Line Card

NVIDIA Data Center Platform

Accelerate every workload.

Rapid developments and continuous breakthroughs in AI are fueling transformative change, spanning all industries and revolutionizing the workflows of scientists, engineers, creators, and more. On top of the demand for accelerated computing to power traditional AI applications—such as machine learning, deep learning, natural language processing, and computer vision—a new use case has emerged that's unlocking a frontier of opportunities—generative AI. The NVIDIA data center platform is the world's leading accelerated computing and generative AI 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 generative AI training and inference to scientific computing and virtual desktop infrastructure (VDI) applications, with a diverse range of GPUs, from the highest performing 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.

GPU Portfolio: NVIDIA Hopper™ and Ada Lovelace Architectures

Solution Category	GPU	Networking Solutions	Deep Learning Training and Data Analytics	Deep Learning Inference	HPC / AI	NVIDIA Omniverse™ / Render Farms	Virtual Workstation	Virtual Desktop (VDI)	AI Video	Far-Edge Acceleration
Compute	**GH200**	QTM2 SPTM4	■	■	■					
	H100	QTM2 SPTM4	■	🟩	■					
Graphics and Compute	**L40S**	QTM1 SPTM3	🟩	🟩	🟩	🟩	🟩		🟩	
	L40	SPTM3				■	■			
Small Form Factor (SFF) Compute and Graphics	**L4**	SPTM3		🟨		🟨	■	■	■	■

Price-performance comparison within each solution category (Compute, Graphics and Compute, SFF Compute and Graphics) and workload column.

■ Best

🟩 Better

🟨 Good

🟦 NVIDIA Quantum-1 IB switch plus BlueField-2 DPUs or NVIDIA ConnectX®-6/6 Dx SmartNICs

🟦 NVIDIA Quantum-2 IB switch plus BlueField-3 DPUs or ConnectX-7 SmartNICs

🟦 NVIDIA Spectrum™-3 Ethernet switch plus BlueField-2 DPUs or ConnectX-6/6 Dx SmartNICs

🟦 NVIDIA Spectrum-4 Ethernet switch plus BlueField-3 DPUs or ConnectX-7 SmartNICs

GPU Portfolio: NVIDIA Ampere Architecture

Solution Category	GPU	Networking Solutions	Deep Learning Training and Data Analytics	Deep Learning Inference	HPC / AI	NVIDIA Omniverse™ / Render Farms	Virtual Workstation	Virtual Desktop (VDI)	AI Video	Far-Edge Acceleration	AI-on-5G
Compute	A100	QTM1 SPTM3	■	■	■						AX800 | A100X ■
	A30	SPTM3		▣	▣						A30X □
Graphics and Compute	A40	SPTM3				■	▣		▣		
	A10	SPTM3		▣		▣	■	▣	■	▣	
	A16	SPTM3					□	■	▣		
Small Form Factor (SFF) Compute and Graphics	A2	SPTM3		□			■	■	■	■	

Price-performance comparison within each solution category (Compute, Graphics and Compute, SFF Compute and Graphics) and workload column.

■ Best

▣ Better

□ Good

🟪 A100X/A30X converged accelerators

🟦 NVIDIA Quantum-1 IB switch plus BlueField-2 DPUs or NVIDIA ConnectX®-6/6 Dx SmartNICs

🟦 NVIDIA Spectrum™-3 Ethernet switch plus BlueField-2 DPUs or ConnectX-6/6 Dx SmartNICs

NVIDIA Inference Portfolio

GPU	NLP/LLM				Image/Video Generative AI	Recsys	Graph / Vector Database	Computer Vision	AI Video
	Up to 5B	6B to 65B	66B to 175B	> 175B					
GH200	■	■	▨	□	▨	■	■		
HGX H100 (8-way)	■	■	■	■	▨	▨	▨		
L40S	▨	▨	□		■			▨	▨
L4	□				□			■	■

Price-performance comparison relative across each entire workload column. This chart should be used in conjunction with measured data for targeted workloads.

■ Best
▨ Better
□ Good

NVIDIA Training Portfolio

GPU	NLP/LLM				Image/Video Generative AI	Recsys
	Up to 5B	6B to 65B	66B to 175B	> 175B		
GH200	■	■	■	🟩	🟩	■[1]
HGX H100 (8-way)	■	■	■	■	■	🟩
L40S	🟩	🟩	🟩	🟨	🟩	🟨

[1] Comparison for 256 GPU + CPU NVLink connected DGX GH200 system.

Price-performance comparison relative across each entire workload column. This chart should be used in conjunction with measured data for targeted workloads.

■	Best
🟩	Better
🟨	Good

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

NVIDIA.