A Comparative Study of GPU, TPU, and IPU Architecture for AI Workloads

Yunisha Basnet
Student ID: 123456789

April 10, 2025

Abstract

In today's world, Artificial Intelligence (AI) has drastically increased the demand in every field, especially in high-performance computing. Traditional CPUs are no longer sufficient to handle the complexity and scale of AI models. This paper explores and compares three specialized processors – : Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Intelligence Processing Units (IPUs) –, detailing their architectures, performance characteristics, and their unique roles in the advancement of artificial intelligence.

1 Introduction

The increase of machine learning and AI has a major shift in computer science. With the increase in data and the intensity of computational programming, the general purpose of the CPUs is to struggle to keep up. To solve the problem, industry leaders have developed high-tech solutions to solve the issue. They have developed specialized processors that deal with AI workloads. GPUs, TPUs, and IPUs each have different advantages, be they flexibility, performance, or efficiency. In this paper, there is a detailed comparison of, advantages and disadvantages of all three types of architecture, pointing out that multiple processor types are necessary for the diversity of AI applications.

2 Methodology

This study uses documentation and benchmarks from trusted sources such as AWS, Google Cloud, and Graphcore. These sources provide information on the architectural features, processing efficiency, and scalability of GPUs, TPUs, and IPUs. Comparative evaluation is based on how well each processor handles AI workloads in terms of speed, energy consumption, and application specificity.

3 Graphics Processing Unit (GPU)

A Graphics Processing Unit (GPU) is a highly parallel process architecture composed of processing elements and a memory hierarchy. It is also known as a GPU, an electronic circuit that can perform a mathematical calculation at high speed. A GPU helps in applications that require working with large datasets, such as graphics rendering, machine learning, and video editing. It can work on multiple tasks with large data values at a time. However, it was not originally the same as it is today. The GPU was initially designed to perform a specific task at a time, such as controlling image display. In the 1940s and 1950s, dot matrix screens were used for displays before GPUs. Vector and raster displays followed. The CPU handled processing for the early graphics controllers, which were not programmable. A 3D imaging project aimed to produce pixels more quickly, which led to the GPU. The first GPUs, which combined lighting, transformation, and rendering engines on a programmable chip, appeared by the late 1990s, primarily for CAD and gaming. GPUs are vital for modern AI, powering tasks like image recognition and language processing. Their capacity to manage many calculations at once makes them ideal for training and running complex models. During training, GPUs speed up the process by performing parallel computations to adjust model parameters. In the inference phase, they enable fast, real-time predictions, supporting applications like self-driving cars and chatbots. GPUs are essential for AI because they speed up both training and inference, making it faster and more efficient to develop and deploy models. As AI continues to grow in complexity, the demand for powerful GPUs will continue to rise. GPU can be used in a wide range of compute-intensive applications, including large-scale finance, defense, and research activities. Some of the applications where GPUs can be used are gaming, professional visualization, machine learning, blockchain, and simulation. The GPUs that are best for AI depend on the task that is being worked on. For example, a GPU with a large amount of memory may be better suited for inferring large AI models, while a GPU with a high clock

speed may be better suited for low-latency inference serving.

4 Tensor Processing Unit (TPU)

Tensor Processing Units are application-specific integrated circuits designed by Google to accelerate machine learning workloads. They use custom chips developed by Google to accelerate tensor computations, especially for neural networks. They are optimized for matrix multiplications and tightly integrated with TensorFlow. Similar to the control processing unit, the TPU was developed in response to the need for deep learning and artificial intelligence workloads. Even GPUs have trouble processing large amounts of data. In 2016, Google released the first-generation TPU, also known as TPU v1. Its purpose was to speed up inference tasks in TensorFlow applications. TPUs were designed as application-specific integrated circuits (ASICs), designed with precision for high-throughput matrix operations, a fundamental part of neural network computations, in contrast to GPUs, which are general-purpose parallel processors. With the release of TPU v2 (2017), TPU v3 (2018), and TPU v4 (2021), Google added support for training, improved interconnect bandwidth, and improved performance and memory. TensorFlow, a highly effective AI accelerator designed specifically for large-scale machine learning models, is the result of Google's strategy of tightly coupling hardware and software. TPUs are grouped into Pods for large-scale AI tasks, with slices for even greater scalability. From TPU v4 onward, a 4×4×4 cube topology improves performance. Newer versions, such as v5p and v6e, also include Spare Cores for recommendation models. ICI resiliency ensures stable performance by rerouting around hardware faults. Each TPU version brings improvements for efficient AI training and inference. Rapid growth in artificial intelligence has directly influenced the evolution of TPU architecture.

5 Intelligence Processing Unit (IPU)

An intelligence processing unit consists of many individual cores, called tiles, allowing high parallel computation. It is released by Intel, which is like SmartNIC, and seeks to improve the processing, networking, and storage by creating space in the CPU. A programming networking device designed to enable cloud and communication service providers which reduce the overhead and improve up performance of the CPU released by Intel in 2021. The primary goal of an IPU is to enable customers to better utilize the resources in a secure, programmatic, and stable solution that enables them to balance processing and storage. Graphcore, a UK-based company, introduced the first commercial IPU in 2016. Their goal was to build processors that could overcome the memory and compute bottlenecks of GPUs when executing AI workloads, especially for training complex models. Each IPU contains hundreds of independent processor tiles that work in parallel, with distributed memory close to each compute unit. This architecture avoids memory bottlenecks and enables models to execute more efficiently, especially those with fine-grained control flow, like recurrent networks and transformers. IPUs are designed specifically for AI tasks. Their structure allows better model performance with lower latency and energy consumption. They are widely used in research and enterprise environments for natural language processing, computer vision, and reinforcement learning. IPUs offer a new paradigm of computing that complements GPUs and TPUs in the AI ecosystem.

6 Comparison

As artificial intelligence evolves, the processors that power its workloads become more specialized. Originally designed for graphics rendering, GPUs are now widely used in AI because of their parallel processing capabilities. They are versatile and support a diverse array of applications, but their general-purpose design can result in inefficiencies and high energy consumption when performing specific AI tasks. In contrast, Google-designed TPUs are optimized for tensor computations and excel at deep learning workloads. They provide faster and more energy-efficient training and inference, but are less flexible and primarily limited to neural network operations. IPUs, which were first introduced by Graphcore, offer an alternative method by utilizing thousands of independent processing tiles with local memory. This allows for fine-grained parallelism for models with irregular computation patterns, like those found in reinforcement learning and natural language processing. As a new class of AI hardware designed specifically for natural language tasks, Language Processing Units (LPUs) have recently arrived. LPUs perform exceptionally well in sequential processing, which makes them ideal for language-based models such as LLMs. As these specialized processors become more popular, the development of AI hardware is changing from general-purpose solutions to task-specific accelerators that optimize efficiency and performance in particular AI domains.

7 Results

Processor	Core Count	Efficiency	Flexibility	AI Suitability
GPU	Thousands	Moderate	High	High
TPU	Few (Custom ASIC)	High	Low	Very High
IPU	Hundreds (Tiles)	High	Medium	High

Table 1: Basic Comparison of GPU, TPU, and IPU

Table 2: Overall Architectural Comparison of GPU, TPU, and IPU

Feature / Metric	**GPU (Graphics Processing Unit)**	**TPU (Tensor Processing Unit)**	**IPU (Intelligence Processing Unit)**
Developer	NVIDIA, AMD	Google	Graphcore
Purpose	General-purpose (graphics, AI, HPC)	Neural network acceleration	AI-specific for sparse, dynamic models
Architecture Type	Thousands of parallel cores (SIMD)	Matrix units, systolic array (ASIC)	Hundreds of tiles (MIMD), distributed memory
Performance (AI)	High for training and inference	Very high for tensor-heavy models	High for NLP, GNN, RL models
Precision Support	FP32, FP16, INT8, BF16	INT8, BF16	FP16, FP32
Flexibility	Very high, general-purpose support	Limited to specific AI tasks (TensorFlow)	Moderate, AI-specific with model flexibility
Energy Efficiency	Moderate	High	High
Memory Access	GDDR/HBM (off-chip)	On-chip with high-bandwidth access	On-chip, local memory per tile
Programming Support	CUDA, PyTorch, TensorFlow, OpenCL	TensorFlow (Google ecosystem)	Poplar SDK, PyTorch integration
Best Use Cases	CNNs, GANs, image/video tasks, gaming	CNNs, Transformers, cloud-scale ML	NLP, GNNs, reinforcement learning
Availability	Widely available, consumer + enterprise	Google Cloud only	Limited to research/enterprise
Ecosystem Maturity	Very mature	Mature in Google ecosystem	Emerging

8 Discussion

GPUs are flexible and widely available, making them ideal for general-purpose AI workloads. However, they consume more energy for specific tasks. TPUs outperform GPUs in tensor-heavy operations but are limited to Google's ecosystem. IPUs offer fine-grained parallelism and excel in natural language processing and reinforcement learning.

The introduction of task-specific hardware like LPUs signals a move towards domain-specific architectures in AI. These processors optimize performance for particular model types, increasing efficiency and reducing energy consumption.

9 Conclusion

As AI applications grow more complex, the need for specialized processors becomes apparent. GPUs provide versatility, TPUs deliver high performance in tensor operations, and IPUs bring efficient execution for structured AI tasks. Each processor addresses unique needs, and together they form a diverse ecosystem that supports the rapid advancement of AI.

10 References

1. Amazon Web Services. (n.d.). What is a GPU? Retrieved April 21, 2025, from `https://aws.amazon.com/what-is/gpu/`

2. Bredun, R. (n.d.). What is TPU and how/why it works. Medium. Retrieved April 21, 2025, from `https://medium.com/@ruslanbredun007/what-is-tpu-and-how-why-it-works-9a0a4a59399e`

3. Google Cloud. (n.d.). GPU for AI. Retrieved April 21, 2025, from `https://cloud.google.com/discover/gpu-for-ai`

4. Google Cloud. (n.d.). System architecture: TPU VM. Retrieved April 21, 2025, from `https://cloud.google.com/tpu/docs/system-architecture-tpu-vm`

5. Graphcore. (n.d.). About the IPU. Retrieved April 21, 2025, from `https://docs.graphcore.ai/projects/ipu-programmers-guide/en/latest/about_ipu.html`

6. Ramkumar, H. (2023, April 23). Comparing GPU vs TPU vs LPU – The battle of AI processors. Medium. `https://medium.com/@harishramkumar/comparing-gpu-vs-tpu-vs-lpu-the-battle-of-ai-processo`

7. Trenton Systems. (n.d.). What is an IPU? Retrieved April 21, 2025, from

https://www.trentonsystems.com/en-us/resource-hub/blog/what-is-an-ipu

Thank you!!!