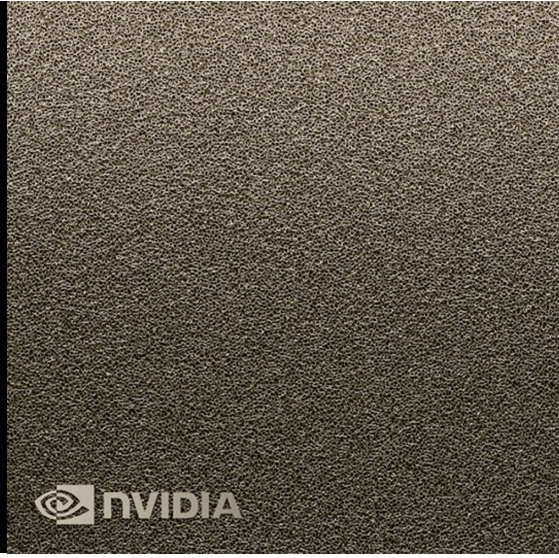
データシート

NVIDIA DGX B300

AI ファクトリーのパフォーマンスに新たな基準を設定
トレーニングから推論までをカバー



あらゆる企業に前例のないパフォーマンスを提供

AI の導入はさまざまな業界で短期間で指数関数的な成長を見せており、企業が AI の変革にアプローチする方法に根本的な変化が起こっていることを示しています。これまで懐疑的に見ていて技術革新の採用に消極的だった組織も AI を活用するためにデータ センターに適切なインフラストラクチャを整備し、適切な人材でチームを強化しようと急いでいますが、これらの組織の多くは、AI の導入が計画したほど簡単ではないことに気づいています。

生成 AI の可能性は変革的ですがこのような企業はこういった技術の採用と拡張に際して、いくつか共通の課題に直面しています。その中には、統合の複雑さに対処できる適切なソリューションを見つけること専門知識の大きなギャップを埋めること、エネルギー消費とコストを管理することなどが含まれます。これらの組織はハイパースケーラーと同じように拡張と運用を行うための設備が整っていないことに気づいています。

NVIDIA DGX SuperPOD の構成要素である NVIDIA DGX™ B300 は、AI 推論の計算要件を満たすようにカスタマイズされた専門の AI インフラストラクチャソリューションです。フルスタックのソフトウェアを活用することで、AI を合理的な方法で展開する企業の負担を軽減します。NVIDIA Blackwell Ultra GPU を搭載した DGX B300 は、推論で 144 petaFLOPS、トレーニングで 72 petaFLOPS を実現します。すべて最新のデータセンターに円滑に収納できるように設計された新しいフォームファクターで提供され、NVIDIA MGX™ および従来のエンタープライズ ラックと互換性があります。DGX B300 を使用することで、あらゆる企業がこれまでにないレベルの効率で多様な AI ワークロードのトレーニングと推論を実行できるようになります。

リアルタイム AI の強力な存在

DGX B300 はリアルタイム推論機能に飛躍的な進歩をもたらし、あらゆる規模の企業が、これまでハイパースケーラーに限定されていた AI のパフォーマンスを活用できるようにします。NVIDIA Blackwell Ultra GPU と NVIDIA ConnectX-8 ネットワーキングを搭載した世界初の完全統合システムであり、最適化された NVIDIA Mission Control ソフトウェアを搭載する DGX B300 は前世代と比較して 11 倍の推論パフォーマンスと 4 倍のトレーニング パフォーマンスを実現します。

主な特徴
NVIDIA DGX B300

- NVIDIA Blackwell Ultra GPU を搭載
- 2.3 TB の GPU メモリスペース
- 72 petaFLOPS のトレーニング パフォーマンス
- 144 petaFLOPS の推論パフォーマンス
- NVIDIA ネットワーキング
- デュアル Intel Xeon プロセッサ
- NVIDIA DGX BasePOD™ と NVIDIA DGXSuperPOD™ の基盤
- NVIDIA AI Enterprise と NVIDIA Mission Control ソフトウェアを活用

NVIDIA DGX B300 Technical Specifications

GPU	NVIDIA Blackwell Ultra GPU
GPU Memory	2.3 TB
Performance	144 PFLOPS FP4 Inference* 72 PFLOPS FP8 Training*
NVIDIA® NVSwitch™	2x
NVIDIA NVLink Bandwidth	14.4 TB/s Aggregate Bandwidth
System Power Usage	～14kW
CPU	Dual Intel® Xeon® 6776P Processors
Networking	8x OSFP ports serving 8x NVIDIA ConnectX-8 VPI > Up to 800Gb/s NVIDIA InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField-3 DPU > Up to 400Gb/s NVIDIA InfiniBand/Ethernet
Management Network	1GbE onboard NIC with RJ45 1GbE RJ45 Host Baseboard Management controller (BMC)
Storage	OS: 2x 1.9TB NVMe M.2 Internal Storage: 8x 3.84TB NVMe E1.S
Software	NVIDIA DGX OS / NVIDIA Mission Control / NVIDIA Base Command Manager / NVIDIA AI Enterprise Supports Red Hat Enterprise Linux / Rocky / Ubuntu
Rack Units (RU)	10RU
Operating Temp	10C° - 35C°
Support	Three-year Business-Standard Hardware and Software Support

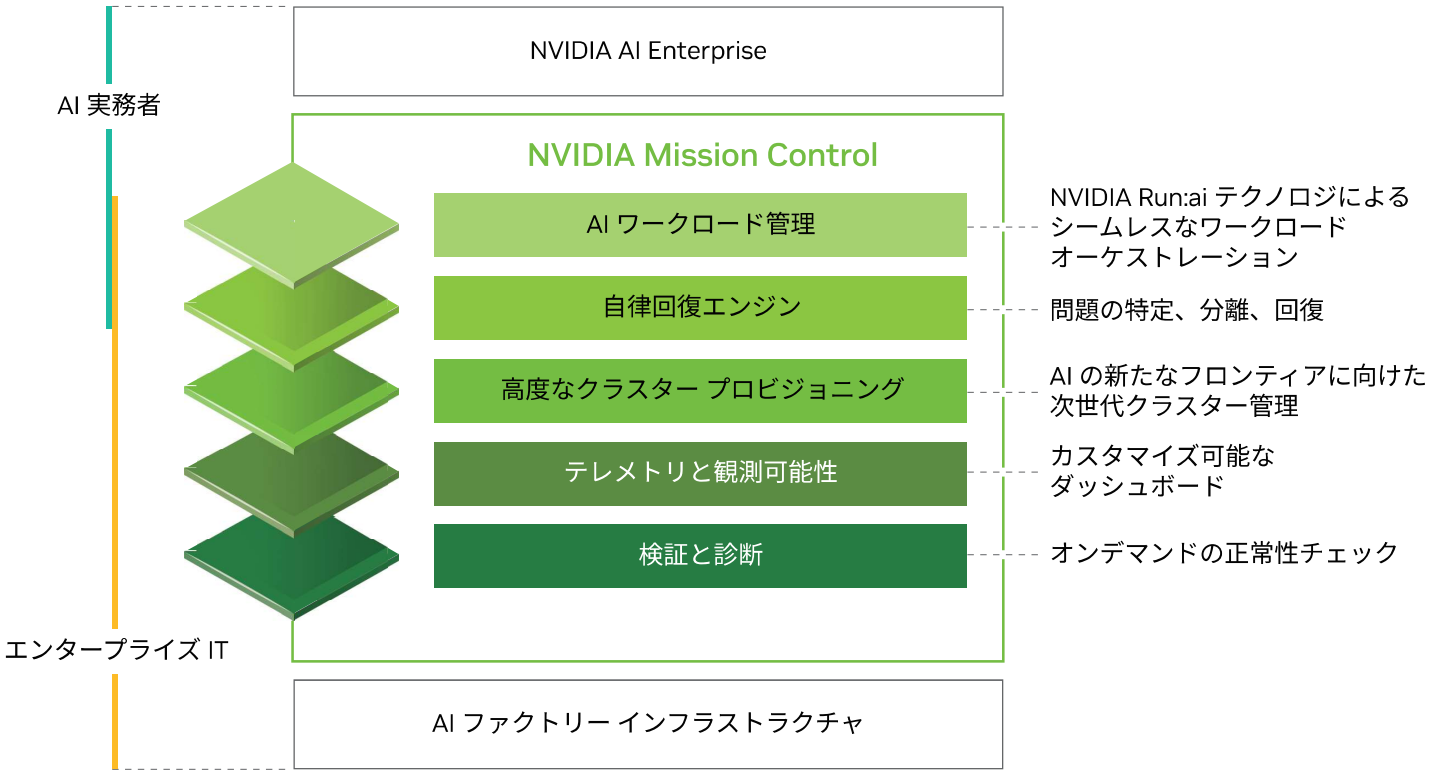
*Shown with sparsity.

現代のデータセンターのための設計図

DGX B300 は、NVIDIA MGX ラックに収まるように再設計されたシャーシを導入し、最新のデータ センターにおける互換性と拡張性を確保します。DGX B300 の空冷設計では、既存のデータ センター インフラストラクチャへの統合が容易になります。顧客は初めて、バスバーと電源ユニット (PSU) のオプションを選択できる柔軟な電力アーキテクチャを利用できるようになり、既存のインフラと持続可能性の目標に最適なものを選択できるようになりました。この新しい DGX 設計は、高速化されたコンピューティング インフラストラクチャの最適設計の設計図として機能し、AI インフラストラクチャを大規模に構築およびデプロイする柔軟な基盤を提供します。
最先端の設計と実用的な保守性を組み合わせることで、DGX B300 は、適応性と効率性に優れ、将来を見据えた AI インフラストラクチャの新たな基準を確立します。

モデルを実行し、基本的なタスクを自動化する NVIDIA Mission Control

NVIDIA Mission Control は、世界最高水準の運用チームのスキルをソフトウェアとして提供し、開発者のワークロードからインフラ、施設に至るまで、AI ファクトリー運用のあらゆる側面を強化します。推論とトレーニングに即時の俊敏性をもたらす一方で、インフラストラクチャの耐障害性にフルスタックのインテリジェンスを提供します。Mission Control により、あらゆる企業がハイパースケールの効率性で AI を実行できるようになり、AI の実験が加速します。さらに、NVIDIA AI Enterprise は、AI の開発とデプロイを効率化するソフトウェア スイートを提供し、NVIDIA DGX システム上で実行するように最適化されています。NVIDIA NIM™ マイクロサービスを使用して最適なモデルのデプロイを行い、スピード、使いやすさ、管理性、セキュリティを提供します。



NVIDIA DGX ソフトウェア スタック

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
お問合せ
https://www.hpctech.co.jp
info@hpctech.co.jp