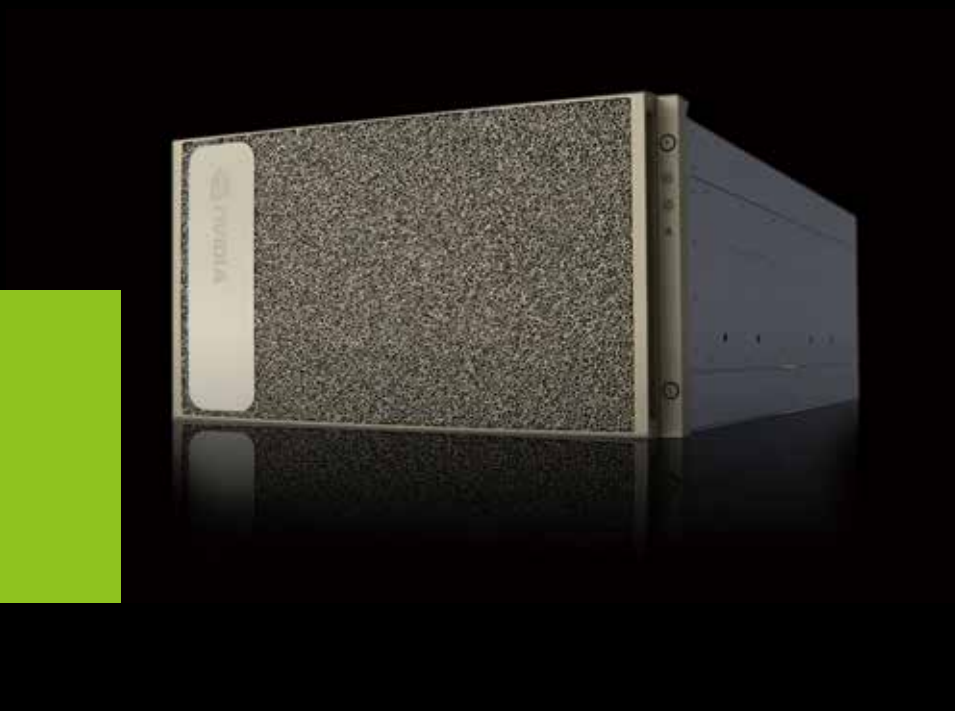

NVIDIA DGX A100

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。

	NVIDIA DGX A100 640GB	NVIDIA DGX A100 320GB
GPUs	8x NVIDIA A100 80 GB GPUs	8x NVIDIA A100 40 GB GPUs
GPU Memory	640 GB total	320 GB total
Performance	5 petaFLOPS AI 10 petaOPS INT8	
NVIDIA NVSwitches	6	
System Power Usage	6.5 kW max	
CPU	Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)	
System Memory	2 TB	1 TB
Networking	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 2x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 1x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 30 TB (8x 3.84TB) U.2 NVMe drives	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15 TB (4x 3.84TB) U.2 NVMe drives
Software	Ubuntu Linux OS	
System Weight	271.5 lbs (123.16 kgs) max	
Packaged System Weight	359.7 lbs (163.16 kgs) max	
System Dimensions	Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) max Length: 35.3 in (897.1 mm) max	
Operating Temperature Range	5-30 °C (41-86 °F)	

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40MB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200Gb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。

6 倍のトレーニング性能

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

172 倍の推論性能

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

13 倍のデータ分析性能

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パートナーを通じて、統合されたすぐに展開可能な製品を提供します。これらのソリューションで、データセンターへのAI導入がよりシンプルにITの高速化を実現します。

詳細については https://www.hpc.co.jp/ai-deeplearning/product/nvidia-dgx-a100/ をご覧ください。

NVIDIA Partner Network（NPN）に認定されました

HPCシステムズはNVIDIA社のパートナー認定制度 "NVIDIA Partner Network (NPN)" においてHigh Performance Computing (HPC) ならびに Deep Learning の ELITE PARTNER に認定されています。また、DGX製品の販売資格である「Advanced Technology Program (ATP)」 を保有しています。

NVIDIA ELITE PARTNER

HPC SYSTEMS