Engineering AI Accelerators: GPUs, TPUs, and NPUs

Yongjian Zhang

Abstract

Artificial intelligence requires specialized hardware to perform the large number of mathematical calculations needed for training and inference. This paper examines three major types of AI accelerators: Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Neural Processing Units (NPUs). Each processor is designed to improve AI performance in different ways. GPUs are widely used due to their parallel processing performance and flexibility. TPUs are specialized for machine learning workloads and matrix operations. NPUs are designed for efficient processing on laptops and cell phones. This paper explains how each accelerator works, the advantages and disadvantages of each accelerator, and how each plays a role in modern AI systems. Overall, specialized AI hardware allows artificial intelligence to operate faster and more efficiently compared with relying on CPUs alone.

Introduction

Artificial intelligence is now involved in many aspects of everyday life. Applications range from customer service on websites to infotainment systems in cars. Many modern AI systems rely on specialized processors such as Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Neural Processing Units (NPUs). These processors can be found in devices ranging from everyday laptops and computers to some of the most advanced AI servers in the world. They can help run AI tasks either locally on a device or as part of cloud-based computing systems. These processors accelerate the mathematical computations used to train and run AI models. They accelerate AI workloads by performing large numbers of mathematical operations in parallel. Each processor has its own strengths and is suited for different workloads, such as machine learning, on-device AI, and graphics processing [1]–[3].

Fundamentals of AI Acceleration

Artificial intelligence requires specialized hardware because AI models perform a large number of mathematical computations. AI relies on matrix multiplication and other mathematical operations that can be performed simultaneously, so specialized accelerators such as GPUs, NPUs, and TPUs can perform these workloads more efficiently than CPUs alone [1], [3].

Artificial intelligence refers to computer systems designed to perform tasks such as recognizing patterns, making predictions, and processing language. AI systems use neural networks, which perform millions of mathematical operations during training and inference. These tasks can be computationally demanding for CPUs, especially as AI models increase in size and complexity. AI acceleration hardware uses many specialized processing units to perform calculations at the same time. This increases the speed and efficiency of AI training and inference [1], [3].

Parallel computing breaks large computational problems into smaller operations that can be performed simultaneously. This is especially useful for AI because neural networks require large numbers of matrix and vector calculations. Performing many of these calculations at the same time increases the speed of both training and inference [5].

AI workloads are generally divided into two major stages: training and inference. Training is the process in which an AI model learns from large amounts of data and adjusts its parameters based on patterns in that data. After being exposed to large amounts of data, models can recognize patterns and make predictions based on new information. Inference occurs after training, when a trained model uses new input data to generate predictions or outputs. Inference is used in real-world applications such as autonomous vehicles, finance, research, and large language models [4].

The performance of AI accelerators can be measured using many different metrics. One common measure of processing performance is FLOPS, or floating-point operations per second. FLOPS measures the computational throughput of a processor, but a higher FLOPS rating does not always result in better real-world performance. Factors such as memory bandwidth and the type of workload can also limit the performance of an AI accelerator [15].

Graphics Processing Units (GPUs)

GPUs were primarily used for graphics processing and gaming until the 2000s, when their ability to solve complex math problems became useful in the AI world. They commonly work alongside CPUs, handling large amounts of parallel processing. Some GPUs are integrated with CPUs, while more intensive systems use discrete GPUs as separate components. GPUs handle AI tasks efficiently due to their parallel processing capabilities and ability to perform large amounts of mathematical computation [2], [3].

GPUs accelerate AI tasks due to their ability to employ parallel processing and scale to large computing systems. For highly parallel workloads, GPUs can achieve greater computational throughput than CPUs. This results in GPUs being used for both AI training and inference [6].

GPUs process AI calculations through a combination of matrix multiplication, data parallelism, and high throughput. During training, a model receives input data, makes predictions, compares them with expected outputs, and adjusts its weights through backpropagation. GPUs accelerate this process by performing matrix calculations for many pieces of data in parallel. During inference, a trained model uses new input data to generate outputs, and GPUs can process these inputs quickly. Models can also be optimized for efficiency and accuracy [7].
Here is how the whole process works:

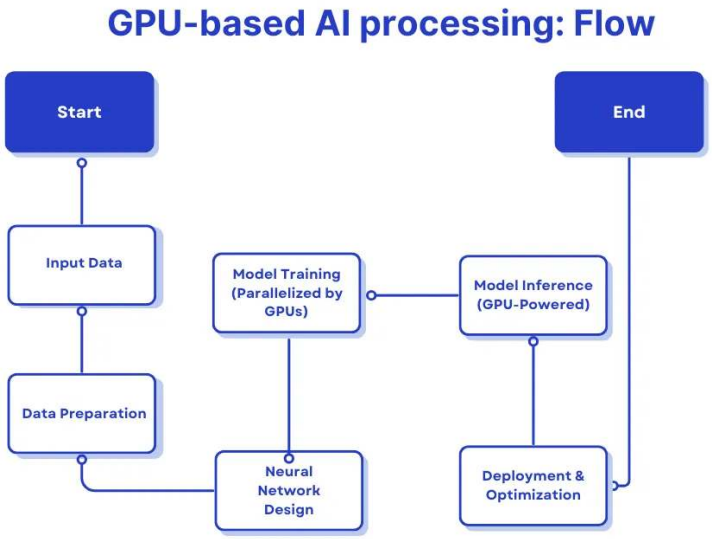

Figure 1. General AI training and inference workflow. Source: Adhikari [7].

GPUs have several advantages due to their architecture. Companies choose GPUs because of the many benefits they provide, most notably their high performance and parallel processing capabilities. GPUs can provide significantly greater computational performance than CPUs for highly parallel workloads. Their parallel processing capabilities make them well suited for deep learning, image recognition, and natural language processing. GPUs can also provide high computational performance per watt for parallel workloads, which becomes important when systems are scaled up. GPUs are also highly scalable because they can be deployed in large clusters [6], [8].

With these advantages also come disadvantages. GPUs are highly energy intensive and require large amounts of electricity to operate, while AI data centers can also require significant amounts of water for cooling. This increases operating costs and can contribute to environmental impacts, including carbon emissions associated with data-center electricity use [9].

Overall, GPUs have been dominant in the industry for training AI models due to their high performance, flexibility, and ability to perform parallel calculations. However, some drawbacks such as high cost and power consumption have led to the development of alternative, more specialized AI accelerators.

Tensor Processing Units (TPUs)

TPUs are application-specific integrated circuits (ASICs) designed by Google to improve machine learning workloads. They are specifically designed to perform matrix operations, making them well suited for machine learning. Machine learning workloads can be run on TPUs using frameworks such as PyTorch and JAX. TPUs are designed as matrix processors specializing in neural network workloads. They are not made for running word processors or basic processes, but can efficiently handle matrix operations in neural networks [10].

The main task of TPUs is to perform matrix processing, which is made up of a combination of operations. TPUs use a Matrix Multiplication Unit (MXU) to perform matrix operations. The MXU contains a grid of multiply-accumulate units arranged in a systolic array architecture. The TPU first loads parameters from high-bandwidth memory (HBM) into the MXU. As multiplication is performed, results are passed between multiply-accumulate units, and the final result is the sum of multiplication operations between data and parameters. This allows TPUs to achieve high computational throughput for neural network calculations. During matrix multiplication, the TPU does not need repeated access to memory because data and intermediate results can move between processing units [10].

Another feature that improves TPU performance is its use of lower-precision number formats for machine learning calculations. Each TensorCore has one or more matrix-multiply units (MXUs), and the number of TensorCores depends on the version of the TPU. MXUs provide most of the computational power within a TensorCore, and each MXU can perform a large number of multiply-accumulate operations per cycle. The multiplication uses bfloat16 inputs, while the accumulation of the results is performed using the FP32 number format. bfloat16 is a 16-bit floating-point format that requires less memory than FP32. It uses less memory, reducing the amount of memory required to store and transfer values. This helps TPUs perform machine learning calculations more efficiently [10].

TPUs can be connected together to deal with machine learning workloads that are too large or computationally demanding for a single chip. This is called a TPU pod, which is a set of TPUs connected together over a network. The number of TPU chips in a pod depends on the TPU version. They use high-speed connections to communicate, allowing workloads to be distributed across multiple TPU chips. This is particularly useful for very large AI models [10].

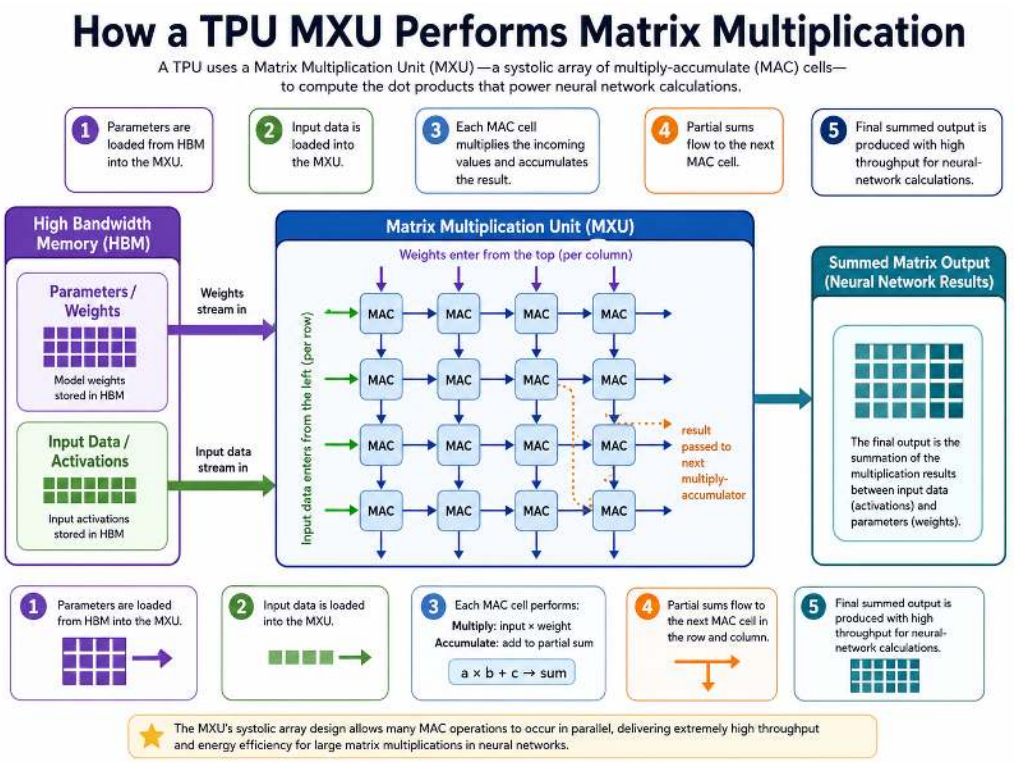

Figure 2. TPU matrix multiplication and MXU data flow. Diagram based on Google Cloud [10].

TPUs come with advantages and disadvantages. TPUs are very good at what they are made for: matrix-heavy neural network workloads in AI systems. They are built for specialized tasks, unlike more general-purpose processors. TPUs are also easy to scale thanks to TPU pods. However, TPUs are less flexible than more general-purpose processors, and their performance depends on whether a workload is well suited to the TPU architecture [10], [11].

Neural Processing Units (NPUs)

Neural processing units (NPUs) are specialized processors designed to accelerate artificial intelligence and machine-learning workloads. Unlike CPUs and GPUs, NPUs are built for AI workloads such as neural networks that use scalar, vector, and tensor math. NPUs are usually used in heterogeneous computing systems that combine multiple processors. Some systems use standalone NPUs, but many devices use NPUs paired with other processors [12], [13].

NPUs are designed to mimic aspects of how neurons process information. NPUs are optimized for parallel processing and use specialized processing cores capable of performing a variety of AI tasks. They have specialized built-in modules for performing multiplication, addition, activation functions, 2D data operations, and decompression. These modules accelerate operations used in neural networks, such as dot products, matrix multiplication, and convolution [12].

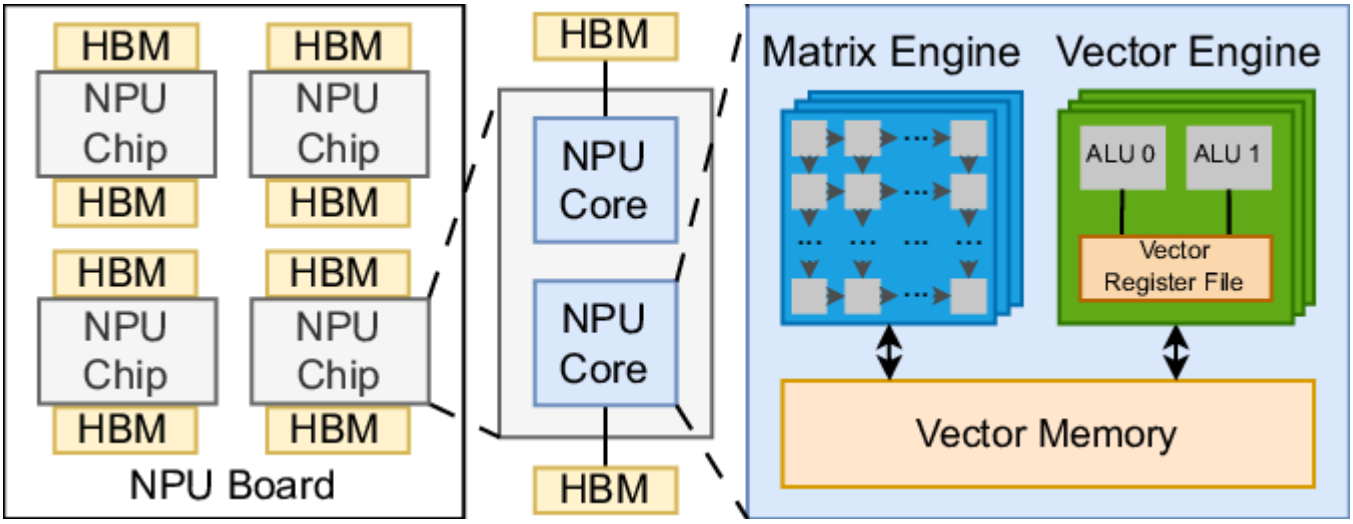

Figure 3. System architecture of a typical NPU board. Source: Xue et al. [14].

NPUs are useful for AI due to their high efficiency and low power consumption. NPUs use specialized processing units and memory that is placed close to the processing cores. This reduces the number of times data has to move between the processor and memory. Some NPUs use lower-precision number formats, which use less memory while maintaining sufficient accuracy for many AI workloads. This allows NPUs, which are tailored to accomplish these processes, to continuously perform neural network calculations while using less energy than less specialized processors [12].

NPUs are commonly paired with other processors that operate together within a heterogeneous computing system. This structure allows each processor to focus on the tasks it is best suited for. These systems often combine an NPU with a CPU and a GPU. Heterogeneous computing can improve efficiency, performance, and battery life. The NPU handles AI tasks, while the CPU deals with general tasks and the GPU handles graphics and many highly parallel workloads. This is especially useful in laptops and smartphones, where NPUs can perform AI tasks while using less power [12], [13].

NPUs come with both advantages and disadvantages. One of their biggest advantages is low power consumption and high efficiency when running neural network workloads. They are also well suited for real-time multimedia processing and can handle inputs such as images, video, and speech. However, NPUs are specialized for AI tasks, making them less flexible than more general-purpose processors. In consumer devices, NPUs are commonly integrated alongside CPUs and GPUs. For larger AI data centers, GPUs may be more suitable because they provide greater flexibility for large-scale workloads [12], [13].

Conclusion

AI accelerators are important because modern AI systems require large amounts of mathematical computation that CPUs may handle less efficiently on their own. GPUs, TPUs, and NPUs each take a different approach to accelerating AI workloads. GPUs are powerful and flexible processors that are mostly used for AI training and inference. TPUs are specialized for machine learning and matrix operations. NPUs are focused on efficient AI processing in devices

such as laptops and phones. Each processor has its own advantages and disadvantages depending on performance, energy efficiency, and the type of workload being performed. As artificial intelligence continues to develop, specialized hardware will continue to play an important role in making AI systems faster and more efficient.

References

[1] A. Koksal, “AI Hardware Accelerators Explained: NPUs, TPUs, and GPUs,” HP Tech Takes. [Online]. Available: https://www.hp.com/us-en/tech-takes/components/explainer/ai-hardware-accelerators-npu-tpu-gpu-guide.html. Accessed: Aug. 29, 2026.
[2] QNAP Marketing Team, “CPU, GPU, NPU, TPU: What Are They?,” QNAP Blog, Mar. 18, 2024. [Online]. Available: https://blog.qnap.com/en/cpu-gpu-npu-tpu-what-are-they/. Accessed: Aug. 29, 2026.
[3] S. Doyle, “AI 101: GPU vs. TPU vs. NPU,” Backblaze, Aug. 2, 2023. [Online]. Available: https://www.backblaze.com/blog/ai-101-gpu-vs-tpu-vs-npu/. Accessed: Aug. 29, 2026.
[4] Cloudflare, “AI inference vs. training: What is AI inference?,” Cloudflare Learning Center. [Online]. Available: https://www.cloudflare.com/learning/ai/inference-vs-training/. Accessed: Aug. 29, 2026.
[5] IBM, “What is parallel computing?,” IBM Think. [Online]. Available: https://www.ibm.com/think/topics/parallel-computing. Accessed: Aug. 29, 2026.
[6] R. Merritt, “Why GPUs Are Great for AI,” NVIDIA Blog, Dec. 4, 2023. [Online]. Available: https://blogs.nvidia.com/blog/why-gpus-are-great-for-ai/. Accessed: Aug. 29, 2026.
[7] R. C. Adhikari, “How GPU-Based AI Processing Works?,” Medium, Jan. 21, 2025. [Online]. Available: https://medium.com/@RC.Adhikari/how-gpu-based-ai-processing-works-ffa29803fdcb. Accessed: Aug. 29, 2026.
[8] A. Shugarts, “Why GPUs Are Essential for AI Workloads,” Rafay, June 3, 2025. [Online]. Available: https://rafay.co/the-kubernetes-current/why-gpus-are-essential-for-ai-workloads. Accessed: Aug. 29, 2026.
[9] T. Marwala, “Rethinking Tech and Why GPUs Are Not the Future of AI Training,” United Nations University, Apr. 2, 2025. [Online]. Available: https://unu.edu/article/rethinking-tech-and-why-gpus-are-not-future-ai-training. Accessed: Aug. 29, 2026.
[10] Google Cloud, “TPU architecture,” Google Cloud Documentation, updated Aug. 26, 2026. [Online]. Available: https://docs.cloud.google.com/tpu/docs/system-architecture-tpu-vm. Accessed: Aug. 29, 2026.
[11] OpenMetal, “TPU vs GPU: Pros and Cons,” OpenMetal Documentation, updated Nov. 5, 2025. [Online]. Available: https://openmetal.io/docs/product-guides/private-cloud/tpu-vs-gpu-pros-and-cons/. Accessed: Aug. 29, 2026.
[12] J. Schneider and I. Smalley, “What is a neural processing unit (NPU)?,” IBM Think. [Online]. Available: https://www.ibm.com/think/topics/neural-processing-unit. Accessed: Aug. 29, 2026.
[13] D. Malladi and P. Lawlor, “What is an NPU? And why is it key to unlocking on-device generative AI?,” Qualcomm OnQ Blog, Feb. 1, 2024. [Online]. Available: https://www.qualcomm.com/news/onq/2024/02/what-is-an-npu-and-why-is-it-key-to-unlocking-on-device-generative-ai. Accessed: Aug. 29, 2026.

[14] Y. Xue, Y. Liu, L. Nai, and J. Huang, “Hardware-Assisted Virtualization of Neural Processing Units for Cloud Platforms,” in Proc. 57th IEEE/ACM Int. Symp. Microarchitecture (MICRO), 2024. [Online]. Available: https://arxiv.org/abs/2408.04104. Accessed: Aug. 29, 2026.
[15] Google Cloud, “AI accelerator performance and benchmarking,” Google Cloud Documentation. [Online]. Available: https://docs.cloud.google.com/docs/ai-ml/accelerator-performance-benchmarking. Accessed: Aug. 29, 2026.