

# Neuromorphic Brain-Inspired Computing with Hybrid Neural Networks

Zongyuan Cai<sup>1,\*</sup>, Xinze Li<sup>2,\*</sup>,†<sup>1</sup> Shenyang Ligong University, Shenyang, China<sup>2</sup> Beijing University of Chemical Technology, Beijing, China

\*Corresponding author e-mail: zongyuan.cai@qq.com, 2016013010@mail.buct.edu.cn

†These authors contributed equally.

**Abstract**—Neuromorphic brain-inspired computing is believed to solve the bottleneck of traditional Von Neumann architecture computers and may promote the development of the next-generation of high-performance computer architectures. Therefore, in recent years, brain-inspired computing has received extensive attention. Some large-scale brain-inspired research projects have yielded some results, such as further increasing computer capabilities in data processing and machine learning. However, there is a lack of widely used artificial neural network based on computer science and neuroscience-inspired models and algorithms. This low compatibility between software and hardware reduces the computational programming efficiency and becomes an obstacle to the development of brain-inspired computing. This paper introduces the concept of neuromorphic brain-inspired computing and then introduces the research results of Neurogrid, Spinnaker, TrueNorth, Loihi and Tianjic, respectively. Finally, the paper shares the views on the future development trend of the field of neuromorphic brain-inspired computing.

**Keyword:** *Neuromorphic Brain-Inspired Computing; Hybrid Neural Networks*

## I. INTRODUCTION

In recent years, the development of brain-inspired computing is full of opportunities and challenges. With the development of computer hardware and the huge growth of data, it heralds the coming of the era of big data and artificial intelligence. Since AlphaGo beat Lee Se-dol at Go without human knowledge a few years ago [1], the GPT-3 model now has billions of parameters and can talk to us almost like humans [2]. People believe that the era when robots can serve human beings completely autonomously is no longer far away. However, the reality is that the technology, Artificial General Intelligence (AGI) [3], we're pursuing has hit a bottleneck.

Increasingly large amounts of data are called the new 'oil', and they are used to model artificial intelligence algorithms. However, the amount of data is exploding exponentially, its form is becoming more and more diversified, and the demand for real-time data processing is also growing rapidly. Using such a large amount of data to train algorithmic models requires more powerful and less power-consuming computing equipment. Thus, how to improve the computing power of the hardware and optimize the artificial intelligence algorithm model has become a problem. There are many problems in the

current popular Artificial Neural Networks (ANNs) algorithm, such as the lack of interpretation of the deep learning algorithm model process, the lack of model robustness, the lack of large-scale data used to train the model, and too much reliance on intensive features, etc. On the hardware side, the size of transistors inside microprocessors has shrunk to near the physical limit of atomic size and Moore's Law has come to an end. In the traditional Von Neumann architecture, which separates the memory from the computing unit, the communication delay between the storage unit and the computing unit has become the bottleneck of its performance, resulting in the 'memory wall' problem [4].

To improve computing power, scientists have been exploring an alternative route: neuromorphic brain-inspired computing. At present, there is no unified standard for neuromorphic brain computation, so it is an up-and-coming frontier technology to study. The main body of this paper introduces the concept of neuromorphic brain-inspired computing. According to the development of time and technology, the research results of Neurogrid, Spinnaker, TrueNorth, Loihi and Tianjic on large-scale neuromorphic brain-inspired computing are introduced, respectively. Finally, the future development trend of neuromorphic brain-inspired computing is shared.

## II. NEUROMORPHIC BRAIN INSPIRED COMPUTING

Neuromorphic computing is also known as neuromorphic engineering. The concept, proposed by Carver Mead, is based on using large scale integration (VLSI) systems with electronic analog circuits to simulate the neurobiological structure of the nervous system [5]. Neuromorphic Electronic Systems refer to the biological brain with Electronic circuits. Brain-inspired Computing is an idea that uses the basic principles of brain science to improve computers. Its ultimate goal is to develop computing comparable to that of the human brain at the level of AGI [3]. Brain-inspired computing, as a computing system, includes algorithms, chips and systems. Therefore, brain-inspired computing is inspired by the biological brain and builds a new computing paradigm by simulating the nervous system of the biological brain.

The biological brain has always been considered as a very complex biological structure with high computational power, low power consumption, and reliability. If we compare the

human brain to be a ‘computer’, its characteristics and advantages include: learning by interacting with the outside world independent (without explicit programming), highly fault-tolerant (tolerate a large number of the death of neurons and does not affect the basic function), high parallelism ( $10^4$  neurons), high connectivity ( $10^4$  synapses), low operation frequency (100 Hz), low speed communication (a few meters per second) and low power consumption (about 20 watts) [6]. Therefore, scientists hope to imitate the biological brain and develop a computer with low power consumption, strong robustness and the ability to handle a variety of complex tasks. Although the working principle of the brain are still largely unknown, neuroscience-inspired algorithms such as ANNs are excellent at deep learning [7].

To simulate the brain is to simulate nerve cells, among which neurons are the most basic structural and functional units of the nervous system. It includes dendrite, soma, synapse and axon. Neurons are excitable cells, and they process and transmit information through electrical and chemical signals in the nervous system. Neurons are stimulated to produce action potentials, also known as impulses, to communicate with each other. Therefore, Neuromorphic Brain-Inspired Computing used transistors to simulate the functions of biological neurons and synapses and to simulate the brain to realize neuron and synaptic computing through pulse-driven communication. Later, it rapidly developed to include event-driven nature of computing as discrete "impulses". Today, the advantages and limitations of pulse-driven computing, especially learning with pulses, are widely studied. Inspired by the pulse pattern of human brain activities, spiking neural networks (SNNs) exhibits high biological fidelity, rich Spatio-temporal information encoding and event-driven particularity, becoming a major neural computing paradigm with high energy efficiency in the process of dynamic sequential information processing [8].

### III. HYBRID NEURAL NETWORKS

SNNs is a commonly used algorithm model for neural mimicry computation. It is a computational model known as the third generation of neural networks that takes neurons as the basic processing elements [9]. The pulse neurons used in it are mainly of ‘integration-emission’ type and exchange information through pulses. Neurons in SNNs do not transmit information every propagation cycle, as in typical multilayer perceptual Networks, but only when the membrane potential reaches a threshold. When the membrane potential reaches a threshold, the neuron fires and generates a signal that propagates to other neurons, which in turn increase or decrease its potential. Nowadays, there are many pulse coding methods of SNNs. Pulse coding is the process of converting a multi-bit signal into a single bit pulse. Rate coding is that the rate of pulse is proportional to the strength of the original signal, and the data of the original signal represents the rate of the pulse signal. In a certain period, the more the pulse rate is, the stronger the original signal is. There are other ways, such as latency coding: coding based on the pulses sequence and the first arriving pulses are stronger.

Unlike ANNs, SNNs includes time as an explicit dependency in their calculations. At any given moment, one or more neurons may send a single-bit pulse through a directional

connection called a synapse to a neighboring neuron, which may travel for a non-zero time. Neurons have local state variables, their evolution, and temporal rules for the generation of impulses. Thus, the network is a dynamic system in which individual neurons interact with each other through impulses [10]. Simply put, SNNs considers the internal dynamics of each neuron, that is, the neurodynamics, and the output of each neuron is an impulse train. However, ANNs ignores the internal dynamics of each neuron, and the output of each neuron is a state value. So SNNs can be thought of as a dynamic combination of ANNs and internal neurons. Therefore, the model made under the ANNs framework can be transformed into the SNNs model.

Because of their respective advantages, the integration of ANNs and SNNs is called hybrid neural networks (HNNs). While HNNs retains the basic characteristics of neural networks, ANNs and SNNs can work together to perform complex tasks. In computer vision, researchers use ANNs model to extract edge contrast in images and further use SNNs model for processing to achieve low power consumption and high performance. Therefore, it is promising for efficient implementations on specific domain hardware platforms [11]. For example, the Tianjic chip model is composed of a non-pulsed Neural networks (ANNs) and a pulsed Neural networks (SNNs), also its chip architecture is composed of an SNNs Neuromorphic (multi-core) architecture and ANNs supports [12].

### IV. THE DEVELOPMENT OF BRAIN-INSPIRED COMPUTING

#### A. Neurogrid

Neurogrid, a neuromorphic system designed to simulate large-scale neural models in real time, was developed at Stanford University in 2004. Neurogrid is a hybrid digital-analog system that uses electronic circuits to simulate neurons to maximize the number of synaptic connections, a choice that also maximizes energy efficiency. To improve throughput, it also interconnects the neural arrays using a tree network [13]. Each circuit board in Neurogrid integrates 16 neural computing cores, simulating  $256 \times 256$  neurons and million level neurons by a single-core, and its communication part uses FPGA digital communication. Also, the Neurogrid’s power consumption is approximately 3W. As an early digital-to-analog hybrid chip, Neurogrid was less programmable, but it laid the foundation for later brain-inspired computing.

#### B. SpiNNaker

SpiNNaker is the acronym of ‘Spiking Neural Network Architecture’. It is the leading computing platform of the EU Brain Project. Since 2005, the University of Manchester in the United Kingdom began to develop the ARM-based multi-core computing Architecture. It is a massively parallel computer designed specifically to integrate one million ARM processors into a single system capable of modeling up to one billion neurons and a trillion synapses. Accelerating our understanding of the brain through its ability to support a massive pulsing neuron system in real-time [14].

A Spinnaker chip contains 18 ARM processor cores and a 128 (Mbyte) memory chip, with shared access between

processors. A single core can simulate 1000 neurons, which can be flexibly programmed to achieve a variety of models, the power consumption of a single chip is less than 1 (W). Forty-eight of these chips are packaged on a circuit board, which adds up to 864 cores per board. For ease of extension, 24 circuit boards can be mounted in a 19-inch card frame, and five card frames can be stacked in a standard 19-inch cabinet [15].

Different from Neurogrid, SpiNNaker is an all-digital design with low power consumption ARM and a chip integrated with multiple ARM and a circuit board integrated with multiple chips, which provides a good idea for the later development of a larger brain-inspired computing platform.

### C. TrueNorth

TrueNorth is a brain-inspired computing project developed by IBM with funding from the U.S. Defense Advanced Research Projects Agency. The ultimate goal is to develop hardware that breaks the Von Neumann system and fundamentally breaks away from the traditional system design, which requires attention to real-time operation, low power consumption and scalability of the whole system.

The TrueNorth chip has 4,096 neural computing cores, containing 1 million digital neurons and 256 million synapses. Also, it's connected by a two-dimensional, grid-like routing topology. It implements a non-von Neumann low-power (65 MW), highly parallel, scalable, and fault-tolerant architecture.

IBM has also developed a suite of asynchronous-synchronous hybrid circuits and a complete work procedure for TrueNorth to build an event-driven, low-power synaptic chip. The TrueNorth chip is also fully configurable in terms of connectivity and neural parameters and can be customized for a wide range of cognitive and sensory sensing applications. The use of TrueNorth-based systems in multiple applications, including visual object recognition, has a higher performance and an order of magnitude lower power consumption than the same algorithm running on the von Neumann architecture. Also, the TrueNorth make it easier for designers to develop new brain-inspired architectures and systems [16].

As a computing platform, the learning function is critical. TrueNorth does not have on-chip online learning capabilities, so the neural network training needs to be done offline. Therefore, a massively parallel computing platform based on central processing unit (CPU) or graphics processing unit (GPU) is needed to download the trained neural network topology and parameters to the chip for execution. This restriction dramatically simplifies chip design, but it also limits the chip's ability to adapt dynamically, requires human intervention or a system restart to make any changes [6].

### D. Loihi

Intel in 2015 began the study of neuromorphic computing, hoping to neuroscience, study new computer architecture. In 2017, Intel can launch a self-learning neural chip Loihi. Loihi chip uses Intel's 14-nanometer process to build, contains 2.07 billion transistors and 33 (MB) of SRAM across its 128 neuromorphic cores and three x86 cores. Each single core supports up to 1,024 neurons [10].

Intel has launched an extended cluster based On Loihi chips

from Pohoiki Beach to the latest Pohoiki Springs, integrating 768 Loihi neural simulation research chips, containing 100 million neurons. The chips are installed in five standard server-sized cases, expand the Loihi Chip by more than 750 times and operate at less than 500 watts of power. Intel has developed support programming API: providing unified extensible programming in heterogeneous computing model is convenient for developers, reducing the complexity of application transplant.

Recently a lot of research on neural mimicry calculation using the Loihi. Some researches develop the Keras programming interface and the compiler NxTF derivation, so developers can flexibly develop SNNs algorithm quickly on the Loihi. Through the interface and compiler optimization, depth of convolution SNNs can be mapped to multicore Intel Loihi structure. For a MobileNet model with a 28-layer 4M parameter input size of 128 x 128. Besides, they adopt the combination of transfer learning, computational neuroscience, and deep learning principles. For example, the SOEL learning system can carry out the gesture recognition on Loihi and quickly learn new gestures online from real-time data streams [17].

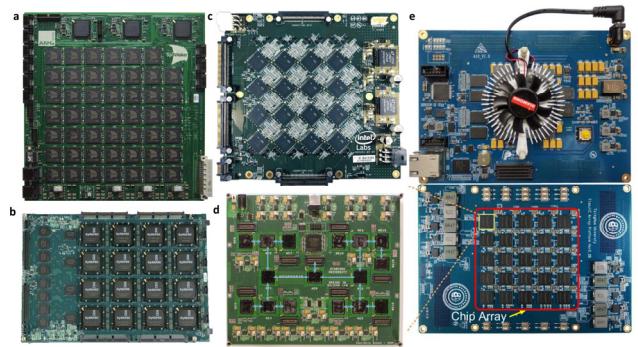


Fig. 1 | Large-scale brain-inspired computing systems.

a, SpiNNaker. b, TrueNorth. c, Loihi. d, Neurogrid. e, Tianjic.

### E. Tianjic

Unlike the brain-inspired computing platform mentioned above, Tianjic is a brain-inspired computing with Unified Architecture for Hybrid Paradigm, which is also one of the China's brain engineering achievements in recent years. At many strategy meetings organized by the Ministry of Science and Technology and the National Natural Science Foundation of China, discussions among Chinese scientists have reached a consensus that understanding the neural basis of human cognition is a universal goal of neuroscience and should be a central pillar of China's brain engineering. Among them, brain-inspired intelligence technology is one of the wings of 'One Body, Two Wings', the core strategy of China's Brain engineering [18]. Brain-Inspired Center of Tsinghua University was founded in September 2014 as a young research institute with the aim of studying how the human brain processes information and solving the bottlenecks of current artificial intelligence. Tianjic is a core product of the Center for Brain-inspired Computing of Tsinghua University, aiming to build a universal brain-inspired computing system architecture to

support the research and exploration of biological cognitive mechanisms. In 2017, Tianjic II was released, and soon after, Tsinghua University developed Tianjic's first generation software toolchain. Finally, Tsinghua University built the first brain-inspired demonstration platform in 2019.

Tianjic wants to create a general-purpose computing platform for both computer science and neuroscience. The platform should have good compatibility and can support mainstream computer science-based ANNs as well as neuroscience-inspired models and algorithms. What's more, Tianjic is a unified architecture for hybrid paradigm which absorbs many advantages of predecessors mentioned above, breaks the restriction that the current AL chip only supports ANNs or SNNs, opens up a new cross-paradigm computing chip approach of ANNs and SNNs fusion and achieves innovation under the existing technology. Tianjic chip supports heterogeneous fusion of ANNs and SNNs neurons and heterogeneous fusion of neural networks. This is also known as HNNs. While HNNs retains the basic characteristics of neural networks, ANNs and SNNs can work together to perform complex tasks. It is promising for efficient implementations on specific domain hardware platforms. The Tianjic Chip model is composed of a non-pulsed Neural networks (ANNs) and spike neural networks (SNNs), and its Chip architecture is composed of an SNNs Neuromorphic (multi-core) architecture and ANNs support [12].

In order to improve the compatibility between software and hardware of Tianjic system, and enhance programming flexibility and development efficiency, Tsinghua University also first put forward the 'Neuromorphic Complexity' and a brain-inspired computing system hierarchy that is decoupled from software and hardware which relaxed the requirement for hardware completeness. It proposed the corresponding system hierarchy, including a Turing complete software abstract model and a general abstract neuromorphological system architecture. The hardware completeness and compilation feasibility of this system is proved by theoretical demonstration and prototype experiment. The application scope of brain-inspired computing system is extended to support general computing [19].

The chip uses a 28 (nm) process and is comprised of 156 FCores, including 40,000 neurons and 10 million synapses. Each of neurons is input by 256 axons. For its performance, with its distributed on-chip memory and decentralized multi-core architecture, Tianjic offers more than 610 (GB) per second of internal storage bandwidth and generates an effective peak performance of 1.28 megawatts per second when running at 300 (MHz) in ANNs models. In SNNs models, where synaptic operations are commonly used to benchmark the chip, Tianjic achieved an effective peak performance of about 650 (GB) per second of synaptic operations (GSOPs) per watt [20].

The practice of deep learning shows that the size of neurons is a critical factor for performance improvement. Tianjic's network scaling capability is powerful in order to reach large-scale neurons. The architecture takes the scalable grid connections of TrueNorth and Loihi, such as the core-chip-board-system-cloud server node-large data center scaling approach, but it is more flexible than them.

To demonstrate the practicality of Tianjic's system,

Tsinghua University has also developed a robotic platform for brain-inspired computing. It can perform multiple complex tasks simultaneously involving multimodal perception, high-level decision making and autonomous motion. An unmanned bicycle is developed based on this platform, which simultaneously completes various tasks including object tracking, obstacle avoidance, voice command recognition, balance control and decision-making in various real environments. This study demonstrates the application of cross-paradigm neuromorphic computing systems in robotic platforms, which can not only support large-scale and diverse networks, but also be developed and improved over time to facilitate the development of online learning [21].

## V. CONCLUSION

In conclusion, in recent years, brain-inspired computing has developed rapidly and is moving towards the direction of large neuronal scale and cross-paradigm. Brain-inspired computing is still far from practical application in industry. However, these challenges also provide new research directions and opportunities for researchers. For example, optimizing SNNs models, solving the bottleneck of large-scale training, enhancing generalization ability, robustness, and autonomous learning, etc., the HNNs mentioned above is also a potential research direction.

In addition to hardware, the ecosystem of software systems will also determine the development of brain-inspired computing platforms. Develop an interactive brain-inspired operating system that supports the management of neuron hardware resources in the system and can run ANNs, SNNs, or HNNs algorithms, which will facilitate the testing of the entire platform, and facilitate developers to build a brain-inspired computing platform ecology jointly.

With a healthy application development ecosystem in place, many more fields can be integrated with brain-inspired computing in the future. For example, the low-latency vision applications, because the SNNs can be processed at the microsecond level, the combination of SNNs and ANNs can help develop computer vision which is particularly useful for applications such as autonomous driving that need to monitor the surrounding road conditions in real time. Also, it can be applied to brain-computer interfaces. Brain-inspired computing can process brain pulse signals from the brain-computer interface with low power consumption and high performance, then decode them to open up the communication between the human brain and the computer. After that, a hybrid intelligent system of brain-computer fusion is formed.

Finally, the development of a new technology cannot be separated from opportunities and challenges. We need patience and confidence. In the future, the Unified Architecture for Hybrid Paradigm with Neuromorphic Brain-Inspired Computing will possibly be one of the best choices in neuromorphic computing fields.

## REFERENCES

- [1] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.

- [2] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *arXiv*, May 2020, [Online]. Available: <http://arxiv.org/abs/2005.14165>.
- [3] B. Goertzel, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," *J. Artif. Gen. Intell.*, vol. 5, no. 1, pp. 1–48, Dec. 2014, doi: 10.2478/jagi-2014-0001.
- [4] W. A. Wulf and S. A. McKee, "Hitting the memory wall," *ACM SIGARCH Comput. Archit. News*, vol. 23, no. 1, pp. 20–24, Mar. 1995, doi: 10.1145/216585.216588.
- [5] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, May 1990, doi: 10.1109/5.58356.
- [6] D. Scal-, T. E. P. Sec-, and G. Kasparov, "Gu Zonghua, Pan Gang Zhejiang University," pp. 10–20, 2015.
- [7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [8] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, 2019, doi: 10.1038/s41586-019-1677-2.
- [9] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, Dec. 1997, doi: 10.1016/S0893-6080(97)00011-7.
- [10] M. Davies *et al.*, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018, doi: 10.1109/MM.2018.112130359.
- [11] G. Wang, S. Ma, Y. Wu, J. Pei, R. Zhao, and L. Shi, "End-to-End Implementation of Various Hybrid Neural Networks on a Cross-Paradigm Neuromorphic Chip," *Front. Neurosci.*, vol. 15, no. February, pp. 1–13, Feb. 2021, doi: 10.3389/fnins.2021.615279.
- [12] L. Deng *et al.*, "Tianjic: A unified and scalable chip bridging spike-based and continuous neural computation," *IEEE J. Solid-State Circuits*, vol. 55, no. 8, pp. 2228–2246, 2020, doi: 10.1109/JSSC.2020.2970709.
- [13] B. V. Benjamin *et al.*, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, 2014, doi: 10.1109/JPROC.2014.2313565.
- [14] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker Project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, May 2014, doi: 10.1109/JPROC.2014.2304638.
- [15] S. B. Furber, "Brain-inspired computing," *IET Comput. Digit. Tech.*, vol. 10, no. 6, pp. 299–305, 2016, doi: 10.1049/iet-cdt.2015.0171.
- [16] F. Akopyan *et al.*, "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015, doi: 10.1109/TCAD.2015.2474396.
- [17] B. Rueckauer, C. Bybee, R. Goetsche, Y. Singh, J. Mishra, and A. Wild, "NxTF: An API and Compiler for Deep Spiking Neural Networks on Intel Loihi," vol. 1, no. 1, pp. 1–21, Jan. 2021, [Online]. Available: <http://arxiv.org/abs/2101.04261>.
- [18] M. Poo, J. Du, N. Y. Ip, Z.-Q. Xiong, B. Xu, and T. Tan, "China Brain Project: Basic Neuroscience, Brain Diseases, and Brain-Inspired Computing," *Neuron*, vol. 92, no. 3, pp. 591–596, Nov. 2016, doi: 10.1016/j.neuron.2016.10.050.
- [19] Y. Zhang *et al.*, "A system hierarchy for brain-inspired computing," *Nature*, vol. 586, no. 7829, pp. 378–384, 2020, doi: 10.1038/s41586-020-2782-y.
- [20] J. Pei *et al.*, "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, Aug. 2019, doi: 10.1038/s41586-019-1424-8.
- [21] Z. Zou *et al.*, "A hybrid and scalable brain-inspired robotic platform," *Sci. Rep.*, vol. 10, no. 1, p. 18160, Dec. 2020, doi: 10.1038/s41598-020-73366-9.