

Rebooting Computing in Post Moore Era

Yuchao Yang* and Ilia Valov*

Traditional digital computers have physically separated memory and computing units, where data movements cost extensive time and energy consumption. As Moore's law slows down and memory-intensive tasks get prevalent, such von Neumann computing architecture becomes increasingly capacity- and power-limited. In order to meet the requirement for increased computing capacity and efficiency in the post-Moore era, emerging computing architectures, such as in-memory computing and neuromorphic computing, have been extensively pursued and become important candidates for new-generation brain inspired computers.

In consideration of the memory access bottleneck, an important trend of evolution in computing technology is data-centric architecture, such as near-memory computing and in-memory computing, which are particularly promising for memory-intensive workloads, such as neural networks. By incorporating memory die and processing die together using advanced IC package, the data movements between memory and processing units can be significantly reduced and the memory bandwidth bottleneck can be mitigated. Instead, compute periphery can be put inside an embedded memory chip, therefore enhancing energy efficiency. The data movement can be further reduced by performing computing inside memory devices, including conventional storage class memory, SRAM and advanced nonvolatile memories (ReRAM, PCM, MRAM, etc.). Currently, compute-in-memory (CIM) architectures are gaining steam, and there still exist different CIM chip designs depending on the specific requirements on throughput, energy efficiency and compute precision, etc., which are under active developments.


In addition to near-memory and in-memory computing, a more radical architecture shift is neuromorphic computing. The idea is supported by several national neuromorphic computing projects worldwide, and a couple of CMOS based neuromorphic chips have been developed.^[1] Emerging devices with high integration density and rich dynamics, such as memristors, have

also been exploited for the construction of neuromorphic systems. The physical embodiments of memristors correspond to various resistive switching devices based on different mechanisms, which endow the memristors with rich nonlinear dynamics, but theoretically all memristors can be described as a set of differential equations that indicate how the internal state variables determine device characteristics and how external electrical stimulations influence these state variables. The increases in the number of state variables and internal dynamics have dramatically enriched the dynamics and functionality of memristors (Figure 1). Further exploration and engineering of such dynamics are essential for highly efficient information processing applications.^[2]

The above developing trends in new computing technologies in post Moore era are well covered in this special issue. Emerging memories (memristive devices, spintronics, and electronics based on 2D materials) constitute promising building blocks for realizing artificial synapses and neurons as well as construction of hardware neural networks (2200068). Moreover, intelligent systems interacting with the environment need to process huge amount of data produced from sensors, which suffer from the separation of sensors from the memory and processor. The rise of in-sensor computing, where artificial sensory synapses and neurons are important building blocks, facilitates efficient signal processing at the edge.

Although memristors have been widely investigated as artificial synapses, the majority of memristors need a forming process, which brings about extra energy consumption and peripheral circuits. Thus, forming-free memristors are highly desirable. 5 nm thick TiO₂ memristors (2200001) and 1.5 nm thick HfO₂ memristor array (2200053) are demonstrated separately, exhibiting initially low resistance states and analog switching characteristics. Since tuning resistive switching characteristics is of great importance for neuromorphic applications, Xiao et al. (2100244) investigated the strain effect of HZO-based memristors systematically, which can simulate learning

Y. Yang
National Key Laboratory of Science and Technology on Micro/Nano
Fabrication
School of Integrated Circuits
Peking University
Beijing 100871, China
E-mail: yuchaoyang@pku.edu.cn

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.202200161>.

© 2022 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202200161

Y. Yang
Center for Brain Inspired Chip
Institute for Artificial Intelligence
Peking University
Beijing 100871, China

Y. Yang
Center for Brain Inspired Intelligence
Chinese Institute for Brain Research (CIBR), Beijing
Beijing 102206, China

I. Valov
Research Centre Jülich
Peter Gruber Institute
D-52425 Jülich, Germany
E-mail: i.valov@fz-juelich.de

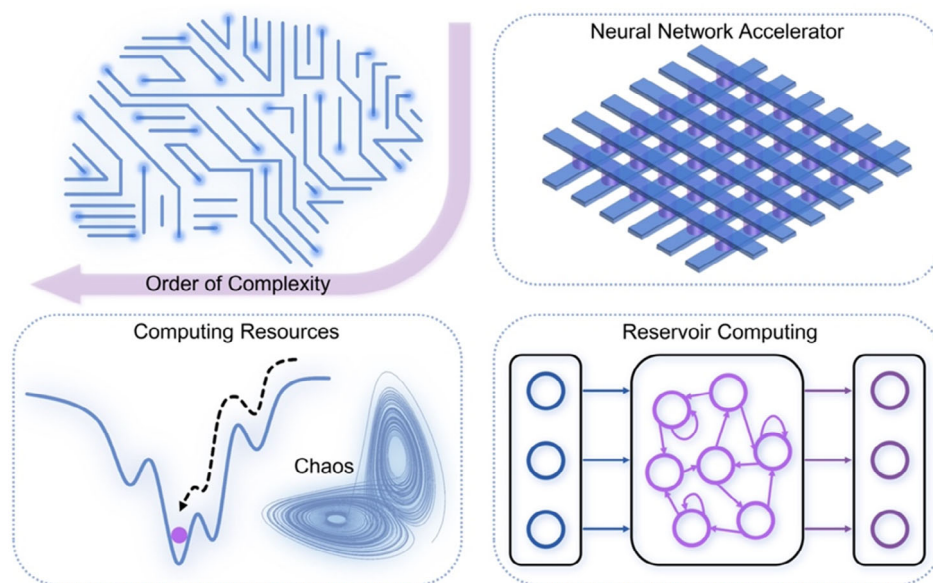


Figure 1. Increased order of complexity forms a substrate for increased computing capacity and efficiency.

behaviors and the switching mechanism stems from joint participation of ferroelectricity and oxygen vacancy migration. Ferroelectric HZO is also explored for capacitive synaptic array with high linearity (2100258). Array-level simulation shows $20\text{--}200\times$ lower energy consumption compared with resistive crossbar array counterparts. For artificial visual applications, an optoelectronic synapse based on three-terminal flexible memory phototransistors is demonstrated (2100257), which exhibits high responsivity, high detectivity and emulates the visual nociceptive behavior.

In addition, artificial neurons play an important role in encoding the information, where oscillation neurons are investigated for rate coding. An adaptive neuron is demonstrated with VO_2/HfO_2 memristor, where the firing threshold and firing rate can be modulated by changing the conductance of the HfO_2 layer. This adaptive neuron possesses spatiotemporal encoding capabilities via the correlated neural firing patterns (2100264). An artificial sensory neuron can be constructed by a serial connection of piezoresistive sensor and VO_2 volatile memristor. Based on the voltage dividing effect and the intrinsic thermal sensitivity of VO_2 , multimodal haptic/temperature patterns are detected, encoded, fused and recognized (2200039).

In the construction of CIM systems, a synaptic crossbar array can naturally implement vector-matrix multiplication with high efficiency and parallelism, whereas the non-idealities of the array may affect the network performance. The current approaches of tolerating noise effects for both training and inference phases, and solutions from hardware-software co-design, circuit, algorithm and system levels are discussed (2200029). Influence of programming variation on inference accuracy is also investigated with realistic system limitations. Approaches to mitigating such non-idealities via architecture-aware training are evaluated (2100199). An off-chip training method of one selector one resistor (1S1R) crossbar array considering nonlinearity and wire resistance is proposed, which significantly improves the inference accuracy and reduced the running time to 1% of

HSPICE-based simulation (2100256). Yang et al. (2200032) proposes a hardware-friendly distributed computing algorithm based on cellular neural network. A distributed design considering the hardware limitations (e.g., overhead and limited bit precision) is also discussed. A mix-precision continual learning model is deployed on a hybrid analogue-digital hardware system with a ReRAM chip (2200026), which circumvents the requirement for high-precision weights during inference. Combined with in-situ fine-tuning method, high classification accuracies are achieved and the energy consumption is reduced by ~ 200 times. Non-idealities can also be exploited to facilitate computing, where stochasticity in memristor programming is utilized to produce random matrices for random convolutional-pooling. Combined with echo state networks, energy efficient spatiotemporal signal classification is realized (2200027).

In particular, hardware neural networks based on synaptic arrays and peripheral circuits from PCB suffer from RC delay and signal loss, thus the tight integration of synaptic array and peripheral circuits on chip is essential. Cai et al. (2200014) designed system-on-chip with integrated ReRAM tiles and a RISC-V processor. The ReRAM tile is designed as an independent IP, flexible for different ReRAM technologies and expandable for larger-scale neural networks. In the work of Shin et al. (2200034), a spiking restricted Boltzmann machine is implemented on an on-chip trainable spiking neural network chip with 6T2R PCM cell, where pattern training, inference, and regeneration are demonstrated.

Looking forward, we believe these novel computing concepts and architectures will find more and more applications, as some key materials and device technologies get mature, yet reconfigurability will be an important consideration in future product delivery. The eventual application will also necessitate developments of new design automation tools, compilers, etc. We hope this special issue focusing on “Computing Technology in Post Moore Era” will further promote the discussion in the community and look forward to continued thriving and progresses of

this field. We would like to thank the authors for their novel contributions and the editorial team of “Advanced Intelligent Systems” for their efforts in organizing this special issue.

- [1] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, D. S. Modha, *Science* **2014**, 345, 668.
- [2] S. Kumar, X. Wang, J. P. Strachan, Y. Yang, W. D. Lu, *Nat. Rev. Mater.* **2022**, <https://doi.org/10.1038/s41578-022-00434-z>.



Yuchao Yang received his Ph.D. degree from Tsinghua University in 2010. After that he joined University of Michigan, Ann Arbor as a postdoctoral research fellow and was promoted to a senior research fellow in 2013. He joined Peking University as Assistant Professor in 2015, and is now Professor with tenure at School of Integrated Circuits, Peking University. His research interests include memristors, neuromorphic computing, and in-memory computing chips.



Ilia Valov is a principle investigator at the Research Centre Juelich and RWTH-Aachen University, Germany. In 2006, he becomes Dr. rer. nat (Ph.D.) with summa cum laude at the Institute of Physical Chemistry, Justus-Liebig-University, Germany in the field of physical chemistry of solids, defect chemistry and solid-state electrochemistry. Since 2009 he works on fundamental processes at an atomic and nano-scale. His research interests and activities now are concentrated on electrochemical and, in general, physicochemical phenomena at the nano and sub-nanoscale, such as mass and charge transport, point defects, surfaces and interfaces with a focus on resistive switching memories, memristive devices and energy conversion and electro-catalysis.