

## RESISTIVE MEMORY

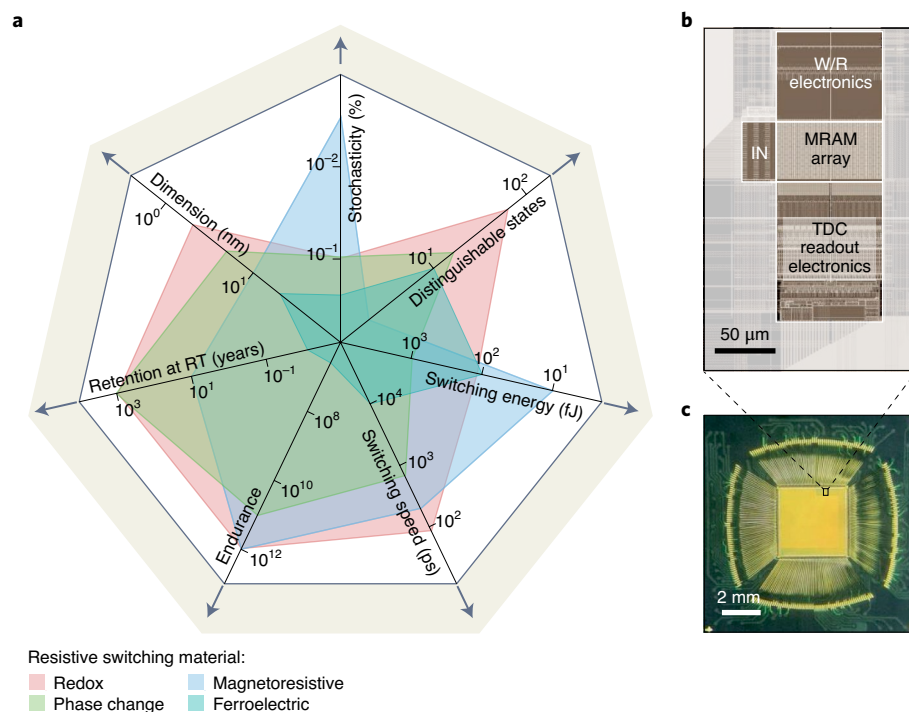
## Efficient AI with MRAM

In-memory computing chips based on magnetoresistive random-access memory devices can provide energy-efficient hardware for machine learning tasks.

Qiming Shao, Zhongrui Wang and J. Joshua Yang

In the past decade, artificial intelligence (AI) has undergone unprecedented development, introducing ground-breaking applications such as face recognition, language translation and industrial automation. But the continued advance of AI creates critical challenges in traditional digital hardware. Computers currently have to shuttle massive amounts of data between off-chip memory and processing units, a limitation known as the von Neumann bottleneck. At the same time, Moore's law, which has fuelled the development of digital electronics for decades, is running out of gas. Fundamental changes to computing hardware are needed. One potential solution is resistive memories (or memristors). When non-volatile resistive memory cells are grouped into a crossbar array, they can perform multiply-accumulate operations — the most expensive and frequent operations in AI — by directly using Ohm's law for multiplication and Kirchhoff's current law for summation. As a result, data are both stored and processed in the same location, which essentially removes the energy and time overheads incurred by expensive off-chip memory access for data fetching in digital hardware. Resistive memory cells are also simple capacitor-like structures, providing excellent manufacturability, stackability and scalability.

Different non-volatile resistive memories leverage different physical mechanisms for resistance modulation, and thus offer unique strengths and weaknesses when used for in-memory computing<sup>1</sup>. Traditional floating gate transistors are three-terminal resistive memories<sup>2</sup>; their charge injection programming requires large voltages and leads to slow write speeds and limited endurance. Emerging two-terminal and three-terminal resistive memories exploit redox reactions, phase transitions, ferroelectricity or magnetoresistance (Fig. 1a)<sup>1</sup>. Redox reactions and phase transitions often rely on the formation and rupture of conducting channels — providing excellent analogue conductance — and have recently been integrated into large-scale



**Fig. 1 | In-memory computing based on resistive memories.** **a**, Selected properties of four types of emerging random-access memory for applications in multiply-accumulate operations in artificial neural networks. RT, room temperature. **b**, Micrograph and layout of the  $64 \times 64$  MRAM crossbar array with peripheral circuits. W/R, write/read; TDC, time-to-digital converter. **c**, Photograph of an unpackaged MRAM chip. Panels adapted with permission from: **a**, ref. <sup>1</sup>, Springer Nature Ltd; **b,c**, ref. <sup>9</sup>, Springer Nature Ltd.

computing-in-memory systems to accelerate different machine learning tasks<sup>3,4</sup>. But many of them suffer from relatively large programming energy and variability. Ferroelectric resistive memory stores bits with the polarization of ferroelectric domains, the scalability of which may be limited by domain size<sup>1,5</sup>.

Magnetoresistive random-access memory (MRAM) is based on magnetic domain flipping. This means minimal atom displacement in the switching process, which should in turn provide good endurance and repeatability<sup>1,6,7</sup>. Recently, an MRAM-based in-memory chip with an energy efficiency of 5.1 tera operations




per second (TOPS) per watt, which is notably better than state-of-the-art digital alternatives, was reported<sup>8</sup>. However, this approach, like other MRAM systems, suffers from low resistance and small dynamic range. Writing in *Nature*, Donhee Ham, Sang Joon Kim and colleagues now report an MRAM-based in-memory computing chip that overcomes the low-resistance challenge by replacing current summation with resistance summation<sup>9</sup>. The approach essentially uses potential (voltage) signals instead of the commonly used current signals and thus greatly reduces the energy consumption, especially in arrays with low-resistance cells. It also has a

smaller footprint than a purely digital implementation based on XNOR logical gates, and offers non-volatility, which is important for low-energy edge applications.

The researchers — who are based at Samsung Electronics and Harvard University — use a time-to-digital converter to read out the weighted resistance of an entire column of the chip (Fig. 1b). Different column resistances will yield different delays in reaching the stable voltage, and they obtain the resistance information by sampling the time when the voltage reaches the reference voltage. As a result, their chip (Fig. 1c), which is based on MRAM cells developed for storage-class memories, exhibits an excellent energy efficiency of 262 TOPS per watt in performing vector-matrix multiplications. They use their chip to experimentally classify Modified National Institute of Standards and Technology (MNIST) handwritten digits, accelerating both a binarized multilayer perceptron and a classical VGG-8 neural network model; the classification performance of this analogue in-memory computing approach is on par with that of the software baseline.

Four MRAM chips are also successfully employed in the VGG-10 feature extractor of a SqueezeDet neural network for end-to-end human face detection in real time, highlighting the potential of the approach for low-power face authentication at the edge.

As the technology evolves, a cross-layer design that seamlessly optimizes devices, circuits, systems and algorithms is critical for the development of brain-like AI hardware<sup>6,7</sup>. The work of Ham, Kim and colleagues is an illustration of the potential of such an approach. To push the technology further, other emerging MRAM devices, such as spin-orbit torque MRAM and magnetoelectric MRAM, could be used to reduce currents because of their large-resistance cells. At the same time, and at the algorithm level, tailoring machine learning methods according to the underlying hardware could be used to maximize both system performance and efficiency. □

Qiming Shao <sup>1</sup>, Zhongrui Wang <sup>2</sup> ✉ and J. Joshua Yang <sup>3</sup> ✉

<sup>1</sup>Department of Electronic and Computer

Engineering, Hong Kong University of Science and Technology, Hong Kong, China. <sup>2</sup>Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, China. <sup>3</sup>Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA.

✉ e-mail: zrwang@eee.hku.hk; jjshuay@usc.edu

Published online: 17 February 2022  
<https://doi.org/10.1038/s41928-022-00725-x>

#### References

1. Wang, Z. et al. *Nat. Rev. Mater.* **5**, 173–195 (2020).
2. Guo, X. et al. In *2017 IEEE International Electron Devices Meeting (IEDM)* 6.5.1–6.5.4 (2017); <https://doi.org/10.1109/IEDM.2017.8268341>
3. Yao, P. et al. *Nature* **577**, 641–646 (2020).
4. Ambrogio, S. et al. *Nature* **558**, 60–67 (2018).
5. Jerry, M. et al. In *2017 IEEE International Electron Devices Meeting (IEDM)* 6.2.1–6.2.4 (2017); <https://doi.org/10.1109/IEDM.2017.8268338>
6. Grollier, J. et al. *Nat. Electron.* **3**, 360–370 (2020).
7. Shao, Q. et al. Preprint at <https://arxiv.org/abs/2112.02879> (2021).
8. Deaville, P., Zhang, B., Chen, L.-Y. & Verma, N. In *ESSCIRC 2021—IEEE 47th European Solid State Circuits Conference (ESSCIRC)* 75–78 (2021); <https://doi.org/10.1109/ESSCIRC53450.2021.9567807>
9. Jung, S. et al. *Nature* **601**, 211–216 (2022).

#### Competing interests

The authors declare no competing interests.