NVIDIA DGX-2
最も複雑な AI の問題に挑む世界最強のディープラーニング システム

AI の要求に応えるスケーリングへの挑戦

ビジネスや研究分野の差し迫った課題に対応するため、ディープ ニューラル ネットワークは、その規模と複雑性を急速に拡大しています。今日の最先端 AI ワークロードをサポートするために必要なコンピューティング能力は、従来のデータセンター アーキテクチャが持つ能力をはるかに超えています。ますます大規模なアクセラレーテッド コンピューティング クラスターが開発され、データセンターの拡大が極限まで進められたことから、モデルの並列性の利用をさらに進めている最新技術の前に GPU 間バンド幅の壁が立ちはだかっています。この壁を打ち破り、世界を変革できるようなインサイトをすばやく得るには、ほぼ無制限の AI コンピューティング スケーリングを提供する新しいアプローチが求められます。

過去に例を見ないトレーニング能力

AI はますます複雑化し、かつてないレベルの計算処理能力が必要とされています。NVIDIA® DGX-2™ は、世界最先端の GPU を 16 基備えた世界初の 2 petaFLOPS システムとして、これまでトレーニング不可能だった最新のディープラーニング モデルを高速化します。画期的な GPU スケールを使用して、1 ノードで 4 倍大きなモデルのトレーニングを行うことができます。レガシー x86 アーキテクチャと比較すると、DGX-2 と同等の能力で ResNet-50 のトレーニングを行うには、デュアル Intel Xeon Gold CPU を備えたサーバーが 300 台必要になり、270 万ドルのコストがかかります。

NVIDIA NVSwitch —革新的な AI ネットワーク ファブリック

最先端の研究では、モデルの並列性を自由に活用する必要があり、これまでにないレベルの GPU 間バンド幅が必要になります。NVIDIA は、このニーズに対応するために NVSwitch を開発しました。ダイヤルアップが超高速ブロードバンドにまで進化したように、NVSwitch も将来を見据えたネットワーク ファブリックを提供しています。NVIDIA DGX-2 では、モデルの複雑性とサイズが従来のアーキテクチャの制約を受けることはありません。DGX-2 は、前世代より 24 倍高速になった 2.4 TB/ 秒のバイセクション バンド幅を提供します。このネットワーク ファブリックを使用した並列モデル トレーニングをぜひ体験してください。この新しい相互接続による「スーパーハイウェイ」は、一度に 16 基の GPU にわたる分散トレーニングのパワーを活用するようなモデルに無限の可能性をもたらします。

システム仕様

GPU	16X NVIDIA® Tesla V100
GPU メモリ	総計 512GB
性能	2 petaFLOPS
NVIDIA CUDA® コア	81920
NVIDIA Tensor コア	10240
NVSwitche	12
最大消費電力	10 kW
CPU	Dual Intel Xeon Platinum 8168, 2.7GHz, 24 コア
システムメモリ	1.5TB
ネットワーク	8X 100Gb/sec Infiniband/100GigE Dual 10/25/40/50/100 GbE
ストレージ	OS: 2X 960GB NVME SSD 内部ストレージ: 30TB (8x 3.84TB) NVME SSD
ソフトウェア	Ubuntu Linux OS (詳細は次ページ)
システム重量	163.3 kg
システムサイズ	全高: 440mm 全幅: 482mm 奥行: 795mm (フロント ベゼルなし) 834mm (フロントベゼルあり)
梱包重量	181.4kg
梱包サイズ	D1016mm x W610mm x H904mm* *パレット含む
運用温度範囲	5℃ - 35℃

AI をまったく新たなレベルに拡大

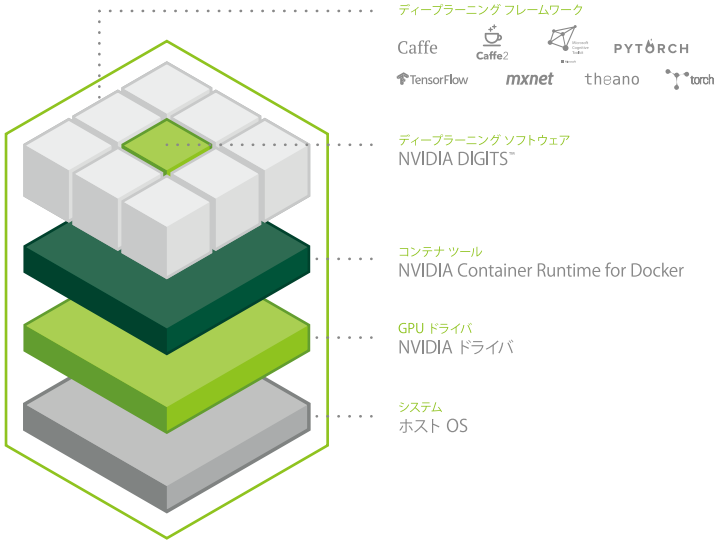
近年、企業にとって AI のパワーを迅速に展開することはビジネス上の必須事項になっており、コストと複雑性を増大させることなく AI の規模を拡大することが必要になっています。NDIVIA は DGX-2 を開発し、迅速な展開と簡素な運用をより大規模に行うための DGX ソフトウェアを統合しました。DGX-2 は、AI のスケールアップを最速で行うことができるソリューションとしてすぐに使用でき、仮想化のサポートと相まって、お客様がエンタープライズ グレードのプライベート AI クラウドを独自に構築できるようにします。容易な拡張を目的とした専用のアーキテクチャと高速展開モデルを利用して、お客様は、インフラストラクチャの構築にかかる時間を削減し、インサイトを得るためにより多くの時間を費やすことができます。

エンタープライズ グレード AI 基盤

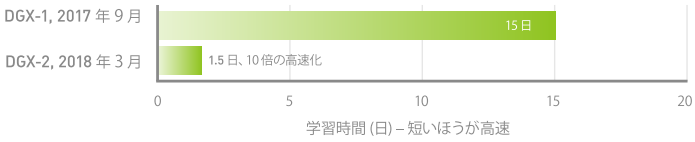
ビジネスにとって AI プラットフォームが決定的に重要な意味を持つなら、信頼性、可用性、保守性 (RAS) を念頭に置いて設計された AI プラットフォームが必要です。DGX-2 は、厳格に AI の 24 時間運用を目指して設計されたエンタープライズ グレードのシステムです。また、RAS を目的として、計画外のダウンタイムを削減し、保守性を無駄なく高め、継続的な運用を維持します。

調整や最適化に費やす時間を減らし、新たな発見のための時間を増やしてください。NVIDIA のエンタープライズ グレードのサポートは、時間のかかるハードウェアとオープン ソース ソフトウェアのトラブル シューティング作業からお客様を解放します。すべての DGX システムにおいて、ソフトウェア、ツール、NVIDIA 専門スタッフからなる総合的なソリューションが、すばやい利用開始、迅速なトレーニング、スムーズな運用を実現します。

NVIDIA DGX-2 ソフトウェア スタック



1 年未満で 10 倍の高速化



Performance gain through hardware and software improvements across the stack
Workload: FairSeq, 55 epochs to accuracy. PyTorch training performance.

DGX-2 は NVIDIA Advanced Technology Program (ATP)パートナーを通じて提供されます。

HPCシステムズ株式会社
www.hpc.co.jp

詳細については www.nvidia.co.jp/DGX-2 をご覧ください。