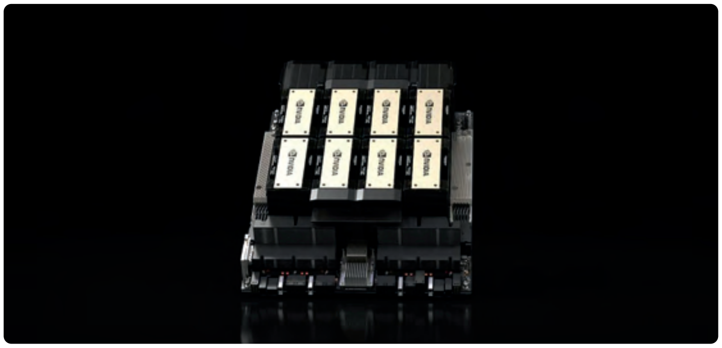
NVIDIA Tensor Core GPUs - Blackwell & Hopper

The Gold Standard for AI Computing Infrastructures

NVIDIA HGX B200



NVIDIA HGX H200

Powering the Next Generation of AI

Artificial Intelligence has transformed the way businesses operate by automating tasks, generating insights, enabling innovation, and increasing productivity. As AI becomes adopted by the mass market, increased advancements will require more advanced hardware.

NVIDIA Tensor Core GPUs are the gold standard GPU for AI computations, featuring NVIDIA Blackwell and NVIDIA Hopper. Designed specifically to execute the calculations found in AI and neural networks, whether its large scale training, lightning fast AI inferencing, or heavy HPC workloads, NVIDIA Tensor Core GPUs get the job done.

With the launch of NVIDIA Blackwell, should you still consider Hopper in your data center deployment?

AI Training

Blackwell is approximately 3x faster in AI training, but NVIDIA Hopper has respectable performance and can be tasked to train AI models at a lower cost than Blackwell.

AI Inferencing

Blackwell's high throughput, more memory bandwidth, and faster interconnect deliver approximately 30x improvement, making it indisputable for AI Inferencing over Hopper.

HPC

NVIDIA Hopper shines in FP64 and FP64 Tensor Core cost-to-performance, making it optimal for high precision HPC workloads like weather modeling, simulation, and analytics.

NVIDIA Blackwell & NVIDIA Hopper Specifications

🟦 - Blackwell Architecture 🟩 - Hopper Architecture

GPU Name	NVIDIA B200	NVIDIA B100	NVIDIA H200	NVIDIA H200 NVL
Form Factor	SXM	SXM	SXM	PCIe
FP64	40 teraFLOPS	30 teraFLOPS	34 teraFLOPS	34 teraFLOPS
FP64 Tensor Core	40 teraFLOPS	30 teraFLOPS	67 teraFLOPS	67 teraFLOPS
FP32	80 teraFLOPS	60 teraFLOPS	67 teraFLOPS	67 teraFLOPS
FP32 Tensor Core	2.2 petaFLOPS	1.8 petaFLOPS	989 teraFLOPS	989 teraFLOPS
FP16/BF16 Tensor Core	4.5 petaFLOPS	3.5 petaFLOPS	1979 teraFLOPS	1979 teraFLOPS
INT8 Tensor Core	9 petaOPs	7 petaOPs	3958 teraOPs	3958 teraOPs
FP8/FP6 Tensor Core	9 petaOPs	7 petaOPs	3958 teraOPs	3958 teraOPs
FP4 Tensor Core	18 petaFLOPS	14 petaFLOPS	-	-
GPU Memory	192GB HBM3e	192GB HBM3e	141GB HBM3e	141GB HBM3e
Memory Bandwidth	Up to 8TB/s	Up to 8TB/s	4.8TB/s	4.8TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Multi-Instance GPUs	Up to 7 MIGs @23GB	Up to 7 MIGs @23GB	Up to 7 MIGs @16.5GB	Up to 7 MIGs @16.5GB
Interconnect	NVLink 1.8TB/s	NVLink 1.8TB/s	NVLink 900GB/s	NVLink bridge 900GB/s
Options	NVIDIA DGX NVIDIA HGX	NVIDIA HGX Drop-In Replacement	NVIDIA DGX NVIDIA HGX	NVIDIA Certified System with 1-8 GPUs
NVIDIA AI Enterprise Included	DGX - Included HGX - Optional	HGX - Optional	DGX - Included HGX - Optional	PCIe - Optional

Deploying Hopper & Blackwell Together

NVIDIA Hopper H200s can be deployed at a moments notice; stand up a computing infrastructure today with Hopper at a lower cost compared to future Blackwell deployments. You can scale your computing further with future Blackwell deployments for increased inferencing and training performance.

Depending on the workload, reap the benefits and strengths of both systems. Simultaneously deploy both Hopper for HPC and AI training and Blackwell for AI Inferencing, and tackle the entire AI workflow.