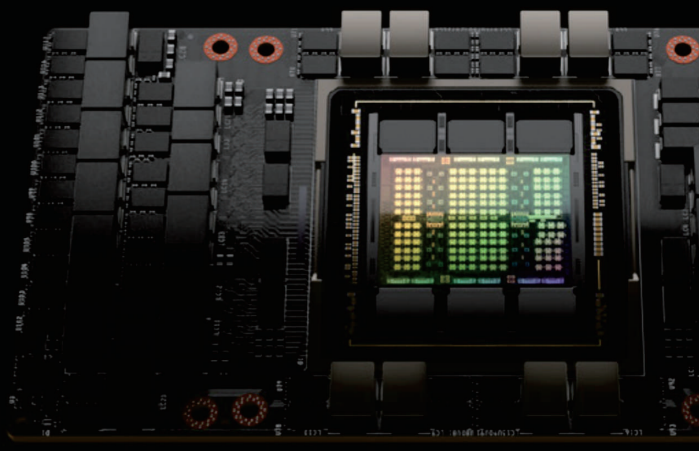
DATASHEET



NVIDIA H100 TENSOR CORE GPU

Unprecedented performance, scalability, and security for every data center.

高速化されたコンピューティングのための桁外れの飛躍

NVIDIA H100 Tensor Core GPUは、あらゆるワークロードに対して、かつてないほどのパフォーマンス、スケーラビリティ、およびセキュリティを提供します。NVIDIA® NVLink® Switch Systemにより、最大256個のH100 GPUを接続してエクサスケールでのワークロードを加速することができ、専用のTransformer Engineは1兆個ものパラメータを持つ言語モデルをサポートします。H100は、NVIDIA Hopper™アーキテクチャの画期的なイノベーションを使用して、業界をリードする会話型AIを実現し、大規模な言語モデルを前世代よりも30倍高速化します。

エンタープライズからエクサスケールまで、ワークロードを安全に高速化

NVIDIA H100 GPUは、第4世代のTensor CoreとFP8精度のTransformer Engineを搭載し、大規模な言語モデルで最大9倍の高速トレーニングと驚異の30倍もの推論高速化により、市場をリードするNVIDIAのAIリーダーシップをさらに拡張します。さらに、ハイパフォーマンス・コンピューティング（HPC）のアプリケーション向けに、H100はFP64の1秒あたりの浮動小数点演算（FLOPS）において3倍の速さを実現し、ダイナミックプログラミング（DPX）命令を追加することで最大7倍の性能を発揮します。第2世代のマルチ・インスタンスGPU（MIG）、内蔵のNVIDIAコンフィデンシャル・コンピューティングおよびNVIDIA NVLinkスイッチシステムにより、H100はエンタープライズからエクサスケール規模までのあらゆるデータセンターのワークロードすべてを安全に加速させます。

H100は、ハードウェア、ネットワーキング、ソフトウェア、ライブラリ、最適化されたAIモデルおよびNVIDIA NGC™カタログのアプリケーションといった構成要素をまとめ上げる包括的なNVIDIAデータセンター・ソリューションの一部となります。データセンター向けの最も強力なエンド・ツー・エンドのAIおよびHPCプラットフォームを代表するこのソリューションは、研究者が実世界で結果を出し、運用によってさまざまなソリューションを大規模に生み出すことを可能にします。

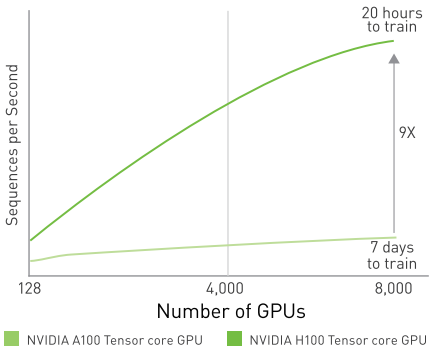
SPECIFICATIONS

	H100 SXM	H100 PCIe
FP64	**30 TFLOPS**	**24 TFLOPS**
FP64 Tensor Core	**60 TFLOPS**	**48 TFLOPS**
FP32	**60 TFLOPS**	**48 TFLOPS**
TF32 Tensor Core	**1,000 TFLOPS***	**800 TFLOPS***
BFLOAT16 Tensor Core	**2,000 TFLOPS***	**1,600 TFLOPS***
FP16 Tensor Core	**2,000 TFLOPS***	**1,600 TFLOPS***
FP8 Tensor Core	**4,000 TFLOPS***	**3,200 TFLOPS***
INT8 Tensor Core	**4,000 TOPS***	**3,200 TOPS***
GPU memory	**80GB**	**80GB**
GPU memory bandwidth	**3TB/s**	**2TB/s**
Decoders	**7 NVDEC 7 JPEG**	**7 NVDEC 7 JPEG**
Max thermal design power (TDP)	**700W**	**350W**
Multi-Instance GPUs	**Up to 7 MIGS @ 10GB each**	
Form factor	**SXM**	**PCIe dual-slot air-cooled**
Interconnect	**NVLink: 900GB/s PCIe Gen5: 128GB/s**	**NVLink: 600GB/s PCIe Gen5: 128GB/s**
Server options	**NVIDIA HGX™ H100 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX™ H100 with 8 GPUs**	**Partner and NVIDIA-Certified Systems with 1–8 GPUs**

* Shown with sparsity. Specifications 1/2 lower without sparsity.

Up to 9X Higher AI Training on Largest Models

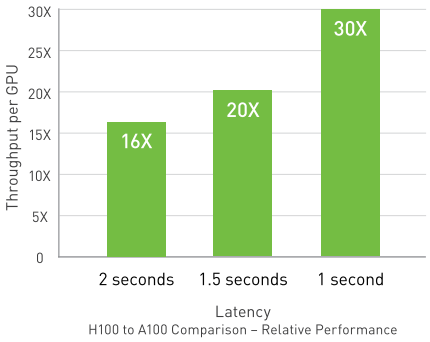
Mixture of Experts (395 Billion Parameters)



Projected performance subject to change. Training Mixture of Experts (MoE) Transformer Switch-XXL variant with 395B parameters on 1T token dataset | A100 cluster: HDR IB network | H100 cluster: NVLink Switch System, NDR IB

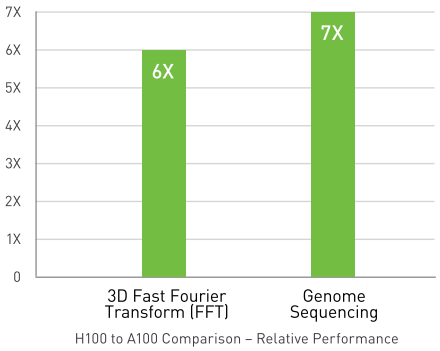
Up to 30X Higher AI Inference Performance on Largest Models

Megatron Chatbot Inference (530 Billion Parameters)



Projected performance subject to change. Inference on Megatron 530B parameter model chatbot for input sequence length=128, output sequence length =20 | A100 cluster: HDR IB network | H100 cluster: NDR IB network for 16 H100 configurations | 32 A100 vs 16 H100 for 1 and 1.5 sec | 16 A100 vs 8 H100 for 2 sec

Up to 7X Higher Performance for HPC Applications



Projected performance subject to change. 3D FFT (4K^3) throughput | A100 cluster: HDR IB network | H100 cluster: NVLink Switch System, NDR IB | Genome Sequencing (Smith-Waterman) | 1 A100 | 1 H100

NVIDIA Hopperの技術的ブレークスルー

世界最先端チップ

最先端のTSMC 4Nプロセスを使用した800億個のトランジスタはNVIDIAの加速コンピューティングのニーズに合わせてカスタマイズされたもので、それらで構築されたH100はこれまでに作られた世界で最も先進的なチップです。データセンター規模でAI、HPC、メモリ帯域、インターコネクト、および通信を加速するための大きく進歩したことが特徴です。

トランスフォーマーエンジン

Transformer Engineで使われるソフトウェアとHopper Tensor Coreテクノロジーは、世界で最も重要なAIモデルの構成要素であるトランスフォーマーによって構築されたモデルのトレーニングを加速するために設計されています。Hopper Tensor Coreは、FP8とFP16の混合精度を適用して、トランスフォーマーのAI計算を劇的に高速化することができます。

NVLINKスイッチシステム

NVLinkスイッチシステムは、PCIe Gen5の7倍以上の帯域幅である、1 GPUあたり双方向で900 GB/sの速さで、複数のサーバでのマルチGPU入出力（IO）のスケーリングを可能にします。
最大256台のH100のクラスタに対応し、NVIDIA Ampereアーキテクチャでの InfiniBand HDRの9倍もの帯域幅を実現します。

NVIDIAコンフィデンシャル・コンピューティング

NVIDIAコンフィデンシャル・コンピューティングは、Hopperに内蔵されたセキュリティ機能で、NVIDIA H100は世界初の機密演算機能付きアクセラレータとなります。ユーザーは、H100 GPUの卓越したアクセラレーションにアクセスしながら、使用中のデータやアプリケーションの機密性と整合性を保護することができます。

第2世代マルチ・インスタンスGPU（MIG）

Hopperアーキテクチャの第2世代MIGは、仮想化環境におけるマルチテナント、マルチユーザ構成をサポートし、GPUを適正サイズのインスタンスへと安全に分割し、7倍以上の安全なテナントに対してサービス品質（QoS）を最大化することができます。

DPX命令

HopperのDPX命令は、ダイナミック・プログラミング・アルゴリズムを高速化し、CPUに比べて40倍、NVIDIA AmpereアーキテクチャGPUに比べて7倍の速さとなります。これにより、疾病診断、リアルタイム・ルーティング最適化、グラフ分析などの時間が劇的に短縮されます。

NVIDIA H100 CNX コンバージド・アクセラレータ

NVIDIA H100 CNXは、単一の独自のプラットフォームの中にNVIDIA H100の威力と**NVIDIA ConnectX®-7** スマートネットワークインターフェースカード（SmartNIC）の高度なネットワーキング機能を統合させています。このコンバージェンスにより、企業データセンターでの分散型AIトレーニングやエッジでの5G処理など、GPUによるIO集中型ワークロードに比類のない性能を提供します。

エンタープライズ・レディ

深層学習、HPC、データ分析のために構築されたこのプラットフォームは、あらゆる主要な深層学習フレームワークを含む2,700以上のアプリケーションを加速させます。さらに、AIとデータ分析用のエンド・ツー・エンドでクラウドネイティブなソフトウェアスイートであるNVIDIA AI Enterpriseは、VMware vSphereによるハイパーバイザーベースの仮想インフラにおいて、H100上で実行することが認定されています。これにより、ハイブリッドクラウド環境におけるAIワークロードの管理およびスケーリングが可能になります。

HPCTECH

NVIDIA ELITE PARTNER

株式会社 HPCテック
本　　社：〒103-0006 東京都中央区日本橋富沢町 7-13
TEL:03-5643-2681　FAX:03-5643-2682
大阪営業所：〒532-0011 大阪市淀川区西中島4丁目5-1
TEL 06-6195-6464　FAX 06-6195-6468
https://www.hpctech.co.jp
sales@hpctech.co.jp