NVIDIA DATA CENTER PLATFORM

Accelerating Every Workload

Modern applications are transforming every business. From AI for better customer engagement, to data analytics for forecasting, to advanced visualization for product innovation, the need for accelerated computing is rapidly increasing. 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 from the highest performance to entry-level, all powered by a single unified architecture. For optimal performance, it's essential to identify the ideal GPU for a specific workload. Use this as a guide to those workloads and the corresponding NVIDIA GPUs that deliver the best results.

Choose the Right Data Center GPU

WORKLOAD	DESCRIPTION	NVIDIA A100 SXM | PCIe	NVIDIA A30	NVIDIA A2	NVIDIA A40	NVIDIA A16	NVIDIA A100X	NVIDIA A30X
		Highest Perf Compute	Mainstream Compute	Entry-Level Compact AI	Highest Perf Graphics	Optimized for VDI	Highest Perf Converged Accelerator	Mainstream Converged Accelerator
		Recommended Number of GPUs or Converged Cards per Server						
Deep Learning (DL) Training and Data Analytics	For the absolute fastest model training and analytics	SXM PCIe 4 or 8 GPUs (PCIe | SXM) > 80GB: Bn+ parameter models (DLRM, GPT-3)					1-2 cards for multi-node training	
DL Inference	For batch and real-time inference	SXM PCIe 1-2 GPUs (PCIe) | 4-8 GPUs (SXM) with multi-instance GPU (MIG) > 80GB: large batch size constrained models (RNN-T)	2–4 GPUs with MIG	1-4 GPUs				
High-Performance Computing (HPC) / AI	For Higher Education Research and scientific computing centers	SXM PCIe 2-4 GPUs (PCIe) | 4 GPUs (SXM) with MIG	2–4 GPUs with MIG				1-2 cards for multi-node workloads	
Render Farms	For batch and real-time rendering				4–8 GPUs			
Graphics	For the best graphics performance on professional VDI			1-4 GPUs for entry-level virtual workstations*	2–4 GPUs for midrange to high-end virtual workstations*	2–4 GPUs for highest virtual desktop and workstation user density**		
Cloud Gaming	4K resolution / Android			1-4 GPUs for mobile Android	4–8 GPUs for (4K resolution)			
Enterprise Acceleration	For mixed workloads, including graphics, ML, DL, analytics, training, and inference	PCIe 1-2 GPUs with MIG for compute workloads	1-2 GPUs with MIG for compute workloads	1-4 GPUs for balanced workloads*	1-2 GPUs for graphics-intensive workloads*			1 card for compute acceleration with software-defined infrastructure
Edge Acceleration	For differing use cases and deployment locations	PCIe 1-2 GPUs with MIG	1-2 GPUs with MIG	1-4 GPUs for inference and video workloads	1-4 GPUs for graphics-intensive workloads & AR / VR*		1 card for AI-on-5G with heavy workloads	1 card for AI-on-5G with average workloads
5G vRAN	For low-latency GPU-network communication							1-2 cards
AI-Based Security	For GPU-powered network processing							1 card

* NVIDIA RTX Virtual Workstation (vWS) software license required for virtual workstation workloads.
** NVIDIA Virtual PC (vPC) software license required for VDI workloads.

NVIDIA