Analysis of the Principal, Fabrication and Application of AI Chips

Tiancheng Pu*

Department of Electrical Engineering, Case Western Reserve University, Cleveland, Ohio, USA

*Corresponding author: txp315@case.edu

Abstract. As a matter of fact, artificial intelligence (AI) technology has developed by leaps and bounds with the rapid development computation ability and improvements of machine learning scenarios (e.g., deep learning neural network). On this basis, its applications have appeared in various fields in recent years. In order to support complex artificial intelligence computing, AI chips have gradually appeared in people's vision, as well as have played an essential role in the future development of life. With this in mind, based on the development of AI chips in recent years, this paper discusses the definition and classification, principles and design, preparation, and application of AI chips. According to the analysis, the current limitations in terms of the state-of-art facilities will be demonstrated. At the same time, this study predicts AI chips' future development prospects, so that readers can have a more comprehensive understanding of AI chips and the field of artificial intelligence.

Keywords: AI; chips; chips fabrication.

1. Introduction

In the field of semiconductors, the first ideas for integrated circuits, or chips, were conceived when it was discovered that transistors had faster transmission speeds and even lower heat loss than vacuum tubes. In 1958, Jack Kilby designed the first integrated circuit by connecting several devices on a silicon substrate [1]. Based on the development of chip technology, chips can be designed to be smaller and have faster transmission speeds. Intel co-founder Gordon Moore found that the number of transistors in a fixed area of an integrated circuit grows exponentially over time, which is called Moore's law. However, nowadays, as transistor size shrinks to a certain level, transistor leakage occurs, and this bottleneck makes it impossible for chip development to follow Moore's law [2].

Meanwhile, the emergence of the field of artificial intelligence seems to open up a new path for chip development. With the rapid growth of the video and gaming industries, GPUs are breaking through. Researchers found that the logic of parallel computing paired well with AI's algorithms for analyzing big data, then they used GPUs in AI computing to replace CPUs for greater efficiency. After 2015, the emergence of more AI industries greatly increased the demand for the chip's computing power, so more AI chips dedicated to AI computing were designed [3]. At the same time, the large number of markets brought by commercialization of AI chips in recent years requires chips to achieve computation with low power consumption and low cost, which also poses a considerable challenge for AI chip manufacturing [4]. In the following part, this paper will elaborate on the Basic Description, Principle and design, Fabrication, Application, and Future outlooks of AI chips.

2. Basic Descriptions of AI chips

In the case of the slow development of semiconductor integrated circuits, the main means of chip improvement has changed from manufacturing to design. At the same time, the large number of applications of AI technology also put forward the need for a specific chip that is highly parallel and predictable. In this case, the researchers laid out and optimized the transistors on the chip by relying on the highly parallelized, specialized computations required by the AI algorithm, and stored the entire AI algorithm on the chip specifically for processing high-speed AI commands with minimal power consumption [4]. Compared with traditional CPUs, AI chips perform calculations in parallel

with much higher efficiency and are more diverse, and their design can be changed by different application requirements [5].

AI chips are mainly divided into three types: graphics processing units (GPU), field-programmable gate arrays (FPGA), and application-specific integrated circuits (ASIC). Different AI chips will be used to perform training or inference according to their different task requirements. Graphics processing units were originally designed to process images, which can be simplified by parallel computation. With the rise of AI, GPUs are increasingly being used to train AI systems. Although GPUs are AI chips, they are still essentially designed for general-purpose computing. Compared to GPUs, field-programmable gate arrays and application-specific integrated circuits are used more for inference tasks. FPGAs come with logic blocks so that they can still be reconfigured after production to better accommodate different algorithms. ASICs use fixed line connections and are specifically designed for specific algorithms. Because of the specificity of ASIC, its efficiency is generally greater than FPGA. However in the long run, with the development of AI chips, FPGAs, which are more applicable chips, will become mainstream [5].

In terms of hierarchy, AI chips are also divided into three types. The most powerful is the server-level AI chip, which is mainly used for high-end applications in data centers, and its volume is also the largest. The second is the most well-known AI chip for laptops. The Mobile chip, which mainly used for inference, is the least powerful and smallest of the three, and it is usually integrated with the CPU in a chip so that it can be used in mobile devices [4].

3. Principle and Design

The principle and design of AI chip implementation cannot be separated from Deep neural networks (DNNs), which is the AI algorithm responsible for the vast majority of computation today. DNN uses a machine learning model called supervised learning, which includes two operation steps: training and inference. For the training process, a large amount of training data is imported for the AI to analyze in order to build a new algorithm framework for the neural network model [6]. This step requires implementing the same calculation millions of times, which has a high demand on the hardware equipment. For inference process, AI uses the constructed AI algorithm framework to analyze the new data and classify the data consistent with the previous training database [5]. At the same time, some specific technologies underpin AI Computing principles, such as Parallel Computing, Low-Precision computing, and Domain-Specific Languages.

Parallel computing is one of the biggest improvements in AI chips compared to traditional chips, because DNNS are identical and run simultaneously, which means that the operational data does not affect each other. To increase the speed, a large number of multiplication-accumulate circuits (MAC) are applied to each AI chip design to enable the chip to perform matrix multiplication more efficiently through parallel computation. At the same time, connecting each AI chip in parallel can further increase its operating speed. Parallel Computing can be subdivided into Data parallelism and Model parallelism. Data Computing refers to dividing a database into many small branches and assigning these small branches to each executive element on a single chip or multiple parallel AI chips for computation. Model parallelism is the division of the model into different parts so that data can be processed in different execution units of a single chip or multiple AI chips in parallel. The former is more widely used. Dividing the database into many branches for computation in a limited range can achieve equivalent model accuracy without increasing the amount of computation, but there is an upper limit to this technique: when the number of branches reaches a certain value, more computation becomes inevitable. This limitation makes parallel computing not the best direction for developing AI chip technology [5].

Higher bit data types (such as 64-bit, 32-bit) can produce more possible values, which allows the data type to represent a larger range of data or to represent data with higher precision within a limited range. However, since DNN after training is not affected by errors, and some parameters in DNN can be determined within a small range of values before operation, most AI operations can achieve nearly

the same result using only low data types, such as 8-bit or 16-bit, in both the training and inference steps. This loss of precision in exchange for higher computing speed and efficiency is also important in AI computing principles, and it also reduces the number of transistors and energy required to perform the same operation [5].

Domain-Specific Languages (DSLs) have significantly helped improve the computational efficiency of specialized chips such as AI chips. Although common computer programming languages such as Python can also perform calculations, their logic is more inclined to be easy for people to understand, and the efficiency cannot meet the high-intensity operations of AI chips. In contrast, DSLs are designed for efficient programming and computing on a specific chip. For example, Google's DSL for AI chips, called TensorFlow, runs code more efficiently than normal programming languages [5, 6]. A sketch of AI chip is shown in Fig. 1.

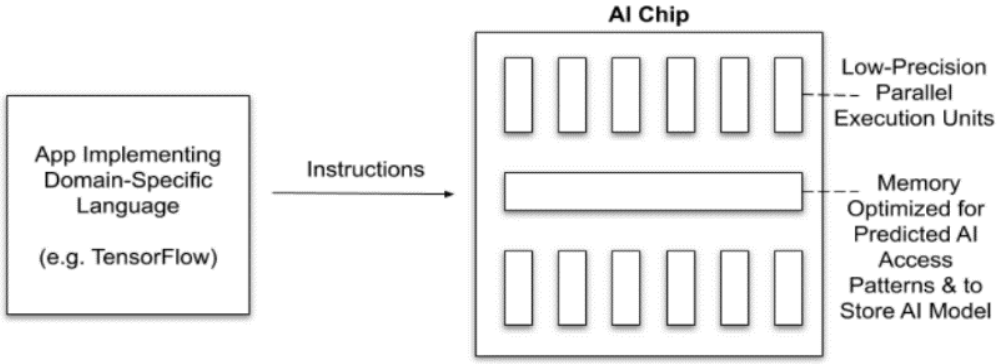

Fig. 1 A sketch of the AI chip.

4. Fabrication

Similar to all advanced semiconductor chips, the manufacturing process of AI chips is complex and can be divided into three broad categories: design, manufacturing, and assembly and packaging [7]. At present, most chips are manufactured in foundry mode, and the manufacturer outsources the task of designing chips to a design company such as United Microelectronics Corporation (UMC), which completes the design and sends it back to the manufacturer after passing the simulation test of the chip. Manufacturers simply focus on the fabrication of chips. On the front end, AI chips can now be manufactured on semiconductor wafers surfaces. These semiconductor wafers are made of high-purity polysilicon that is about 98% pure from silicon dioxide. Silicon wafers are produced by rolling, cutting, grinding, polishing and other processes, which have a typical diameter of 10 to 30 cm and a thickness of up to 1 mm [2, 8]. Lithography, etching, thin film, doping, chemical mechanical planarization (CMP) and other technologies have also been applied in the manufacturing of silicon wafers. Manufacturing technology determines the number of transistors that can be contained on a wafer. The manufacturing process of chips has made a qualitative leap from 500nm at the end of the 20th century to less than 10nm today. FinFET technology is used for processes smaller than 22nm; Today's most advanced "Extreme Ultraviolet" lithography technology can reduce the size of the chip to 7nm [2]. The imprinted wafer is then coated with a metal layer and a precise circuit pattern is etched with a precision instrument. This step will be repeated to create many tiny individual chips on the surface of silicon wafers. For the back end, each completed wafer is detected and tested by professional equipment to ensure its quality and filter out defective products. The chips are then cut from wafers and placed on the circuit board and wrapped in ceramic or plastic to prevent damage [7].

When it comes to AI chip manufacturing materials, more than silicon can be used. Although the silicon process is the most mature, other semiconductor materials, such as SiC and GaN, have also been applied in their appropriate fields, and some of their performance is even better than silicon. The figure shows the Comparison of physical properties of three different semiconductor materials (GaN, GaAs, and Si) [2].

5. Application

According to different situations in different fields, AI chips continue to improve and optimize, largely meeting the needs of artificial intelligence applications, among which the more prominent are mobile terminals and automatic driving. The primary goal of mobile AI chips is to ensure high computing energy efficiency, and then according to the different use patterns of different products, there will be additional requirements for AI chips such as low latency and low cost. Take the most common smartphone as an example, which contains many sensors and requires a certain degree of versatility of the AI chip in order to handle multiple types of tasks at the same time. At the same time, because the power supply of mobile devices is relatively limited, it also has a certain limit on the computing power consumption of AI chips. At present, research is devoted to the development of dedicated ASIC chips or low-power DSP to meet the needs of AI computing in mobile terminals [9].

In the field of autonomous driving, autonomous vehicles need to be equipped with a large number of electronic devices including cameras, radars, sensors, which generate tens of thousands of data information per second; In addition, this field needs chips to achieve low latency and high accuracy to ensure the safety of passengers, so AI chips used in this field must have extremely high computing power. In this case, relying on the cloud computing method has been unable to meet the demand, and it is necessary to install an on-board AI chip with high computing power and fast response ability for the car. At present, the mainstream way in this field is to use CPU+GPU chips, whose CPU is responsible for logic processing, and GPU is responsible for parallel computing and image processing. In addition, ASIC chips used in this field are also being developed [9].

On the other hand, edge intelligence is a big driving force for the development of AI today, which refers to the provision of AI chips specifically for edge devices, so that they can have artificial intelligence functions. Implementing edge AI can be divided into three categories: On-device intelligence, Near-device intelligence, and Far-device intelligence. On-device intelligence refers to the addition of AI chips to sensors and processing units so that they have neural networks for intelligent operation. In-sensor processing is a good example, since sensors contain neural circuits that can complete commands with a certain degree of intelligence. Near-device intelligence refers to the calculation of intelligent hardware which is outside the device. One example is the drone with real-time camera function. It transmits the image information captured by the drone to an additional processing station, which calculates the data through intelligent hardware analysis and issues commands to the drone. Far-device intelligence is based on intelligent hardware in the cloud or server for data processing, and generally intelligent hardware is Far away from the device (seen from Fig. 2) [4, 10]. There are many benefits brought by edge artificial intelligence, such as improving the computing efficiency of edge devices, improving the security of AI data processing, and saving energy consumption.

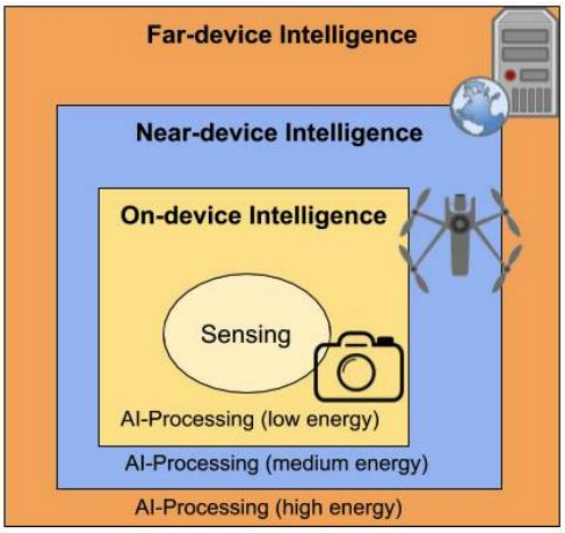

Fig. 2 A sketch of the different AI devices.

6. Limitations and Prospects

Contemporarily, the AI industry is developing rapidly. The development trend of AI chips in the future can be divided into four directions. First, the chip field will develop with higher computing power and lower power consumption. In recent years, with the development of the Internet of Things industry, the influence of edge intelligence has gradually increased, resulting in the concept of combining artificial intelligence and the Internet of Things. A large amount of data is collected and stored through the cloud and edge nodes, and this model of AI calculation by the cloud or edge forms a more advanced artificial intelligence, and only hardware devices with lower power consumption can meet this demand, which has become the target of future AI chip development. Secondly, general AI chips will become the mainstream of development. general AI chips are not limited to specific fields, and their data calculation can be used in all aspects of inference. Such AI chips can accelerate the calculation of general artificial intelligence, which has been set as the ultimate goal of related research. Third, AI will accelerate its integration with other disciplines in the future. Relying on biology, brain science and other related disciplines, AI chips can apply their knowledge to build new computing models to achieve bionic intelligence. At the same time, the introduction of AI will also accelerate the continuous development of various disciplines. The fourth future direction is to complete the transition from artificial intelligence to human-computer hybrid intelligence. Human-computer hybrid intelligence refers to modeling human behavior and cognition and introducing it into artificial intelligence system, so as to improve the performance of AI computing system, make it more suitable for the concept of artificial intelligence, and solve complex and difficult problems more efficiently [9].

7. Conclusion

To sum up, based on the discussion of the classification, principle, fabrication, application, and outlook of AI chips, it is not difficult to see that AI chips have gradually become the indispensable cornerstone of People's Daily life; Its powerful computing efficiency supports artificial intelligence computing in various fields, thus bringing convenience to huamn beings‘ lives. In the future, the computing and processing power of AI chips will become stronger to meet different needs, and it can be combined with various fields to accelerate their development, which has an inestimable prospect.

References

[1] Łukasiak L, Jakubowski A. History of semiconductors Cornell University. Retrieved from: https://djena.engineering.cornell.edu/hws/history_of_semiconductors.pdf.

[2] Li Y, He J, Xie Z, Jiang D. Frontier Development of chips design and production. Procedia Computer Science, 2018, 139: 554–560.

[3] Mellit A, Kalogirou S A. MPPT-based artificial intelligence techniques for photovoltaic systems and its implementation into field programmable gate array chips: Review of current status and future perspectives. Energy, 2014, 70: 1-21.

[4] James A P. The why, what, and how of artificial general intelligence chip development. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(2): 333–347.

[5] Khan S, Mann A. AI chips: What they are and why they matter. CSEP, 2020.

[6] Viswanathan S M. AI Chips: New Semiconductor Era. International Journal of Advanced Research in Science, Engineering and Technology, 2020, 7(8): 14687-14694.

[7] Inside an AI chip. MacroPolo. Retrieved from: https://macropolo.org/digital-projects/supply-chain/ai-chips/inside-an-ai-chip/.

[8] IC designs. University of California, Berkeley. Retrieved from: http://bwrcs.eecs.berkeley.edu/Classes/icdesign/ee141_f01/Notes/chapter2.pdf.

[9] Wang Y, Zhang L, Zhang W. The Development and Application of Artificial Intelligence Chips. 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE, 2022: 689-696.

[10] Furano G, Meoni G, Dunne A, et al. Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities. IEEE Aerospace and Electronic Systems Magazine, 2020, 35(12): 44-56