Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230358

Artificial-Intelligence integrated circuits: Comparison of GPU, FPGA and ASIC

Yujie Wang

Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK

ho19623@bristol.ac.uk

Abstract. In the recent years, the boom in technology industries has been greatly accelerated by the development of artificial intelligence (AI). AI, which is based on machine learning (ML), can only be developed rapidly because of the continuously increasing computational capacity of AI processors. Compared to general-purpose processors (GPPs), AI processors have specially designed architectures to accelerate the operations of AI applications, such as convolution, matrix, and massive parallel computing. The objectives of this paper are: (1) to illustrate the differences between general-purpose processors and AI processors; (2) to summarise the characteristic three mainstream AI processors: GPU, FPGA and ASIC, and draw a comparison among them. It shows that GPUs provide very competitive performance with high power consumption; FPGAs can offer high efficiency at low cost; and AISCs provide the highest performance with the lowest power consumption, but cost the most.

Keywords: AI, Integrated Circuits, GPU, FPGA, ASIC.

1. Introduction

Machine learning (ML) is the fundamental principle of artificial intelligence (AI). ML is based on algorithms turning massive data into models, which requires a powerful computing ability of integrated circuits. With the explosive growth of big data industry, the performance of traditional computing architecture can no longer keep up with the growth of data.

Nowadays, most computers use general-purpose processors as their central processing units, which are designed for general purposes, but not for specific applications. AI processors, however, refer to those processors that have been specially designed for AI algorithms. The architectures of AI processors comprise neuromorphic processing units which are designed based on machine learning and artificial neural network. These processors are fast and can analyse human behaviour and conduct calculations based on it. AI processors are regarded as the next breakthrough in the integrated circuit industry. According to a report published by Allied Market Research, the global artificial intelligence chip market size is expected to increase from $8.02 billion in 2020 to $194.90 billion in 2030, growing at a compound annual growth rate (CAGR) of 37.41% from 2021 to 2030 [1]. There are three mainstream AI processors that are currently applied in the field of AI, which are GPUs, FPGAs, and ASICs. This paper will summarise and compare the characteristics of the three mainstream AI processors. It will provide a reference for hardware designers to choose the right processor for the AI system and thus achieving a balance between performance, power consumption and cost.

DOI: 10.54254/2755-2721/4/20230358

2. General-purpose processors

General-purpose processors (GPP) are processors that are not designed for any specific application. For instance, the central processing unit (CPU) of a computer is a general-purpose processor because it is designed for general computing applications. The main structure of a CPU usually comprises control units (CU), arithmetic logic units (ALU), random access memory (RAM), and cache, as shown in Figure 1. The data is only calculated in the ALU (which forms the "core"), while the functions of other modules are to ensure that instructions will be executed in order. This traditional structure is suitable for general programming calculations. However, the operation of machine learning does not depend on complex instructions but massive parallel computing, and the traditional structure cannot meet the needs of AI applications [2]. Therefore, researchers are focusing on a new processor structure to improve the hardware performance of AI applications.

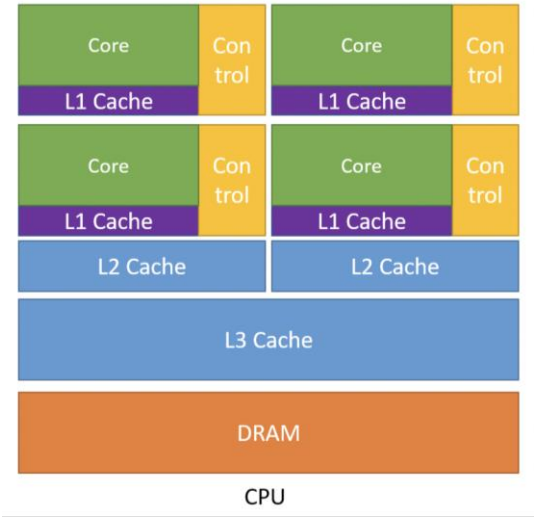

Figure 1. Architecture of a CPU.

3. Artificial intelligence processors

Technically, AI processors refer to processors that are specially designed for AI algorithms [3]. The cores of mainstream AI processors use Multiply-accumulate (MAC) acceleration array [2] to accelerate operations such as convolutions and matrices, significantly improving the calculating speed. Also, due to their specially designed structure, more parallel computing can be conducted to accelerate the training phase of ML. Moreover, some AI processors (FPGAs, ASICs) can be customized to meet the special needs of the applications, which optimizes the speed and power-consumption of the AI processors. There is a great trend towards processor specialization to improve performance, among which GPU, FPGA and ASIC become the mainstream hardware to implement AI applications.

3.1. Graphics processing units (GPUs)

GPUs are used chiefly to accelerate real-time 3D graphics applications, such as games. With the development of technology, GPUs have become more programmable than before, allowing them to accelerate many more applications that are far beyond traditional graphics rendering. Today's GPUs can be divided into three categories: GPUs for gaming, GPUs for content creation, and most importantly, GPUs for machine learning. The reason that GPUs can handle machine learning technology is due to its highly parallel computing architecture. Figure 2 shows the structure comparison of a CPU and GPU. It shows that most of the CPU's area is occupied by CU, cache, and DRAM, while the number of cores is very limited. However, the GPU has a tremendous number of cores but much smaller CUs and cache. This structure makes GPUs suitable for processing large numbers of simple calculation in parallel, such as the training phase in ML.

DOI: 10.54254/2755-2721/4/20230358

Because of the characteristics of large-scale parallel computation, GPUs play a huge role in the field of deep learning (DL), which is a branch of ML. Deep learning relies on neural networks—networks that are highly like the human brain—and the purpose of such networks is to analyse massive amounts of data at high speeds. It is believed that in the era of artificial intelligence, the GPU is no longer a graphics processor in the traditional sense, but a dedicated AI processor with powerful parallel computing capabilities.

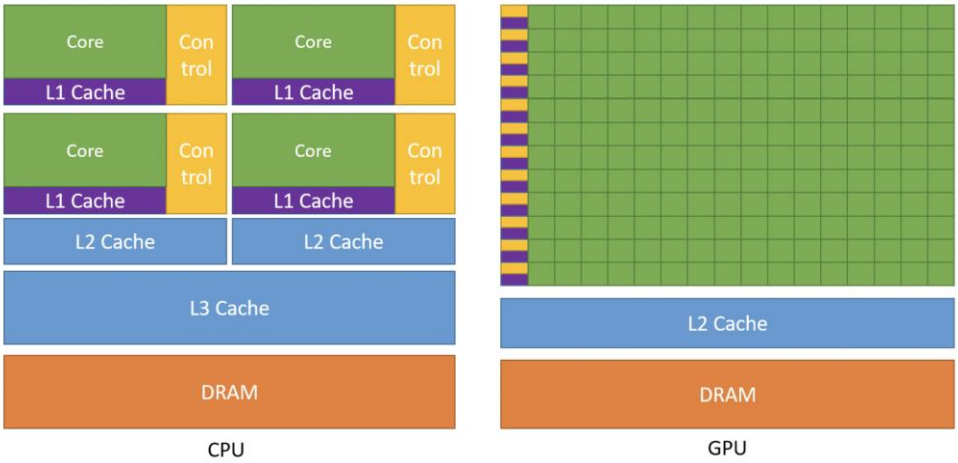

Figure 2. The structures of CPU and GPU.

3.2. Field-programmable gate arrays (FPGAs)

FPGAs are semi-custom integrated circuits with reconfigurable fabrics. The circuitry inside the FPGA chip is not hard-etched. The function of FPGAs can be reconfigured by programming in hardware description language (HDL). Due to the specific computation pattern of artificial neural networks, for example, convolutional neural networks (CNN), general-purpose processors cannot meet the requirements of CNN implementation. Therefore, various accelerators based on FPGAs, GPUs, and ASICs have been adopted to improve the performance of CNN designs. Among these approaches, FPGA-based accelerators have been attracted by scientists because of their advantages of high performance, low power consumption, and reconfiguration capability [4].

The reason for FPGAs' being efficient is their architecture without instructions. Traditional CPUs use the von Neumann structure. In this structure, the execution units may execute arbitrary instructions. Therefore, components like instruction memories, decoders, operators, and branch units (BU) are required. However, the function of each FPGA's component can be customised during the reconfiguration phase. Therefore, no instructions are required during running, which can significantly reduce the power consumption and enhance the overall performance [3]. Also, because specific FPGA circuits can be designed by programming and reorganising circuits, and together with its ability to perform parallel computing, it may take only one clock cycle to complete a particular operation. Therefore, though CPUs have much higher frequencies, they need multiple clock cycles to conduct specific calculations, and thus their overall performance might not be as good as FPGA.

3.3. Application-specific integrated circuits (ASICs)

ASICs are full-custom integrated circuit chips that are designed to meet specific needs. Compared to FPGA, an equivalent ASIC usually has higher performance and lower energy consumption. This is because the reconfigurable feature of FPGA requires large resources in the FPGA for on-chip routing and wiring [3]. However, ASICs are fully customized and, therefore, do not need to sacrifice internal resources for reconfigurability. Nurvitadhi et al. [5] experimented to compare the performance of equivalent FPGA and ASIC accelerate Binarized neural networks (BNNs). BNNs are recently proposed optimized variants of deep neural networks (DNNs), which have dramatically improved efficiency due to a reduction in their memory and computational demand. In this experiment, the 20-nm Aria 10 FPGA

DOI: 10.54254/2755-2721/4/20230358

and 14-nm ASIC are used to accelerate BNN. The result shows that ASIC has about 4.5 times higher performance and about 8 times less energy consumption than FPGA. Another research team [6] conducted an experiment measuring the gap between a 90-nm CMOS SRAM-programmable FPGA and a 90-nm CMOS standard-cell ASIC. The result shows that the standard-cell ASIC is 3.4 to 4.6 times faster than its equivalent FPGA.

4. Comparison between AI processors

The reason for the emergence of AI processors is to achieve a balance between a system's performance, power consumption, and cost. GPUs have a large amount of ALU to support massive parallel calculations. FPGAs can be reprogramed to fit the needs of specific algorithms, which are more flexible and efficient. ASICs are specially designed dedicated processors, which provide even higher performance and power efficiency. Each kind of processor has its design threads, which makes them suitable for different fields.

4.1. Performance

Due to two limitations, the performance of FPGAs is relatively low compared to GPU and ASICs. Firstly, the configurable logic blocks (CLB), which is the basic unit of the FPGA, has limited computing power. Secondly, the reconfigurable structure will consume large resources on the chip, which will sacrifice the performance of the FPGA.

Between GPUs and ASICs, ASICs have higher performance than GPUs. ASICs can be specially designed for specific requirements and systems and usually have a vast performance improvement. One of the most representative AI ASIC is Google's tensor processing unit (TPU), which uses 128×128,16-bit matrix multiply units (MXU) for matrix multiplication to accelerate machine learning. Wang et al. conducted an experiment which compares the acceleration of 6 real-world DL networks on TPU V3 and the NVIDIA TESLA V100 GPU [7]. The result shows the speedup of TPU over GPU ranging from 3 times (DenseNet) to 6.8 times (SqueezeNet).

4.2. Power consumption

The power consumption of GPUs is very high. There are thousands of ALUs, and billions of transistors integrated inside a GPU, and in every clock cycle, they are keeping changing and discharging, which consumes massive electricity. FPGAs consume much less power than GPUs for the same performance. FPGAs can be reconfigured to meet the power efficiency requirements. It can also accommodate multiple functions at the same time to achieve more power efficiency. For instance, the NVIDIA TESLA V100 GPU integrated 640 tensor cores and 5120 CUDA cores and has a maximum power consumption of 250W [8], while Samsung's SmartSSD Drive FPGA only has a maximum power consumption of 30W [9]. In an experiment of CSV parsing, running on both FPGA and GPU, the SmartSSD drive FPGA shows 25 times increase in performance/power over Tesla V100 GPU [9].

ASICs consume significantly less power than FPGAs under the same manufacturing process, due to ASICs fully customised architecture. The experiment conducted by Kuon, I., & Rose, J. [6] shows a 90-nm CMOS standard-cell ASIC consumes 87 times less static power and 14 times less dynamic power than an equivalent 90-nm CMOS FPGA.

4.3. Cost

FPGAs have a cost advantage over ASICs due to FPGA's reconfiguration capability. Firstly, the circuitry of FPGA can be reconfigured by hardware description language (HDL) to realize different functions without purchasing another processor. The architecture of the ASIC is fixed, which means ASICs cannot be reused. Secondly, FPGAs can be programmed and applied directly after purchasing, while it will take three months to a year to manufacture ASICs after they are designed. This feature can earn companies time to launch their products. Thirdly, the initial cost of FPGA is neglectable. As shown in Figure 2, there is almost no Non-Recurring Engineering (NRE cost) at the beginning, while the NRE cost for ASICs is very high. However, the slope of ASIC is flatter, which means in large volume

DOI: 10.54254/2755-2721/4/20230358

manufacturing, the cost per ASIC becomes less than FGPA [10]. The pricing of GPUs ranges wildly, from hundreds of dollars to tens of thousands of dollars per GPU. However, due to its generality, no cost needs to be used in programming, designing, or manufacturing.

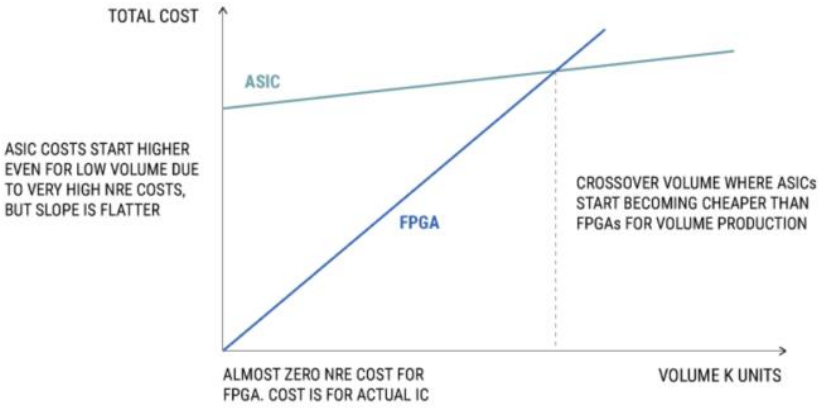

Figure 3. FPGA and ASIC cost analysis.

4.4. Sum-up

The results of this section are concluded in Table 1. GPUs have high performance, high power consumption, and an acceptable price. FPGAs have relatively lower performance than GPUs and ASICs, but they consume less power than GPUs and have a lower cost due to their reconfigurability. FPGAs are suitable for scenarios where the need for customisation is strong but the demand is not large. ASICs have the highest performance, lowest cost, and highest cost. They are suitable for scenarios with high demand and high requirements for processor customisation.

Table 1. Comparison between Three AI Processors.

	Performance	**Power consumption**	**Cost**
GPU	High	High	Medium
FPGA	Medium	Medium	Low
ASIC	Very high	Low	Very High

5. Conclusion

Artificial Intelligence technology has made continuous breakthroughs over the past few years. As the foundation of AI technology, AI processors have an important strategic position and promising market prospects. The general-purpose processor (GPP) is suitable for general programming calculations, while the AI processor has less generality but greater parallel computing ability to meet the needs of AI applications. The comparison results of three AI processors — GPU, FPGA and ASIC show that the GPU has high performance, high power consumption and acceptable cost. The FPGA has relatively medium performance, medium power consumption and low cost. The ASIC has the highest performance, and lowest power consumption, but also the highest cost. Each processor has its suitable application field and should be applied to reach the balance between the performance, power consumption and cost of a system. This paper lacks the discussion of the brain-like processor, which is also considered a promising direction for future AI processors. Nowadays GPUs are still dominating the AI processor market. However, FPGAs and ASICs have wide development space and will be the main research direction in the future.

References

[1] Savekar, A., & Sachan, S. (2022). Artificial Intelligence Chip Market Size, Share | Analysis-203

DOI: 10.54254/2755-2721/4/20230358

0. Allied Market Research. Retrieved 20 July 2022, from https://www.alliedmarketresearch.com/artificial-intelligence-chip-market.

[2] He, Y. (2021). Application of Artificial Intelligence in Integrated Circuits. Journal Of Physics: Conference Series, 2029(1), 012090. DOI: https://doi.org/10.1088/1742-6596/2029/1/012090

[3] Li, B., Gu, J., & Jiang, W. (2019). Artificial Intelligence (AI) Chip Technology Review. 2019 International Conference On Machine Learning, Big Data And Business Intelligence (MLBDBI). DOI: https://doi.org/10.1109/mlbdbi48998.2019.00028

[4] Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., & Cong, J. (2015). Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks. Proceedings Of The 2015 ACM/SIGDA International Symposium On Field-Programmable Gate Arrays, 161-170. DOI: https://doi.org/10.1145/2684746.2689060

[5] Nurvitadhi, E., Sheffield, D., Sim, J., Mishra, A., Venkatesh, G., & Marr, D. (2016). Accelerating Binarized Neural Networks: Comparison of FPGA, CPU, GPU, and ASIC. 2016 International Conference On Field-Programmable Technology (FPT), 77-84. DOI: https://doi.org/10.1109/fpt.2016.7929192

[6] Kuon, I., & Rose, J. (2007). Measuring the Gap Between FPGAs and ASICs. IEEE Transactions On Computer-Aided Design Of Integrated Circuits And Systems, 26(2), 203-215. DOI: https://doi.org/10.1109/tcad.2006.884574

[7] Wang, Y. E., Wei, G. Y., & Brooks, D. (2019). Benchmarking TPU, GPU, and CPU platforms for deep learning. arXiv preprint arXiv:1907.10701.

[8] NVIDIA V100 | NVIDIA. NVIDIA. (2022). Retrieved 28 July 2022, from https://www.nvidia.com/en-us/data-center/v100/.

[9] FPGA vs. GPU Acceleration: Considering Performance/Power. Blog.bigstream.co. (2022). Retrieved 28 July 2022, from https://blog.bigstream.co/fpga-vs.-gpu-acceleration-considering-performance-and-power.

[10] Singh, R. (2022). FPGA vs ASIC: Differences between them and which one to use? | Numato Lab Help Center. Retrieved 29 July 2022, from https://numato.com/blog/differences-between-fpga-and-asics/