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

Solution Category	GPU	DL Training & Data Analytics	DL Inference	HPC / AI	Omniverse/ Render Farms	Virtual Workstation	Virtual Desktop (VDI)	Mainstream Acceleration	Far Edge Acceleration	AI-on-5G
Compute	H100*	PCIE SXM — Best	PCIE SXM — Best	PCIE SXM — Best				PCIE — Best		
	A100	PCIE SXM A100X — Better	PCIE SXM — Better	PCIE SXM A100X — Better				PCIE A100X — Better		A100X — Better
	A30		PCIE — Good	PCIE — Good				PCIE — Good		A30X — Good
Graphics / Compute	L40		Best		Best	Best		Best		
	A40				Better	Better		Better		
	A10		Good		Good	Best	Better	Good	Better	
	A16					Good	Best			
Small Form Factor Compute/ Graphics	A2		Best			Best	Best	Best	Best	
	T4		Better			Better	Better	Better	Better	

To learn more about NVIDIA data center GPUs, visit
nvidia.com/data-center

NVIDIA.