Conventional Von Neumann and Neuromorphic Architecture of AI Chips

Jingwei Qu

Dipont Huayao Collegiate School Kunshan, Suzhou, China 215300

m40318@hznu-edu.cn

Abstract. In the background of the shortage of computing power in the training AI field, the current solutions and future outlooks will be shown if they tackle this dilemma. From the continuation of the traditional computer structure, Von Neumann structure**,** the three types of AI chips GPU, FPGA, and ASIC will be introduced and state their developments and drawbacks. Besides, a brand new solution gaining inspiration from our brain will also be discussed and introduce the fundamental electronic component to accomplish the goal. The principle of how a nerve fires will also be illustrated and based on this catch a glimpse of the neural network. Besides, The principle of memristors and with the support of crossbars, how could they be used in manufacturing AI Chips will be explained. Finally, the latest research in this field. The advantages and challenges will also be involved. Finally, a comparison of these solutions will be proposed.

Keywords: GPU, FPGA, ASIC, Brain-like Chips, Memristors, Crossbar.

1. Introduction

With the development of generative AI or autonomous driving, etc, the requirement of computing power to train AI is getting greater and greater, while the growth of computing power under Moore's Law is getting slower and slower. And because of the traditional Von Neumann architecture, more and more meaningless energy is consumed in computation, and consequently, the energy consumption for cooling computing devices also rises. This review will focus on organizing and comparing the conventional or future promising resolutions about acceleration in the training AI field. GPU's (Graphic Processing Unit) development of constant evolution in transistors count and die size, and the scalable GPU architecture will be described. The high parallel structure and computing in memory will further illustrate the development of GPU for acceleration. Besides, other AI chips like FPGA and ASIC will also be involved. However, the energy efficiency ratios of these solutions are at low levels and the giant energy consumption is predicted that 50% of electricity will be used to compute and cool down the equipment in 2030. As a result, memristors A brand-new solution for the problem has been proposed, as the previous solution has not been completely transformed. Neuromorphic systems, which are composed of neurons with synapses are able to change dynamically, serve as the basic building blocks of this new solution. The term "neuromorphic" was introduced by Mead in 1990, in which he highlighted the advantages of biologically-inspired information processing over conventional digital computing in cases where input data are ill-conditioned. The ultimate goal of this brain-inspired approach is to use biological guidelines of neural computation to model large-scale biological neural systems at high speed,

Therefore, Memristor, a new electrical component, has been proposed for the neural network, which can store and retrieve information by modifying the resistance of its conducting channel in a persistent manner based on the history of current that has flowed through it, making it capable of handling convolutional algorithms. The successfully developed Memristor chips by Qinghua University have proved the huge competitive edge of the Memristor chip, which is only 3% of the energy of advanced application-specific integrated circuit (ASIC) systems will consume [1]. However, memristor still faces several challenges that need to be addressed, particularly the need to improve its accuracy; otherwise, the final results of several layers of calculation could turn out to be completely different.

2. Method

Under the continuation of the traditional computer structure, the computing power of acceleration hardware is still mainly represented by three types of chips, namely GPU, FPGA, and ASIC. GPU is currently the mainstream, for instance, from the large Language model, Chat GPT, is trained and runs on thousands of NVIDIA GPUs, and runs generative AI services used by more than 100 million people.

2.1. GPU

Despite the limitations of Moore's Law, unlike CPUs, GPUs are still growing in performance by 1.5 times per year [2]. Originally, the GPU was only used for 3D graphics which the tasks are nature parallel applications. Vertices, polygons, and pixels that are handled are all independent of data and can be processed concurrently. Therefore, the number of processor cores has an essentially linear relationship with processing speed. In 2006, to achieve the current state of near-photorealistic 3D graphics, a major revolution in graphics architecture took place. It was the Compute Unified Device Architecture or CUDA, a software framework created by NVIDIA that facilitates the easy programming of GPUs for various applications. After that, GPUs have been applied in a much wider range and dramatically increased their numbers of cores, especially from the introduction of the unifiedshader [2]. As NVIDIA, AMD, and other companies continue to advance their support for GPU massively parallel architecture, GPUs for general computing (such as GPGPU, GENERAL PURPOSE GPU) have become an important means of accelerating parallel applications. NVIDIA has released a series of increasingly parallel-structured GPUs, from the A100 in 2021 to the latest H200 with 80 billion transistors. However, based on what we have mentioned previously, its design was originally intended to cope with massively parallel computation in image processing. As a result, some challenges, for instance, power and energy, memory bandwidth and energy and programmability are waiting to be confronted [3]. Besides, compared to FPGA, the hardware structure cannot be flexibly configured and the energy efficiency of running algorithms is also lower.

2.2. FPGA

Field Programmable Gate Arrays(FPGA) is a further development product based on programmable devices such as PAL, CPLD and etc. The advantages it offers include programmability, great performance, and low power consumption. Compared to CPUs and GPUs, it provides significant performance and energy savings, but at the expense of increased user strain. By burning these links into the FPGA configuration file, users can specify how these gates are connected to the memories. This type of burning is an ongoing process. The FPGA can be set up to function as an MCU microcontroller, for instance. The same FPGA can be configured as an audio codec by the user by editing the configuration file after use. As a result, it addresses both the drawbacks of the original programmable device's restricted number of gate circuits as well as the lack of flexibility of customized circuits. The ability of FPGA to simultaneously process tasks and data allows for more noticeable efficiency gains while handling particular applications. Multiple clock cycles may be needed by a general-purpose CPU for a particular operation; while an FPGA can reorganize the circuit through programming and directly generate a dedicated circuit, consuming only a small amount or even one clock cycle to complete the operation. Additionally, a lot of hardware control in underlying operation technologies that are challenging to implement using ASICs or general-purpose processors can be simply implemented using FPGA because to its versatility. This feature gives the algorithm greater room to be optimized and implemented functionally. In addition, FPGA's (photolithography mask manufacture cost) one-time cost is substantially less than ASIC's. The reconfigurable features of FPGA chips can be utilized when the need for chips has not yet reached a significant size, the deep learning algorithm has not yet become stable, and ongoing iterative improvements are required. One of the greatest solutions is to use artificial intelligence chips that are semi-customized. Furthermore, FPGA has intrinsic architectural advantages in terms of power usage. To participate in the operation,

the execution unit (such as the CPU core) in the typical Feng structure needs an instruction memory, a decoder, arithmetic units for different instructions, and branch jump processing logic. However, each logic unit of FPGA functions is already determined during reprogramming (that is, burning in), and no instructions or shared memory are required, which can greatly reduce the power consumption per unit execution and improve the overall energy consumption ratio [4]. Nonetheless, there are additionally numerous other restrictions as well: Firstly, the processing power of the basic unit is constrained. The FPGA has a lot more very fine-grained fundamental units, but its computational capacity is far less than that of the ALU modules of the CPU and GPU [5]. Secondly, the percentage of computational resources is rather minimal [5].

2.3. ASIC

Application Specific Integrated Circuit is an integrated circuit designed for specialized purposes. It refers to the integrated circuits designed and manufactured in response to specific user requirements and the needs of specific electronic systems. ASIC is characterized by specific user-oriented needs, ASIC is divided into fully customized and semi-customized, the highlight of which is dedicated, and tailored so the implementation of faster. In a word, it is a chip that cannot be bought on the market. Apple's A-series processor is a typical ASIC. The performance gains of ASICs are very noticeable. For deep learning-related model training and inference applications, NVIDIA's Tesla V100, for instance, can offer tensor computing of up to 125 teraflops, with data processing speed 12 times faster than its GPU series, which was introduced in 2014 [5]. and another example from Intel, a leading research ASIC is the neuromorphic research chip Loihi, which achieves a 10,000x improvement in efficiency and 1,000x improvement in speed relative to CPUs for specialized applications [6]. Additionally, because of the need to introduce cognitive interaction capabilities in certain areas like driverless cars, smart appliances and so on, and real-time and data privacy reasons, by using the use of ASIC design approach, the AI chips may be fully customized, allowing the deep learning algorithms' performance, power consumption, and area metrics to be optimized. Nevertheless, the specific use of ASIC could also be a disadvantage for itself. ASIC requires a longer development cycle, the risk is greater, and once there is a problem, the whole piece of waste.

3. Result

Brain-like chips abandon the traditional von Neumann structure and instead simulate neuromorphic architecture design. Hence it is crucial to understand how nerves work and relate.

3.1. The process of a nerve firing

Neurons transmit messages through a precise sequence of events. It has been illustrated in Figure 1. Initially, a neuron's resting state has a slightly negative charge due to ion distribution across its selectively permeable membrane. Upon stimulation, neurotransmitters from neuron A's terminal buttons activate neuron B by fitting into its dendrite receptors. If the threshold is met, neuron B's membrane allows positive ions in, creating an action potential that propagates rapidly. The electrical impulse triggers neurotransmitter release from neuron B, potentially starting the process anew in subsequent cells. This all-or-none firing ensures consistent neural responses, highlighting the neuron's binary response mechanism.

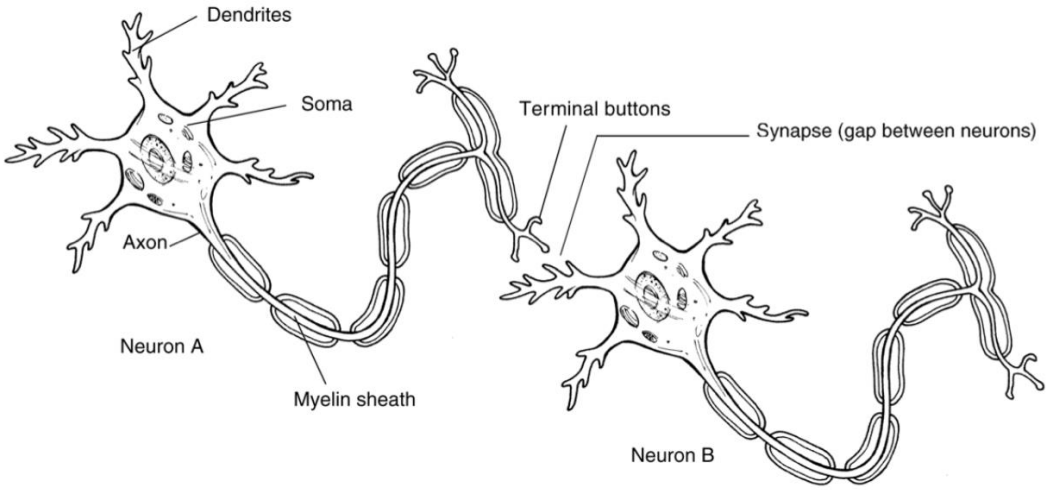


Figure 1. Structure and Connection between Two Neurons

3.2. Neural network

There are numerous layers, and within each layer are numerous neurons. First-layer neurons function as sensors, taking in information or data. Additionally, neurons in the last layer function as output reactors. Every neuron has a specular weight that is connected to the neuron in the layer below. It implies that every line has a distinct weight. Therefore, by applying a specular function, the neurons in the latter layer can be retrieved from the neurons in the former layer.

3.3. Memristor

Based on these, the Neuromorphic architectures have been elicited. These hardware devices seek to base their operation on the principles of how neurons work. Their objective is to overcome distributed computing issues by utilizing physiologically inspired ideas including inhibition, short or long potentiation, activation thresholds, and weighted connections. The promise of intrinsic low power consumption and fault-tolerant functioning realized directly in hardware is presented by neuromorphic systems. These systems are intended for use in distributed and embedded computing tasks, where the massive scaling of current architectures does not provide a long-term solution [7]. In 1970, Leon O. Chua had proposed a missing electronic component called Memristor based on the fourth basic circuit element [8].

Thirty-seven years later, in 2008, HP scientists produced this kind of electrical component. Its structure is very simple. A layer of semiconductor film is added between two electrodes. It consists of two layers, one layer is doped with cations and the other is not. Based on semiconductor physics, the holes formed by cations give the semiconductor good conductivity, while the other layer is not doped so the conductivity is very poor. At the time, a voltage is applied to both ends of the electrode. Under the action of the voltage, the cations move in the undoped direction, so the doped area becomes big and the resistance of the entire film becomes smaller. If a reverse voltage is applied, the pressure bar will move back. Therefore, for forward current, the resistance becomes smaller, and for reverse current, the resistance becomes larger. Once the current stops, the cations remain at the current position, so the resistance value also stops at the current position [9] which makes them suitable for non-volatile memory applications. Due to their ability to mimic synaptic plasticity and perform complex computations efficiently and low power consumption, memristors have the potential to revolutionize the field of electronics by enabling new types of computing architectures and memory technologies that are more energy-efficient and capable of performing brain-like computations. With these characteristics of memristors, a structure called cross bar fulfills the function and potential of memristors for neuromorphic architectures. A crossbar, in the context of electronics and computing, refers to a grid-like structure where horizontal wires intersect with vertical wires, forming a matrix

of junctions. These junctions can be used for various purposes such as connecting different components, implementing memory arrays, or performing computations in neuromorphic systems. An example is provided below in Figure 2.

$$\begin{bmatrix} 0.1 & 0.2 & 0.3 \\ 0.4 & 0.5 & 0.6 \\ 0.7 & 0.8 & 0.9 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 3.2 \\ 5 \end{bmatrix}$$

Figure 2. Matrix Calculation for Training AI

Regarding this matrix calculation, a simple memristor array has 3 horizontal wires and 3 vertical wires, with a total of nine intersection points. Just apply a voltage to each of the three vertical wires representing the value of the vector, and finally measure the total current on the horizontal wires to complete the calculation of these vectors. The crossbar architecture allows for parallel processing and efficient communication between neurons, mimicking the connectivity and functionality of biological neural networks. By leveraging the dense and interconnected nature of crossbar arrays, researchers can design energy-efficient and high-performance neuromorphic systems for tasks such as pattern recognition, machine learning, and cognitive computing [10]. As a result, lots of research teams try to use this in the calculation of machine learning. In 2015, IBM used a 12*12 array to realize three-letter recognition [11]. In 2017, Wu Huaqiang's team at Tsinghua University used a 128*8 array to realize face recognition [12]. However, because memristors use physical laws to perform simulation calculations, the calculation accuracy is a big test. Small changes in current and voltage, and calculations through hundreds or thousands of layers of neural networks will make the results very different. The latest in 2023, the team of Wu Huaqiang of Tsinghua University made optimizations from several levels. The first optimization is the accuracy of the device itself. They designed a mille-feuille structure. 2. Added a current limiting circuit. 3. Introduced a set of error models, striving to eliminate errors in the calculation of each layer. Based on this series of engineering optimizations, they created a complete storage and computing integrated chip, which contains two memristor arrays, drivers for calculation and programming, digital-to-analog converters, and input and output buffers. They installed this chip on a car, allowing the car to adapt to new scenarios simply through offline on-chip learning. The most important thing is that as a 130-nanometer chip, it has the same recognition rate as a traditional 28-nanometer chip [1], and the inference speed is 20 times faster [1], and the inference calculation performance is 1,000 times faster [1]. Nonetheless, AI chips of memristors are still in research progress.

4. Conclusion

With the rapid development of artificial intelligence in recent years, the required AI arithmetic power is getting higher and higher. To cope with this arithmetic crisis, GPUs integrate more transistors through a highly parallel structure to improve arithmetic, which is currently the most mature and stable solution. FPGAs and ASICs are more optimized for the underlying circuit design to get rid of the traditional computer structure, and customized functions are used to achieve increased computational speed and solve the arithmetic crisis. But all of this does not solve the real problem, von Neumann architecture under the arithmetic wall and storage wall of the boundary has not been broken, and with the growing demand for AI arithmetic, the energy consumed by the power will also rise. Memristor AI chips may break this bottleneck, but it still faces many challenges and is still in the research stage, which needs to be continuously optimized and improved to reach the commercial stage. Therefore, the current development should vigorously develop brain-like chips based on amnesia, and really change the chip structure from the underlying design instead of continuing to sew on the traditional structure.

References

[1] Zhang W , Yao P., Gao B., Liu Q., Wu D., Zhang Q., Li Y., Qin Q., Li J., Zhu Z., Cai Y., Wu D., Tang J., Qian H., Wang Y., Wu H., "Edge learning using a fully integrated neuro-inspired memristor chip," Science 381, 6663 (2023).

[2] Baji, T. "Evolution of the GPU Device Widely Used in AI and Massive Parallel Processing," IEEE Electron Devices Technology and Manufacturing Conference Proceedings of Technical Papers 978,1,5386,3712 (2018).

[3] Stephen W. K., William J. D, Brucek K, Michael G, David, "G GPUs and the future of Parallel computing," IEEE Computer Society, 0272,1732 (2011).

[4] Sorna M V, "AI Chips: New Semiconductor Era," International Journal of Advanced Research in Science, Engineering and Technology 7, 8 (2020).

[5] Li B., Gu J., Jiang W., "Artificial Intelligence (AI) chip technology review," International Conference on Machine Learning Big Data and Business Intelligence (MLBDBI). 978,1,7281, 5094 (2019).

[6] Samuel K. M., "Intel's Neuromorphic System Hits 8 million Neurons, 100 Million Coming by 2020," (2019).

[7] Robert A. N., Richard M. V., and Sean E. S., "A Mini Review of Neuromorphic Architectures and Implementations" IEEE TRANSACTIONS ON ELECTRON DEVICES," 63, 10 (2016).

[8] Leon O. C., "Memristor-The Missing Circuit Element," IEEE Transactions on circuit theory, 18, 5, (1971).

[9] Dmitri B. S, Gregory S. S, Duncan R. S& R. S W, "The missing memristor found," Nature, 453-1, (2008).

[10] Xia Q, J. J. Y., "Memristive crossbar arrays for brain-inspired computing," Nature Materials, 18, 309, 323(2019).

[11] Mei P, F., Mei B. D., Han G. C., An K. K. L, and Ding B. S., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors Letter," Nature, 10(1038), 14441 (2018).

[12] Yao P., Wu H., Gao B., Sukru B. E., Huang X., Zhang W., Zhang Q., Deng N., Shi L., Wong H. S. P. and Qian H., "Face classification using electronic synapses," Nature Communication 8, 15199 (2017).