



Systematic Analysis of TPU – The Game changer of Machine Learning

Zixuan Ye

Dulwich College, London, SE21 7LG, United Kingdom
yemichael1746@gmail.com

Abstract. In the era of rapid advancements in artificial intelligence (AI) and deep learning, the choice of hardware processors plays a critical role in determining the efficiency of model training and inference. This paper talks about the two leading processors: the general purposed Graphics Processing Unit (GPU) and the specialized Tensor Processing Unit (TPU) and compares their uses in AI deep learning. Whilst NVIDIA's GPUs were originally designed for rendering graphics, they were successfully repurposed for training an AI model. This paper explains how tensor cores were introduced and allows much faster parallel computations. However, GPU's versatile architecture leads to significant inefficiencies for deep learning tasks, often wasting energy and having a high-power consumption. Therefore, this shows that GPUs are not the best choice for AI deep learning. Then, the essay discusses Google's TPUs, which were built specifically for AI. This specialized design makes them much faster and more energy-efficient for training large AI models. The text describes how TPUs have improved over several generations and shows how the TPU delivers superior performance-per-watt for large scale training and inference.

Keywords: GPU, TPU, AI Hardware, Specialized Hardware, Computational Efficiency

1 Introduction

In the last decade, artificial intelligence has undergone a rapid transformation, shifting from a theoretical concept into the main driver of our technological revolution. This evolution is largely dependent on the maturation of deep learning, a type of machine learning that simulates our human brain's neural networks to spot patterns within huge amounts of data. It has led to some significant achievements such as ChatGPT, facial recognition, autonomous vehicles and so on [1].

However, these achievements come at a considerable cost. Unlike traditional programming, which relies on explicitly specifying each instruction, deep learning models learn inductively from the huge amount of data that you feed in, therefore a huge amount of computational power is required. Training a model using neural networks requires trillions of mathematical operations, mostly matrix multiplication and tensor calculations [2].

Consequently, this processing demand has ignited a race among companies to develop the most powerful processor for AI. While you may think that GPUs, especially those made by NVIDIA, are no doubt dominating in this area, what if I were to tell you that there is a type of chip, called the TPU chips, which are far more powerful and efficient, and are specialized just for deep learning?

2 Basic Theory Analysis of GPU

2.1 Introduction to GPU

To get this current shift, we must first understand what originally makes deep learning possible. The key catalyst was the repurposing of the GPU. GPUs are originally designed to render complex scenes for video games and computer graphics. Think of it not as a single chip, but as a highly specialized parallel computing platform [3]. Its internal structure is a group of components built together for massive parallel processing.

2.2 Components of a GPU

The Fig.1 shows an AD102 GPU. It includes 12 GPCs, 72 TPCs, 144 SMs, 18432 CUDA Cores, 144 RT Cores, 576 Tensor Cores and 576 texture Units [4].

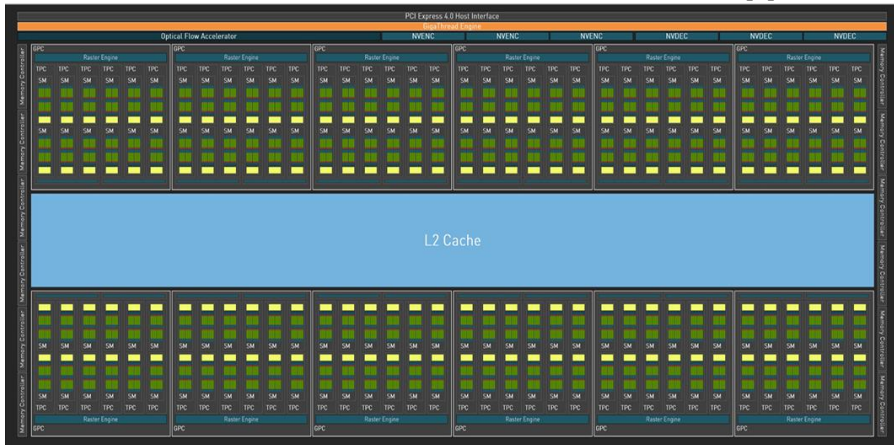


Fig. 1. AD102 GPU [4]

First of all, let's start with the Graphics Processing Cluster (GPC) first. It is the highest-level organizational unit within a GPU die. It acts as a semi-autonomous processing engine that can handle major tasks independently. It is responsible for managing a set of texture processing clusters and coordinating graphics rendering or computing tasks within its domain. It contains key graphics specific hardware like the Raster Engine, which converts 3D vector graphics into a raster image of pixels. Since we've brought on to the Texture Processing Cluster (TPC), lets explain what it is. The TPC is an intermediate module cluster that resides within a GPC. It is a critical

organizational block that groups together processing and texturing resources. The TPC groups Streaming Multiprocessor (SMs) with texturing resources and each TPC usually has 2 SMs. SMs are the absolute heart of parallel computation and the fundamental processing unit of the GPU [5]. All parallel computation happens inside the SMs. The number of SMs in a GPU determines how powerful the GPU can be.

2.3 Breakdown of a SM

The SM is a highly complex unit designed to execute hundreds of threads concurrently. Its key components include: Warp Schedulers where instructions are being scheduled for warps and dispatches them to the execution units; Dispatch Units which deliver the scheduled instructions to the various cores for execution; and a Register File and Shared Memory which are two extremely fast on-chip memories [6]. The SM's strong parallel processing power also comes from its array of specialized cores including the CUDA Core, Tensor Core, RT Core, and Texture Unit.

2.4 Tensor Core and its job in deep learning

Out of these four cores, the Tensor Core is the most critical component used for training an AI model. It is a specialized processing unit within the NVIDIA GPUs, first introduced with the Volta architecture in 2017. Tensor cores are designed specifically to accelerate matrix multiplication and convolution operations. They work by performing mixed precision calculations, using FP16 for input and accumulating results in FP32. This resulted in a significant boost in computational speed and efficiency while maintaining necessary accuracy for model training and inference. This allows GPUs with Tensor Cores to process large scale matrix computations much faster than traditional GPUs, leading to substantial reduction in training times, often 3times or more, for complex neural networks such as the ResNet-50 and BERT [7].

2.5 Disadvantages of a GPU

However, even though NVIDIA has such a powerful beast in its GPUs, the GPU itself is not inherently optimized for AI workloads. Its architecture, originally designed for graphics rendering, leads to significant inefficiencies when applied to deep learning tasks. A large portion of the GPU's hardware resources, often 30-40%, remains idle during neural network computations. Its architectural mismatch results in two primary disadvantages

First of all, a low hardware utilization for AI, the GPU's versatility means that not all its transistors are dedicated to the matrix multiplications and accumulations that form the core of deep learning algorithms. This leads to underutilization of the silicon when running AI models. Secondly, the high-power consumption of a GPU is also another major drawback, especially outside of data centers. For example, gaming GPUs like the RTX 4090 that we all know, can have a TDP of 450W, while data center GPUs like the H100 can consume up to 700W per unit [8]. This results in a poor performance-per-watt ratio for AI tasks. In a compelling comparison, a pure GPU solution for an L4 autonomous vehicle could consume around 150W, drastically reducing the vehicle's range compared to a more specialized, efficient alternative. This highlights a critical

limitation for mobile and edge applications where power is constrained. As a result, using a GPU for AI is not the most efficient or effective processor.

3 Basic Theory Analysis of TPU

3.1 Introduction to TPU

On the other hand, Google's development of the TPU represents a significant shift from relying on a general purposed processor to designing a specialized hardware that focuses only on artificial intelligence. This journey began in 2013 when Google faced a critical challenge: its data centers powered by traditional CPUs and GPUs were struggling to efficiently handle the growing computational demands of neural networks, particularly in terms of power consumption and performance. This led to their development of the first TPU [9].

3.2 Generations of TPU

The first-generation TPU, announced in 2016, was a groundbreaking innovation. It was designed primarily for inference tasks, the process of using a trained model to make predictions. The TPU was manufactured on a 28nm process, it operated at 700MHz and consumed 28-40W of power. Its key architectural advantage was its use of a systolic array for matrix multiplications, which minimized data movement and memory access, leading to massive improvements in efficiency. Compared to contemporary GPUs, it offered a 30 times higher performance improvement and was nearly 80 times more power efficient. This was achieved by focusing on low precision INT8 operations, which were sufficient for many inference tasks, and by eliminating unnecessary general purposed components like complex cache systems and branch prediction.

Recognizing that AI models were rapidly evolving and required more than just inference capabilities, Google's later TPU generations expanded their focus to include model training. The TPU v2, introduced in 2017, was a major step forward. It incorporated 16GB of High Bandwidth Memory with a bandwidth of 600GB/s, a substantial upgrade from the DDR3 memory used in v1. This allowed it to handle the massive datasets required for training complex models. Crucially, v2 introduced support for Google's bfloat16 floating-point format, which offered a wider numerical range than FP16 while using less memory than FP32 [10]. This made it viable for both training and inference, enabling Google to train models like the AlphaGo Zero. Furthermore, Google introduced the concept of a "TPU Pod," interconnecting 256 v2 chips to create a supercomputer capable of delivering 11.5 petaflops of performance.

The TPU v3, released in 2018, doubled down on this approach by further scaling the number of multipliers, increasing clock speeds, and raising performance to 123 teraflops per chip. However, this increased power also means more heat released when running, with its thermal design power reaching 220W; liquid cooling systems are required to cool it down.

The TPU v4, launched in 2021, was manufactured on a more advanced 7nm process. It featured 32GB of HBM memory with a bandwidth of around 1.2TB/s and adopted a

Mesh interconnection architecture that allowed for even larger pods of up to 4096 chips. This generation also introduced architectural optimizations like support for sparsity, allowing the chip to skip calculations involving zero values, which could double computational efficiency in models with 50% sparse weights [11]. Around this time, Google also formalized a more specialized strategy, learning from previous generations that a one-size-fits-all chip was not optimal for large-scale data center operations. This led to the creation of the TPU v4i, a variant specifically optimized for inference. It was designed with a lower TDP of around 175W to allow for air cooling, making it easier to deploy globally. It also uses a large Common Memory of 128MB for faster processing of intermediate tensors and rapid switching between multiple models, addressing the need for multitasking in production environments.

By 2023, Google's strategy evolved into offering a differentiated product line with the TPU v5 series. The TPU v5e was focused on cost-efficiency for inference and medium-sized model training, featuring 197 TFLOPS at bfloat16 and 16GB of HBM memory, while the TPU v5p was designed for maximum performance in training the largest models, with 459 TFLOPS at bfloat16 and 95GB of HBM memory. Google claimed that the TPU v5e offers nearly 2 times faster training and 2.5 times faster inference than TPU v4 at half the cost. This has also attracted lots of other companies such as Apple and Anthropic to start to use the TPU chips for training their own AI models.

In 2024, Google released the TPU v6, codenamed Trillium, which provided a 4.7x increase in peak compute performance per chip compared to v5e, reaching 918 TFLOPS at bfloat16 with 32GB of HBM memory and 1.64TB/s bandwidth. It also improved energy efficiency by 67% over v5e and supported pods of up to 256 units.

The latest generation, the 7th-gen TPU named "Ironwood", was announced in April 2025 and represents the culmination of these lessons, fully heralding what Google calls the "age of inference." Ironwood is the first TPU designed exclusively for inference. Its most significant innovation is being the first to support FP8 calculations, a precision ideal for inference that doubles the computational throughput over FP16. A single Ironwood chip boasts a massive 192GB of HBM (six times the capacity of the previous v6 generation), and a pod of 9,216 interconnected chips can achieve a staggering 42.52 exaflops at FP8, surpassing the world's largest supercomputers. Furthermore, it doubles the power efficiency of the last generation, directly addressing the critical data center constraints of energy consumption and cost [12].

The impact of Google's TPU development is huge. It has drastically improved cost and energy efficiency in data centers, enabled advanced AI features in services from search and translate to Google Photos, and strengthened Google's cloud offerings by providing these powerful accelerators to external customers.

3.3 Disadvantages and improvements to TPU

Even though Google's TPUs are great at large-scale AI model training and deliver exceptional performance for deep learning workloads, their highly specialized architecture limits their versatility. A critical drawback is their narrow focus on tensor operations, which makes them less adaptable to the rapidly evolving AI landscape. As AI development accelerates, spanning diverse applications like generative AI, edge

computing, and multimodal models, TPUs' specialization in a single domain may become its drawback against competing hardware such as NVIDIA's H200 GPUs or AMD's Instinct MI300 series which offers broader flexibility. Additionally, TPUs generate significant heat as it operates and a fully loaded Ironwood TPU rack can consume over 100 kilowatts, necessitating advanced liquid cooling systems that make them impractical for consumer devices or edge applications.

To address the limitations of Google's TPUs, several solutions are being used. Their specialized architecture, optimized for TensorFlow, is being mitigated to broaden software compatibility, such as TensorFlow's integration with other frameworks like PyTorch via ONNX, enabling TPUs to support a wider range of AI workloads. To counter their narrow focus on tensor operations, Google is enhancing TPU flexibility through updates like TPU v5e, which supports more diverse AI tasks, including generative AI models. For thermal and power challenges, advanced liquid cooling systems are already in use, and Google's data centers use energy-efficient designs to manage the high-power consumption of TPU racks. To think of it more abstractly, Google could involve developing hybrid TPU-GPU clusters that dynamically allocate workloads based on task requirements, combining TPU's efficiency with GPU versatility [13]. Additionally, modular TPU edge chips with lower power profiles could be designed, incorporating phase-change cooling materials to enable compact, efficient deployment in edge devices, thus expanding their applicability beyond data centers.

4 Conclusion

The battle for the best AI processor is a clash of specialists against generalists. Google's Tensor Processing Units are like race cars, designed specifically to tackle the complex math behind AI, solving them incredibly fast and energy efficient for large-scale training tasks. In contrast, NVIDIA's GPUs are the all-rounders; originally built for graphics. Their power for parallel computing is backed by a strong and steady software ecosystem that makes them the default choice for a wide range of applications beyond just AI.

While TPUs boast raw performance for specific jobs, their adoption is limited by a less flexible software environment compared to NVIDIA's deeply entrenched platform, which continues to evolve with powerful new chips. This competition, not just between NVIDIA and Google but lots of other companies, ensures that the future of AI hardware won't be dominated by a single winner but will instead be shaped by a dynamic, ongoing clash of innovation.

References

1. Malik, H., Chaudhary, G., Srivastava, S.: Digital transformation through advances in artificial intelligence and machine learning. *J. Intell. Fuzzy Syst.* 42(2), 615–622 (2022)
2. Thakur, A., Konde, A.: Fundamentals of neural networks. *Int. J. Res. Appl. Sci. Eng. Technol.* 9(VIII), 407–426 (2021)

3. Stopper, D., Roth, R.: Massively parallel GPU-accelerated minimization of classical density functional theory. *J. Chem. Phys.* 147(6), 1–12 (2017)
4. Vuduc, R., Choi, J.: A brief history and introduction to GPGPU. In: *Modern Accelerator Technologies for Geographic Information Science*, pp. 9–23. Springer, Boston, MA (2013)
5. Yu, V.W., Govoni, M.: GPU acceleration of large-scale full-frequency GW calculations. *J. Chem. Theory Comput.* 18(8), 4690–4707 (2022)
6. Yu, V.W., Govoni, M.: GPU acceleration of large-scale full-frequency GW calculations. *J. Chem. Theory Comput.* 18(8), 4690–4707 (2022)
7. Andreucci, F.: Performance analysis of GPU-to-GPU communications in distributed FFTs in Quantum ESPRESSO. Dissertation, SISSA (2025)
8. Harris, J.A., Liu, R., Martins de Oliveira, V., et al.: GPU-accelerated all-atom particle-mesh Ewald continuous constant pH molecular dynamics in Amber. *J. Chem. Theory Comput.* 18(12), 7510–7527 (2022)
9. Rohm, K., Manas-Zloczower, I.: A micromechanical approach to TPU mechanical properties: framework and experimental validation. *Mech. Mater.* 180, 104627 (2023)
10. Shamoii, P.: Guide to selecting the best hardware (GPU vs. TPU) for training a large semantic segmentation model. Preprint (2024)
11. Garza, D.E.T., Trevino, D.E., Yilmaz, M.: Contemporary artificial intelligence accelerator cache design perspectives. In: *Proc. Int. Conf. on Computational Science and Computational Intelligence (CSCI 2024)*, pp. 192–205. Springer, Cham (2024)
12. Shi, Z., Land, S., Locatelli, A., et al.: Understanding likelihood over-optimisation in direct alignment algorithms. *arXiv:2410.11677* (2024)
13. Jouppi, N., Kurian, G., Li, S., et al.: TPU v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In: *Proc. 50th Annual Int. Symp. on Computer Architecture (ISCA 2023)*, pp. 1–14. ACM, New York (2023)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

