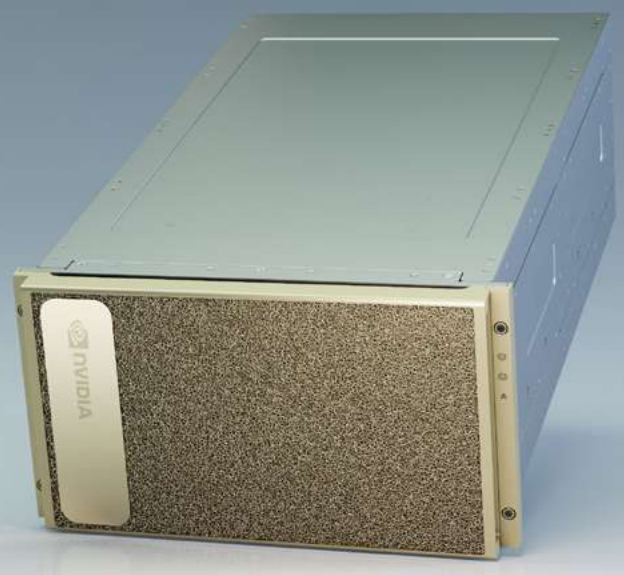
NVIDIA DGX A100

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5petaFLOPS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(MIG)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能力を実現し、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ます。総計640ギガバイト(GB)までのGPUメモリが利用できるので、大規模なトレーニングジョブのパフォーマンスが最大3倍に向上し、MIGインスタンスのサイズが2倍になります。DGX A100は、単純で小さなジョブだけでなく、大規模かつ非常に複雑なジョブにも対応します。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

システムの仕様

	NVIDIA DGX A100 640GB
GPU	NVIDIA A100 80 GB GPU x 8
GPUメモリ	総計 640 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 Rome7742、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	2 TB
ネットワーク	シングルポート Mellanox ConnectX-6 VPI 200 Gb/秒 HDR InfiniBand x 8 デュアルポート Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/ 秒 Ethernet x 2
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: 30 TB(3.84 TB x 8) U.2 NVMe ドライブ
ソフトウェア	Ubuntu Linux OS その他: Red Hat Enterprise Linux CentOS
重量	123.16 kg(最大)
梱包重量	163.16 kg(最大)
サイズ	全高:264.0 mm 全幅:482.3 mm 奥行:897.1 mm
運用温度範囲	5ºC〜30ºC

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。

最速での解決

8基のNVIDIA A100 Tensorコア GPUを搭載するNVIDIA DGX A100は、比類の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と同じように動作しながらも1秒あたりの浮動小数点演算回数(FLOPS)が前世代の20倍のAIを実現するTensor Float 32(TF32)という新しい精度に対応しています。最大の特長は、コードを変更することなくこの高速化を実現できる点です。またFP16を活用したNVIDIAの自動混合精度機能を使用すれば、A100ではコードを1行追加するだけで、さらに2倍の性能が得られます。

A100 80GB GPUは、高帯域幅メモリが40GB(HBM)から80GB(HBM2e)に倍増し、GPUメモリ帯域幅がA100 40GB GPUを30%上回る、世界初の毎秒2テラバイト超を実現しています。DGX A100は第3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倍の速度を持つ新しい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。

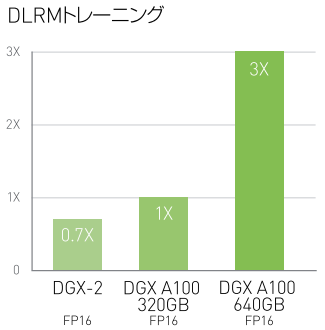
世界で最も安全なエンタープライズ向けAIシステム

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。

NVIDIA Mellanoxによるデータセンターの比類なきスケーラビリティ

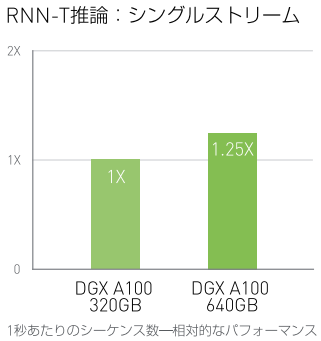
DGXシステムの中で最速のI/Oアーキテクチャを備えたNVIDIA DGX A100は、NVIDIA DGX Super POD™のような大規模なAIクラスターのための基本構成要素となり、企業は拡張性の高いAIインフラストラクチャの計画を策定できます。DGX A100は、クラスタリング用に8つのシングルポートNVIDIA Mellanox® ConnectX®-6 VPI HDR InfiniBandアダプターと、ストレージとネットワーキング用に最大2つのデュアルポートConnectX-6 VPI Ethernetアダプターを備えており、いずれも毎秒200 Gbの性能を発揮します。大規模なGPU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。

大規模モデルでのAIトレーニング性能が最大3倍向上



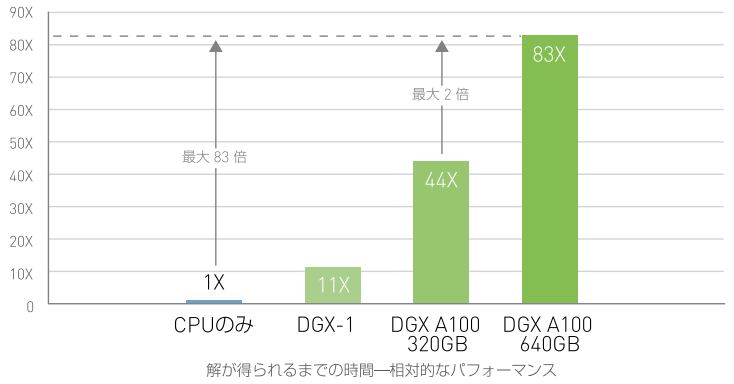
HugeCTR フレームワークでのDLRM、精度 = FP16 | DGX A100 640 GB(バッチサイズ = 48) x 1 | DGX A100 320 GB(バッチサイズ = 32) x 2 | DGX-2(V100 32 GB の16倍)(バッチサイズ = 32) x 1
GPU の台数に標準化した速度向上率

AI推論スループットが最大1.25倍向上



MLPerf 0.7 RNN-T:(1/7) MIG スライスで測定
フレームワーク:TensorRT 7.2、データセット = LibriSpeech、精度 = FP16

ビッグデータ分析ベンチマークでCPUより最大83倍、DGX A100 320 GBより2倍性能が向上



ビッグ データ分析ベンチマーク | 10 TB データセットで分析リテール クエリ 30 個、ETL、ML、NLP | CPU:Intel Xeon Gold 6252 2.10 GHz の 19 倍、Hadoop | DGX-1(V100 32 GB の 8 倍)の 16 倍、RAPIDS/Dask | DGX A100 320 GB の 12 倍および DGX A100 640 GB の 6 倍、RAPIDS/Dask/BlazingSQL GPU の台数に標準化した速度向上率

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、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(NPN)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、データセンターへの AI 導入を簡素化かつ迅速化できます。

GDEPソリューションズ株式会社
東京都文京区本郷三丁目34番3号 本郷第一ビル8階
TEL:03-5802-7050 FAX:03-5842-4123
https://gdep-sol.co.jp