Analysis of the Configuration for State-of-art AI Chips and Applications

Yangrui Lin*

Department of optical engineering, Sichuan University, Chengdu, China

*Corresponding author: 2021141450020@stu.scu.edu.cn

Abstract. As a matter of fact, the artificial intelligence chips provide core hardware computing support for the rapid development and application of artificial intelligence technology in recent years. In reality, they are the key driving force for the efficient, rapid application and energy-saving operation of artificial intelligence technology and are also the strategic commanding heights of the era of artificial intelligence. The purpose of this study is to analyze and summarize the architecture and practical application of the most advanced artificial intelligence chips. On this basis, this study will first analyze the development process of AI chips. At the same time, the classification, main architecture and working principle of AI chips are introduced. In addition, the study explains the application of AI chips in the existing mainstream market. Finally, this study will analyze the limitations of its research as well as look forward to the future. Overall, these results shed light on guiding further exploration of AI chips.

Keywords: Artificial intelligence; AI chips; deep learning.

1. Introduction

Modern electronic products and equipment have brought a lot of convenience to people's lives and improved their quality of life in entertainment, communication, security, scientific research, etc. Among them, the development of modern microelectronics technology has brought about changes in people's daily life, work and interaction. In the past few decades, thanks to the continuous evolution of Moore's Law, the feature size of the devices on the chip has been shrinking, resulting in the continuous improvement of computing power, bringing faster computing, lower prices and lower energy consumption of electronic products. According to Gordon Moore's theory, the number of transistors in the chip will increase by 100% every two years. After the optimization of David House, thanks to more fast transistors, the performance growth rate of the chip will be 1.3 times the original. But today, 50 years after Gordon Moore's theory was successfully put forward, silicon COMS technology is approaching its basic physical limits, which makes the continuation of Moore's Law very challenging. On this basis, the emergence of artificial intelligence provides the possibility to solve the problem that the performance gain of electronic products and equipment is too dependent on the reduction of device feature size [1]. The deep learning algorithms mainly used in artificial intelligence require a large number of matrix multiplication operations, which has high requirements for large-scale parallel computing power and cannot be met by CPU and traditional computing architecture. Therefore, it is necessary to customize a special chip for artificial intelligence applications, which is called AI chip.

Carver Mead began to study artificial intelligence chips from the California Institute of Technology and neuron simulation systems in the 1980s [2]. He used analog circuits to imitate the structure of the biological nervous system, which is commonly known as the early exploration period of neural networks. After more than 40 years of efforts, various artificial intelligence chips have come out, mainly including graphics processing units (GPUs), field programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), etc. The following are the important time nodes in the exploration and development of AI chips [3-8].

In the 1980s and 1990s (the early exploration of neural networks), researchers began to use computers to simulate neural networks, although the hardware at that time was far from meeting today's needs. In the early 2000s (the rise of GPUs), people began to use graphics processing units

(GPUs) to accelerate neural network computing. Companies such as NVIDIA have promoted the development of this field. The parallel processing power of the GPU is suitable for the matrix operations required for deep learning. In the early 2010s (the revival of deep learning), with the improvement of algorithms and the availability of a large amount of data, deep learning began to rise. This leads to the demand for higher computing power and specialized hardware. In 2016 (Google's TPU period), Google launched the Tensor Processing Unit (TPU), an accelerator specially designed for its TensorFlow machine learning framework. TPU marks an important milestone for AI-specific hardware. Since 2017 (the period of AI chip diversification), many companies have begun to invest in the development of AI-specific hardware. From cloud computing centers to edge devices, from self-driving cars to medical diagnosis, various application needs have promoted the diversification of AI chip design. Modern and future development (the period of evolution to edge computing) With the development of IoT and edge computing, the demand for low-power and high-performance AI chips is also increasing. Many companies are developing AI chips that can be directly embedded in devices.

Table 1. Advantages and disadvantages of the AI chips

Technical Architecture	Advantage	Disadvantage
GPU	It is a general-use manager, which is more flexible than CPU programming and has stronger parallel computing power. And have a mature development environment	The price and power consumption are higher.
TPU	It has high performance and energy efficiency, and is optimized for tensor operations, and can be closely integrated with some machine learning frameworks.	It is less flexible, may have limited applicability, and may have higher costs and reliance on specific frameworks.
NPU	It is designed for neural networks, suitable for edge computing and has parallel processing power.	It may not perform well for non-AI tasks or special AI algorithms and is more difficult to program.
FPGA	It is semi-determined, which can program and configure the core hardware layer, with lower power consumption than the GPU.	The hardware programming language is difficult to learn. Compared with ASIC, it has a certain degree of electronic tube redundancy, and the power consumption and cost have further compression space.
ASIC	It can be prepared for special tasks, so as to achieve the effect of low cost, low power consumption and high performance.	Poor chip versatility, programmable, high difficulty in architecture design, and large investment in the early stage

2. The Classification of AI Chips

The following table 1 shows some advantages and disadvantages of the AI chips [3]. The classify AI chips by architecture can be described as follows:

GPU: graphics processing unit, with highly parallel characteristics, suitable for large-scale matrix operations, used for many deep learning tasks.

TPU: Tensor processing unit, a hardware optimized for tensor operation launched by Google.

NPU: Neural processing unit, a processor specially designed for neural network computing.

FPGA: Field programmable gate array, which can be reprogrammed to adapt to specific AI algorithms.

ASIC: A chip that applies a specific integrated circuit and is customized for a specific application or algorithm.

The main structure of the AI chips are given as follows:

Core computing units: Including multiple arithmetic logic units (ALU) and floating-point units (FPUs) for performing a large number of parallel calculations.

Memory hierarchy: This includes caches, register files and other storage units to quickly access data.

Data bus and interconnection structure: used to transfer data between various parts of the chip.

Neural network processing unit: Some AI chips will include a specific neural network processing unit (NPU) to accelerate the calculation of neural networks.

Control unit: used to guide and coordinate the operation of various parts on the chip.

Accelerator: In order to improve the performance of specific tasks, some AI chips will also include some special accelerators, such as tensor processing units (TPUs).

Input/output interface: including the necessary interface to communicate with other system components, such as PCIe interface, memory interface, etc.

3. AI Chips' Working Principle

The working principle of AI chip is mainly based on the concept of neural network and parallel computing. The following are the basic principles of how these chips work:

Neural network simulation: AI chip is designed to simulate the neural network structure of the human brain. Through a large number of processing units (similar to neurons) and connections (similar to synapses), they can perform large-scale parallel calculations [4].

Parallel processing: Unlike traditional CPUs, AI chips usually include a large number of processing units that can run in parallel. This allows them to handle multiple tasks at the same time, especially when performing a large number of the same mathematical operations, such as matrix multiplication required in deep learning.

Specific hardware acceleration: Many AI chips include hardware accelerators optimized for specific types of computing (such as tensor operations). These accelerators can greatly improve the efficiency of running specific algorithms.

Storage optimization: Because AI computing usually involves a large amount of data, AI chips usually include a specially designed storage structure to minimize data access latency and ensure that data is always available in the processing unit.

Programmability and flexibility: Many AI chips also include programmable elements that allow them to adapt to changing AI algorithms and model requirements.

4. The Application of AI Chips

The core of the artificial intelligence chip is the implementation of the neural network algorithm. Deep Neural Network (DNN) has made major breakthroughs in many artificial intelligence fields such as natural language processing, machine vision, speech recognition, medical image analysis, etc. [5, 6]. This part will focus on the mainstream application of artificial intelligence chips.

IDC (Internet Data Center) is a specialized facility for enterprises or organizations to store, manage and distribute large amounts of data. They are the core components of building the infrastructure of cloud services, websites, mobile applications, etc. IDC realizes centralized storage and network connection, thus providing a safe and reliable data storage environment, preventing data leakage, damage or illegal access, and ensuring the speed and stability of data transmission [3]. At present, the mainstream AI chips used in data center (IDC) are: NVIDIA GPU series, Google TPU (Tensor Processing Unit), Nervana Neural Network Proces Sor (NNP), Intel Xeon Scalable Processors, AMD Radeon Instinct, Graphcore IPU (Intelligence Processing Unit), Cerebras Systems' Wafer-Scale Engine (WSE), Habana Labs (acquired by Intel).

Autonomous driving technology has a high demand for AI chips because it requires processing a large amount of real-time data and making quick and accurate decisions. This mainly requires high-performance computing on AI chips to process data from multiple sensors (such as cameras, radars, lidar, etc.) in real time [7]. At the same time, in order to use the vehicle for a long time, AI chips are needed to achieve efficient energy utilization with low energy consumption. Since any calculation error can lead to serious security problems, the chip must be very reliable. AI chips also need to have multitasking capabilities to handle multiple tasks at the same time, such as object detection, path planning, control algorithms, etc. Moreover, in the automatic driving scenario, the delay is unacceptable, so the chip needs to be able to make decisions at the millisecond level. Finally, the chip needs to meet the vehicle specifications, so it is difficult to design [8]. At present, the mainstream AI chips used for autonomous driving are: NVIDIA DRIVE, Tesla FSD Chip, Mobileye (belonging to Intel), Qualcomm Snapdragon Ride

Because many edge devices and IoT applications require real-time response, such as security monitoring and medical monitoring, edge computing and the Internet of Things (IoT) require their matching AI chips to be real-time, so that data can be processed quickly on the device. At the same time, because edge devices and IoT usually rely on battery power, low-power AI chips are necessary. And many IoT devices are small in size, so they need a small and efficient chip design. In order to reduce the risk of being intercepted, data needs to be processed on the device rather than transmitted to the cloud. Sometimes IoT devices may not have a continuous network connection, so AI chips should be able to process data locally without a network. The existing mainstream AI chips for edge computing and the Internet of Things are: NVIDIA Jetson series, Google Edge TPU: this is Google, Intel Movidius Myriad, ARM ML processor, Qualco Mm Snapdragon

In consumer electronic products, AI chips need to process user interaction data in real time to realize popular functions such as voice assistant, facial recognition and gesture control, and provide a smooth user experience. At the same time, the AI chip can enhance the performance of the camera, such as automatic scene detection, portrait mode and super-resolution technology. Consumer electronic products such as smartphones and watches often rely on battery power, so low-power AI chips are needed to maintain longer usage time. The existing mainstream AI chips for consumer electronics are: Apple Bionic and Neural Engine, Qualcomm Snapdragon, Huawei Kirin and NPU (Neural Process) Ing Unit), Samsung Exynos with AI capabilities: Samsung Exynos, MediaTek Helio with APU (AI Processing Unit), NVIDIA Tegra, Google Tensor

Medical images (e.g., MRI, CT, X-ray and ultrasound) usually require high-precision analysis. AI can help identify abnormalities, disease markers and lesions, and provide doctors with more accurate diagnostic advice. In many cases, such as surgery or monitoring patients in the intensive care unit, the AI system needs to respond in real time. With the rise of gene sequencing and personalized medicine, a large amount of data needs to be processed and analyzed, and AI chips can accelerate this process. Many portable medical devices and household medical devices require small, low-power AI chips. The medical field has very high requirements for the stability and accuracy of equipment, because it is directly related to the safety of patients. Existing mainstream AI chips for medical diagnosis: NVIDIA GPUs, Google Cloud TPUs, Intel Movidius and Xeon Scalable processors, FPGAs (Fi Eld Programmable Gate Arrays)

The monitoring system needs to identify and analyze objects, faces, vehicles, behaviors, etc. In the image in real time, as well as abnormal detection. With the popularity of 4K and higher-resolution cameras, AI chips need to be able to process a large amount of image data. Due to bandwidth and data privacy considerations, many analyses need to be carried out on cameras or local servers, not in the remote cloud. Many cameras are battery-powered or work in environments with limited bandwidth, so they need energy-efficient AI chips. The monitoring system may include multiple cameras, which require AI chips to support the processing of multiple video streams. Existing mainstream AI chips for intelligent video surveillance: NVIDIA Jetson series, Intel Movidius Myriad, Huawei Ascend, Qualcomm Vision Intelligence P Latform, Ambarella SoCs, Hikvision and Dahua custom chips

Many smart home devices use voice assistants, such as Amazon's Alexa, Google Assistant and Apple's HomePod, so they need AI chips for real-time voice recognition. AI chips recognize objects and faces, because smart cameras and doorbells need to identify residents, pets or visitors in the home. AI chips need to support end-to-end encryption and other security features to protect users' privacy. In order to respond to speed and data privacy, many data processing needs to be carried out on the device, not sent to the cloud, so it needs to have the function of edge computing. Because many smart home devices are powered by batteries, they need efficient and low-power AI chips. The AI chip needs to have the ability to understand the scene. For example, the intelligent lighting system needs to understand the use mode of the room and automatically adjust the brightness. Existing mainstream AI chips for smart home: Amazon's Custom Silicon, Google Edge TPU, Apple A series and M series, Qualcomm QCS series, NXP i.MX 8M series, M ediaTek AIoT platform, Intel Movidius Myriad

5. Limitations and Prospects

AI chips, especially those designed for deep learning or other AI applications, have many advantages over traditional computing chips, but there are also some shortcomings: the research and development, manufacturing and testing of dedicated AI chips require high costs, which may increase the price of end products. At the same time, although many AI chips are optimized for low power consumption, high-performance AI computing may still lead to high power consumption, and a large amount of computing will generate a large amount of heat, which requires more effective thermal management. Because dedicated AI chips are usually optimized for specific AI tasks, some AI chips have challenges in compatibility and versatility, and the new hardware platform may lack complete software tools and framework support, so they may not be suitable for other conventional computing tasks.

There are two main development trends of AI chips in the future. The neural morphic chip refers to the subversion of the classic von Neumann computing architecture, using electronic technology to simulate the operating rules of the proven biological brain, so as to build a chip similar to the biological brain [9]. The neuromorphic chip imitates the parallel processing of the brain. The neuromorphic chip can process a large amount of information at the same time and improve computing efficiency, thus breaking through the bottleneck of Memory Wall. The chip imitates the redundancy mechanism of the brain, has a decentralized nuclear architecture, and has a strong fine-grained interconnection ability. Even if some neurons fail, the neuromorphological system can usually continue to work. Unlike traditional chips, some neuromorphological systems can learn and adapt without a large amount of training data, so they have better learning ability.

The reconfigurable computing chip is also called software-defined chip [10]. It is mainly aimed at the problem that the efficiency and flexibility of AI chips are difficult to balance and the computing accuracy required by different tasks. The chip developed and designed can dynamically configure its internal structure at the hardware level to meet the needs of specific applications. The design idea of the reconfigurable computing chip is that the software and hardware can be programmed, so that flexibility and ultra-high energy efficiency ratio can be achieved. At the same time, it can be reprogrammed to adapt to technological development or meet new application needs, thus extending the service life of the product. Developers can also use reconfigurable computing chips for rapid prototype verification of hardware design, and then decide whether to convert it into customized hardware such as ASIC (special integrated circuit).

6. Conclusion

To sum up, at present, the AI chip industry has just started and is booming, and a large number of related projects are being implemented and commercialized. This paper investigates and analyzes the meaning, classification, development process, technical principles, product application, shortcomings and prospects of AI chips. This research is aim to help readers better understand this industry with

huge industrial value and strategic position. It is also hoped to help engineers and scholars who are interested in this field to expand their horizons and grasp the macro layout.

References

[1] Ren Y, Pan J, Liu J, et al. Research progress of artificial intelligence chips. Micro-nano electronics and intelligent manufacturing, 2019, 1(02):20-34.

[2] Mead C. Neuromorphic electronic systems. Proceedings of the IEEE, 1990, 78(10): 1629-1636

[3] An B. Analysis of the development status and prospects of AI chips. Micro-nano electronics and intelligent manufacturing, 2020, 2(01): 91-94.

[4] China Institute of Electronic Technology Standardization. Artificial Intelligence Standardization White Paper (2018 Edition). Retrieved from: http://www.cesi.ac.cn/images/editor/20180124/20180124135528742.pdf.

[5] Lecun Y, Bengio Y, Hinton G, et al. Deep learning. Nature, 2015, 521(7553): 436-444.

[6] Keller J, Perez O. Improving MCTS and neural net- work communication in computer go. Worcester Poly- technic Institute, 2016.

[7] Yin S Y. Overview of artificial intelligence chip. Micro/nano Electronics and Intelligent Manufacturing, 2019, 2: 7-11

[8] Tan H H, Yu K. The challenge and prospect of edge AI chip. Artificial Intelligence, 2018, 2: 113-121.

[9] Zhao Z P. New progress of nanoelectronics and neu- romorphic chips. Micronanoelectronic Technology, 2018, 55(1): 1-5.

[10] Tsinghua University. Artificial Intelligence Chip Technology White Paper. Retrieved from: https://www.tsinghua.edu.cn/publish/thunews/9659/2018/20181217102627644168087/20181217102627644168087_.html.