データシート

NVIDIA

NVIDIA DGX B200

トレーニング、ファインチューン、
推論のための統合 AI プラットフォーム

次世代の AI を支える

AI は、タスクの自動化、顧客サービスの強化、知見や洞察の創出、イノベーションの実現により、ほぼすべてのビジネスに変革をもたらしつつあります。AI はもはや未来のコンセプトではなく、ビジネスのあり方を根本的に変える現実のものとなっています。ただし、AI ワークロードの増加に伴い、ほとんどの企業において、利用できるよりもはるかに多くの計算処理能力が必要になり始めています。AI を活用するために、企業は安全で信頼性が高く、効率的なハイパフォーマンス コンピューティング、ストレージ、ネットワーク機能を必要としています。

NVIDIA DGX™ B200 は **NVIDIA DGX プラットフォーム**に追加された最新モデルです。この統合 AI プラットフォームは、NVIDIA Blackwell GPU と高速インターコネクトを最大限に活用することで生成 AI の次のステップを定義づけるものとなります。8 基の Blackwell GPU で構成された NVIDIA DGX B200 は、膨大な 1.4 テラバイト (TB) の GPU メモリと毎秒 64 テラバイト (TB/秒) のメモリ帯域幅で比類なき生成 AI パフォーマンスを実現し、あらゆる企業における AI ワークロード処理に最適です。

開発から展開までを支える単一のプラットフォーム

AI ワークフローが高度化するにつれて、トレーニングからファインチューンや推論に至る AI パイプラインの全ての段階で、企業が大規模なデータセットを処理する必要性も高まっています。そのためには、膨大な計算処理能力が必要となります。NVIDIA DGX B200 により、企業はワークフローを高速化するために構築された単一の統合プラットフォームで開発者を支援することができます。次世代の生成 AI のために強化された NVIDIA DGX B200 で、企業は日々の業務や顧客体験向上のために AI を取り入れることができます。

主な特徴
NVIDIA DGX B200

- 8 基の NVIDIA Blackwell GPU で構築
- 1.4TB の GPU メモリ空間
- 72 ペタ FLOPS のトレーニング パフォーマンス
- 144 ペタ FLOPS の推論パフォーマンス
- NVIDIA ネットワーキング
- デュアルの第 5 世代 Intel® Xeon® スケーラブル プロセッサ
- NVIDIA DGX BasePOD と NVIDIA DGXSuperPOD の基盤
- NVIDIA AI Enterprise と NVIDIA Base Command™ ソフトウェアが含まれています

NVIDIA DGX B200 Technical Specifications

GPU	8x NVIDIA Blackwell GPUs
GPU Memory	1,440GB total, 64TB/s HBM3e bandwidth
Performance	72petaFLOPS FP8 training and 144petaFLOPS FP4 inference
NVIDIA® NVSwitch™	2x
NVIDIA NVLink Bandwidth	14.4TB/s aggregate bandwidth
System Power Usage	~14.3kW max
CPU	2x Intel® Xeon® Platinum 8570 Processors 112Cores total, 2.1GHz (Base), 4GHz (Max Boost)
System Memory	2TB, configurable to 4TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField-3 DPU > Up to 400Gb/s InfiniBand/Ethernet
Management Network	10Gb/s onboard NIC with RJ45 100Gb/s dual-port ethernet NIC
Storage	Host baseboard management controller (BMC) with RJ45 OS: 2x 1.9TB NVMe M.2 Internal storage: 8x 3.84TB NVMe U.2
Software	NVIDIA AI Enterprise – Optimized AI Software NVIDIA Base Command – Orchestration, Scheduling, and Cluster Management DGX OS / Ubuntu – Operating System
Rack Units (RU)	10 RU
System Dimensions	Height:444, Width:482.2, Length:897.1 (mm)
Enterprise Support	ハードウェアとソフトウェアの 3年間の エンタープライズ Business-Standard サポート 年中無休のエンタープライズサポートポータル アクセス 現地営業時間中のライブエージェントサポート

究極の AI パフォーマンス

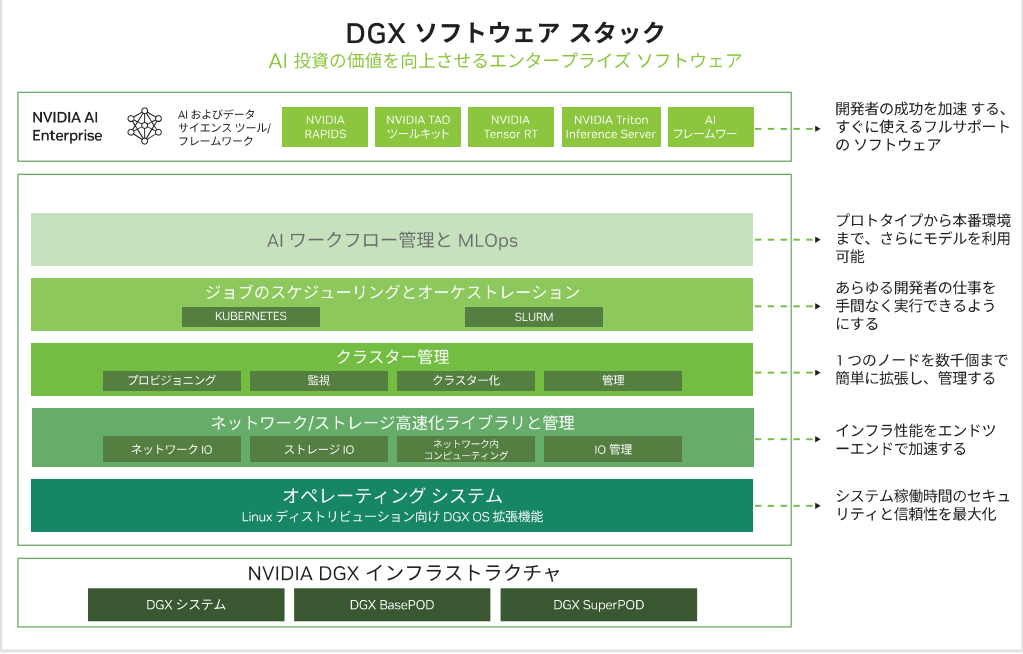
NVIDIA は、企業が直面する最も複雑な AI 問題に対処するための、世界で最もパワフルな次世代スーパーコンピューターの設計に取り組んでいます。NVIDIA DGX B200 は NVIDIA アクセラレーテッド コンピューティング プラットフォームの最新製品であり、その取り組みを示すものです。革新的な NVIDIA Blackwell アーキテクチャによる高度なコンピューティングの進化により、NVIDIA DGX B200 は NVIDIA DGX H100 と比較してトレーニング性能が 3 倍、推論性能が 15 倍となります。NVIDIA DGX POD™ リファレンスアーキテクチャの基盤である NVIDIA DGX B200 は NVIDIA DGX BasePOD™ と NVIDIA DGX SuperPOD™ を支える優れた高速性と拡張性を提供し、手間のかからない AI インフラ ソリューションで最高レベルのパフォーマンスを実現します。

実証済み標準インフラ

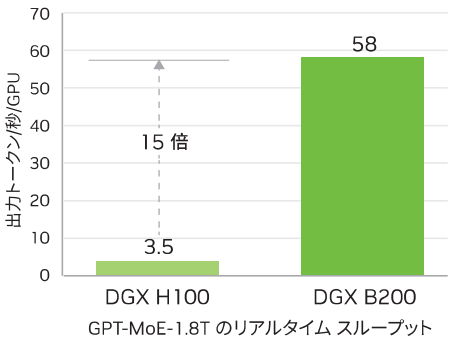
NVIDIA DGX B200 は、NVIDIA Blackwell GPU を搭載した世界初のシステムであり、大規模言語モデルや自然言語処理など、世界で最も複雑な AI 問題を処理するための画期的なパフォーマンスを提供します。NVIDIA DGX B200 は、完全に最適化されたハードウェアおよびソフトウェア のプラットフォームです。NVIDIA AI ソフトウェアのフル スタックを備え、多様なサードパーティのサポートを受けられる充実したエコシステムを利用でき、さらに NVIDIA プロフェッショナル サービスにより専門家からのアドバイスを受けることができます。組織は AI を利用し、最大かつ最も複雑なビジネス問題を解決できます。

NVIDIA Base Command 搭載

NVIDIA Base Command は DGX プラットフォームを強化し、NVIDIA ソフトウェアがもたらすイノベーションを企業が最大限活用できるようにします。企業は、エンタープライズグレードのオーケストレーションとクラスター管理、コンピューティング、ストレージ、ネットワークのインフラを高速化するライブラリ、AI ワークロード向けに最適化されたオペレーティング システムを含む実証済みのプラットフォームで、DGX インフラの可能性を最大限まで引き出すことができます。また、DGX インフラには、AI の開発と展開を効率化するために最適化された一連のソフトウェア、NVIDIA AI Enterprise も含まれています。



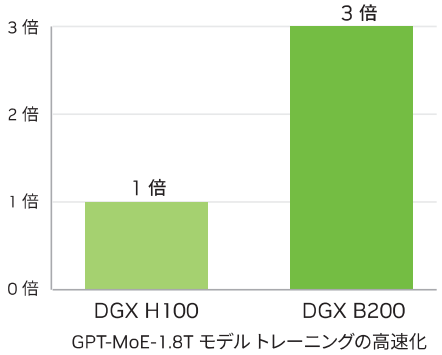
リアルタイム大規模言語モデル推論



GPT-MoE-1.8T のリアルタイム スループット

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。トークン間のレイテンシ (TTL) = 50ms リアルタイム、最初のトークンのレイテンシ (FTL) = 5,000ms、入力シーケンス長 = 32,768、出力シーケンス長 = 1,028、8x 8ウェイ DGX H100 GPU 空冷と 1x 8ウェイ DGX B200 空冷の比較、GPU ごとのパフォーマンス比較。

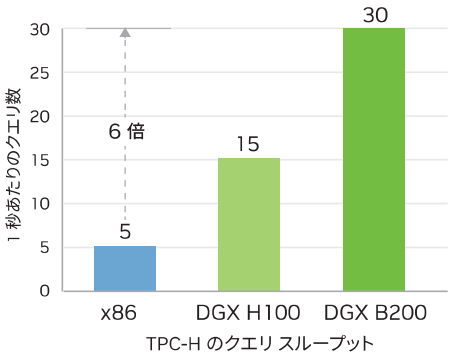
強化された AI トレーニング パフォーマンス



GPT-MoE-1.8T モデル トレーニングの高速化

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。32,768 GPU スケール、4,096x 8ウェイ DGX H100 空冷クラスター : 400G IB ネットワーク、4,096x 8ウェイ DGX B200 空冷クラスター : 400G IB ネットワーク。

高速データ処理



TPC-H のクエリ スループット

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。TPC-H Q4 クエリから派生した、Snappy / Deflate 圧縮によるデータベース結合クエリ。1x x86、1x H100 GPU、1x Blackwell シングル GPU。

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