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 A40	NVIDIA T4 Tensor Core GPU
		Recommended number of GPUs per workload			
Deep Learning Training	For the absolute fastest model training time	8–16 GPUs > 80GB: For largest models (DLRM, GPT-2 over 9.3Bn parameters in one node)	4-8 GPUs		
Deep Learning Inference	For batch and real-time inference	1 GPU with Multi-Instance GPU (MIG) > 80GB: 7 MIGs at 10GB each for large batch-size-constrained models (RNN-T)	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 > 80GB: For largest datasets and high-memory-bandwidth applications	1-4 GPUs with MIG for higher education and research use cases		
Render Farms	For batch and real-time rendering			4–8 GPUs	
Graphics	For the best graphics performance on professional virtual workstations			1–8 GPUs for mid-range and high-end professional graphics and NVIDIA RTX™ workloads or simulation	1–8 GPUs for mid-range virtual workstations for professional graphics
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 > 80GB: data analytics with largest datasets	1-4 GPUs with MIG for compute-intensive, single-GPU workloads	2–4 GPUs for mid-range to high-end virtual workstations for professional graphics and compute workloads	4–8 GPUs for balanced workloads
Edge Acceleration	For deploying AI to the edge with multiple use cases and locations		1 GPU with MIG	2–4 GPUs with virtual workstation for graphics-intensive workloads, including AR and VR	1–8 GPUs for inference and video-code-intensive (e.g., intelligent video analytics, industrial inspection) workloads
KEY FEATURES		> 624 teraFLOPS* of mixed-precision tensor operations for AI training > 312 teraFLOPS* of TF32 for single-precision AI training > 1,248 teraOPS* of INT8 performance for AI inference > 19.5 teraFLOPS of double-precision performance > 40GB HBM2 (1.6TB/s) or 80GB HBM2e memory (2TB/s) > 600GB/s** NVIDIA® NVLink® interconnect bandwidth > Up to 7 MIG instances per GPU > 250W (PCIe), 400 W (SXM4 via NVIDIA HGX™ A100) options > Delivered performance for top apps: 100% (SXM4), 90% (PCIe)		> 48GB GDDR6 GPU memory > Fastest rendering > NVIDIA Ampere Architecture RT Cores, Tensor Cores > Largest 3D models and professional RTX graphics with virtual workstations > 300W > 37.4 teraFLOPS of single-precision FP32 performance > 73.1 teraFLOPS of ray-tracing performance > 149.6* teraFLOPS of TF32 for single-precision AI training > 112GB/s NVIDIA NVLink (bidirectional) interconnect bandwidth	> 16GB memory > 130 teraOPS of INT8 inference performance > 8.1 teraFLOPS of single-precision performance > Dedicated video decode and encode engines > 70W power > Low-profile form factor

www.nvidia.com/en-us/data-center/a100/
www.nvidia.com/en-us/data-center/a40

* With sparsity
** SXM GPUs via HGX A100 server boards, PCIe GPUs via NVLink Bridge for up to two GPUs