

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	Training and Data Analytics	Inference	HPC/AI	NVIDIA Omniverse™ / Render Farms	Virtual Workstations	Virtual Desktops (VDI)	AI Video	Far-Edge Acceleration
Compute	GH200	QTM2 SPCX	Best	Best	Best					
	NVIDIA HGX™ H200	QTM2 SPCX	Best	Best	Best					
	HGX H100	QTM2 SPCX	Best	Good	Best					
	H100 NVL	QTM2 SPCX	Good	Good	Good					
Graphics and Compute	L40S	QTM2 SPCX	Good	Good	Good	Best	Good	Good	Good	Good
Small Form Factor (SFF) Compute and Graphics	L4	SPC3		Good		Good	Best	Best	Best	Best

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

 Best
 Good

 NVIDIA Quantum-2 InfiniBand switch plus ConnectX®-7 NICs and/or NVIDIA® BlueField®-3 DPUs

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

 Spectrum SN5600 Ethernet switch plus BlueField-3 SuperNICs

NVIDIA Inference Portfolio

GPU	Natural Language Processing (NLP) / Large Language Models (LLMs)				Image/Video Generative AI	Recommender Systems	Graph/Vector Database	Computer Vision	AI Video
	Up to 5 Billion (B) Parameters	6B-65B Parameters	66B-175B Parameters	Over 175B Parameters					
GH200	■	■	■	■	■	■	■		
HGX H200 (8-way)	■	■	■	■	■	■	■		
HGX H100 (8-way)	■	■	■	■	■	■	■		
H100 NVL	■	■	■	■	■				
L40S	■	■	■		■			■	■
L4	■				■			■	■

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



NVIDIA Training Portfolio

GPU	NLP/LLM				Image/Video Generative AI	Recommender Systems	Graph/Vector Database
	Up to 5B Parameters	6B-65B Parameters	66B-175B Parameters	Over 175B Parameters			
GH200	Best	Best	Best	Good	Good	Best	Best
HGX H200/H100 (8-way)	Good	Best	Best	Best	Best	Best	Good
H100 NVL	Good	Good	Good	Good	Good	Good	Good
L40S	Good	Good	Good	Good	Good	Good	

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



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

© 2024 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, BlueField, ConnectX, HGX, Hopper, Omniverse, and Spectrum are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 3267213. MAY24

