NVIDIA H100 Tensor Core GPU

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。

Explore the Technology Breakthroughs of NVIDIA HOPPER

NVIDIA H100 Tensor Core GP

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。

Transformer Engine

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 Switch System

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 Confidential Computing

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。

Second-Generation Multi-Instance 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 Instructions

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。

SPECIFICATION COMPARISON

Product Name	H100 SXM	H100 PCIe
FP64	34 TFLOPS	26 TFLOPS
FP64 Tensor Core	67 TFLOPS	51 TFLOPS
FP32	67 TFLOPS	51 TFLOPS
FP32 Tensor Core	989 TFLOPS*	756 TFLOPS*
BFLOAT16 Tensor Core	1,979 TFLOPS*	1,513 TFLOPS*
FP16 Tensor Core	1,979 TFLOPS*	1,513 TFLOPS*
FP8 Tensor Core	3,958 TFLOPS*	3,026 TFLOPS*
INT8 Tensor Core	3,958 TOPS*	3,026 TOPS*
GPU Memory	80GB HBM3	80GB HBM2e
GPU Memory Bandwidth	3.35TB/s	2TB/s
Multi-instance GPUs	Up to 7 MISs @ 10GB each	Up to 7 MISs @ 10GB each
Interconnect	NVLink 900GB/s, PCIe Gen5:128GB/s	NVLink 600GB/s, PCIe Gen5:128GB/s
Form Factor	SXM	PCIe
Max TDP Power	Up to 700W	300-350W

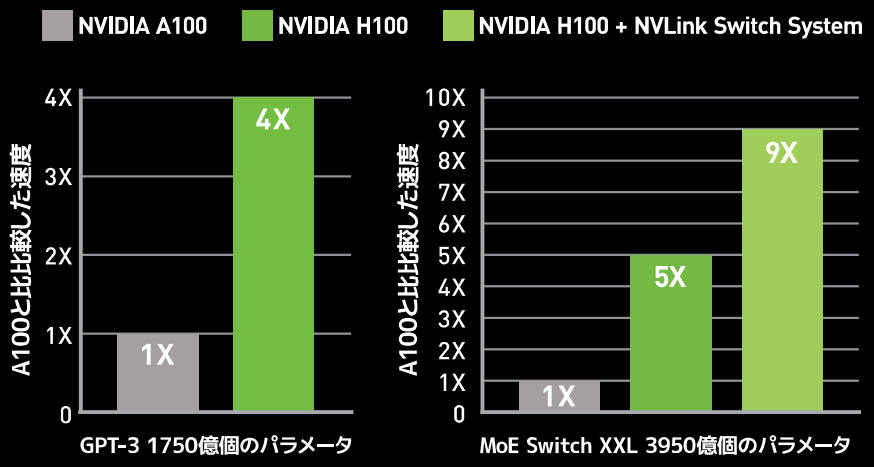
* With sparsity.

HPC TECH

NVIDIA Elite Partner

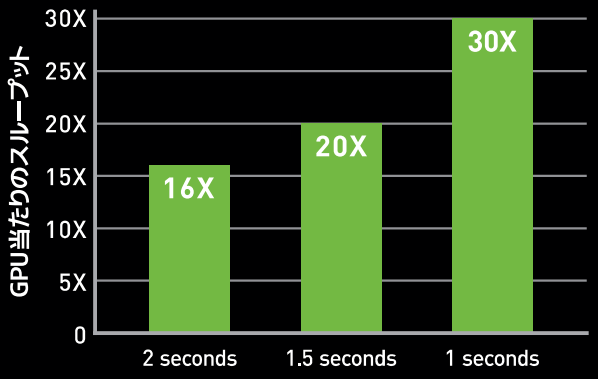
NVIDIA A100 vs NVIDIA H100 vs NVIDIA H100 + NVLink Switch System

GPT-3のAI学習を最大9倍高速化



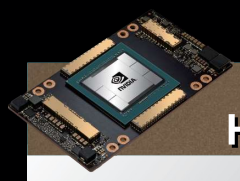
予想されるパフォーマンスは変更される可能性があります。GPT-3 1750億個 トレーニング A100 クラスタ: HDR IB ネットワーク, H100 クラスタ：NDR IB ネットワーク | 3,950 億パラメーターと 1 T トークンのデータセットを使ったトレーニング Mixture of Experts（MoE）Transformer Switch-XXL のバリアント, A100 クラスタ: HDR IBnetwork、H100 クラスタ：NDR IB ネットワークと NVLink Switch System の組み合わせ

大規模モデルを使う AI 推論を最大30倍高速化



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 8 基 H100 の 2 秒で比較

NVIDIA H100 RACKMOUNT SERVER - FLAGSHIP MODEL



NVIDIA H100-SXM High-End GPU Server

MAX 4GPU 2CPU

Up to 4TB Memory
6 SLOT BAY | 4U Rackmount
U.2 NVMe | PCI-Express 5.0

HPCT RS4X42-4GN

GPU	NVIDIA H100 SXM5 x4 80GB HBM3 Memory
CPU	Intel Xeon Platinum 8480+ x2 2.0GHz, 56core Total 112core
RAM	DDR5-4800 64GB x32 2,048GB
SSD	3.84TB NVMe x1
OS	Linux

MAX 8GPU 2CPU

Up to 3TB Memory
16 SLOT BAY | 8U Rackmount
U.2 NVMe | PCI-Express 5.0

HPCT RS8E42-8GN

GPU	NVIDIA H100 SXM5 x8 80GB HBM3 Memory
CPU	AMD EPYC 9654 x2 2.4GHz, 96core Total 192core
RAM	DDR5-4800 64GB x24 1,536GB
SSD	3.84TB NVMe x1
OS	Linux

NVIDIA H100-PCIe High-End GPU Server

MAX 4GPU 2CPU

Up to 2TB Memory
8 SLOT BAY | 4U Rackmount
U.2 NVMe | PCI-Express 5.0

HPCT WRSX42-4GP

GPU	NVIDIA H100 PCIe x4 80GB HBM2e Memory
CPU	Intel Xeon Platinum 6442Y x2 2.6GHz, 24core Total 48core
RAM	DDR4-4800 64GB x16 1,024GB
SSD	3.84TB NVMe x1
OS	Linux

MAX 8GPU 2CPU

Up to 3TB Memory
24 SLOT BAY | 4U Rackmount
U.2 NVMe | PCI-Express 5.0

HPCT RS4E42-8GP

GPU	NVIDIA H100 PCIe x8 80GB HBM2e Memory
CPU	AMD EPYC 9334 x2 2.7GHz, 32core Total 64core
RAM	DDR4-4800 64GB x24 1,536GB
SSD	3.84TB NVMe x1
OS	Linux

上記モデルは一例です。お客様の用途に合うようカスタマイズいたします。
構成・価格はお問い合わせください。

BrightComputing 正規代理店　NVIDIA エリートパートナー　A2ZEON 日本総代理店　ANSYS Discovery Live 代理店

株式会社 HPCテック
www.hpctech.co.jp

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
MAIL: info@hpctech.co.jp

お問い合わせ