Journal of Semiconductors
JOS iopscience.iop.org/jos
www.jos.ac.cn

Embracing the era of neuromorphic computing

Yanghao Wang, Yuchao Yang, Yue Hao, and Ru Huang

Citation: Y H Wang, Y C Yang, Y Hao, and R Huang, Embracing the era of neuromorphic computing[J]. *J. Semicond.*, 2021, 42(1).
View online: https://doi.org/10.1088/1674–4926/42/1/010301

Articles you may be interested in

Reconfigurable computing: a promising microchip architecture for artificial intelligence
Journal of Semiconductors. 2020, 41(2), 020301 https://doi.org/10.1088/1674–4926/41/2/020301

Architecture, challenges and applications of dynamic reconfigurable computing
Journal of Semiconductors. 2020, 41(2), 021401 https://doi.org/10.1088/1674–4926/41/2/021401

A survey of FPGA design for AI era
Journal of Semiconductors. 2020, 41(2), 021402 https://doi.org/10.1088/1674–4926/41/2/021402

Preface to the Special Issue on Reconfigurable Computing for Energy Efficient AI Microchip Technologies
Journal of Semiconductors. 2020, 41(2), 020101 https://doi.org/10.1088/1674–4926/41/2/020101

The application of halide perovskites in memristors
Journal of Semiconductors. 2020, 41(5), 051205 https://doi.org/10.1088/1674–4926/41/5/051205

Light emission of heavily doped AlGaN structures under optical pumping
Journal of Semiconductors. 2018, 39(4), 043002 https://doi.org/10.1088/1674–4926/39/4/043002

Follow JOS WeChat public account for more information

COMMENTS AND OPINIONS

Journal of Semiconductors
(2021) 42, 010301
doi: 10.1088/1674-4926/42/1/010301

Embracing the era of neuromorphic computing

Yanghao Wang[1], Yuchao Yang[1, 2, †], Yue Hao[3, †], and Ru Huang[1, 2, †]

[1]Department of Micro/nanoelectronics, Peking University, Beijing 100871, China

[2]Center for Brain Inspired Chips, Institute for Artificial Intelligence, Peking University, Beijing 100871, China

[3]School of Microelectronics, Xidian University, Xi'an 710071, China

Citation: Y H Wang, Y C Yang, Y Hao, and R Huang, Embracing the era of neuromorphic computing[J]. *J. Semicond.*, 2021, 42(1), 010301. http://doi.org/10.1088/1674-4926/42/1/010301

In recent years, deep learning has made tremendous achievements in computer vision, natural language processing, man-machine games and so on, where artificial intelligence can reach or go beyond the level of human beings. However, behind so many glories, some serious challenges exist in the bottom hardware, hindering the further development of Artificial Intelligence. While the remarkable Moore's Law becomes slower and computing consumption on von Neumann bottleneck can no longer be afforded, current accelerator chips are difficult to deal with demanding massive data, especially in some power-limited scenes. These significant challenges lead to a natural upsurge for exploring new computing paradigms, i.e. a computational scientific revolution[1]. Such computing paradigm is not expected to replace the von Neumann architecture that has worked well in the past, but forms an important compliment to the previous architecture that can no longer handle with more and more emerging computing problems and applications. e.g. those in big data and artificial intelligence.

Candidates for the new computation paradigm include in-memory computing, quantum computing and neuromorphic computing, which can respectively solve some important problems more successfully than classical computing systems, although they have demonstrated only limited scope of application and accuracy to date. Among them, if we want to follow up the victory that deep learning has won and further build a general, efficient and brain-like intelligence, it is suggested to develop a paradigm of neuromorphic computing, which combines architecture, algorithms, circuits and devices tightly. From this view, deep learning is only a precursor to the approaching era of neuromorphic computing.

It has been about three decades since Carver Mead got inspiration from human brain and first proposed the concept of neuromorphic computing[2]. It takes advantage of analog signals to imitate electrical properties of synapses and neurons as basic computing elements, and assembles them to functional systems following simplified brain operating rules. Our brains utilize spikes to transmit and process information, running on the edge of chaos, so they have incredibly rich computational dynamics, as well as powerful capabilities for spatiotemporal integration. Since the introduction of neuromorphic computing, many impressive exploratory works have been completed, like IBM's TrueNorth[3] and Intel's Loihi[4]. However, a research consensus has not been established regarding neuromorphic computing yet. From the device perspective, obviously synapses and neurons composed by multiple transistors are costly, which restricts further scaling up. Fortunately, some emerging devices such as memristors can imitate synapses and neurons directly with its inner physical dynamics in single cells, thus holding great prospect in neuromorphic hardware. These devices can be compatible with current semiconductor technology, and can be used for construction of both deep learning accelerators and neuromorphic computing systems (Fig. 1). From the algorithm perspective, spike-based neural network models are immature compared with state of the art artificial neural networks on existing benchmarks and tasks[5]. Nevertheless, it should be noticed that existing effective algorithms are all suitable for classic computing systems, and the advancement of neuromorphic computing necessitates its own algorithms and benchmarks. Thus, there is an incommensurable way between these two computing paradigms.

Neuromorphic devices are memristive devices essentially that can change resistances through internal physical states and external electrical stimulations, which naturally correspond to synapses with adjustable weights. It has been proved that various emerging devices based on ion migration, phase transition, spin and ferroelectricity can obtain excellent modulation effects. For deep learning accelerators, ideal neuromorphic devices should have high state precision, low variation, long retention, linearity, as well as large dynamic range. However, current neuromorphic devices cannot combine all aspects of the abovementioned performances. For example, memristors based on ion migration have inevitable variations, and devices based on phase transition suffer from conductivity drift. In some interesting cases, these imperfections can be used as computing resource instead. The nonlinearity in conductance modulation can accelerate simulated annealing process in transiently chaotic neural network for the solution of various optimization problems[6]. Moreover, the stochasticity in devices conductance can be utilized as a random matrix in direct feedback alignment, reducing the training cost of neural networks (Fig. 1)[7].

For spike-driven neuromorphic computing involving the coding and representation of time information, neuromorphic devices should have capabilities to process sequential

Correspondence to: Y C Yang, yuchaoyang@pku.edu.cn; Y Hao, yhao@xidian.edu.cn; R Huang, ruhuang@pku.edu.cn

Received 30 DECEMBER 2020.

©2021 Chinese Institute of Electronics

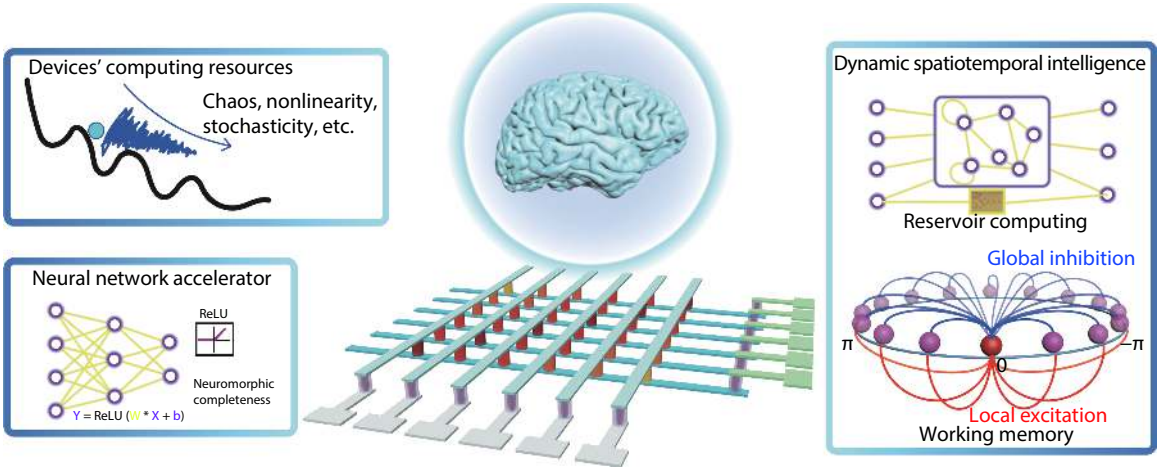

Fig. 1. (Color online) A possible roadmap of neuromorphic computing.

spikes and behave distinctly. Spike timing dependent plasticity (STDP), which attached computational significance to synapses, can be locally realized by a pair of connected volatile and nonvolatile memristors[8]. Furthermore, the leaky integrate and fire dynamics, which are symbolic computing functions of neurons, can be realized by a volatile device and its intrinsic capacitance[9]. The inner dynamics of devices, especially in Mott phase transition, can bring powerful computational functions, such as chaotic neurons[10] and high-order dynamics[11]. These devices are naturally suitable for the implementation of spiking neural networks, and it will be appealing to further find out how far the computing complexity can go ultimately relying on device dynamics. Such efforts may transform the simplicity of computing elements in Turing machine framework to the sufficient complexity in neuromorphic computing framework from practice.

Recently, it has been proved that a machine with neuromorphic completeness can solve any Turing-computable problems through approximation[12]. Its basic computing operations are vector-matrix multiplication and threshold. A crossbar array integrated by neuromorphic devices can calculate the vector matrix multiplication with great efficiency, which is the most computationally intensive part. This computing method relies on Ohm's law and Kirchhoff's law, and the nonvolatile nature of memristors can help avoid frequent memory access. However, there are still three major problems to be settled. First, current external circuits, such as analog/digital converters, are not efficient enough, which may eat into the advantages neuromorphic devices bring. It is therefore necessary to design specialized analog/digital convertors for a specific class of applications in neuromorphic computing. Second, there is no real spatial architecture driven by data flow for neuromorphic devices yet. Third, it is recognized that neuromorphic device array has little possibility of hardware multiplexing, since the storage is always waiting to be used. It is often accompanied by mixed-precision quantization, when mapping a specific neural network on limited hardware resources. It is thus suggested to develop EDA tools for automated deployment of different neural network models, where different layers are arranged for hardware multiplexing.

As neuromorphic computing can have higher efficiency on some tasks and has Turing equivalence with existing computing paradigms, the unique superiority of neuromorphic computing itself is still unclear. Some studies have made preliminary explorations. Since spike-based representations can efficiently encode time, the neuromorphic hardware can detect the synchronization of spike sequences in fine time scale among noisy signals[9]. Furthermore, volatile and oscillating devices can be assumed as neuron groups for reservoir computing in automatic generation of patterns[13, 14]. The oscillating devices can also realize microwave neural processing and broadcasting with great robustness[15]. More complex functions in brain, such as consciousness, emotion and attention, are still important research topics in computational neuroscience and their mechanisms remain elusive. Among others, working memory is a dynamic mode of information storage and processing in the brain, which is assumed as the basis for future advanced functions like attention and can be implemented on neuromorphic hardware (Fig. 1). Proper symmetric distribution of synaptic weights can form a continuous attractor neural network, where the neuronal population coding is always representing states in a continuous curve or plane. Therefore, it has significant superiority in processing the dynamic spatiotemporal information, compared with classical discrete storage. By introduction of working memory into neuromorphic hardware, the computing paradigm may expand the integration of storage and computing at device level to a structured storage at the system level. Furthermore, it is promising to replace a mathematically complex function, such as attention in Transformer[16], with inner dynamics of single devices. The exploration on dynamic spatiotemporal intelligence is beneficial for efficient combination of algorithms with physical devices, similar with what our brain is doing (Fig. 1).

In the more than Moore era, it is meaningful to make a transition in computing paradigm, figure out tightly entangled theories, methods and standards and set benchmarks. As for neuromorphic computing, we believe the emerging neuromorphic devices will trigger a radical shift in computing paradigm eventually. The new computing paradigm may first play a role in selected areas, e.g. edge computation with ultra low power consumption, but eventually lead to more capable computing systems and higher intelligence as well as vast applications.

Acknowledgements

This work was supported by the National Key R&D Pro-

gram of China (2017YFA0207600), National Outstanding Youth Science Fund Project of National Natural Science Foundation of China (61925401), PKU-Baidu Fund Project (2019BD002), National Natural Science Foundation of China (92064004, 61927901, 61421005, 61674006), and the 111 Project (B18001). Y. Y. acknowledges the support from the Fok Ying-Tong Education Foundation, Beijing Academy of Artificial Intelligence (BAAI), and the Tencent Foundation through the XPLORER PRIZE.

References

[1] Kuhn T S. The structure of scientific revolutions. Chicago: University of Chicago Press, 1996
[2] Mead C. Neuromorphic electronic systems. Proc IEEE, 1990, 78, 1629
[3] Merolla P A, Arthur J V, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345, 668
[4] Davies M, Srinivasa N, Lin T H, et al. Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro, 2018, 38(1), 82
[5] LeCun Y. Deep learning hardware: Past, present, and future. IEEE International Solid-State Circuits Conference, 2019, 12
[6] Yang K, Duan Q, Wang Y, et al. Transiently chaotic simulated annealing based on intrinsic nonlinearity of memristors for efficient solution of optimization problems. Sci Adv, 2020, 6(33), eaba9901
[7] Lu Y, Li X, Yan L, et al. Accelerated local training of CNNs by optimized direct feedback alignment based on stochasticity of 4 Mb C-doped $Ge_2Sb_2Te_5$ PCM chip in 40 nm node. IEEE International Electron Devices Meeting, 2020
[8] Wang Z, Joshi S, Savel'ev S E, et al. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. Nat Mater, 2017, 16(1), 101
[9] Duan Q, Jing Z, Zou X, et al. Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks. Nat Commun, 2020, 11(1), 1
[10] Kumar S, Strachan J P, Williams R S. Chaotic dynamics in nanoscale NbO_2 Mott memristors for analogue computing. Nature, 2017, 548(7667), 318
[11] Kumar S, Williams R S, Wang Z. Third-order nanocircuit elements for neuromorphic engineering. Nature, 2020, 585(7826), 518
[12] Zhang Y, Qu P, Ji Y, et al. A system hierarchy for brain-inspired computing. Nature, 2020, 586(7829), 378
[13] Moon J, Ma W, Shin J H, et al. Temporal data classification and forecasting using a memristor-based reservoir computing system. Nat Electron, 2019, 2(10), 480
[14] Torrejon J, Riou M, Araujo F A, et al. Neuromorphic computing with nanoscale spintronic oscillators. Nature, 2017, 547(7664), 428
[15] Talatchian P, Romera M, Tsunegi S, et al. Microwave neural processing and broadcasting with spintronic nano-oscillators. IEEE International Electron Devices Meeting, 2018, 27.4.1
[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv: 1706.03762, 2017

Yuchao Yang received his PhD from Tsinghua University. He is now an Assistant Professor in Department of Micro/nanoelectronics and serves as Director of Center for Brain Inspired Chips at Peking University. His research interests include memristors, neuromorphic computing, and in-memory computing.

Yue Hao is currently a Professor of Microelectronics and Solid State Electronics with Xidian University, Xi'an, China. His current interests include wide and untra-wide bandgap materials and devices, advanced CMOS devices and technology, semiconductor device reliability physics and failure mechanism, and organic electronics. Prof. Hao is a senior member of IEEE and member of the Chinese Academy of Sciences.

Ru Huang is currently a professor and vice president of Peking University. She is an elected academician of Chinese Academy of Science and IEEE Fellow. Her research interests include nano-scaled CMOS devices, ultra-low-power new devices, new device for neuromorphic computing, emerging memory technology and device variability/reliability. She is the Vice President of IEEE Electron Devices Society (EDS), the elected BoG member and the Chair of IEEE EDS SRC Region 10.