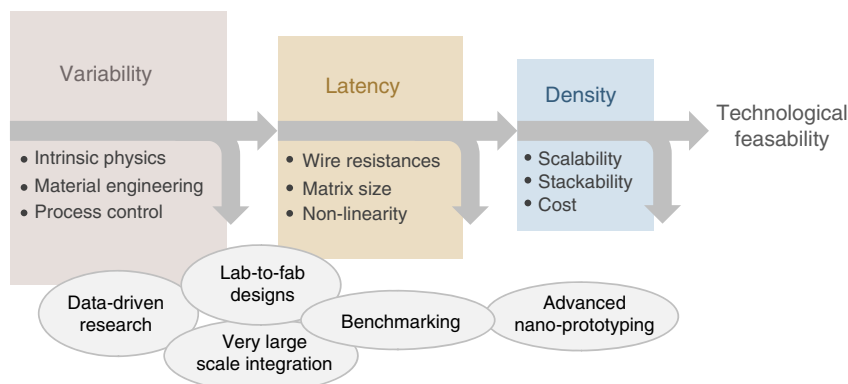


Fig. 2 Roadmap for manufacturing challenges and possible approaches to accelerate progress. Understanding the underlying factors behind variability can be enabled by data-driven research through lab-to-fab designs and very large-scale integration of memristive matrices with traditional digital access circuitry. Benchmarking (performance metrics, standardized device/matrix sizes, methods of testing, etc) will ensure comparable results between groups. Ultimately, once variability and latency issues are tackled, the technology development will benefit from advanced nano-prototyping techniques, such as extreme ultraviolet lithography, for cost-effective scalability and stackability

number of manufacturing steps and state-of-the-art cleanroom equipment. The characterization of large matrices is resource intensive as well, involving custom data acquisition set-ups. The solution is the integration of memristor matrices with the digital read/write circuitry which requires foundry material compatibility and sustained academia-industrial partnerships. Appropriate performance benchmarking amongst distinct materials, standardized device/matrix sizes and methods of testing are also needed to ensure reproducible results across different labs. A repository of these large datasets would strengthen the research capabilities of the community, enabling accurate device modeling and system-level simulations.

In the coming years, memristive neuromorphic hardware will likely flourish in select embedded applications based on medium-sized matrices suitable for cost-effective training off-site and pre-deployment. Complex systems would take longer to reach commercial maturity since they require larger memristive matrices with lower density of imperfections appropriate for fast on-site continuous learning. Ultimately though, the balance between system-level performance vs. manufacturing cost will be what drives widespread adoption.

Received: 26 July 2018 Accepted: 12 November 2018

Published online: 10 December 2018

References

1. Ceze, L. et al. Nanoelectronic neurocomputing: status and prospects In *2016 74th Annual Device Research Conference (DRC)* (IEEE, Newark, DE, USA, 2016).
2. Stathopoulos, S. et al. Multibit memory operation of metal-oxide bi-layer memristors. *Sci. Rep.* **7**, 17532 (2017).
3. Goux, L. et al. Ultralow sub-500nA operating current high-performance TiN/Al₂O₃/HfO₂/Hf/TiN bipolar RRAM achieved through understanding-based stack-engineering. In *2012 Symposium on VLSI Technology (VLSIT)*, pp. 159–160 (IEEE, Honolulu, HI, USA, 2012).
4. Choi, B. J. et al. High-speed and low-energy nitride memristors. *Adv. Funct. Mater.* **26**, 5290–5296 (2016).
5. Chen, A. and Lin, M. R. Variability of resistive switching memories and its impact on crossbar array performance. In *2011 International Reliability Physics Symposium (IRPS)* (IEEE, Monterey, CA, USA, 2011).
6. Pi, S., Lin, P. & Xia, Q. Cross point arrays of 8 nm × 8 nm memristive devices fabricated with nanoimprint lithography. *J. Vac. Sci. Technol. B* **31**, 06FA02 (2013).
7. Boybat, I. et al. Neuromorphic computing with multi-memristive synapses. *Nat. Commun.* **9**, 2514 (2018).
8. Zhang, L. et al. High-drive current (>1MA/cm²) and highly nonlinear (>103) TiN/amorphous-Silicon/TiN scalable bidirectional selector with excellent reliability and its variability impact on the 1S1R array performance. In *2014 International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2014).
9. Pi, S. et al. Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension. *Nature Nanotechnology*, <https://doi.org/10.1038/s41565-018-0302-0>, 2018.
10. Baek, I. G. et al. Realization of vertical resistive memory (VRRAM) using cost effective 3D process. In *2011 IEEE International Electron Devices Meeting (IEDM)*, pp. 737–740 (IEEE, Washington, DC, USA, 2011).
11. Xu, C. et al. Design implications of memristor-based RRAM cross-point structures. In *2011 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, Grenoble, France, 2011).
12. Luo, Q. et al. 8-Layers 3D vertical RRAM with excellent scalability towards storage class memory applications. *Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2017).
13. Liang, J. et al. Effect of wordline/bitline scaling on the performance, energy consumption, and reliability of cross-point memory array. *ACM J. Emerg. Technol. Comput. Syst.* **9**, 9 (2013).
14. Peng, X. et al. Cross-point memory design challenges and survey of selector device characteristics. *J. Comput. Electron.* **16**, 1167–1174 (2017).
15. Hsieh, M. et al. Ultra high density 3D via RRAM in pure 28nm CMOS process. In *2013 IEEE Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, USA, 2013).

Acknowledgements

We thank Brian Hoskins for useful discussions and the editorial team for constructive feedback. We also wish to acknowledge the support from the Royal Society and the Engineering and Physical Sciences, Research Council (EPSRC) grant EP/R024642/1.

Author contributions

G.C.A., A.K., and T.P. conceived the idea of the paper and contributed in writing the manuscript.

Additional information

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018