Superchargi

Now availa

Datasheet

Center Prod

The GPU for Generative AI and HPC

The NVIDIA H200 GPU supercharges generative AI and high-performance computing (HPC) workloads with game-changing performance and memory capabilities. As the first GPU with HBM3E, the H200's larger and faster memory fuels the acceleration of generative AI and large language models (LLMs) while advancing scientific computing for HPC workloads.

Read the P

Llam nce

1.9X Faster

1.6X Faster

High-Performance Computing

110X Faster

Benefits

Higher Performance With Larger, Faster Memory

Based on the NVIDIA Hopper™ architecture, the NVIDIA H200 is the first GPU to offer 141 gigabytes (GB) of HBM3e memory at 4.8 terabytes per

rence

of AI,
dress a
ds. An AI
iver the
est TCO
assive user

eed by up to
en handling

Explore NVIDIA's AI Inference Platform >

Preliminary specific
Llama2 13B: ISL 128,
GPU BS 64 | H200 S
GPT-3 175B: ISL 80,
x8 H200 SXM GPUs BS 128
Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1x GPU BS 8 | H200 SXM 1x GPU BS 32.

Supercharge High-Performance Computing

Memory bandwidth is crucial for HPC applications as it enables faster data transfer, reducing complex processing bottlenecks. For memory-intensive HPC applications like simulations, scientific research, and artificial intelligence, the H200's higher memory bandwidth ensures that data can be accessed and manipulated efficiently, leading up to 110X faster time to results compared to CPUs.

Preliminary specifications. May be subject to change.
HPC MILC- dataset NERSC Apex Medium | HGX H200 4-

evels. This
s
within the
). AI
systems
o more eco-
dge that
mmunity

Preliminary specific
Llama2 70B: ISL 2K,
GPU BS 8 | H200 SX

omputing

Accelerating AI Acceleration for Mainstream Enterprise Servers With H200 NVL

NVIDIA H200 NVL is ideal for lower-power, air-cooled enterprise rack designs that require flexible configurations, delivering acceleration for every AI and HPC workload regardless of size. With up to four GPUs connected by NVIDIA NVLink™ and a 1.5x memory increase, large language model (LLM) inference can be accelerated up to 1.7x, and HPC applications achieve up to 1.3x more performance over the H100 NVL.

Enterprise-Ready: AI Software Streamlines Development and Deployment

NVIDIA H200 NVL comes with a five-year NVIDIA AI Enterprise subscription. This software accelerates AI development and deployment for production-ready generative AI solutions, including computer vision, speech AI, retrieval augmented generation (RAG), and more. NVIDIA AI Enterprise includes NVIDIA NIM™, a set of easy-to-use microservices designed

Specifications

NVIDIA H200 GPU

	H200 SXM[1]	H200 NVL[1]
FP64	34 TFLOPS	30 TFLOPS
FP64 Tensor Core	67 TFLOPS	60 TFLOPS
FP32	67 TFLOPS	60 TFLOPS
TF32 Tensor Core[2]	989 TFLOPS	835 TFLOPS
BFLOAT16 Tensor Core[2]	1,979 TFLOPS	1,671 TFLOPS

GPU Memory Bandwidth		
Decoders		
Confidential		
Max Thermal Power (TDP)		urable)
Multi-Instan		B each
Form Factor		Dual-slot air-cooled
Interconnect	NVIDIA NVLink™: 900GB/s PCIe Gen5: 128GB/s	2- or 4-way NVIDIA NVLink bridge: 900GB/s per GPU PCIe Gen5: 128GB/s
Server Options	NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs	NVIDIA MGX™ H200 NVL partner and NVIDIA-Certified Systems with up to 8 GPUs
NVIDIA AI Enterprise	Add-on	Included

[1] Preliminary specifications. May be subject to change.
[2] With sparsity.

View Datasheet

Rendez-vous sur le site Internet NVIDIA de votre région pour connaître le contenu, les tarifs et les partenaires revendeurs locaux.

Continuer

NVIDIA

Sign

US Sign In

NVIDIA and our third-party partners use cookies and other tools to collect and record information you provide as well as information about your interactions with our websites for performance improvement, analytics, and to assist in our marketing efforts. By clicking "Accept All", you consent to our use of cookies and other tools as described in our Cookie Policy. You can manage your cookie settings by clicking on "Manage Settings". Please see our Privacy Policy for more information on our privacy practices.

Data Center C

Data Center P

Deep Learnin

Energy Efficie

GPU Cloud Co

MLPerf Bench

NGC Catalog

NVIDIA-Certif

NVIDIA Data C

NVIDIA Data C

Qualified System Catalog

Where to Buy

Follow Data Center