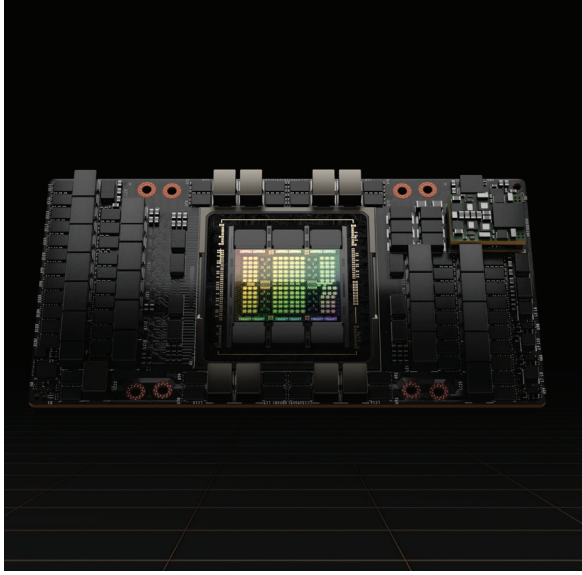


NVIDIA H100 Tensor Core コア GPU

比類なきパフォーマンス、拡張性、セキュリティを
あらゆるデータセンターに

アクセラレーテッド コンピューティングを桁違いに高速化

NVIDIA H100 Tensor Core GPU コア GPU は、比類のないパフォーマンス、拡張性、セキュリティをあらゆるワークロードに提供します。NVIDIA®NVLink®Switch System との組み合わせで、最大256 基の H100 GPU を接続して、エクサスケールのワークロードを高速化します。さらに専用のTransformer Engineで、数兆個のパラメーターを持つ言語モデルにも対応します。H100 は、NVIDIA Hopper TMアーキテクチャが実現する画期的な変革を利用して業界最高水準の対話型 AI を提供し、大規模言語モデルの処理を前世代の 30 倍以上の速さで実行します。

エンタープライズ AI への対応は?

NVIDIA H100 Tensor Core GPUには、NVIDIA AI Enterpriseソフトウェアスイートのエンタープライズサポートを含む5年間のソフトウェアサブスクリプションが付属しており、非常に優れた性能によってAI 活用を簡素化します。これにより、企業や組織は、AIチャットボット、レコメンデーションエンジン、ビジョンAIなど、H100で高速化するAIワークフローを構築するために必要なAIフレームワークおよびツールが活用できるようになります。**NVIDIA AI Enterprise ソフトウェア サブスクリプション**やこれに関連する NVIDIA H100 サポートの詳細については、こちらをご覧ください。

エンタープライズからエクサスケールまで、さまざまな規模のワークロードをセキュアに高速化

NVIDIA H100 GPU は、第 4 世代の Tensor コアと FP8 精度の Transformer Engine を搭載し、大規模言語モデルを使うトレーニングにおいて前世代比で最大 9 倍高速化、推論の速度を最大 30 倍と驚異的に高めることで、AI 市場における NVIDIA のリーダーシップをさらに確かなものにします。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についても、FP64 による FLOPS(1 秒間に実行できる浮動小数点演算の回数)を 3 倍に高め、新しい DPX 命令は、動的計画法のアルゴリズムを最大 7 倍まで高速化します。第 2 世代のマルチインスタンス GPU(MIG)、組み込みの NVIDIA コンフィデンシャル コンピューティング、NVIDIA NVLink Switch System を装備する H100 は、エンタープライズからエクサスケールまで、さまざまな規模のデータ センターのあらゆるワークロードを安全に高速化します。

あらゆるワークロードをあらゆる場所で加速する

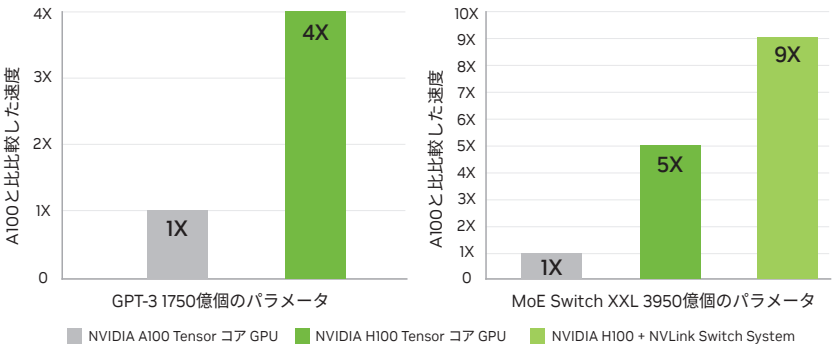
NVIDIA H100は、NVIDIA データセンター プラットフォームにとって不可欠な存在です。AI、HPC、データ分析のために構築されたこのプラットフォームは、3,000以上のアプリケーションを高速化します。データセンターからエッジまであらゆる場所で利用でき、劇的なパフォーマンス向上とコスト削減の両方を実現します。

規格			
	H100 SXM	**H100 PCIe**	**H100 NVL[1]**
FP64	34 TFLOPS	26 TFLOPS	68 teraFLOPS
FP64 Tensor コア	67 TFLOPS	51 TFLOPS	134 teraFLOPS
FP32	67 TFLOPS	51 TFLOPS	134 teraFLOPS
TF32 Tensor コア	989 TFLOPS[2]	756 TFLOPS[2]	1,979 teraFLOPS[2]
BFLOAT16 Tensor コア	1,979 TFLOPS[2]	1,513 TFLOPS[2]	3,958 teraFLOPS[2]
FP16 Tensor コア	1,979 TFLOPS[2]	1,513 TFLOPS[2]	3,958 teraFLOPS[2]
FP8 Tensor コア	3,958 TFLOPS[2]	3,026 TFLOPS[2]	7,916 teraFLOPS[2]
INT8 Tensor コア	3,958 TOPS[2]*	3,026 TOPS[2]	7,916 TOPS[2]
GPU メモ リ	80GB	80GB	188GB
GPU メモリ帯域幅	3.35TB/s	2TB/s	7.8TB/s[3]
デコーダー	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG	14 NVDEC 14 JPEG
最大熱設計電力 (TDP)	Up to 700W (configurable)	300-350W (configurable)	2x 350-400W (configurable)
マルチインスタンス GPUs	Up to 7 MIGs @ 10GB each	Up to 7 MIGs @ 10GB each	Up to 14 MIGs @ 12GB each
フォームファクター	SXM	PCIe > dual-slot > air-cooled	2x PCIe > dual-slot > air-cooled
相互接続	NVLink > 900GB/s PCIe > Gen5: 128GB/s	NVLink > 600GB/s PCIe > Gen5: 128GB/s	NVLink > 600GB/s PCIe > Gen5: 128GB/s
サーバーオプション	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	Partner and NVIDIA Certified Systems with 2-4 pairs
NVIDIA AI Enterprise	Add-on	Included	Included

1 Preliminary specifications. May be subject to change. Specifications shown for 2x H100 NVL PCIe cards paired with NVLink Bridge.
2 With sparsity.
3 Aggregate HBM bandwidth.

技術突破

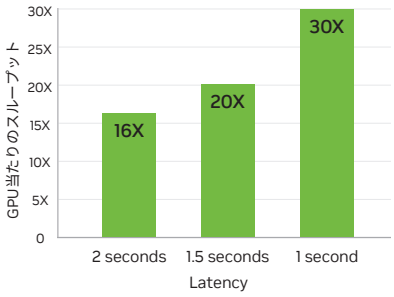
GPT-3のAI学習を最大9倍高速化



予想されるパフォーマンスは変更される可能性があります。GPT-3 1750億個 トレーニングA100クラスタ: HDR IBネットワーク, H100クラスター: NDR IB ネットワーク | 3,950 億パラメーターと 1T トークンのデータセットを使ったトレーニング Mixture of Experts (MoE) Transformer Switch-XXL のバリアント, A100 クラスター: HDR IBnetwork、H100クラスタ: NDR IBネットワークとNVLink Switch Systemの組み合わせ

大規模モデルを使うAI 推論を最大30倍高速化

Megatron チャットボット推論
(5300億個のパラメータ)



入力シーケンス長を128、出力シーケンス長を 20 としたときの Megatron 5,300 億パラメーター モデル チャットボットによる推論 | A100クラスター:HDR IB ネットワーク | H100 クラスター:16 基 H100 構成の NDR IB ネットワーク | 32 基 A100と16 基 H100 の1秒および 1.5 秒での比較 | 16 基 A100と□基H100の2秒でん比較

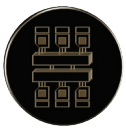
NVIDIA Hopper がもたらす画期的なテクノロジー

NVIDIA H100 Tensor コア GPU

NVIDIA のアクセラレーテッド 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最先端の TSMC 4N プロセスで製造されており、800 億のトランジスタを集積した 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。この飛躍的な進歩により、データセンター規模で AI、HPC、メモリ帯域幅、相互接続、通信が高速化されます。

Transformer Engine

Transfomer Engine は、Transformer モデルのトレーニングと推論を高速化するために特別に設計されたソフトウェアと NVIDIA Hopper Tensor コア テクノロジを組み合わせています。Hopper Tensor コアは、FP8とFP16 が混在する精度を適用することで、Transformer の AI 演算を大幅に高速化できます。



NVLink Switch System

NVLink Switch System は、マルチ GPU の入出力 (IO) を GPU 当たり 900 GB/秒の双方向帯域幅で複数のサーバーにスケールさせます。これは PCIe Gen5 の帯域幅の 7 倍以上です。このシステムは、最大256 基のH100 をサポートし、NVIDIA Ampere アーキテクチャの Infini and HDR の 9 倍の帯域幅を提供します。

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 に組み込まれたセキュリティ機能です。これにより、NVIDIA H100 は世界初のコンフィデンシャル コンピューティング機能付きのアクセラレーターとなっています。使用中のデータとアプリケーションの機密性と完全性を保護しながら、H100 GPU による比類のない高速化も獲得できます。

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、マルチユーザー構成をサポートし、7 つの安全なテナントに対してサービス品質 (QoS) を最大化するために、GPU を分離した適切なサイズに安全に分割します。

DPX 命令

Hopper のDPX 命令は、動的プログラミング アルゴリズムの処理を CPU と比較して 40 倍、NVIDIA Ampere アーキテクチャ GPU と比較して 7 倍に高速化します。これにより、病気の診断、リアルタイムでの経路の最適化、グラフ分析に必要な時間が大幅に短縮されます。

H100 と NVIDIA AI プラットフォームを導入

NVIDIA AI は、本番環境向け AI の包括的なオープン プラットフォームであり、NVIDIA H100GPU を基盤とします。NVIDIA のアクセラレーテッド コンピューティング インフラストラクチャ、インフラストラクチャの最適化と AI の開発、展開のためのソフトウェア スタック、市場投入を速めるためのアプリケーション ワークフローが、このプラットフォームに含まれます。無料のハンズオン ラボを利用して、**NVIDIA AI と NVIDIA H100 を NVIDIA LaunchPad** でお試しください。

始めるには?

NVIDIA H100 Tensor コア GPU の詳細については、
https://www.nvidia.com/ja-jp/data-center/h100/