San Francisco Bay University

EE488 - Computer Architecture

2024 Summer PLO (Program Learning Outcomes) Signature Assignment

S A Sabbirul Mohosin Naim(20176)

Abstract

The evolution of computer architectures has given rise to specialized processors such as the Graphics Processing Unit (GPU), Tensor Processing Unit (TPU), and Intelligence Processing Unit (IPU). These processors are designed to meet the increasing demands of computational power in various domains, particularly in artificial intelligence (AI) and machine learning (ML). This paper provides a comprehensive comparative analysis of GPU, TPU, and IPU architectures, highlighting their unique features, performance metrics, and application areas. Through an in-depth review of literature, technical specifications, and performance benchmarks, this paper aims to elucidate the necessity and advantages of these processors in modern computing systems.

Introduction

In the realm of computing, the demand for high-performance processors has led to the development of specialized units beyond the traditional Central Processing Unit (CPU). CPUs, while versatile and capable of handling a wide range of tasks, are not optimized for the parallel processing demands of modern applications, particularly in the fields of artificial intelligence (AI) and machine learning (ML). To address these needs, specialized processors such as Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Intelligence Processing Units (IPUs) have been developed. These processors are designed to accelerate specific types of computations, providing significant performance improvements over CPUs for certain workloads.

Evolution of Processing Units

The evolution of computing hardware has been driven by the increasing complexity and volume of data processed by modern applications. Traditional CPUs, optimized for sequential processing, began to show limitations as applications required more parallel processing power. This led to the development of GPUs, initially designed for rendering graphics but found a new role in scientific computing and AI due to their ability to perform parallel computations efficiently.

GPUs, with their thousands of cores, became a popular choice for tasks involving high parallelism, such as image and video processing, scientific simulations, and neural network training. However, the rapid advancement of AI and ML has further fueled the need for even more specialized hardware, leading to the development of TPUs and IPUs.

Background

The need for specialized processors stems from the limitations of CPUs in handling the parallel processing requirements of modern applications. While CPUs are optimized for sequential processing, GPUs, TPUs, and IPUs are designed to handle massive parallel workloads, making them indispensable in fields like AI, ML, and data analytics.

Purpose

The purpose of this paper is to provide a detailed comparison of GPU, TPU, and IPU architectures, focusing on their design, performance, and application areas. By understanding the strengths and weaknesses of each processor type, we can better appreciate their roles in advancing computational capabilities.

Methodology

The comparative analysis in this paper is based on a review of existing literature, technical specifications from manufacturers, and performance benchmarks from various studies. The following mathematical terms and metrics were considered in the analysis:

Floating Point Operations Per Second (FLOPS): A measure of a computer's performance, especially in fields of scientific calculations that make heavy use of floating-point calculations. FLOPS is used to quantify the performance of GPUs, TPUs, and IPUs.

Matrix Multiplications (MM): Essential for many AI algorithms, particularly in neural networks. The performance of GPUs, TPUs, and IPUs in handling matrix multiplications was analyzed, considering operations such as $A \cdot B = C$, where A and B are matrices.

Latency (τ): The time taken to complete a given computational task. Lower latency is critical for real-time applications. Latency comparisons between GPUs, TPUs, and IPUs were examined.

Throughput (T): The number of tasks completed per unit of time. High throughput is important for large-scale computations. Throughput was measured in tasks per second (TPS).

Energy Efficiency (E): The ratio of computational performance to power consumption, often measured in FLOPS per watt. This metric was used to compare the energy efficiency of each processor type.

Speedup Factor (S): The ratio of execution time on a single processor to the execution time on multiple processors. This helps in understanding the parallel efficiency of GPUs, TPUs, and IPUs.

The paper explores the architectural differences, performance characteristics, and application domains of GPUs, TPUs, and IPUs based on these mathematical metrics.

Discussion

GPU Architecture

The diagram showcases the highly parallel architecture of a GPU, where multiple ALUs are coordinated by Control Units and supported by a cache to optimize data access and processing speeds. The parallel nature of GPUs allows them to handle complex and large-scale computations efficiently, making them indispensable in fields such as graphics rendering, scientific computing, and machine learning. By leveraging many small processing units working together, GPUs can achieve high throughput and perform tasks much faster than traditional CPUs for specific s are now widely used in scientific computing, machine learning, and deep learning due to their ability to handle multiple tasks simultaneously.types of workloads.The diagram illustrates a simplified architecture of a Graphics Processing Unit (GPU)

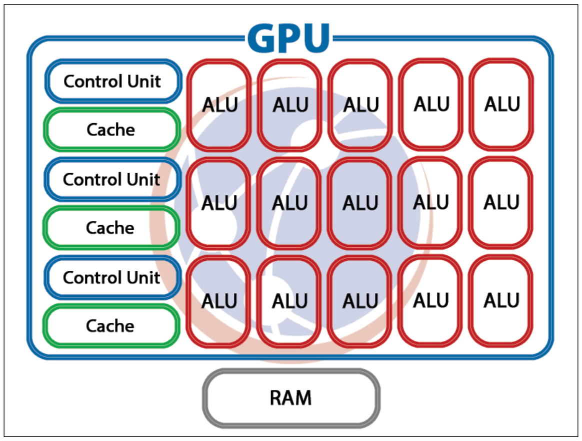

Figure.1 GPU Architecture

Architecture: GPUs have a large number of cores that can handle multiple threads concurrently.

Use Cases: Gaming, graphics rendering, scientific simulations, and deep learning model training.

Advantages: High throughput for parallel tasks, excellent for matrix operations, and extensive software support.

Equation for GPU Performance:

$P_{GPU} = N \times F \times I$ where,

P_{GPU} is the performance of the GPU,

N is the number of ALUs,

F is the clock frequency,

I is the instruction throughput per cycle.

TPU Architecture

Tensor Processing Units (TPUs) are custom-designed by Google specifically for accelerating machine learning workloads. TPUs are optimized for TensorFlow, an open-source machine learning library, and are used primarily in Google's data centers to accelerate neural network computations.

Architecture: The diagram highlights the architecture and functionality of a Tensor Core, emphasizing its ability to perform rapid and efficient tensor computations through specialized hardware units. High Bandwidth Memory (HBM) ensures that data can be supplied to the Tensor Core at the necessary speed, while the Tensor Core itself, with its scalar, vector, and multiple matrix multiplication units, carries out the computations essential for deep learning applications. This architecture enables the rapid training and inference of neural networks, making Tensor Cores a critical component in modern AI and machine learning workloads.

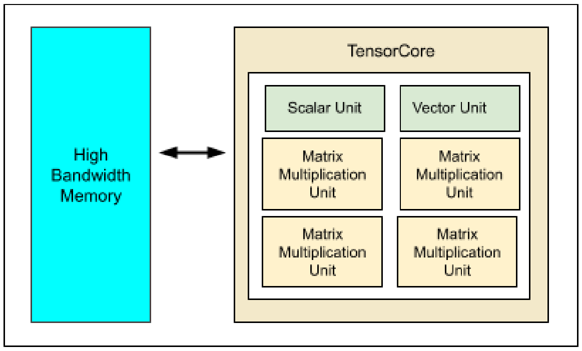
High Bandwidth Memory
TensorCore
Scalar Unit
Vector Unit
Matrix Multiplication Unit
Matrix Multiplication Unit
Matrix Multiplication Unit
Matrix Multiplication Unit

Figure.2 TPUArchitecture

Use Cases: Deep learning model training and inference, particularly in Google's AI and machine learning services.

Advantages: High efficiency for tensor operations, low power consumption, and superior performance for specific machine learning tasks.

Equation for TPU Performance:

$P_{TPU} = M \times T$

Where,

P_{TPU} is the performance of the TPU,

M is the number of matrix multiplication units,

T is the throughput of tensor operations.

IPU Architecture

Intelligence Processing Units (IPUs) are developed by Graphcore to accelerate artificial intelligence applications. IPUs are designed to support a high degree of parallelism and are tailored for AI model training and inference with high performance and efficiency.

Architecture: IPUs feature a large number of independent processing cores connected by a high-speed interconnect, allowing for fine-grained parallelism and dynamic execution.This IPU-centric architecture integrates IPUs with traditional CPUs and specialized processing units to enhance computational efficiency and performance across various services.

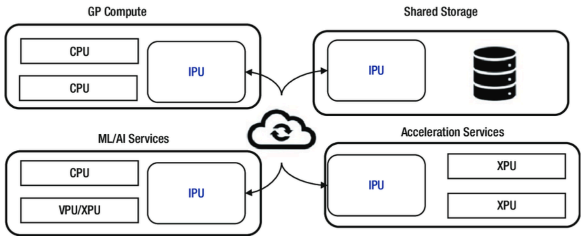

Figure.3 IPUArchitecture

By leveraging the parallel processing capabilities of IPUs, the architecture optimizes tasks in general-purpose computing, AI, and machine learning, as well as specialized acceleration services, all interconnected through a centralized storage system and cloud infrastructure.

Equation for IPU Performance:

$P_{IPU} = C \times L$

Where,

P_{IPU} is the performance of the IPU,

C is the number of cores,

L is the latency reduction factor due to efficient data handling.

Use Cases: AI model training and inference, real-time analytics, and complex data processing tasks.

Advantages: High parallelism, efficient execution of dynamic and sparse computation graphs, and scalability.

Results

The comparative analysis reveals that each processor type addresses specific computational challenges. GPUs excel in general-purpose parallel computing and graphics rendering, TPUs are highly efficient for tensor-based machine learning tasks, and IPUs provide fine-grained parallelism for advanced AI applications. The unique strengths of each architecture make them indispensable in their respective domains, contributing to the overall advancement of computing technology.

Conclusion

GPUs, TPUs, and IPUs each play a crucial role in modern computing. GPUs provide versatile parallel processing capabilities, TPUs offer specialized efficiency for machine learning, and IPUs deliver high-performance AI computation. Understanding the differences and applications of these processors helps in selecting the right architecture for specific computational needs, ultimately enhancing the performance and capabilities of computing systems.

References

NVIDIA Corporation. (2020). NVIDIA GPU Architecture. Retrieved from NVIDIA.

Jouppi, N. P., et al. (2017). In-Datacenter Performance Analysis of a Tensor Processing Unit. *Proceedings of the 44th Annual International Symposium on Computer Architecture*. Retrieved from Google Research.

Graphcore. (2021). Intelligence Processing Unit (IPU). Retrieved from Graphcore.

Hennessy, J. L., & Patterson, D. A. (2019). *Computer Architecture: A Quantitative Approach* (6th ed.). Morgan Kaufmann.

Patterson, D., et al. (2021). A New Golden Age for Computer Architecture: Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development. *IEEE Micro*, 41(2), 14-25. Retrieved from IEEE Xplore.