2021

Three-dimensional hybrid circuits: the future of neuromorphic computing hardware

Peng Lin
Zheijang University

Qiangfei Xia
University of Massachusetts Amherst

Recommended Citation

Lin, Peng and Xia, Qiangfei, "Three-dimensional hybrid circuits: the future of neuromorphic computing hardware" (2021). *Nano Express*. 1203.
https://doi.org/10.1088/2632-959X/ac280e

Nano Express

PERSPECTIVE • OPEN ACCESS

Three-dimensional hybrid circuits: the future of neuromorphic computing hardware

To cite this article: Peng Lin and Qiangfei Xia 2021 *Nano Ex.* **2** 031003

IOP Publishing   Nano Express 2 (2021) 031003   https://doi.org/10.1088/2632-959X/ac280e

NANO EXPRESS

OPEN ACCESS

PUBLISHED
27 September 2021

PERSPECTIVE

Three-dimensional hybrid circuits: the future of neuromorphic computing hardware

Peng Lin[1,*] and Qiangfei Xia[2,*]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, People's Republic of China
[2] Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003, United States of America
[*] Authors to whom any correspondence should be addressed.

E-mail: penglin@zju.edu.cn and qxia@umass.edu

Keywords: neuromorphic computing, three-dimensional circuits, emerging devices

Abstract

Recently there have been intensive research efforts to adopt emerging electronic devices for neuromorphic computing. However, the usage of these devices and arrays mainly was to implement parallel matrix multiplication in the two-dimensional (2D) space. This Perspective discusses the importance and implementation of three-dimensional (3D) hybrid circuits for neuromorphic computing, focusing on the integration density, data communication, and functional connectivity. We believe that 3D neuromorphic systems represent the future of artificial intelligence hardware with much-improved power efficiency and cognitive capabilities.

The revive of the analog computing paradigm results from the unprecedented challenges faced by digital computers in dealing with modern deep neural networks [1]. Emerging electronic devices such as resistive switching memristors, phase-change memories, ferroelectric transistors, and others [2], have opened up a new opportunity to functional and scalable neuromorphic systems that could mimic part of the brain functionality with high energy efficiency and computing throughput [3]. Despite promising early demonstrations, state-of-the-art neuromorphic systems still heavily rely on software to implement neural networks due to the challenges of efficiently implementing the circuits' network architecture.

Even though an integrated chip usually has a few metal layers, the critical data flow during the computing is still two-dimensional (2D). The physical connectivity of devices, essential to a neuromorphic system, is bounded by the limited number of metal layers. Hence, it is critical to develop a 3D computer architecture in which devices are in a 3D space and data flows intra-plane and inter-plane. Taking advantage of the 'extra' design freedom along the third dimension, building more complex neuromorphic circuits and systems makes our designs one step closer to that of a brain's topologies.

A future 3D neuromorphic system, equipped with all these elements (figure 1), forms a new type of system-on-chip design that can: (1) employs large on-chip storage capability to host all the weights of a massive neural network; (2) physically implement large, complex neural circuits to carry out efficient neuromorphic processing, and (3) integrate sensing and computing together to process high volume data without much communication overheads. This perspective discusses how a 3D circuit can change the horizon of existing 2D circuit-based neuromorphic systems. We start with the generic benefits of 3D circuits, including high density and fast circuit speed. We then introduce new developments of 3D brain-like architectures, such as complex, hardwired connections between artificial neurons and synapses and heterogeneous 3D circuits with processing and sensor layers. Finally, we discuss potential issues and solutions of 3D neuromorphic circuits.

Benefits and applications of 3D neuromorphic circuits

The first large-scale application of 3D neuromorphic circuits could be the high-volume 3D synaptic arrays that can host all the weights of a modern deep neural network. The sizes of neural networks are growing rapidly to

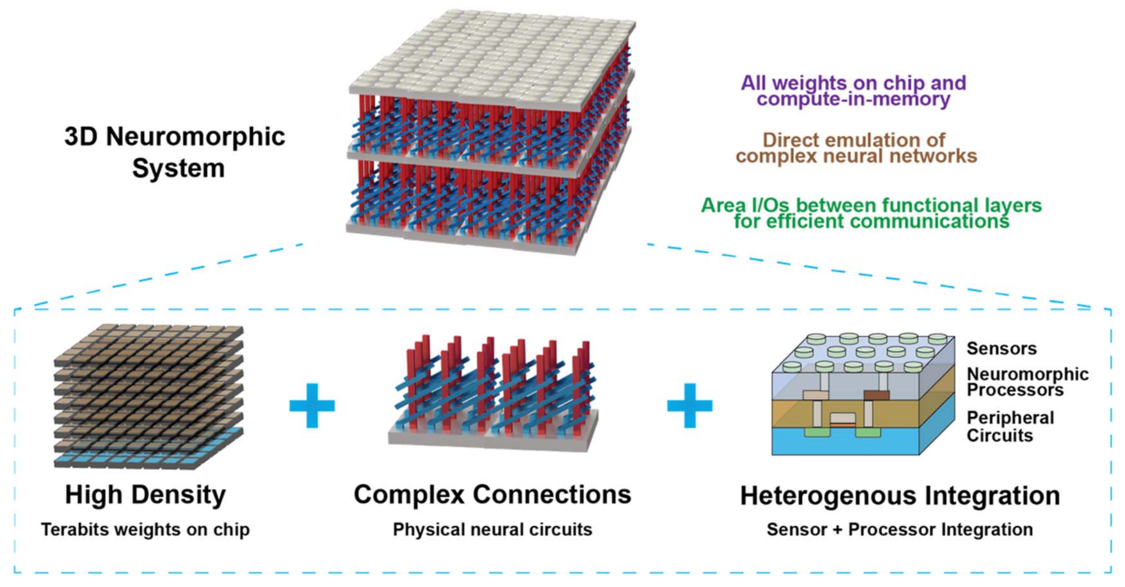
3D Neuromorphic System
All weights on chip and compute-in-memory
Direct emulation of complex neural networks
Area I/Os between functional layers for efficient communications
High Density
Terabits weights on chip
Complex Connections
Physical neural circuits
Sensors
Neuromorphic Processors
Peripheral Circuits
Heterogenous Integration
Sensor + Processor Integration

Figure 1. Bolts and nuts to build a future 3D neuromorphic circuit.

accommodate more extensive input data and provide advanced functionalities such as multi-tasking and high-level perceptions, which calls for significant expansion of on-chip storage. Although 3D FLASH has been mainly developed for high-capacity storage purposes, significant interest is rising to use 3D FLASH as a synaptic array for weight storage and compute-in-memory operations [4]. The state-of-the-art 3D FLASH product, fabricated with a 28 nm technology, can achieve an integration density exceeding 10 Gb mm^{-2} [5], compared to a 2D NAND FLASH with less than 1 Gb/mm2 made at the 14 nm node [6]. Emerging technologies such as memristors and phase change memories could achieve higher packing density due to their simple structure.

The second opportunity for 3D neuromorphic circuits lies in the capability to construct 3D connections between devices, which enables new computing paradigms for complex neural networks. In a standard 3D cell arrangement such as that used in 3D NAND and VRRAM [7], synaptic devices are connected through horizontal word lines and vertical bit lines. These 3D arrays implement a fully connected topology between input and outputs and are typically designed to serve as generic hardware for matrix computations. Neural networks with different topologies are divided into sub-networks that can be mapped to each array block. There is a compromise of speed and efficiency in most cases due to substantial control and communication overhead needed to route data between different physically separated sub-networks. Design and fabrication of custom arrays that can directly emulate the complex connections in neural networks become indispensable. As a proof-of-concept demonstration, an eight-layer 3D array was designed with a new array topology that emulates the local connectivity of convolutional kernels [8]. The unique array design enables pixel-wise parallel convolutions for image classification and video processing applications with much-reduced processing overheads, otherwise impossible in conventional 2D and 3D array designs.

The third category of 3D neuromorphic hardware is those that integrate various functional layers, including sensors, memory, and logic blocks, into one 3D chip [9]. Each functional block can be connected through fine-grained area interfaces, which provides high bandwidth for big data applications. Parallel processing could be highly beneficial for time-sensitive applications such as robotics and autonomous driving. In such a system, both the 'in-memory computing' and 'near-memory computing' concepts could be organically integrated, maximizing the parallelism in the computing while minimizing the data movement in the computing. It is projected that such an arrangement could bring 1000× improvements in the efficiency-speed product [10].

Potential issues and solutions

While 3D circuits have shown early promises in high density and design flexibility for neuromorphic computing, practical implementation of such architecture still requires overcoming challenges in devices, circuits, processes, and more. These topics have been discussed in detail in recent publications [11], [12]. As opposed to conventional 2D circuits, 3D circuits employ multiple device layers fabricated through a series of fabrication processes, some of which are unconventional. It requires a more delicately designed process to meet the thermal budget limit and maximize the fabrication yield. Since different layers are stacked on top of each other, thermal dissipation could be a bottleneck for 3D circuits.

Extensive research and development efforts are required to solve these issues fundamentally. We believe a short-term solution is to relax some non-critical requirements and focus on architecture and designs that maximize the benefits of 3D (such as high density and high parallelism). For example, a brain transmits signal orders of magnitude slower than a modern digital system but can still make instant decisions utilizing its vast parallel processing network. Similarly, we could use the high-density integration and parallelism while sticking to a lower operating frequency. An effective approach to reducing heat generation is to operate the circuits at a substantially lower speed when possible. Another method is to implement sparse neural networks in 3D, in which the overall activity of circuit components is kept relatively low.

Secondly, to maintain the claimed data throughput, the input and output of the 3D arrays should have a high enough bandwidth. This could be a potential issue since peripherals are usually placed on a 2D silicon surface. System-level co-design is needed to coordinate 2D/3D systems to avoid a potential speed bottleneck.

Finally, to use these 3D neuromorphic systems in real-world applications, a systematic design of the software is also required, ranging from the electronic design automation (EDA) tools to the operating systems, providing the compensation and management programs for the 3D circuits. The thermal issues could also be mitigated at the software level using thermal management programs. On the other hand, a significant advantage of the 3D neuromorphic circuit is its connectivity that brings in much-improved computing throughput. A future 3D neuromorphic chip may have a set of neural circuit modules to participate in a wide range of applications. A software abstraction of those 3D neuromorphic modules should be established for developers, and therefore dedicated simulators, compilers, and development environments for end-users should also be available.

In summary, ever since Carver Mead created the concept of neuromorphic engineering over thirty years ago, we have been aiming to build a brain-like computer. With the toolset in shape, it calls for more research efforts and more disruptive approaches. We project that the future of neuromorphic computing is in 3D form—a morphology that mostly resembles the brain. 3D circuits have symbolic similarities to the brain and offer significant benefits in functionality, performance, and scalability, all core needs to go beyond the current AI hardware.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Qiangfei Xia https://orcid.org/0000-0003-1436-8423

References

[1] Gokmen T and Vlasov Y Acceleration of deep neural network training with resistive cross-point devices: design considerations *Front. Neurosci.* **10** 1–13

[2] Wang Z *et al* 2020 Resistive switching materials for information processing *Nat. Rev. Mater.* **5** 173–95

[3] Xia Q and Yang J J 2019 Memristive crossbar arrays for brain-inspired computing *Nat. Mater.* **18** 309–23

[4] Shim W and Yu S 2021 Technological design of 3D NAND-based compute-in-memory architecture for GB-scale deep neural network *IEEE Electron Device Lett.* **42** 160–3

[5] Park J *et al* 2021 A 176-Stacked 512Gb 3b/cell 3D-NAND flash with 10.8Gb/mm^2 density with a peripheral circuit under cell array architecture *2021 IEEE Int. Solid-State Circuits Conf.* 64, 422–3

[6] Lee S *et al* 2016 A 128Gb 2b/cell NAND flash memory in 14 nm technology with tPROG = 640 μs and 800MB/s I/O rate *Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf.* 59, 138–9

[7] Baek I G *et al* 2011 Realization of vertical resistive memory (VRRAM) using cost effective 3D process *Tech. Dig. - Int. Electron Devices Meet. IEDM* 737–40

[8] Lin P *et al* 2020 Three-dimensional memristor circuits as complex neural networks *Nat. Electron.* **3** 225–32

[9] Shulaker M M *et al* 2017 Three-dimensional integration of nanotechnologies for computing and data storage on a single chip *Nature* **547** 74–8

[10] Sabry Aly M M *et al* 2019 The N3XT approach to energy-efficient abundant-data computing *Proc. IEEE* **107** 19–48

[11] Seok J Y *et al* 2014 A review of three-dimensional resistive switching cross-bar array memories from the integration and materials property points of view *Adv. Funct. Mater.* **24** 5316–39

[12] Hudec B *et al* 2016 3D resistive RAM cell design for high-density storage class memory—a review *Sci. China Inf. Sci.* **59** 1–21