# NVIDIA DATA CENTER PLATFORM
# ONE PLATFORM.
# UNLIMITED ACCELERATION.

## The Exponential Growth of Computing

Accelerated computing is being rapidly adopted across industries and large-scale production deployments. Because new compute demands are outstripping the capabilities of traditional CPU-only servers, enterprises need to optimize their data centers—making this acceleration a must-have. The NVIDIA data center platform is the world's leading accelerated computing solution, deployed by the largest supercomputing centers and enterprises. It enables breakthrough performance with fewer, more powerful servers, driving faster time to insights while saving money.

The platform accelerates a broad array of workloads, from AI training and inference to scientific computing and virtual desktop infrastructure (VDI) applications, with a diverse range of GPUs. For optimal performance, it's essential to identify the ideal GPU for a specific workload. A guide to those workloads and the corresponding NVIDIA GPUs that deliver the best results is provided on the next page.

# Choose the Right NVIDIA Data Center GPU for You

| WORKLOAD | DESCRIPTION | NVIDIA A100 Tensor Core GPU SXM4 | NVIDIA A100 Tensor Core GPU PCIe | NVIDIA V100 Tensor Core GPU | NVIDIA T4 Tensor Core GPU | NVIDIA Quadro RTX™ 6000/8000 |
|---|---|---|---|---|---|---|
| | | Recommended number of GPUs per workload | | | | |
| Deep Learning Training | For the absolute fastest model training time | 8–16 GPUs | 4-8 GPUs | 8–16 GPUs | | |
| Deep Learning Inference | For batch and real-time inference | 1 GPU with Multi-Instance GPU (MIG) | 1 GPU with MIG | | 1 GPU | |
| High-Performance Computing (HPC) | For scientific computing centers and higher education and research institutions | 4 GPUs with MIG for supercomputing centers | 1-4 GPUs with MIG for higher education and research use cases | 4 GPUs for high-performance computing centers | | |
| Render Farms | For batch and real-time rendering | | | | | 4–8 GPUs |
| Graphics | For the best graphics performance on professional virtual workstations | | | | 2–8 GPUs for mid-range virtual workstations (e.g., ProE, Autodesk) or mainstream graphics (e.g., CATIA) | 4–8 GPUs for running graphics and simulation applications (e.g., CATIA and live rendering) |
| Enterprise Acceleration | For enterprises running mixed workloads (e.g., graphics, machine learning, deep learning, data science, and analytics) | 1-4 GPUs with MIG for compute-intensive, multiple-GPU workloads | 1-4 GPUs with MIG for compute-intensive, single-GPU workloads | 4 GPUs for compute-intensive, multi-GPU workloads (SXM2) or single-GPU workloads (PCIe) | 4–8 GPUs for balanced workloads | 2–4 GPUs for graphics-intensive workloads |
| Edge Acceleration | For deploying AI to the edge with multiple use cases and locations | | 1 GPU with MIG | | 1–8 GPUs for inference and video-code-intensive (e.g., intelligent video analytics, industrial inspection) workloads | 2–4 GPUs for graphics-intensive workloads (e.g., augmented reality, virtual reality |
| KEY FEATURES | | > 624 teraFLOPS* of mixed-precision tensor operations for AI training<br>> 312 teraFLOPS* of TF32 for single-precision AI training<br>> 1,248 teraOPS* of INT8 performance for AI inference<br>> 19.5 teraFLOPS of double-precision performance<br>> 40 GB HBM2 memory<br>> 600 GB/s** NVIDIA® NVLink® interconnect bandwidth<br>> 1.6 TB/s memory bandwidth<br>> Up to 7 MIG instances per GPU<br>> 250 W (PCIe), 400 W (SXM4 via NVIDIA HGX™ A100) options<br>> Delivered performance for top apps: 100% (SXM4), 90% (PCIe)<br>* With sparsity<br>** SXM GPUs via HGX A100 server boards, PCIe GPUs via NVLink Bridge for up to 2-GPUs | | Specs for NVIDIA V100S:<br>> 32 GB HBM2 memory<br>> 130 teraFLOPS of mixed-precision tensor operations for deep learning<br>> 16.4 teraFLOPS of single-precision performance<br>> 8.2 teraFLOPS of double-precision performance<br>> 300 GB/s NVIDIA NVLink interconnect bandwidth<br>> 1,134 GB/s memory bandwidth | > 16 GB memory<br>> 130 teraOPS of INT8 inference performance<br>> 8.1 teraFLOPS of single-precision performance<br>> Dedicated video decode and encode engines<br>> 70 W power<br>> Low-profile form factor | > 24 / 48 GB memory<br>> 130 teraFLOPS of tensor operations for deep learning<br>> 261 teraOPS of INT8 inference performance<br>> 16 teraFLOPS of single-precision performance<br>> RT Cores for high-performance rendering<br>> 100 GB/s NVIDIA NVLink interconnect bandwidth |

**www.nvidia.com/en-us/data-center/a100/**
**www.nvidia.com/en-us/data-center/v100/**