NVIDIA DATA CENTER PLATFORM

Accelerating Every Workload

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, 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 A40	NVIDIA A10	NVIDIA T4	NVIDIA A16
		Highest Perf Compute	Mainstream Compute	Highest Perf Graphics	Mainstream Graphics	Small Footprint Low Power	Optimized for VDI
		Recommended Number of GPUs per Server					
Deep Learning (DL) Training and Data Analytics	For the absolute fastest model training and analytics	SXM PCIe 4-8 GPUs > 80GB: Bn+ parameter models (DLRM, GPT-2)					
DL Inference	For batch and real-time inference	SXM PCIe 1-2 GPUs w/ multi-instance GPU (MIG) > 80GB: large batch size constrained models (RNN-T)	2–4 GPUs with MIG		4–8 GPUs	4–8 GPUs	
High-Performance Computing (HPC) / AI	For Higher Education Research and scientific computing centers	SXM 1-4 GPUs with MIG	2–4 GPUs with MIG				
Render Farms	For batch and real-time rendering			4–8 GPUs	4–8 GPUs		
Graphics	For the best graphics performance on professional VDI			2–4 GPUs for high-end virtual workstations*	2-8 GPUs for mid-range virtual workstations*	2-8 GPUs for entry-level virtual workstations*	2–4 GPUs for highest virtual desktop user density**
Cloud Gaming	For 4K resolution / Android			4–8 GPUs (4K resolution)	4–8 GPUs (4K resolution)	1-2 GPUs (Android)	
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	1-2 GPUs for graphics-intensive workloads*	1-2 GPUs for graphics-intensive* and compute workloads	1-4 GPUs for balanced workloads*	
Edge Acceleration	For differing use cases and deployment locations	PCIe 1-2 GPUs with MIG	1-2 GPUs with MIG	1-4 GPUs for graphics-intensive workloads & AR / VR*	1-8 GPUs for inference and video workloads	1-8 GPUs for inference and video workloads	

To learn more about Data Center GPUs, visit **www.nvidia.com/ampere**

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

NVIDIA