Beyond von Neumann

Data-centric computation and the scalability limits of current computing systems call for the developments of alternative to von Neumann architecture.

Digital computing has deeply permeated the fabric of the modern society. Its transformative power endowed by the remarkable technological evolution and commercial success begs no question of legitimacy. Notwithstanding, the underlying concept of the computer hardware design that has remained fundamentally unchanged since the days of von Neumann is in need of serious reform. In the current architecture where data moves between the physically separated processor and memory, latency is unavoidable. With no improvement in data transfer rates, the high-speed processor spends more time idle, waiting for data to be fetched from memory. To mitigate this issue within the von Neumann framework a number of solutions including caching, multi-threading, new types of random access memory and near-memory computing, with a processor mingled with memory on a single chip, have been proposed and implemented with varying degrees of success. Although the current architecture is unlikely to be abandoned in the foreseeable future, the growing trend of computational heterogeneity and a gradual shift towards learning computing with a data-centric approach typical of machine learning and deep learning calls for more specialized non-von Neumann platforms. One notable example is the architectures loosely modelled on the human brain structure, which infer a collocation of memory and processing units. In this scenario, the redundancy associated with data traffic could be entirely eliminated provided that computational tasks and data storage are both performed in place in the memory itself. This energy efficient solution, known as in-memory computing, could reduce the computational complexity and mitigate the issue of memory thrashing. Moreover, this approach is in keeping with the requirements of learning-based computing and has been actively explored for applications related to artificial intelligence.

As discussed in a Review in this issue by Abu Sebastian and co-workers, memory devices are essential building blocks of key computational primitives for in-memory computing. Similar to conventional memory, there is no universal solution for computational memory in that both charge-based and resistance-based memory technologies can be employed. For example, SRAM and DRAM are perfectly capable of performing in-memory logic operations while Flash memory is fit for matrix–vector multiplication operations. Another potent technology is phase-change memories (PCM) that have been successfully used to demonstrate the coexistence of storage and computation in a non-von Neumann architecture based on nanoscale PCM devices harnessing the crystallization dynamics[1]. Memristor-based memory devices often referred to as resistive random access memory (RRAM) relying on the formation of conducting filaments for switching between low and high resistance states are particularly attractive for in-memory computing owing to their non-volatile storage capability with a continuum of conductance states. In the context of the application-specific approach to computation, memory-based computational primitives can be used in a variety of tasks ranging from high-precision scientific computing to largely imprecise stochastic computing and everything in-between including deep learning in artificial neural networks (ANNs).

The original attempt to design ANNs in complementary metal–oxide–semiconductor (CMOS) technology has proved unsustainable with respect to energy consumption prompting the need for alternative non-von Neumann solutions for neuromorphic computing. Although, rethinking the hardware design at the device and system levels is a valid tactic, exploring the potential of emerging nanomaterials could enable the much needed departure from the conventional approaches to neuromorphic hardware. Neuromorphic nanoelectronic materials ranging from zero-dimensional, one-dimensional (1D) and two-dimensional (2D) nanomaterials to van der Waals heterostructures and mixed-dimensional heterojunctions have been actively explored for implementation in electronic and optoelectronic synapses. In a second Review in this Focus issue, Vinod Sangwan and Mark Hersam provide a detailed overview of the most prominent examples of nanomaterials for neuromorphic architectures including quantum dots that have been successfully employed in electro-photo-sensitive memristors, RRAM and quantum memristors based on Josephson junctions; 1D nanomaterials, particularly carbon nanotubes enabling the realization of synaptic transistors for unsupervised learning in spiking neural networks (SNNs) and group IV and III–V semiconducting nanowires exhibiting non-volatile memory characteristics. 2D materials, that have been widely covered by *Nature Nanotechnology* as potentially promising candidates for nanoelectronics, can also achieve neuromorphic functionality, particularly in view of improved device scaling and integration with planar wafer technology. For example, in one demonstration monolayer transition metal dichalcogenides (TMDCs) have been made into ultrathin vertical memristors where switching is likely to occur due to point defects[2]. Similar to 1D materials, synaptic transistors can be realized using ionic motion in layered TMDCs and black phosphorus, while phase transition in some TMDs have been harnessed to fabricate vertical RRAM. Moreover, the propensity of 2D materials for scalable processing and their ability to form van der Waals heterostructures can be explored for large-area, flexible and printable neuromorphic circuits.

To get a more complete overview of applications of 2D materials in nanoelectronics beyond neuromorphic computing we refer interested readers to another Review in this Focus issue by Chunsen Liu and colleagues, where the authors analyse the possibility of integrating 2D materials with the existing Si CMOS technology, in-memory computing platforms and matrix computing for ANNs and SNNs applications. By virtue of their atomic thickness 2D materials represent the ultimate limit for downscaling, the milestone that is hardly achievable in the context of the continued MOSFET shrinking. ❐

Published online: 9 July 2020
https://doi.org/10.1038/s41565-020-0738-x

References

1. Sebastian, A. et al. *Nat. Commun.* **8**, 1115 (2017).
2. Ge, R. et al. *Nano Lett.* **18**, 434–441 (2018).