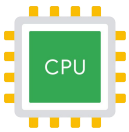
Google Cloud

AI Infrastructure: Cloud GPU

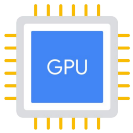
Congratulations on completing the course. This course summary is your review guide.

The performance-optimized, purpose-built hardware essential for efficient AI deployment consists of three core components: networking, storage, and compute.

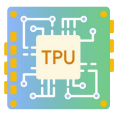
Compute | Storage | Networking



A **central processing unit (CPU)** is the primary component of a computer that executes instructions.



A **graphics processing unit (GPU)** is designed for handling a large number of calculation tasks, making it ideal for tasks like graphics rendering and machine learning.



A **Tensor Processing Unit (TPU)** is a specialized hardware accelerator designed by Google specifically for neural network machine learning tasks, offering significant speed and efficiency improvements over CPUs and GPUs for these specific workloads.

Accelerating frameworks

- CUDA(Compute Unified Device Architecture) is NVIDIA's parallel computing platform and programming model.
- Accelerated Linear Algebra (XLA) is a domain-specific compiler developed originally by Google.
- PyTorch/XLA: This integration allows PyTorch models to leverage XLA's optimizations.
- JAX/XLA: JAX is designed from the ground up with XLA as its core compilation backend.

GPU clusters provisioning options

Google Cloud offers multiple platforms for provisioning GPU-accelerated clusters, each designed to support different workloads, levels of control, and scaling needs. In this lesson, you'll explore five of the most widely used platforms and learn how to match them to the requirements of your AI and ML workloads.

GPU options:

- Pre-training - A4, A3 Ultra, A3 Mega, A3 High, A2
- Fine-tuning - A4, A3 Ultra, A3 Mega, A3 High, A2
- Serving inference - A4, A3 Ultra, A3 Mega, A3 High, A2
- Graphics-intensive workloads - G2, N1+T4
- High performance computing - The best fit depends on the amount of computation that must be offloaded to the GPU.

When making your GPU decision, consider these key factors:

- Distinguish between training or fine-tuning and inferencing.
- The scale directly influence the recommended hardware tier.
- Latency is a deciding factor
- Validate performance for customer scenario

Additional Resources

1. AI Hypercomputer documentation
2. GPUs on compute engine

Google Cloud