Key Technologies and Challenges of Memory Resistors and Brain like Chips in Improving Computing Power

Wenhan Wang *

Electronic Information Engineering, China University of Petroleum (Beijing), Beijing, 100193

* Corresponding Author Email: dndisjs138@outlook.com

Abstract. This article explores the cutting-edge applications of brain like chips and memristors in the fields of artificial intelligence and computer science. Brain like chips, imitating the computational model of brain neural networks, provide new hardware support for machine learning and deep learning with unique parallel processing capabilities and efficient energy consumption ratio. Memristors, as an emerging type of non-volatile memory, have great potential in information storage and processing as resistance can change with changes in passing current and voltage. This article first outlines the basic principles and characteristics of brain like chips and memristors, then analyzes application prospects in the fields of computational science and neuroscience, and finally discusses the current challenges and future development trends. This article aims to provide a comprehensive perspective on the latest developments in brain like chips and memristors for researchers in related fields, and promote the deep integration and application of these two technologies in the field of artificial intelligence.

Keywords: Memristors, Parallelism, Chips, Neural network, Artificial intelligence.

1. Introduction

Since the emergence of neural network algorithms and artificial intelligence, the demand for computing power has reached an unprecedented level and is constantly increasing rapidly. However, traditional computers have reached a bottleneck in development. This creates a huge gap between the current computing power and the computing power required for the development of artificial intelligence. To address the huge gap between demand and reality. New types of brain like chips and memristors that are different from traditional computer structures have emerged. The academic community has created a concept called AIoT, which means AI and Internet of things. Through this concept, people can clearly see the enormous demand for computing power in the development of neural network algorithms. The number of connections in 2023 is close to 20 billion, with a compound growth rate of over 30% over the past decade. It is expected that the number of connections will exceed 30 billion by the end of 2025 **[1]**. As shown in Figure 1. In 1965, the founder of Intel, Moore, proposed Moore's Law, which believed that global computer power would grow at a rate of doubling every 18 months. At the beginning, the performance of computer processors did indeed improve significantly. However, after entering the 21st century, the growth rate of computing power slowed down, and Moore's Law gradually became fail **[2]**.

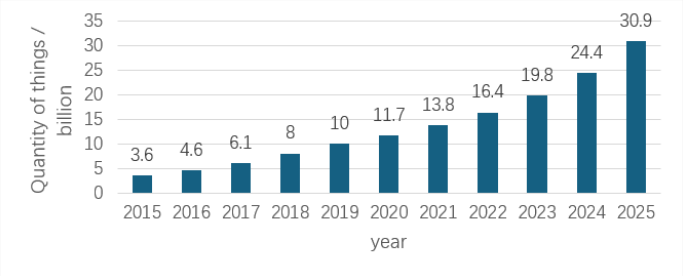

Figure 1. Bar chart of the development of the Internet of Things.

Due to the unique advantage of the von Neumann structure, it allows a large number of engineers from different fields to work together to design a computer, without the need to understand the specific content of other fields. Because of this, the structure of computers has hardly changed for

decades. The reason for the slowdown in computing power growth is that the potential of traditional computer structures has been almost exhausted (the von Neumann bottleneck), and the development of traditional chips has approached the atomic level **[3]**. The development mode of reducing transistor volume and increasing the number of transistors will reach its end. Additionally, as the storage and computing units of traditional chips are separate, this also means that a significant amount of time and energy will be consumed in the transportation of data. Meanwhile, due to limited memory capacity or transmission bandwidth, CPU performance cannot be fully utilized. Despite improvements in memory and communication bandwidth, the growth rate of communication bandwidth remains relatively slow compared to the growth of computing power. Therefore, upgrading computer architecture is imperative. Based on the current problems faced by traditional computers, improving parallelism and integrating storage and computing in Figure 2 are important directions for the development of new computers. The advantage of improving parallelism is that it can fully utilize computing resources and accelerate the running speed of programs. Besides, the advantage of integrated storage and computing is, first save time on data transmission, second Reduce computing costs and third has Greater computing power and higher energy efficiency. Memristors and brain like chips are in line with this solution approach.

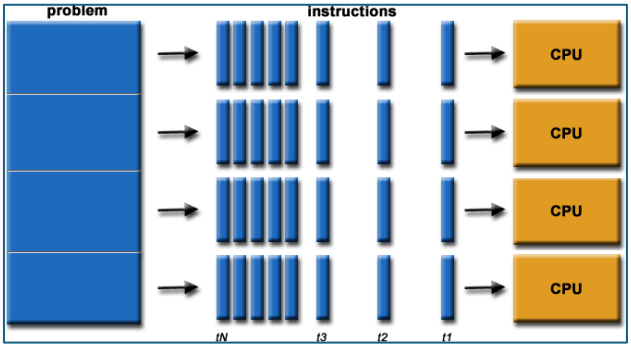

Figure 2. Parallel structure.

2. Method

The development of neuromorphic chips began in the 1980s, and in 1989, Carver Mead from California Institute of Technology proposed the concept of neuromorphic engineering. But in the following decade or so, there was no significant progress in this field. It was not until 2004 that Kwabena Boahe from Stanford University developed the Neurogrid, a brain like chip based on analog circuits that brain like chips began to develop rapidly. In the following twenty years, with the integration and development of fields such as neuroscience, computer science, and nanotechnology, coupled with the enormous demand for computing power raised by the rise of artificial intelligence. Brain like chips have become an important path to improve computer computing power, and have gradually generated multiple branches, including artificial neural network chips, pulse neural network chips, heterogeneous fusion chips, and brain simulation chips **[4].**

The current artificial neural chips have achieved considerable success and have been applied in reality. This is a chip that simulates the structure and function of the human brain's neural network, with the aim of improving computer computing power to assist artificial intelligence in efficient computation and inference. By simulating the connectivity of human brain neurons for decision-making and reasoning. Taking the well-known artificial neural network chip TPU (Tensor processing unit) v4 as an example, this article introduces the characteristics and functions of this chip. TPU v4 is the fourth generation TPU chip launched by Google in 2021. This chip is widely used in edge computing. Due to the emergence of the Internet of Things (IoT), it is almost impossible to simply use cloud computing to process all data. Therefore, a large amount of data needs to be calculated at the edge to reduce the pressure on the cloud and improve the efficiency of data processing. TPU v4 adopts Google's self-designed TPU Matrix Processing Unit (MPU), which consists of 32 cores and

has characteristics such as high throughput, high parallelism, low power consumption, and low latency. In addition, TPU v4 can also connect different devices through the USB interface to complete the calculation, greatly improving the ability of edge computing. The 32 cores owned by TPU v4 have been highly optimized after several iterations, allowing for a large amount of efficient matrix operations in machine learning. This operation method achieves the high throughput, high parallelism, low power consumption, and low latency characteristics of TPU v4. A pod composed of 4096 TPU v4 has a computing power of over 10 18th power floating-point operations per second, even surpassing the fastest computer in the world.

The chip designed based on Pulse Neural Network (SNN) is a type of chip that simulates the information transmission process of neurons in biological neural networks. In a pulse neural network, each neuron receives an input pulse and accumulates for a period of time. When the accumulated voltage exceeds a certain threshold, the neuron will generate a pulse and reset the accumulated voltage to zero. This working method actually simulates information transmission in biological neural networks. Each neuron in the pulse neural network chip has independent input and output pulse lines for information transmission. Therefore, each neuron can perform computation and information transmission simultaneously, with high parallelism, greatly improving computing power **[5]**.Taking IBM's Truenorth chip as an example, the Truenorth chip consists of 4096 cores, each consisting of 256 neurons in sequence, and each neuron can access 256 axons. Each core has 256 x 256 protrusions arranged in a horizontal bar. The axons of neurons can be connected to a single core, for example, for looping connections to own core, or to output pins. If the axon needs to connect to multiple output destinations, additional neurons need to be used to replicate its spikes. The spikes between neurons are transmitted asynchronously, however, in order to maximize efficiency, all neuron states are updated at a rate of 1000 Hz. The computing power of Truenorth chips is not outstanding, only 58gops, far lower than that of typical accelerators. However, its power consumption is also very low, making this chip have a good power consumption ratio. Additionally, due to the use of 2D Mesh on-chip networks, chips can be directly interconnected through LVDS serial ports, providing excellent interconnectivity. At present, pulse neural network chips have some applications in reality, such as the chip launched by the Belgian Microelectronics Center for use in unmanned aerial vehicle collision avoidance radar systems, which uses pulse neural networks. The advantage of this chip is that it has low power consumption and latency, which can significantly improve processing speed and extend battery life. However, although pulse neural network chips have some applications in reality, usage range and scale are relatively small and need further improvement. The reason for this problem is that the computing power of pulse neural network chips is closely related to the number of neurons, and the huge number of neurons makes chip design and production extremely difficult. In addition, the complexity of pulse neural network algorithms also increases the difficulty of application.

The development of heterogeneous fusion chips can be traced back to the late 20th century. In order to solve the problem of insufficient computing power, engineers began to try to integrate different types of processor or controller architectures into one chip, thus giving birth to the concept of heterogeneous fusion chips **[6]**. Unlike the two types of chips mentioned above, heterogeneous fusion chips do not have a specific structure. This type of chip typically consists of various types of controllers and control architectures, and may also include both artificial neural networks and pulse neural networks. The following is a simple structural composition of a typical heterogeneous fusion chip. The central processing unit (CPU) is responsible for the overall control of the chip and processing higher-level tasks with greater difficulty, while the graphics processing unit (GPU) is responsible for handling large-scale simple computing tasks, particularly skilled in computationally intensive tasks such as graphics rendering and video encoding. A digital signal processor (DSP) is responsible for processing digital signals similar to audio and video, improving the efficiency and quality of processing similar signals. A neural network processor (NPU) is responsible for processing neural network computing tasks, which can efficiently perform matrix multiplication and convolution operations, greatly improving the efficiency of heterogeneous fusion chips in processing deep learning tasks. In addition, heterogeneous fusion chips also include other parts such as memory

input/output interfaces **[7].**At present, heterogeneous fusion chips have become one of the important development trends in the chip industry. In the field of mobile communication, heterogeneous fusion chips can achieve more efficient communication and data processing, meeting the performance requirements of devices such as smartphones and tablets. In the field of artificial intelligence, heterogeneous fusion chips can combine various processor architectures such as CPU, GPU, FPGA (Field Programmable Logic Gate Array), etc., providing powerful computing power and flexibility, supporting complex tasks such as deep learning and machine learning.

Memristor, also known as variable resistor, is a special type of circuit device with memory function. Its resistance is not fixed, but determined by the amount of charge flowing through it. Therefore, by measuring the resistance value of the memristor, the amount of charge flowing through it can be determined, thereby achieving the effect of memory charge **[8]**. As shown in Figure 3, the emergence of memristors has perfected the fundamental components of electricity

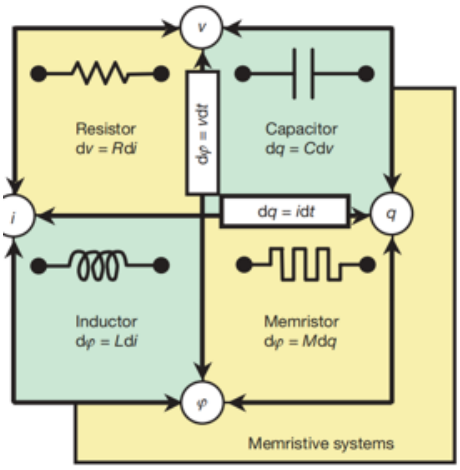

Figure 3. Memristive system.

The working principle of a memristor is based on physical phenomena such as magnetoresistance and electric field modulation. It is usually composed of ferromagnetic or ferroelectric materials, which have special crystal structures and electrical properties. Under the stimulation of external voltage or current, slight deformation occurs inside the material, which affects the resistance value. This change in resistance value can be seen as a form of "memory", and even in the absence of external electric field stimulation, the memristor can maintain its resistance value stable for a long time. Memristors include the following types. Linear memristor: This is the most common type with linear resistance characteristics. When an applied voltage is applied to the memristor, the memristor will change its resistance value based on the magnitude of the voltage. Linear memristors are widely used in analog circuits to adjust resistance values. Nonlinear memristor: The resistance characteristics of this memristor are not linear, and achieve various complex nonlinear functional relationships based on different input signals. Therefore, in the fields of circuit design and computer science, nonlinear memristors have important applications. Digital memristor: specifically designed for storing and processing digital information. Digital memristor can change resistance value through programming to achieve storage and calculation functions. Digital memristors have been widely used in fields such as digital circuits and memory.

The structure and integration process of memristor arrays are important research directions in memristor technology. A memristor array is a two-dimensional or three-dimensional structure composed of multiple memristor units like Figure 4, which can achieve large-scale information storage and processing. The structural design of memristor arrays is crucial for performance and stability. Common memristor array structures include cross arrays, vertical nanowire arrays, and two-dimensional planar arrays. Cross array structures have advantages such as high density, easy scalability, and low power consumption, making memristor more common in practical applications. The vertical nanowire array structure can further improve the integration density and performance of memristors, but the preparation process is relatively complex. The two-dimensional planar array

structure is relatively simple and suitable for preparing large-area, uniform memristor arrays **[9]**.The integration process of memristors is to integrate multiple memristor units together to form a circuit or system with specific functions. Common integration processes include photolithography, peeling, and thin film preparation. The photolithography process is used to create patterns and electrodes for memristor arrays, while the stripping process is used to remove unnecessary materials and form the structure of memristor units. The thin film preparation process is used to grow the material thin film required for memristors on the substrate. In the integration process, attention should be paid to issues such as isolation between memristor units, selection and preparation of electrode materials, quality and uniformity of thin films, etc. Reasonable process parameters and structural design can significantly improve the performance and stability of memristor arrays. The interconnection and data transmission of multiple arrays of memristors are crucial aspects of memristor technology, which determine the information exchange and processing capabilities between memristor arrays and other circuit systems. In order to achieve information exchange between multiple memristor arrays, appropriate interconnection methods need to be adopted. Common interconnection methods include bus interconnection, cross switching, and network interconnection. Bus interconnection achieves communication between arrays through shared buses, cross switches provide direct point-to-point connections, and network interconnection achieves complex communication between arrays by constructing a network topology structure. The interconnection strategy needs to be determined based on specific application scenarios and system requirements. For example, for applications that require high throughput and low latency, a highly parallelized interconnection strategy may be required. Data transmission is a crucial step in exchanging information between memristor arrays and other circuits or systems. Common data transmission methods include serial transmission and parallel transmission. Serial transmission occupies less line resources, but the transmission speed is relatively slow; Parallel transmission can achieve high-speed data transmission, but requires more line resources. To ensure the correctness and reliability of data transmission, it is necessary to design appropriate transmission protocols. These protocols may include synchronous protocols, asynchronous protocols, and error correction mechanisms **[10]**.Memristors use analog storage, which has higher accuracy compared to traditional digital memory. This gives memristors significant advantages in areas that require high-precision weight storage, such as artificial neural networks. Memristors, due to unique properties, can be used to simulate the synaptic behavior of biological neurons. Therefore, memristors are expected to play an important role in the field of neuromorphic computing, providing new hardware implementations for artificial intelligence and machine learning. With the deepening of research on memristors, it is possible to achieve a system wide integrated memristor chip that supports efficient on-chip learning in the future. This will enable memristors to play a greater role in embedded systems, intelligent sensors, and other fields.

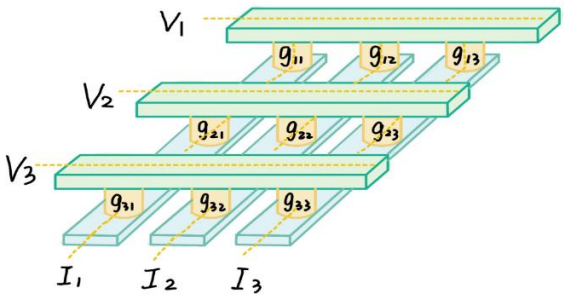

Figure 4. Memristor structure.

3. Result

Memristors, as a new type of non-volatile memory, have great application prospects, but memristors also face some challenges and difficulties. At present, the process for preparing functional thin films for memristors is not yet mature, which makes it difficult to accurately control the quality

of the prepared thin films, thereby affecting the performance of the devices. The behavior of commonly used materials for memristors, such as transition metal oxides (such as HfOx, AlO, etc.) or sulfur compounds under the combined action of ions and electrons, is not fully understood, which limits the further improvement of memristor performance. Although the working principle of memristors has been proposed, there are still multiple explanations for micro conductive mechanisms, which limits the in-depth understanding and optimization of performance. Although the power consumption of memristors has been significantly improved, traditional memristors based on transition metal oxides or sulfur compounds still struggle to meet the needs of low-power memory and computing tasks networks. Here are some possible solutions. Research and develop new materials to improve the energy efficiency and integration of brain like chips. Introducing advanced nanotechnology to reduce chip size and enhance performance. Adopting advanced packaging technology to improve the integration and scalability of chips. Research new interconnection technologies to achieve fast communication and collaborative work among multiple chips. Research the compatibility between memristors and existing semiconductor processes to achieve smooth integration with existing circuits. Develop new interfaces and drivers to improve the compatibility of memristors with other hardware and software. Develop optimization software algorithms for memristors to improve read and write efficiency and data processing speed. Research new programming models to simplify the operation and management of memristors.

4. Conclution

In today's rapidly developing artificial intelligence neural networks and increasing demand for computing power, brain like chips have become an important development direction in the future due to unique advantages of imitating the human brain in design. Memristors are highly anticipated due to high parallelism and integrated storage and computing capabilities. This article introduces and analyzes brain like chips and memristors with great potential.

References

[1] Xu K., “Design and research of high-energy efficient deep neural network acceleration chips [D]”, Xi'an University of Electronic Science and Technology, 2022, 1-10 (2022). (In Chinese)

[2] Xia, Q. and Yang J.J., “Memristive crossbar arrays for brain-inspired computing”, Nat. Mater. 18, 309–323 (2019).

[3] Kumar, S., Wang, X. and Strachan, J.P., “Dynamical memristors for higher-complexity neuromorphic computing”, Nat Rev Mater 7, 575–591 (2022).

[4] Zhao R.J., Fang C. and Xu M.S., “Development Status and Prospects of Artificial Intelligence Chip Industry [J]”, Economic Journal, 2022(11), 75-81 (2022). (In Chinese)

[5] Ambrogio, S., Narayanan, P., Okazaki, A., “An analog-AI chip for energy-efficient speech recognition and transcription”, Nature 620, 768–775 (2023).

[6] Sun Z.Y., Xiao S.Y. and Shen Q., “Design of Satellite Intelligence Platform Based on Heterogeneous Fusion Brain like Chips [J]”, Aerospace Standardization, 2022, (01), 40-43 (2022). (In Chinese)

[7] Zhang F., Zhai J.D. and Chen Z., “Performance Analysis, Optimization, and Application Review for Heterogeneous Fusion Processors [J]”, Journal of Software, 2020, 31(08), 2603-2624 (2020). (In Chinese)

[8] Chen C.L. , Luo C.H. and Liu S. , “Review of research status on memristor based brain computing chips [J]”, Journal of National University of Defense Technology, 2023,45(01), 1-14 (2023). (In Chinese)

[9] Tang C.F. and Hu W., “Low latency and low energy consumption novel sensing amplifier for memory resistor array in memory computing [J/OL]”, Microelectronics and Computers, 2024, (02), 58-66(2024). (In Chinese)

[10] Chen, Z., Lin, Z. and Yang, J., “Cross-layer transmission realized by light-emitting memristor for constructing ultra-deep neural network with transfer learning ability”, Nat Commun 15, 19-30 (2024).