ソリューション概要

ONTAP AI

ネットアップとNVIDIAがデータ パイプラインを
簡易化、加速、統合し、MLとDLに対応

NetApp | NVIDIA

AIインフラの課題

人工知能（AI）、機械学習（ML）、ディープ ラーニング（DL）を活用すると、競争が激化する一方の市場で、不正行為の検出、顧客関係の改善、サプライ チェーンの最適化、革新的な製品とサービスの提供を実現できます。お客様の企業も、AIという新しい手法を駆使してデジタル変革を先導し、優れた競争力を獲得しようとしている組織のひとつなのではないでしょうか。しかし、AIのメリットを最大限に引き出すには、まず、いくつもの大きな課題を克服しなければなりません。

自力による統合環境の構築は複雑です。既製のMLとDLでコンピューティング、ストレージ、ネットワーク、ソフトウェアを組み合わせて統合しようとすると、複雑すぎて導入に長い時間がかかります。貴重なデータ サイエンティストのリソースを、システム統合という作業に無駄に費やす結果となります。

拡張性に優れた予測可能なパフォーマンスの達成は困難を極めます。DL導入のベストプラクティスでは、スモール スタートから始めて徐々に拡張する方法が推奨されています。従来、コンピューティング リソースと直接接続型ストレージ（DAS）は、AIワークフローにデータを提供する手段として使用されてきました。しかし旧来の方法では拡張時に、進行中の処理の中断やダウンタイムが発生することがあります。

データ サイエンティストの生産性に影響するシステム停止：MLとDLのインフラは、相互に依存し合う多数のハードウェアとソフトウェアで構成されるため複雑です。インフラを常に稼働させておくには、AIに関するフルスタックの深い専門知識が必要です。ダウンタイムが発生したりAIのパフォーマンスが低下したりすると、連鎖的に開発者の生産性が低下するだけでなく、運用コストの大幅な増加にもつながります。

ソリューション

AI、ML、DLの可能性を最大限に引き出せるときが来ました。NVIDIA DGX™システムとネットアップのクラウド対応オールフラッシュ ストレージを基盤とするNetApp® ONTAP® AIの実績あるアーキテクチャによって、データ パイプラインを簡易化、加速、統合できます。データ ファブリックを実現することでエッジ、コア、クラウドにわたるデータの流れを確実に合理化し、テストと推論作成に必要な時間を短縮できます。

主なメリット

柔軟な検証済みソリューションでリスクを軽減

- 設計の複雑さを解消し、推測に頼らず迅速に導入
- 事前設定済みソリューションで設定と導入を合理化

最適なパフォーマンスとスケーラビリティを実現

- 小規模構成でスタート、無停止で拡張
- 成果をすばやく引き出せるハイパフォーマンス ソリューション

統合されたデータ パイプラインを構築

- エッジからコア、クラウドまでを統合するパイプラインで、データをインテリジェントに管理
- AIのエキスパートとシンプルなサポート オプションでソリューションの導入を支援

AIワークロードを統合

- インフラのサイロを解消
- ビジネス ニーズに柔軟に対応

NetApp ONTAP AIは、世界初の5ペタフロップAIシステムであるNVIDIA DGX A100システムと、NVIDIA Mellanox高性能イーサネット スイッチが組み込まれた、初めての統合インフラスタックです。AIワークロードの統合、導入の簡易化、投資収益率の向上を実現します。

「ディープ ラーニングは、私たちが関わる市場のほぼすべてを根本から変えています。私たちはさまざまな市場でディープ ラーニングを採用し、できることを実行する技術を推し進めています。NVIDIA DGXシステムとネットアップのオールフラッシュ ストレージを基盤とするNetApp ONTAP AIは、ディープ ラーニングに最適なデータ パイプラインの構築を簡易化し、加速します」

Cambridge Consultants 人工知能部門ディレクター Tim Ensor氏

図1）DGX A100を使用したONTAP AIアーキテクチャ（2ノード、4ノード、8ノード構成）

柔軟な検証済みソリューションでリスクを軽減

AIによる技術革新が急速に進んでいることから、実効性のあるAIインフラを設計しようとすると、さまざまな課題が伴います。ONTAP AIなら、実績のあるリファレンス アーキテクチャを使用して推測に頼らなくてもすぐにAIインフラを構築できます。または、簡単に調達、導入できる事前設定済み統合ソリューションを選択すれば、複雑な設計や管理に煩わされることもありません。

ONTAP AI統合ソリューションには4つの事前設定済みオプションが用意されていて、機能拡張や高度なソフトウェアを選択して利用できます。この統合ソリューションにより、オンサイトでの設置やインシデント発生時のレポートから解決まで、1つの窓口による包括的なサポートが提供されるので、複雑さが大幅に軽減されます。

最適なパフォーマンスとスケーラビリティを実現

DLのトレーニング ルーチンには、大量のコンピューティング能力が必要です。イメージ トレーニングに必要な時間を短縮できれば、コンピューティング コストを全体的に削減しつつ、AIによる技術革新と生産性向上を加速できます。

新しいNVIDIA Ampereアーキテクチャを基盤として構築されたDGX A100システムは、前世代アーキテクチャの最大6倍のトレーニング性能を提供します。分析、トレーニング、推論のためのデータセンター相当のコンピューティング インフラが単一システムに統合されています。DGX A100システムは、CPUシステムに比べて設置面積が25分の1、消費電力が20分の1、コストがわずか10分の1という軽量設計です。

最先端のコンピューティングに投資するには、1秒で何千ものトレーニング イメージを処理できる最先端のストレージが必要です。きわめて負荷の高いDLトレーニング ワークロードの要件を、ハイパフォーマンスのデータ サービス ソリューションで満たせなくてはなりません。

ネットアップのオールフラッシュ ストレージでは、2GB/秒の持続可能なスループット（ピーク時は5GB/秒）を達成できる見込みです。さらに、レイテンシは1ミリ秒を十分に下回り、GP利用率は95%を超えます。NASワークロードの場合、NetApp AFF A800システムは1台で、シーケンシャル読み取りで25GB/秒のスループット、スモール ランダム リードで100万IOPSのパフォーマンスを500マイクロ秒未満のレイテンシで実現します。

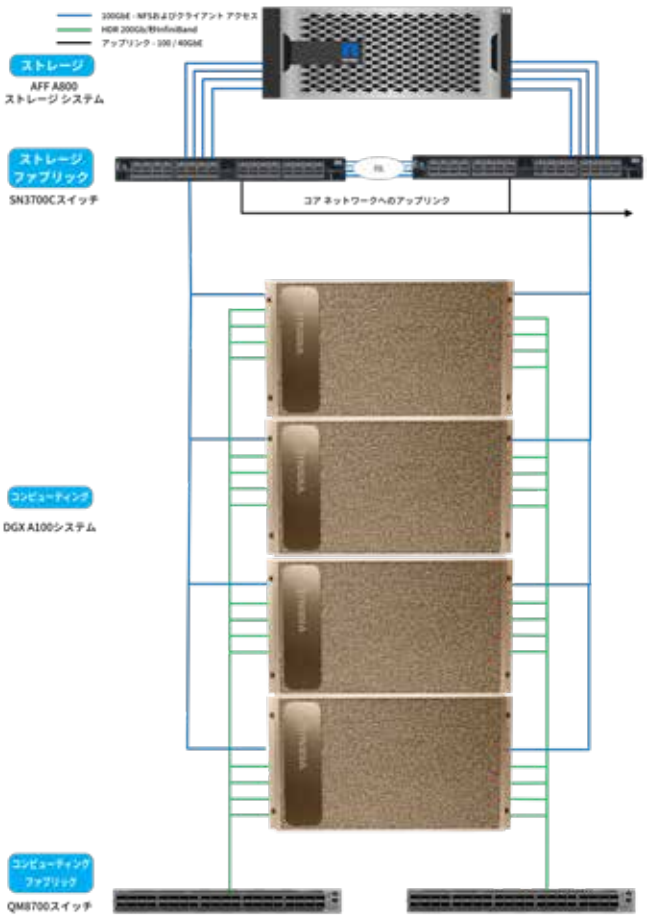

図2）Mellanox Spectrum 100GbEスイッチを使用したONTAP AI 4ノード構成

ネットアップのラックスケール アーキテクチャにより、オールフラッシュ ストレージを使用して数十テラバイトから数十ペタバイトまで拡張可能です。さらに、NetApp ONTAP FlexGroupを使用すれば、最大20PBのグローバル ネームスペースで4,000億を超えるファイルを処理できます。

エッジからコア、クラウドにわたる統合されたデータ パイプラインを構築

ONTAP AIはデータ ファブリックを使用して、データ パイプライン全体のデータ管理を単一プラットフォームで統合します。データが転送中でも使用中でも、保存されている状態でも、同じツールでセキュアに管理、保護することで、コンプライアンス要件を確実に満たします。DL環境に問題が発生した場合は、ネットアップの実績あるサポート モデルを通じてトラブルシューティングを実施し、ガイダンスを提供いたします。

AIワークロードを統合

企業は、十分に利用されていないインフラや、AIワークロードにほとんど活用されていないインフラのサイロを排除できるようになります。DL環境に問題が発生した場合は、ネットアップの実績あるサポート モデルを通じてトラブルシューティングを実施し、ガイダンスを提供いたします。このソリューションは、ビジネス ニーズに柔軟に対応する単一プラットフォームに、分析、トレーニング、推論を統合します。また、レガシー アーキテクチャよりもTCOを低減できます。

ネットアップとNVIDIA：イノベーションを共同で推進

ONTAP AIの中核をなすDGX A100システムは、データセンターAIのための汎用ビルディング ブロックであり、トレーニング、推論、データ サイエンス、その他のハイパフォーマンス ワークロードをサポートします。各DGX A100システムには、8つのNVIDIA A100 TensorコアGPUとデュアル構成の第2世代AMD EPYC™プロセッサが搭載されています。また、最新の高速NVIDIA Mellanox 100 / 200Gbイーサネットと、InfiniBand対応のConnectX-6アダプタ インターコネクトが統合されています。

新しいNVIDIAマルチインスタンスGPU（MIG）テクノロジにより、DGX A100システムをシステムあたり最大56のインスタンスに分割することで、複数の小規模ワークロードを高速化することができます。この高速化により、ONTAP AIでGPUパフォーマンスを効率的に割り当てることができるので、企業のデータ サイエンス チームはより迅速な反復処理や再現の自動化が可能となり、AIプロジェクトをより高い品質で最大3カ月早く実現できるようになります。

業界初のエンドツーエンドNVMeテクノロジを搭載したNetApp AFFシステムは、業界最速の柔軟性に優れたオールフラッシュ ストレージとして、MLおよびDLのプロセスに絶えずデータを提供します。AFF A800システムは、競合他社のソリューションの4倍の速さでDGXシステムにデータを提供できます[1]。

ONTAP AIソリューションには、Mellanox Spectrumイーサネット スイッチが組み込まれており、AI環境に求められる低レイテンシ、高密度、ハイパフォーマンス、省電力を実現します。

1. オールフラッシュ クラスタあたりの最大読み取りスループット（300GB/秒）と、代表的な競合企業のスループット（75GB/秒）を比較

ネットアップが実現するデータ ファブリックによって、業界最高のデータ管理とクラウド統合環境を実現し、重要なデータの管理と保護を実現しつつ、DL環境の構築を加速することができます。ONTAPは、全体的なデータ削減で22:1という圧倒的な削減率を実現し、DASと比べてTCOを最大で54%低減します。

DGX A100システムは、AIとデータ サイエンスのワークロード向けに最適化されたソフトウェアを含むNVIDIA DGXソフトウェア スタックを搭載しています。最大のパフォーマンスを実現することで、企業はAIインフラへの投資をより早く回収できます。

NetApp AIコントロール プレーンは、ネットアップが実現するデータ ファブリックにKubernetesとKubeflowを統合することにより、AIデータ管理を簡易化します。この統合ソリューションは、エッジからコア、クラウドに至るまで、最適なデータ可用性とモビリティを実現します。AIコントロール プレーンが強化されたNetApp DataOps Toolkitは、Pythonライブラリを使用してデータ サイエンティストやデータ エンジニアによる膨大なデータ管理タスクを簡易化します。たとえば、新しいデータ ボリュームのプロビジョニング、データ ボリュームの瞬時のクローニング、データ ボリュームのNetApp Snapshot™コピーの作成により、トレーサビリティとベースライン管理が確保されます。

成功には最適なツールが欠かせません。ONTAP AIでは、Domino Data Lab、Iguazioなどの業界をリードする機械学習運用向け（MLOps）ソフトウェアとの相互運用性も検証済みです。これらのツールを活用することで、AI環境の価値を最大限に引き出し、分析情報をより迅速に得られるようになります。

ソリューションの構成

- NVIDIA DGX A100システム
- NetApp AFF Aシリーズ ストレージ システム（ONTAP 9搭載）
- NVIDIA Mellanox Spectrum SN3700C、NVIDIA Mellanox Quantum QM8700、NVIDIA Mellanox Spectrum SN3700-V
- NVIDIA DGXソフトウェア スタック
- NetApp AIコントロール プレーン
- NetApp DataOps Toolkit

リファレンス アーキテクチャ

ネットアップは、特定業界のユースケースを対象として、ONTAP AIをベースにした以下のリファレンス アーキテクチャを公開しています。

- 医療機関向けONTAP AIリファレンス アーキテクチャ：画像診断
- 自動運転ワークロード向けONTAP AIリファレンス アーキテクチャ：ソリューション デザイン
- 金融サービス ワークロード向けONTAP AIリファレンス アーキテクチャ：ソリューション デザイン

ネットアップについて

ジェネラリストが多い世界で、ネットアップはスペシャリストとしての存在感を示しています。お客様がデータを最大限に活用できるようにすることを1つの目標として、支援に全力を注いでいます。ネットアップは、信頼できるエンタープライズクラスのデータ サービスをクラウドにもたらし、またクラウドのシンプルな柔軟性をデータセンターにもたらします。業界をリードするネットアップのソリューションは、さまざまなお客様の環境や業界最大手のパブリック クラウドに対応します。

クラウド主導のData-Centricなソフトウェア企業であるネットアップは、お客様に最適なデータ ファブリックの構築をサポートし、クラウド対応をシンプルに実現し、必要なデータ、サービス、アプリケーションを適切なユーザにいつでも、どこからでもセキュアに提供できる唯一のベンダーです。www.netapp.com/ja

NVIDIAについて

NVIDIAは、1999年に開発したGPUによってPCゲーム市場の成長を一気に加速させました。この発明は、最新のコンピュータ グラフィックスの定義を変えただけでなく、並列処理にも革命をもたらしました。最近は、GPUを使用したディープ ラーニングによって最新のAI開発ブームに火をつけました。次世代のコンピューティングであるAIでは、GPUがコンピュータやロボット、自動運転車の頭脳として世の中を認識し理解する役目を担います。
詳細については、www.nvidia.comをご覧ください。

NetApp