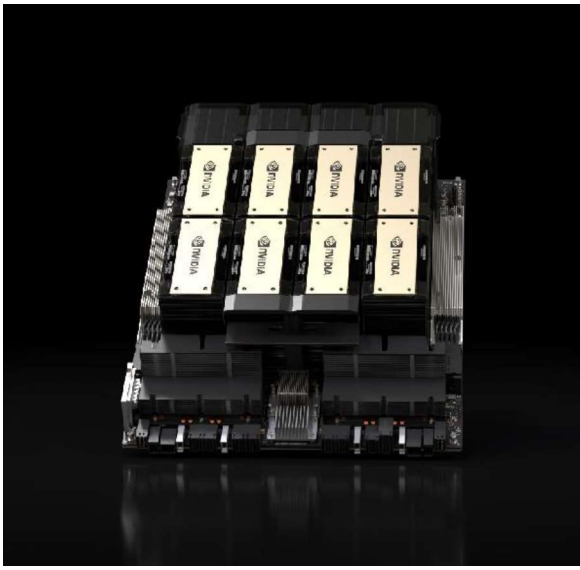
Datasheet

NVIDIA H200 Tensor Core GPU



AIとHPCワークロードをスーパーチャージする

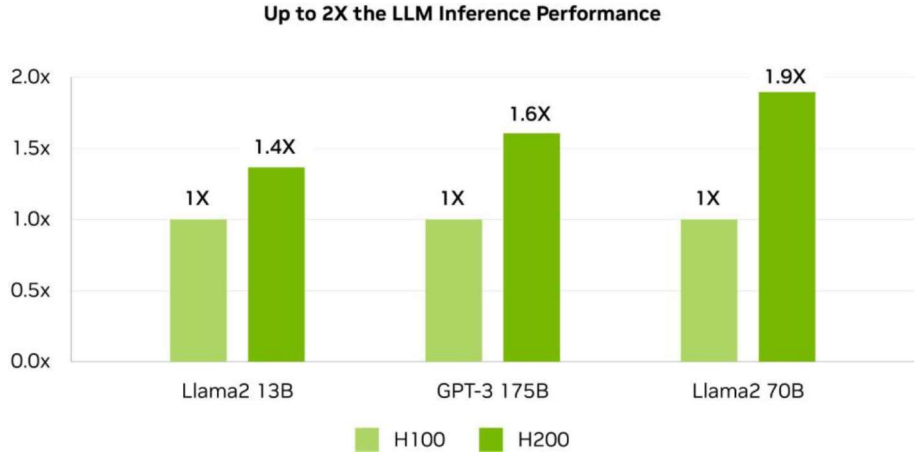
NVIDIA H200 Tensor Core GPUは、生成AIや高性能コンピューティング（HPC）のワークロードを、画期的なパフォーマンスとメモリ機能で強化します。NVIDIA Hopper™ アーキテクチャを基盤としたNVIDIA H200は、初めて141ギガバイト（GB）のHBM3eメモリを搭載し、4.8テラバイト/秒（TB/s）のデータ転送速度を実現します。これはNVIDIA H100 Tensor Core GPUのほぼ2倍の容量であり、メモリ帯域幅も1.4倍です。H200の大容量かつ高速なメモリは、生成AIや大規模言語モデルの処理を加速させ、HPCワークロードにおいても、エネルギー効率を向上させながら、総保有コストを削減します。

Key Features

- 141GB of HBM3e GPU memory
- 4.8TB/s of memory bandwidth
- 4 petaFLOPS of FP8 performance
- 2X LLM inference performance
- 110X HPC performance

高性能なLLM推論でインサイトを引き出す

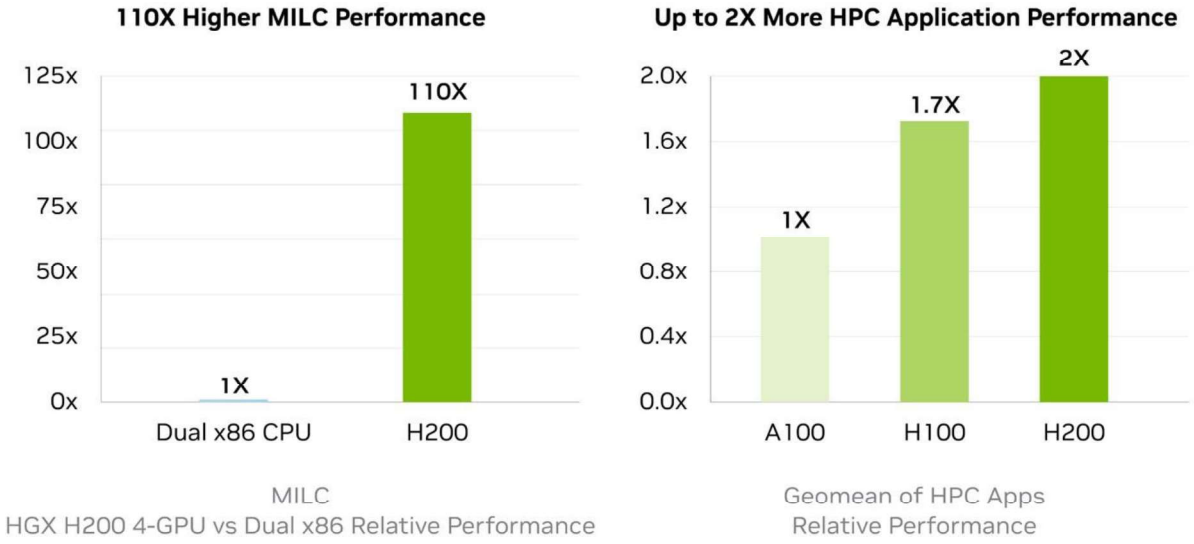
AIの進化し続ける世界では、企業は多様な推論ニーズに対応するために大規模な言語モデルに依存しています。AI推論アクセラレータは、膨大なユーザーベースに対して大規模に展開された際に、最高のスループットを最小の総保有コスト（TCO）で提供する必要があります。H200は、Llama2 70Bのような大規模言語モデルを処理する際に、H100 GPUと比較して推論性能を2倍に向上させます。



Preliminary specifications. May be subject to change.
Llama2 13B: ISL 128, OSL 2K | Throughput | H100 SXM 1xGPU BS 64 | H200 SXM 1xGPU BS 128
GPT-3 175B: ISL 80, OSL 200 | x8 H100 SXM GPUs BS 64 | x8 H200 SXM GPUs BS 128
Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1xGPU BS 8 | H200 SXM 1xGPU BS 32.

高性能コンピューティングを強化する

HPC（高性能コンピューティング）アプリケーションにおいて、メモリ帯域幅は非常に重要です。メモリ帯域幅が広いほど、データ転送が速くなり、複雑な処理のボトルネックが軽減されます。シミュレーションや科学研究、人工知能のようなメモリを多用するHPCアプリケーションにおいて、H200の高いメモリ帯域幅はデータへの迅速なアクセスと効率的な操作を実現し、結果までの時間を110倍も短縮します。



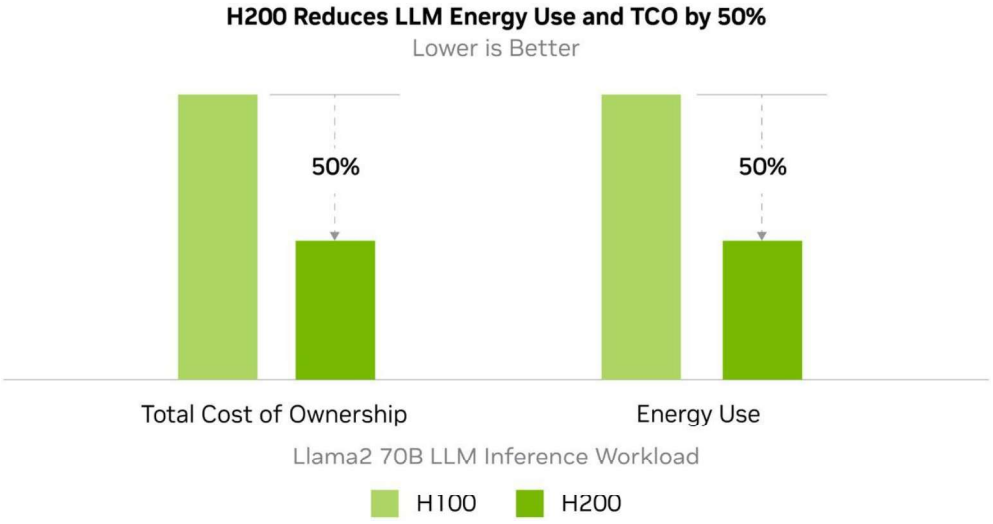
Preliminary specifications. May be subject to change.

HPC MILC- dataset NERSC Apex Medium | HGX H200 4-GPU | dual Sapphire Rapids 8480

HPC Apps- CP2K: dataset H2O-32-RI-dRPA-96points | GROMACS: dataset STMV | ICON: dataset r2b5 | MILC: dataset NERSC Apex Medium | Chroma: dataset HMC Medium | Quantum Espresso: dataset AUSURF112 | 1xH100 SXM | 1xH200 SXM.

エネルギー効率と総保有コスト（TCO）の削減

H200の導入により、エネルギー効率とTCOは新たな次元に到達します。この最先端テクノロジーは、H100 Tensor Core GPUと同じ電力消費範囲内で、比類なきパフォーマンスを提供します。AIファクトリーやスーパーコンピューティングシステムは、より高速でありながら環境に優しいため、AIや科学分野において経済的な優位性をもたらし、これらのコミュニティの発展を加速させます。



Preliminary specifications. May be subject to change.

Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1xGPU BS 8 | H200 SXM 1xGPU BS 32

RYOYO

H200 NVLによる主流企業サーバー向けのAI加速の解放

NVIDIA H200 NVLは、データセンター内でのスペース制約があるお客様に最適な選択肢であり、規模に関わらずすべてのAIおよびHPCワークロードに対して加速性能を提供します。前世代に比べてメモリが1.5倍、帯域幅が1.2倍向上し、数時間で大規模言語モデル（LLM）を微調整し、LLM推論速度を1.8倍に向上させることができます。

企業向け対応：AIソフトウェアが開発と導入を効率化

NVIDIA H200 NVLには5年間のNVIDIA AI Enterpriseサブスクリプションが付属しており、企業向けAIプラットフォームの構築を簡素化します。H200は、コンピュータビジョン、音声AI、RAG（情報検索強化生成）などの生成AIソリューションを生産レベルで迅速に開発・導入できるよう加速します。NVIDIA AI Enterpriseには、企業向け生成AIの導入を加速するための使いやすいマイクロサービスセット「NVIDIA NIM™」が含まれています。これにより、企業レベルのセキュリティ、管理性、安定性、サポートを備えた導入が実現し、パフォーマンス最適化されたAIソリューションによって、より迅速にビジネス価値と実用的なインサイトを提供します。

RYOYO

Technical Specifications		
	H200 SXM[1]	**H200 NVL[1]**
FP64	34 TFLOPS	34 TFLOPS
FP64 Tensor Core	67 TFLOPS	67 TFLOPS
FP32	67 TFLOPS	67 TFLOPS
TF32 Tensor Core[2]	989 TFLOPS	989 TFLOPS
BFLOAT16 Tensor Core[2]	1,979 TFLOPS	1,979 TFLOPS
FP16 Tensor Core[2]	1,979 TFLOPS	1,979 TFLOPS
FP8 Tensor Core[2]	3,958 TFLOPS	3,958 TFLOPS
INT8 Tensor Core[2]	3,958 TFLOPS	3,958 TFLOPS
GPU Memory	141GB	141GB
GPU Memory Bandwidth	4.8TB/s	4.8TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Confidential Computing	Supported	Supported
Max Thermal Design Power (TDP)	Up to 700W (configurable)	Up to 600W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @18GB each	Up to 7 MIGs @18GB each
Form Factor	SXM	PCIe
Interconnect	NVIDIA NVLink™: 900GB/s PCIe Gen5: 128GB/s	2- or 4-way NVIDIA NVLink bridge: 900GB/s 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.

菱洋エレクトロ株式会社
ソリューション事業本部
ソリューション第５ビジネスユニット

【お問い合わせ】
03-3546-6211
nvidia_ws_info@ryoyo.co.jp