SCIENCE CHINA
Information Sciences

• INSIGHT •

June 2018, Vol. 61 060424:1–060424:3
https://doi.org/10.1007/s11432-017-9313-6

Special Focus on Memristive Devices and Neuromorphic Computing

Memcomputing: fusion of memory and computing

Yi LI, Yaxiong ZHOU, Zhuorui WANG & Xiangshui MIAO*

Wuhan National Research Center for Optoelectronics, School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074*, China*

Received 10 October 2017/Accepted 26 October 2017/Published online 3 May 2018

Citation Li Y, Zhou Y X, Wang Z R, et al. Memcomputing: fusion of memory and computing. Sci China Inf Sci, 2018, 61(6): 060424, https://doi.org/10.1007/s11432-017-9313-6

In this world of ubiquitous computing, the contemporary electronic digital computer has become an essential machine that is available anytime and everywhere. Computing systems come in various forms, including desktops, laptops, tablets, mobile phones, smart watches, and other daily life devices. These systems all originate from the earliest huge, heavy EDVAC and UNIVAC machines of the 1940s, and all share a universal architecture conceived by von Neumann, who divided the computing system into five primary groups: a central arithmetic part (CA), a central control part (CC), memory (M) and outside recording medium (R), input, and output. The CA and CC parts evolved into the central processing unit (CPU), whereas the M and R correspond to the high-speed main memory that stores data and instructions (SRAM and DRAM) and external mass storage, respectively.

Based on the organized architecture and single-thread operation principles, the development of computing systems greatly benefited from the invention of the transistor and its subsequent performance boost along with the persistent exponential scaling of feature size, as predicted by Gordon Moore and theoretically formalized by Robert Dennard.

However, this tendency is now encountering unyielding walls. First, Dennard scaling is coming to an end as the capability of 2D lithography approaches the atomic realm. The miniaturization of electronic components is reaching its physical limits. Quantum tunneling in the gate insulator and other reliability issues are also arising. Secondly, today's computers continue to face the von Neumann bottleneck, which refers to the limited data transfer rate between a CPU that is physically separated from its hierarchical memory parts. Although vast caches have been introduced to reduce the performance mismatch in multicore CPUs and GPUs, the overhead data transfers in the bus consume the major proportion of energy in data-intensive computing.

Alternative computing devices and architectures offer keys for overcoming these daunting problems in the era of big data. Memcomputing, as one competitive emerging paradigm (Figure 1), explores the use of memristors or memristive devices, the fourth basic circuit element theoretically proposed by Chua in 1971 and physically demonstrated by HP Labs in 2008 [1]. Due to its desirable attributes, this fancy device is expected to be the building block for future computing technology. Its advantages include its non-volatility and the ability to program its resistance based on the polarity and amplitude of the applied voltage; its resistive switching behavior that can execute logic functions; its operation according to

* Corresponding author (email: miaoxs@hust.edu.cn)

distinct physical principles, as compared with the transistor; and its superior performances, including high switching speed, low power consumption, and 3D integration capability. The first two attributes open an intriguing opportunity for the fusion of memory and computing to develop non-von Neumann architectures. The last two attributes enable the construction of transistor-free, high-speed, energy-efficient, and compact computing systems.

As von Neumann initially conceived, CA most frequently performs the elementary arithmetic operations of addition, subtraction, multiplication, and division, which are implemented by the execution of binary logic.

Similarly, the exploration of a memristive processing unit (MPU) is the first concept in the construction of a memcomputing system. An MPU includes thousands of memcomputing cores consisting of homogenous large arrays of memristive devices. HP Labs took the lead in experimentally implementing stateful IMP and NAND logics with bipolar memristors [2], and proved that an arbitrary Boolean logic could be computed through the sequential iteration of implication and FALSE logic. The term "stateful" means that the logic input and output variables are represented by nonvolatile resistance states other than the volatile charge or voltage in transistor gates, which indicates that the computation results are stored in situ with zero static power and can participate in subsequent computation. In this way, the memory and computation are highly integrated in a single device and the volume of data transfer or memory access can be significantly reduced, which to a certain extent "opens up" the von Neumann bottleneck.

Linn et al. [3] proposed a Boolean algebra with four variables to realize 16 binary logic functions in three steps, which were validated in bipolar and complementary cells. Based on a general logic expression, Li et al. [4] further presented a two-step operation for implementing an arbitrary logic with one memristor in a crossbar architecture. Similar work was demonstrated in the 1T1R configuration involving a transistor into logic operation, which provides a feasible route for building memcomputing cores [5]. However, in these logic-in-memory implementations, the input and output variables are heterogeneous, i.e., voltages at the device terminals are the input variables and the resistance states comprise the output. Other logic solutions such as memristive ratioed logic (MRL) and memristor-aided logic (MAGIC) also show advantages in latency and energy consumption.

Several key criteria for evaluating different memristive logic approaches include functional completeness, computational complexity, cascading capability, and reconfigurability. The term functional completeness indicates whether all possible truth tables can be expressed by a basic set of logical connectives. For instance, {IMP, FALSE}, {NOR}, and {NAND} are complete sets.

Computational complexity indicates the amount of resources required, including spatial complexity and temporal complexity. This can be interpreted as the number of logic gates and logic cycles required to implement a particular function, which influences the system's latency, energy, and area efficiency. The cascading capability is determined by the nature of the input and output variables. For IMP logic, no extra conversion is needed, whereas in some cases, the output resistance state must be converted into voltage signals using buffers for the next-stage computation. Reconfigurability indicates the degree to which the function of the memcores can be changed by altering the control or operation signals [6], which provides high flexibility and intrinsic parallelism.

Much effort has been devoted to the practical application of memcomputing. Based on basic Boolean logics, complex arithmetic blocks have been built, including full adders, multiplication, and look-up tables, which serve as key elements for nonvolatile processors to address data-centric computation. Memristive reconfigurable logic circuits with memristors acting as the configuration bits and switches offer an alternative approach to the FPGA. Recently, researchers demonstrated hyperdimensional computing in 3D memristors based on kernels of multiplication, addition, and permutation for language recognition application. To accelerate convolution computation or matrix vector multiplication, the implementation of the dot-product engine in crossbars is an energy-efficient parallel solution for hardware deep learning and neuromorphic systems.

Although promising applications have been proposed, several formidable challenges hinder the development of memcomputing technology. One of the major challenges is the difficulty of fabricating reliable devices in large-scale arrays. Thus, per-

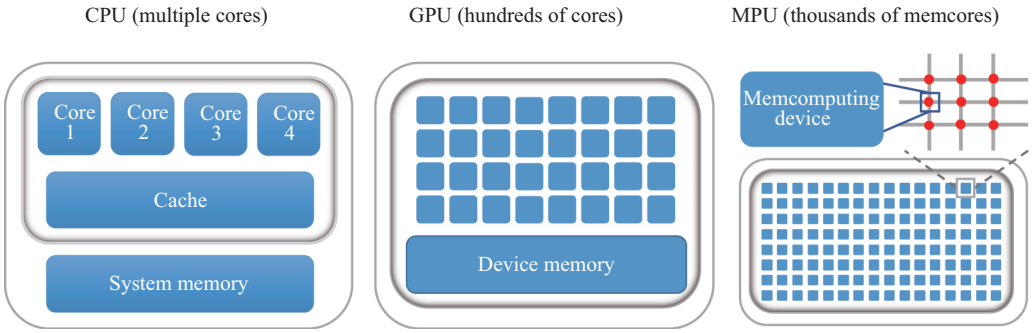

Figure 1 (Color online) Comparison of CPU, GPU, and MPU.

formance robustness in device-to-device and cycle-to-cycle variations should be taken into account in logic algorithm optimization, for instance, developing write-verify and feedback schemes for logic applications. On the other hand, device uniformity and endurance must be further improved by engineering the material and device structure [7]. Filament confinement — a fundamental method for obtaining uniform resistive switching behaviors — has been realized by several means. For example, the generation of oxygen vacancies are easier at the sites of particular foreign dopants, and the enhancement of local electrical fields by the introduction of nanocrystals or nanocone-shaped electrodes can also induce filament formation and disruption. High performance selectors with intrinsic nonlinearity or self-rectifying behaviors will also contribute to the development of large-scale memcores. Moreover, the implementation of memcomputing in 3D architecture is a tempting route for building a monolithic computing system. In addition, the development of foundry-supported material and high-yield processing should be a long-term consideration.

Yet another daunting challenge is the development of a memcomputing architecture with concrete microarchitecture and circuit designs [8]. The flexible dispatching and allocation of memcomputing resources could be a critical consideration. Based on minimized data communication and high parallelism, orders-of-magnitude performance improvement in execution time and energy saving for date intensive applications could be achieved, making them comparable to those of CMOS multicores or GPU platforms.

In summary, memcomputing may open a new avenue for computing approaches that enable the true fusion of memory and computing. Progress will be sustained and accelerated by the co-development of device and architecture technologies.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61674061, 61504045).

References

1 Strukov D B, Snider G S, Stewart D R, et al. The missing memristor found. Nature, 2008, 453: 80–83
2 Borghetti J, Snider G S, Kuekes P J, et al. 'Memristive' switches enable 'stateful' logic operations via material implication. Nature, 2010, 464: 873–876
3 Linn E, Rosezin R, Tappertzhofen S, et al. Beyond von Neumann logic operations in passive crossbar arrays alongside memory operations. Nanotechnology, 2012, 23: 305205
4 Li Y, Zhou Y X, Xu L, et al. Realization of functional complete stateful Boolean logic in memristive crossbar. ACS Appl Mater Interface, 2016, 8: 34559–34567
5 Wang Z R, Su Y T, Li Y, et al. Functionally complete Boolean logic in 1T1R resistive random access memory. IEEE Electron Device Lett, 2017, 38: 179–182
6 Zhou Y X, Li Y, Su Y T, et al. Nonvolatile reconfigurable sequential logic in a HfO_2 resistive random access memory array. Nanoscale, 2017, 9: 6649–6657
7 Lee J, Lu W D. On-demand reconfiguration of nanomaterials: when electronics meets ionics. Adv Mater, 2018, 30: 1702770
8 Zha Y, Li J. Reconfigurable in-memory computing with resistive memory crossbar. In: Proceedings of the 35th International Conference on Computer-Aided Design, Austin, 2016