The Best GPU for Deep Learning

Critical Considerations for Large-Scale AI

Traditionally, the training phase of the deep learning pipeline takes the longest to achieve. This is not only a timeconsuming process, but an expensive one. The most valuable part of a deep learning pipeline is the human element. Data scientists often wait for hours or days for training to complete, which hurts their productivity and the time to bring new models to market.

To significantly reduce training time, you can use deep learning GPUs, which enable you to perform AI computing operations in parallel. When assessing GPUs, you need to consider the ability to interconnect multiple GPUs, the supporting software available, licensing, data parallelism, GPU memory use and performance.

In this guide, you will learn:

Why Are GPUs Important in Deep Learning?

The longest and most resource intensive phase of most deep learning implementations is the training phase. This phase can be accomplished in a reasonable amount of time for models with smaller numbers of parameters but as your number increases, your training time does as well. This has a dual cost; your resources are occupied for longer and your team is left waiting, wasting valuable time.

Graphical processing units (GPUs) can reduce these costs, enabling you to run models with massive numbers of parameters quickly and efficiently. This is because GPUs enable you to parallelize your training tasks, distributing tasks over clusters of processors and performing compute operations simultaneously.

GPUs are also optimized to perform target tasks, finishing computations faster than non-specialized hardware. These processors enable you to process the same tasks faster and free your CPUs for other tasks. This eliminates bottlenecks created by compute limitations.

How to Choose the Best GPU for Deep Learning?

Selecting the GPUs for your implementation has significant budget and performance implications. You need to select GPUs that can support your project in the long run and have the ability to scale through integration and clustering. For large-scale projects, this means selecting production-grade or data center GPUs.

GPU Factors to Consider

These factors affect the scalability and ease of use of the GPUs you choose:

Ability to Interconnect GPUs

When choosing a GPU, you need to consider which units can be interconnected. Interconnecting GPUs is directly tied to the scalability of your implementation and the ability to use multi-GPU and distributed training strategies. Typically, consumer GPUs do not support interconnection (NVlink for GPU interconnects within a server, and Infiniband/RoCE for linking GPUs across servers) and NVIDIA has removed interconnections on GPUs below RTX 2080.

Supporting Software

NVIDIA GPUs are the best supported in terms of machine learning libraries and integration with common frameworks, such as PyTorch or TensorFlow. The NVIDIA CUDA toolkit includes GPU-accelerated libraries, a C and C++ compiler and runtime, and optimization and debugging tools. It enables you to get started right away without worrying about building custom integrations.

Learn more in our guides about PyTorch GPUs, and NVIDIA deep learning GPUs.

Licensing

Another factor to consider is NVIDIA's guidance regarding the use of certain chips in data centers. As of a licensing update in 2018, there may be restrictions on use of CUDA software with consumer GPUs in a data center. This may require organizations to transition to production-grade GPUs.

Algorithm Factors Affective GPU Use

In our experience helping organizations optimize large-scale deep learning workloads, the following are the three key factors you should consider when scaling up your algorithm across multiple GPUs.

- **Data Parallelism** – Consider how much data your algorithms need to process. If datasets are going to be large, invest in GPUs capable of performing multi-GPU training efficiently. For very large scale datasets, make sure that servers can communicate quickly with each other and with storage components, using technology like Infiniband/RoCE, to enable efficient distributed training.

- **Memory Use** – Are you going to deal with large data inputs to model? For example, models processing medical images or long videos have very large training sets, so you'd want to invest in GPUs with relatively large memory. By contrast, tabular data such as text inputs for NLP models are typically small, and you can make do with less GPU memory.

- **Performance of the GPU** – Consider if you're going to use GPUs for debugging and development. In this case you won't need the most powerful GPUs. For tuning models in long runs, you need strong GPUs to accelerate training time, to avoid waiting hours or days for models to run.

Using Consumer GPUs for Deep Learning

While consumer GPUs are not suitable for large-scale deep learning projects, these processors can provide a good entry point for deep learning. Consumer GPUs can also be a cheaper supplement for less complex tasks, such as model planning or low-level testing. However, as you scale up, you'll want to consider data center grade GPUs and high-end deep learning systems like NVIDIA's DGX series (learn more in the following sections).
In particular, the Titan V has been shown to provide performance similar to datacenter-grade GPUs when it comes to Word RNNs. Additionally, its performance for CNNs is only slightly below higher tier options. The Titan RTX and RTX 2080 Ti aren't far behind.

NVIDIA Titan V

The Titan V is a PC GPU that was designed for use by scientists and researchers. It is based on NVIDIA's Volta technology and includes Tensor Cores. The Titan V comes in Standard and CEO Editions.

The Standard edition provides 12GB memory, 110 teraflops performance, a 4.5MB L2 cache, and 3,072-bit memory bus. The CEO edition provides 32GB memory and 125 teraflops performance, 6MB cache, and 4,096-bit memory bus. The latter edition also uses the same 8-Hi HBM2 memory stacks that are used in the 32GB Tesla units.

NVIDIA Titan RTX

The Titan RTX is a PC GPU based on NVIDIA's Turing GPU architecture that is designed for creative and machine learning workloads. It includes Tensor Core and RT Core technologies to enable ray tracing and accelerated AI.

Each Titan RTX provides 130 teraflops, 24GB GDDR6 memory, 6MB cache, and 11 GigaRays per second. This is due to 72 Turing RT Cores and 576 multi precision Turing Tensor Cores.

NVIDIA GeForce RTX 2080 Ti

The GeForce RTX 2080 Ti is a PC GPU designed for enthusiasts. It is based on the TU102 graphics processor. Each GeForce RTX 2080 Ti provides 11GB of memory, a 352-bit memory bus, a 6MB cache, and roughly 120 teraflops of performance.

Best Deep Learning GPUs for Large-Scale Projects and Data Centers

The following are GPUs recommended for use in large-scale AI projects.

NVIDIA Tesla A100

The A100 is a GPU with Tensor Cores that incorporates multi-instance GPU (MIG) technology. It was designed for machine learning, data analytics, and HPC. The Tesla A100 is meant to be scaled to up to thousands of units and can be partitioned into seven GPU instances for any size workload. Each Tesla A100 provides up to 624 teraflops performance, 40GB memory, 1,555 GB memory bandwidth, and 600GB/s interconnects.

NVIDIA Tesla V100

The NVIDIA Tesla V100 is a Tensor Core enabled GPU that was designed for machine learning, deep learning, and high performance computing (HPC). It is powered by NVIDIA Volta technology, which supports tensor core technology, specialized for accelerating common tensor operations in deep learning. Each Tesla V100 provides 149 teraflops of performance, up to 32GB memory, and a 4,096-bit memory bus.

NVIDIA Tesla P100

The Tesla P100 is a GPU based on an NVIDIA Pascal architecture that is designed for machine learning and HPC. Each P100 provides up to 21 teraflops of performance, 16GB of memory, and a 4,096-bit memory bus.

NVIDIA Tesla K80

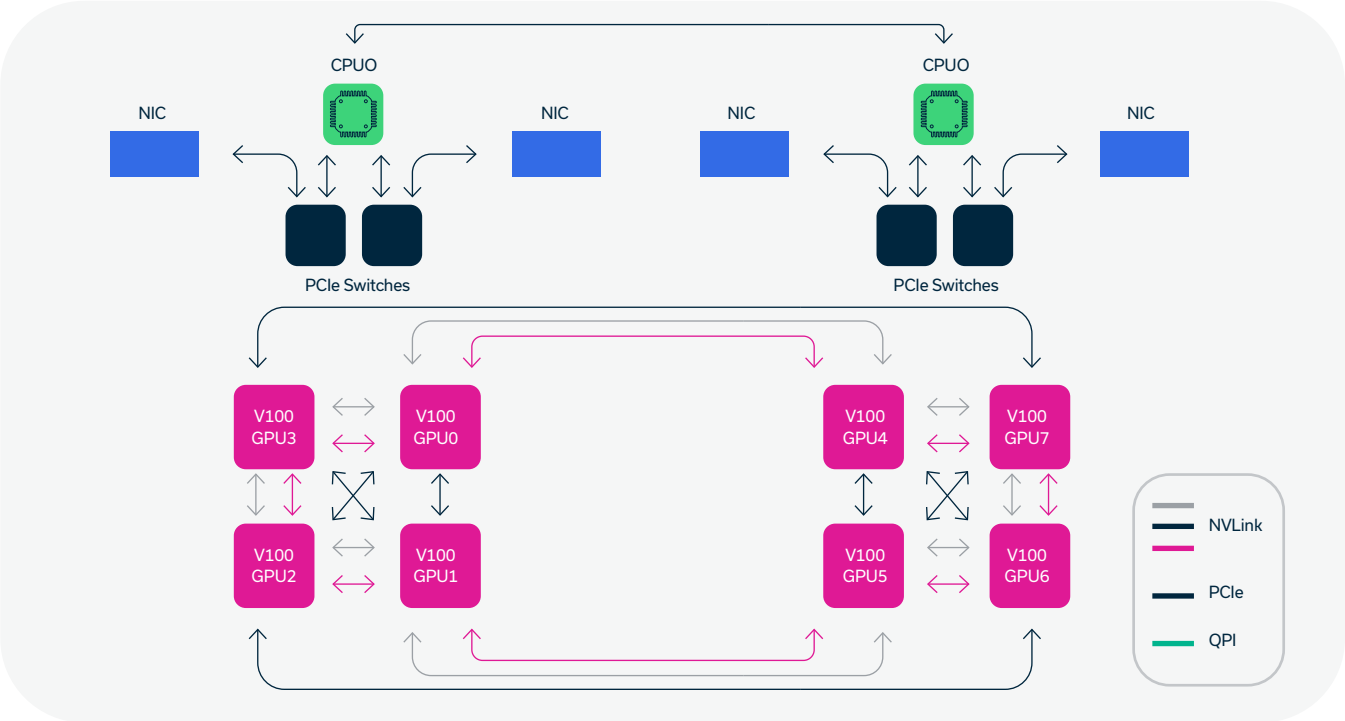
The Tesla K80 is a GPU based on the NVIDIA Kepler architecture that is designed to accelerate scientific computing and data analytics. It includes 4,992 NVIDIA CUDA cores and GPU Boost™ technology. Each K80 provides up to 8.73 teraflops of performance, 24GB of GDDR5 memory, and 480GB of memory bandwidth.

Google TPU

Slightly different are Google's tensor processing units (TPUs). TPUs are chip or cloud-based, application-specific integrated circuits (ASIC) for deep learning. These units are specifically designed for use with TensorFlow and are available only on Google Cloud Platform.

Each TPU can provide up to 420 teraflops of performance and 128 GB high bandwidth memory (HBM). There are also pod versions available that can provide over 100 petaflops of performance, 32TB HBM, and a 2D toroidal mesh network.

DGX for Deep Learning at Scale



The NVIDIA DGX systems are full stack solutions designed for enterprise-grade machine learning. These systems are based on a software stack that is optimized for AI, multi-node scalability, and enterprise-grade support.

You can implement the DGX stack in containers or on bare metal. This technology is meant to be plug-and-play and is fully integrated with NVIDIA deep learning libraries and software solutions. DGX is available for server-class workstations, servers, or pods. Below, the server options are introduced.

DGX-1

The DGX-1 is a GPU server based on the Ubuntu Linux Host OS. It integrates with Red Hat solutions and includes the DIGITS deep learning training application, the NVIDIA Deep Learning SDK, the CUDA toolkit, and the Docker Engine Utility for NVIDIA GPU.

Each DGX-1 provides:

- Two Intel Xeon CPUs for deep learning framework coordination, boot, and storage management
- Up to 8 Tesla V100 Tensor Cores GPUs with 32GB of memory
- 300Gb/s NVLink interconnects
- 800GB/s communication with low-latency
- Single 480GB boot OS SSD and four 1.92 TB SAS SSDs (7.6 TB total) configured as a RAID 0 striped volume

DGX-2

The DGX-2 is the next level up from the DGX-1. It is based on the NVSwitch networking fabric for greater parallelism and scalability.

Each DGX-2 provides:

- Two petaflops of performance
- 2X 960GB NVME SSDs for OS storage and 30TB of SSD storage
- 16 Tesla V100 Tensor Core GPUs with 32GB of memory
- 1.6TB/s low-latency, bi-directional bandwidth
- 1.5TB system memory
- Two Xeon Platinum CPUs for deep learning framework coordination, boot, and storage
- Two high I/O ethernet cards

DGX A100

The DGX A100 is designed to be a universal system for machine learning workloads, including analytics, training, and inference. It is fully optimized for CUDA-X. The DGX A100 can be stacked with other A100 units to create massive AI clusters, including the NVIDIA DGX SuperPOD.

Each DGX A100 provides:

- Five petaflops of performance
- Eight A100 Tensor Core GPUs with 40GB memory
- Six NVSwitches for 4.8TB bi-directional bandwidth
- Two 64-core AMD CPUs for deep learning framework coordination, boot, and storage
- 1TB system memory, 2x 1.92TB M.2 NVME drives for OS storage and 15TB SSD storage

Automated Deep Learning GPU Management With Run:ai

Run:ai automates resource management and workload orchestration for machine learning infrastructure. With Run:ai, you can automatically run as many compute intensive experiments as needed.

Here are some of the capabilities you gain when using Run:ai:

- **Advanced visibility**: create an efficient pipeline of resource sharing by pooling GPU compute resources
- **No more bottlenecks**: you can set up guaranteed quotas of GPU resources, to avoid bottlenecks and optimize billing
- **A higher level of control**: Run:ai enables you to dynamically change resource allocation, ensuring each job gets the resources it needs at any given time.

Run:AI accelerates deep learning on GPU by helping data scientists optimize expensive compute resources and improve the quality of their models.

See Our Additional Guides on Key Artificial Intelligence Infrastructure Topics

We have authored in-depth guides on several other artificial intelligence infrastructure topics that can also be useful as you explore the world of deep learning GPUs.

MLOps

In today's highly competitive economy, enterprises are looking to artificial intelligence in general and machine and deep learning in particular to transform big data into actionable insights that can help them better address their target audiences, improve their decision-making processes, and streamline their supply chains and production processes, to mention just a few of the many use cases out there. In order to stay ahead of the curve and capture the full value of ML, however, companies must strategically embrace MLOps.

See top articles in our MLOps guide:

- Machine Learning Ops: What Is It and Why We Need It
- Machine Learning Automation: Speeding Up the Data Science Pipeline
- Machine Learning Workflow: Streamlining Your ML Pipeline

Kubernetes and AI

This guide explains the Kubernetes architecture for AI workloads and how K8s came to be used inside many companies. There are specific considerations for implementing Kubernetes to orchestrate AI workloads. Finally, the guide addresses the shortcomings of Kubernetes when it comes to scheduling and orchestration of deep learning workloads and how you can address those shortfalls.

See top articles in our Kubernetes for AI guide:

- Kubernetes Architecture – w Understanding Kubernetes Architecture for Data Science Workloads
- The Challenges of Scheduling AI Workloads on Kubernetes