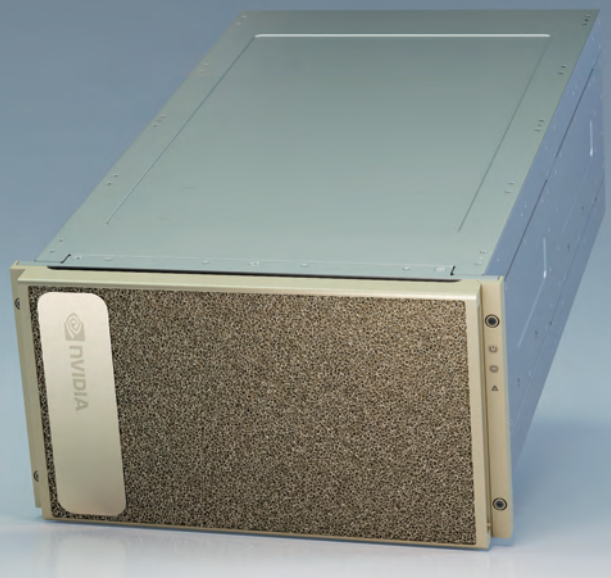
NVIDIA DGX A100
AI インフラストラクチャ向けの ユニバーサル システム



エンタープライズ AI のスケーリングへの挑戦

あらゆるビジネスで、人工知能（AI）を活用した変革が求められています。それは、困難な時代に生き残るためだけではなく、飛躍を遂げるためでもあります。ただし、そのためには、従来のアプローチを改善する AI インフラストラクチャ用のプラットフォームが必要です。これまでは、分析、トレーニング、推論のワークロードごとにサイロ化された低速のコンピューティング アーキテクチャが採用されていましたが、このアプローチでは、複雑さとコストが増大し、スケーリングの速度が制限され、現代の AI には対応できていませんでした。企業、開発者、データ サイエンティスト、研究者に本当に必要なのは、すべての AI ワークロードを統合し、インフラストラクチャを簡素化し、ROI を向上させる新たなプラットフォームです。

あらゆる AI ワークロードに対応するユニバーサル システム

NVIDIA DGX™ A100 は、分析からトレーニング、推論に至るまで、あらゆる AI ワークロードに対応するユニバーサル システムです。6U のフォーム ファクターで 5 petaFLOPS の AI パフォーマンスを発揮し、従来のコンピューティング インフラストラクチャに代わる 1 つの統合システムとして、計算処理密度の新たな水準を確立します。また、NVIDIA A100 Tensor コア GPU に搭載されたマルチインスタンス GPU 機能を利用することにより、コンピューティング パワーをきめ細かく配分するかつてない性能を実現します。これにより、管理者は特定のワークロードに適したサイズのリソースを割り当てられるようになり、シンプルなものや小さなものだけでなく、大規模かつ非常に複雑なジョブも確実にサポートできます。NGC の最適化されたソフトウェアで DGX ソフトウェア スタックが実行され、高密度な計算能力と完全なワークロードの柔軟性を組み合わせることにより、シングル ノードでの展開にも、NVIDIA DeepOps で展開された大規模な Slurm クラスターや Kubernetes クラスターにも最適な選択肢となっています。

NVIDIA DGXperts へのダイレクト アクセス

NVIDIA DGX A100 は、単なるサーバーではありません。DGX の世界最大の実験場である NVIDIA DGX SATURNV で得られた知識に基づいて構築された、ハードウェアとソフトウェアの完成されたプラットフォームです。そして、NVIDIA の何千人もの DGXperts によるサポートを提供します。DGXpert は AI に精通した専門家で、役立つアドバイスや設計に関する専門知識を提供し、AI 変革の加速に向けて支援します。過去 10 年にわたって蓄積してきた豊富なノウハウと経験を活かし、お客様が DGX への投資から最大限の価値を引き出せるようお手伝いします。DGXpert のサポートによって、重要なアプリケーションを迅速に実行し、スムーズな運用を維持し、インサイトを得るまでの時間を飛躍的に短縮することができます。

システムの仕様

GPU	**NVIDIA A100 Tensor コア GPU x 8**
GPU メモリ	**総計 320 GB**
パフォーマンス	**AI で 5 petaFLOPS INT8 で 10 petaOPS**
NVIDIA NVSwitch	**6**
消費電力	**6.5 kW（最大）**
CPU	**Dual AMD Rome 7742、総計 128 コア、2.25 GHz（ベース）、3.4 GHz（最大ブースト）**
システム メモリ	**1 TB**
ネットワーク	**シングルポート Mellanox ConnectX-6 VPI x 8 200 Gb/秒 HDR InfiniBand デュアルポート Mellanox ConnectX-6 VPI x 1 10/25/50/100/200 Gb/秒 Ethernet**
ストレージ	**OS: 1.92 TB M.2 NVME ドライブ x 2 内部ストレージ: 15 TB （3.84 TB x 4）U.2 NVME ドライブ**
ソフトウェア	**Ubuntu Linux OS**
重量	**123 kg**
梱包重量	**143 kg**
サイズ	**全高: 264.0 mm 全幅: 482.3 mm 奥行: 897.1 mm**
運用温度範囲	**5°C ～ 30°C**

最速での解決

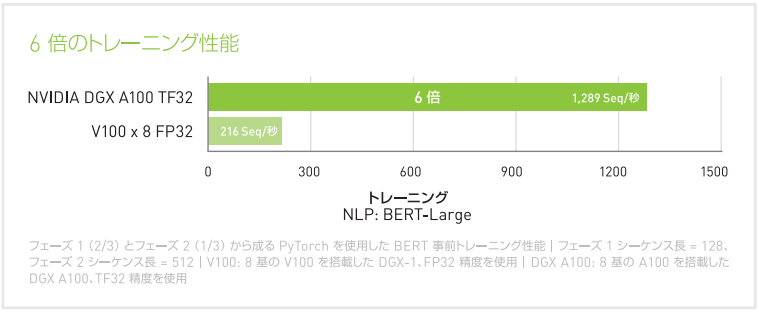
8 つの NVIDIA A100 Tensor コア GPU を搭載する NVIDIA DGX A100 は、これまでにないアクセラレーションを提供し、NVIDIA CUDA-X™ ソフトウェアとエンドツーエンドの NVIDIA データセンター ソリューション スタックに完全に最適化されています。NVIDIA A100 GPU は、FP32 と同じように動作する TF32 という新しい精度を利用して、前世代の 20 倍の演算速度の AI を実現します。そして最大の特長は、コードを変更することなくこの高速化が実現できる点です。NVIDIA の自動混合精度機能を使用すれば、FP16 精度を使用するコードを 1 行追加するだけで、さらに 2 倍の性能が得られます。また、クラス随一の毎秒 1.6 テラバイト（TB/秒）のメモリ帯域幅を備えており、これは前世代と比較すると 70% もの増加となります。さらに、前世代の 7 倍以上となる 40 MB のレベル 2 キャッシュをはじめとするオンチップ メモリを大幅に増強し、計算パフォーマンスを最大化しています。DGX A100 は次世代の NVIDIA NVLink™ を初めて搭載し、GPU 間の直接帯域幅を毎秒 600 ギガバイト（GB/秒）に倍増させています。これは、PCIe Gen 4 のほぼ 10 倍に相当します。他にも、前世代の 2 倍の速度を持つ次世代の NVIDIA NVSwitch も搭載しています。このかつてないパワーによって、最短でソリューションを実現でき、これまで不可能だったり、現実的ではなかったりした課題に取り組めるようになります。

世界で最も安全なエンタープライズ向け AI システム

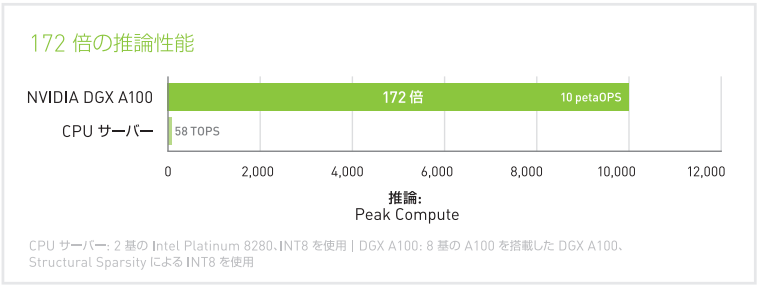
NVIDIA DGX A100 は、あらゆる主要なハードウェアおよびソフトウェア コンポーネントを保護する多層的なアプローチによって、AI を活用する企業において最も堅牢なセキュリティ体制を実現します。ベースボード管理コントローラー（BMC）、CPU ボード、GPU ボード、自動暗号化ドライブ、セキュア ブートなど、幅広いセキュリティ機能が組み込まれているため、IT 部門は脅威の評価や軽減に時間を費やすことなく、AI の運用に集中できます。

Mellanox によるデータセンターの比類なきスケーラビリティ

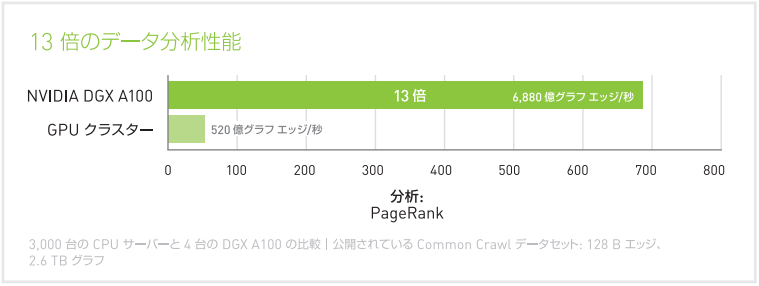
DGX システムの中で最速の I/O アーキテクチャを備えた NVIDIA DGX A100 は、NVIDIA DGX SuperPOD™ のような大規模な AI クラスターのための基本構成要素となり、企業は拡張性の高い AI インフラストラクチャの計画を策定できます。DGX A100 は、クラスタリング用に 8 つのシングルポート Mellanox ConnectX-6 VPI HDR InfiniBand アダプターと、ストレージとネットワーキング用に 1 つのデュアルポート ConnectX-6 VPI Ethernet アダプターを備えており、いずれも毎秒 200 Gb の性能を発揮します。大規模な GPU アクセラレーテッド コンピューティングと、最先端のネットワーキング ハードウェアおよびソフトウェアの最適化を組み合わせることで、数百、数千ノードにまでスケールアップが可能になり、対話型 AI や大規模な画像分類などの難易度の高い課題に対応できます。



フェーズ 1 (2/3) とフェーズ 2 (1/3) から成る PyTorch を使用した BERT 事前トレーニング性能 | フェーズ 1 シーケンス長 = 128、フェーズ 2 シーケンス長 = 512 | V100: 8 基の V100 を搭載した DGX-1、FP32 精度を使用 | DGX A100: 8 基の A100 を搭載した DGX A100、TF32 精度を使用



CPU サーバー: 2 基の Intel Platinum 8280、INT8 を使用 | DGX A100: 8 基の A100 を搭載した DGX A100、Structural Sparsity による INT8 を使用



3,000 台の CPU サーバーと 4 台の DGX A100 の比較 | 公開されている Common Crawl データセット: 128 B エッジ、2.6 TB グラフ

信頼できるデータセンターのリーダー企業と共に構築された実証済みのインフラストラクチャ ソリューション

ストレージとネットワーキングの技術を誇るリーディング プロバイダーとの連携により、NVIDIA が提供しているインフラストラクチャ ソリューションのポートフォリオに、NVIDIA DGX POD™ の最高クラスのリファレンス アーキテクチャが加わりました。これらのソリューションは、NVIDIA パートナー ネットワークを通じて、すぐに導入可能な完全統合型サービスとして提供されるため、より簡単かつ迅速に AI をデータセンターに導入できます。

NVIDIA DGX A100 の詳細については、**www.nvidia.com/ja-jp/data-center/dgx-a100/** をご覧ください。

NVIDIA