NVIDIA DGXシステムを搭載したNetApp AIPod

NetApp Solutions

NetApp
May 03, 2024

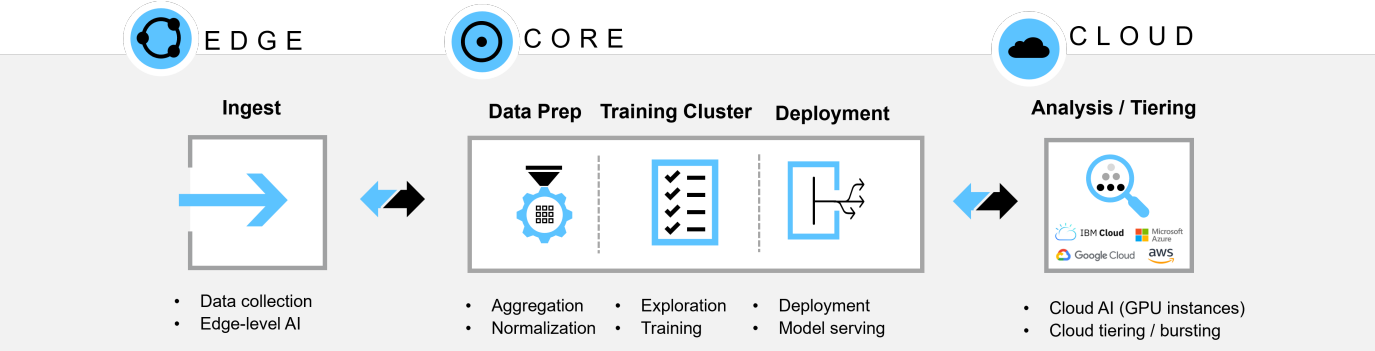
目次

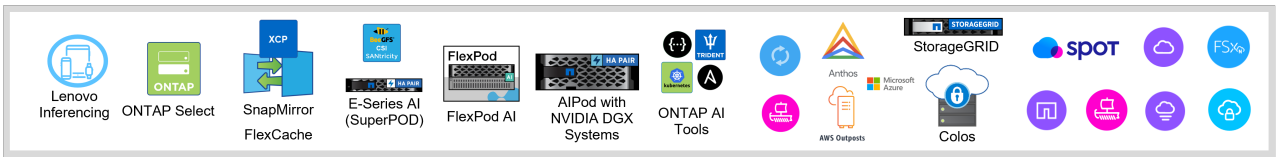
NVIDIA DGXシステム搭載NetApp AIPod -はじめに

NetApp 解決策 エンジニアリング

NVIDIA DGX™システムとNetAppクラウド対応ストレージシステムを搭載したNetApp AIPodは、設計の複雑さと推測を排除することで、機械学習（ML）や人工知能（AI）のワークロード向けのインフラ導入を簡易化します。次世代のワークロードに卓越したコンピューティングパフォーマンスを提供するNVIDIA DGX BasePOD設計を基盤とするAIPodとNVIDIA DGXシステムは、小規模構成から始めてシステムを停止せずに拡張できるNetApp AFFストレージシステムを追加し、エッジからコア、クラウドまでデータをインテリジェントに管理します。次の図に示すように、NetApp AIPodはNetApp AIソリューションの大規模なポートフォリオの一部です。

_ NetApp AIソリューションポートフォリオ_





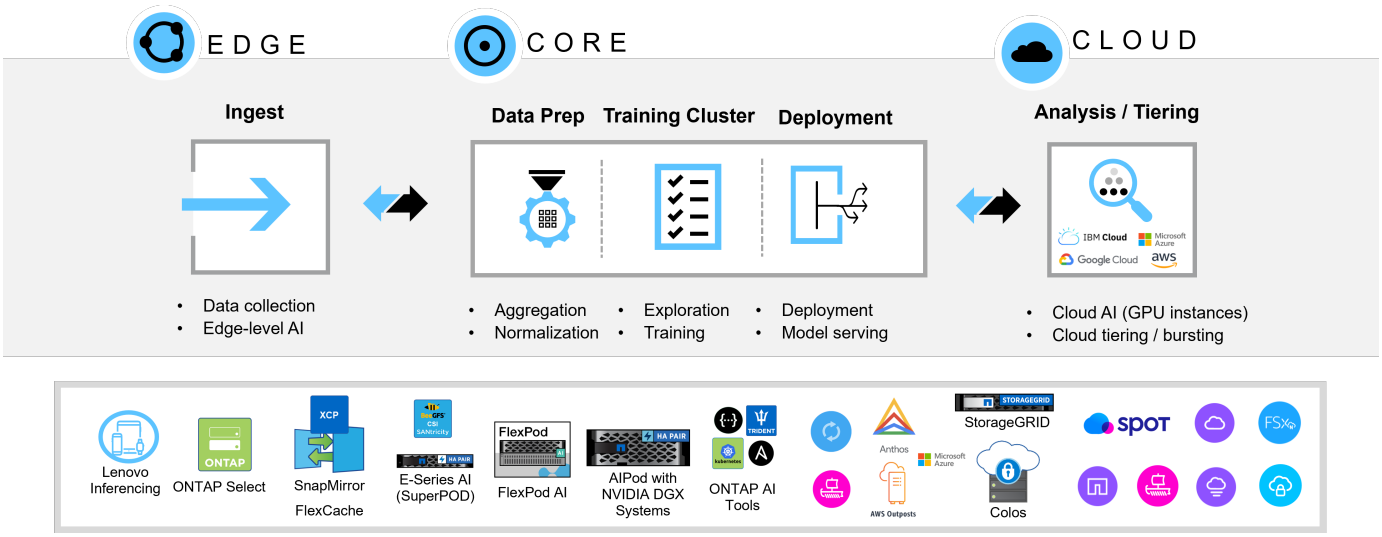
このドキュメントでは、AIPodリファレンスアーキテクチャの主要コンポーネント、システム接続情報、および解決策サイジングガイダンスについて説明します。本ドキュメントは、ML/DLや分析ワークロード向けのハイパフォーマンスインフラの導入に関心をお持ちのNetApp、パートナー様のソリューションエンジニア、お客様の戦略的意思決定者を対象としています。

NVIDIA DGXシステム搭載NetApp AIPod -はじめに

NetApp 解決策 エンジニアリング

NVIDIA DGX™システムとNetAppクラウド対応ストレージシステムを搭載したNetApp AIPodは、設計の複雑さと推測を排除することで、機械学習（ML）や人工知能（AI）のワークロード向けのインフラ導入を簡易化します。次世代のワークロードに卓越したコンピューティングパフォーマンスを提供するNVIDIA DGX BasePOD設計を基盤とするAIPodとNVIDIA DGXシステムは、小規模構成から始めてシステムを停止せずに拡張できるNetApp AFFストレージシステムを追加し、エッジからコア、クラウドまでデータをインテリジェントに管理します。次の図に示すように、NetApp AIPodはNetApp AIソリューションの大規模なポートフォリオの一部です。

_ NetApp AIソリューションポートフォリオ_



このドキュメントでは、AIPodリファレンスアーキテクチャの主要コンポーネント、システム接続情報、および解決策サイジングガイダンスについて説明します。本ドキュメントは、ML/DLや分析ワークロード向けのハイパフォーマンスインフラの導入に関心をお持ちのNetApp、パートナー様のソリューションエンジニア、お客様の戦略的意思決定者を対象としています。

NetApp AIPodとNVIDIA DGXシステム-ハードウェアコンポーネント

NetApp AFFストレージシステム

NetApp AFFの最先端のストレージシステムを使用すると、IT部門は、業界をリードするパフォーマンス、卓越した柔軟性、クラウド統合、業界最高のデータ管理機能によって、エンタープライズストレージの要件を満たすことができます。フラッシュに特化して設計されたAFF システムは、ビジネスクリティカルなデータの高速化、管理、保護に役立ちます。

AFF A900ストレージシステム

NetApp ONTAPデータ管理ソフトウェアを基盤とするNetApp AFF A900は、組み込みのデータ保護機能、オプションのランサムウェア対策機能、最も重要なビジネスワークロードのサポートに必要なハイパフォーマンスと耐障害性を提供します。ミッションクリティカルな運用の中断を解消し、パフォーマンスの調整を最小限に抑え、ランサムウェアの攻撃からデータを保護します。次のメリットがあります。
•業界をリードするパフォーマンス
•妥協のないデータセキュリティ
•シンプルな無停止アップグレード

NetApp AFF A900ストレージシステム_

業界トップクラスのパフォーマンス

AFF A900は、ディープラーニング、AI、高速分析などの次世代ワークロードや、Oracle、SAP HANA、Microsoft SQL Server、仮想アプリケーションなどの従来型エンタープライズデータベースを容易に管理できます。ビジネスクリティカルなアプリケーションを、HAペアあたり最大240万IOPS、100 µ sの低レイテンシで最高速度で実行し、従来のNetAppモデルに比べてパフォーマンスを最大50%向上させます。NFS over RDMA、pNFS、セッショントランキングを使用すると、既存のデータセンターネットワークインフラを使用して、次世代アプリケーションに必要な高レベルのネットワークパフォーマンスを実現できます。
また、SAN、NAS、オブジェクトストレージをユニファイドマルチプロトコルでサポートしているため、拡張も拡張も可能です。また、単一の統合ONTAPデータ管理ソフトウェアで、オンプレミスでもクラウドでも、最大限の柔軟性を実現できます。さらに、Active IQとCloud Insightsが提供するAIベースの予測分析でシステムの健全性を最適化できます。

妥協のないデータセキュリティ

AFF A900システムには、アプリケーションと整合性のあるNetApp統合データプロテクションソフトウェアがすべて含まれています。組み込みのデータ保護機能と最先端のランサムウェア対策ソリューションを提供し、事前の保護と攻撃後のリカバリを実現します。悪意のあるファイルがディスクに書き込まれるのを防ぎ、ストレージの異常を簡単に監視して分析情報を得ることができます。

シンプルな無停止アップグレード

AFF A900は、A700の既存のお客様に、システム停止を伴わないシャーシ内アップグレードを提供します。NetAppでは、高度な信頼性、可用性、保守性、管理性（RASM）機能を使用して、簡単に更新し、ミッションクリティカルな運用の中断を排除できます。さらに、NetApp ONTAPソフトウェアはすべてのシステムコンポーネントのファームウェアアップデートを自動的に適用するため、運用効率がさらに向上し、ITチームの日常業務が簡素化されます。

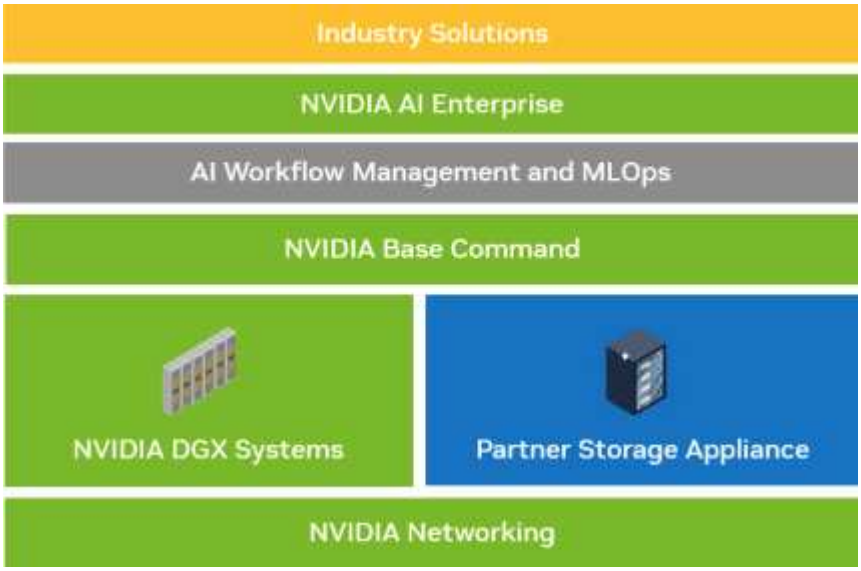
最大規模の環境には、AFF A900システムが最高レベルのパフォーマンスと容量のオプションを提供します。

他のNetAppストレージシステム（AFF A800、AFF C800、AFF A400、AFF C400、AFF A250など）には、より低コストで小規模な環境向けのオプションが用意されています。

NVIDIA DGX BasePOD

NVIDIA DGX BasePODは、NVIDIAのハードウェアおよびソフトウェアコンポーネント、MLOpsソリューション、サードパーティ製ストレージで構成される統合解決策です。NVIDIA製品と検証済みパートナーソリューションを使用したスケールアウトシステム設計のベストプラクティスを活用することで、お客様はAI開発のための効率的で管理しやすいプラットフォームを実装できます。図1は、NVIDIA DGX BasePODのさまざまなコンポーネントを示しています。

*NVIDIA DGX*ベース*POD*解決策



NVIDIA DGX H100システム

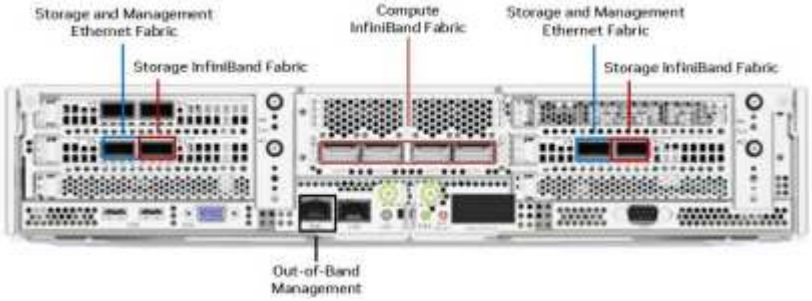
NVIDIA DGX H100™システムは、NVIDIA H100 Tensor Core GPUの画期的なパフォーマンスによって加速されるAIの大国です。

NVIDIA DGX H100システム_

DGX H100システムの主な仕様は次のとおりです。
•NVIDIA H100 GPU×8
•GPUあたり80 GBのGPUメモリ、合計640 GB。
•NVIDIA NVSwitch™チップ4個。
•PCIe 5.0に対応したデュアル56コアインテル®Xeon®Platinum 8480プロセッサー。
•2 TBのDDR5システムメモリ。
•8つのシングルポートNVIDIA ConnectX-7（InfiniBand /イーサネット）アダプタと2つのデュアルポートNVIDIA ConnectX-7（InfiniBand /イーサネット）アダプタを提供するOSFPポート×4

•DGX OS用の1.92TB M.2 NVMeドライブ×2、ストレージ/キャッシュ用の3.84TB U.2 NVMeドライブ×8
•最大出力10.2 kW。
DGX H100 CPUトレイの背面ポートを次に示します。OSFPポートのうち4つは、InfiniBandコンピューティングファブリック用に8つのConnectX-7アダプタを提供します。デュアルポートConnectX-7アダプタの各ペアは、ストレージファブリックと管理ファブリックへのパラレルパスを提供します。アウトオブバンドポートはBMCアクセスに使用されます。

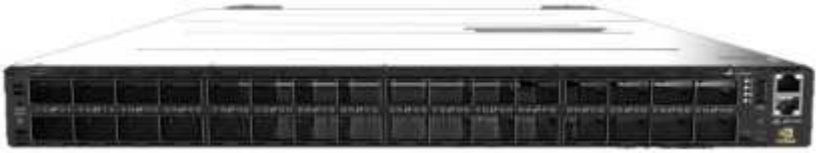
*NVIDIA DGX H100*の背面パネル



NVIDIAネットワーク

NVIDIA Quantum-2 QM9700スイッチ

*NVIDIA Quantum-2 QM9700 InfiniBand*スイッチ



400Gb/秒InfiniBand接続を備えたNVIDIA Quantum-2 QM9700スイッチは、NVIDIA Quantum-2 InfiniBand BasePOD構成のコンピューティングファブリックを強化します。ConnectX-7シングルポートアダプタは、InfiniBandコンピューティングファブリックに使用されます。各NVIDIA DGXシステムは、各QM9700スイッチにデュアル接続されており、システム間に広帯域で低レイテンシの複数のパスを提供します。

NVIDIA Spectrum-3 SN4600スイッチ

*NVIDIA Spectrum-3 SN4600*スイッチ

NVIDIA Spectrum-3 SN4600スイッチは、合計128個のポート（スイッチあたり64個）を提供し、DGX BasePODのインバンド管理用に冗長接続を提供します。NVIDIA SN4600スイッチは、1GbEから200GbEまでの速度を提供できます。イーサネット経由で接続されたストレージアプライアンスには、NVIDIA SN4600スイッチも使用されます。NVIDIA DGXデュアルポートConnectX-7アダプタのポートは、インバンド管理とストレージ接続の両方に使用されます。

NVIDIA Spectrum SN2201スイッチ

*NVIDIA Spectrum SN2201*スイッチ

NVIDIA Spectrum SN2201スイッチは、アウトオブバンド管理用の接続を提供する48ポートを備えています。アウトオブバンド管理は、DGX BasePODのすべてのコンポーネントの統合管理接続を提供します。

NVIDIA ConnectX-7アダプタ

NVIDIA ConnectX-7アダプタ

NVIDIA ConnectX-7アダプタは、25/50/100/200/400Gのスループットを提供できます。NVIDIA DGXシステムは、シングルポートとデュアルポートのConnectX-7アダプタの両方を使用して、400Gb/秒InfiniBandおよび100/200GbイーサネットのDGX BasePOD環境に柔軟性を提供します。

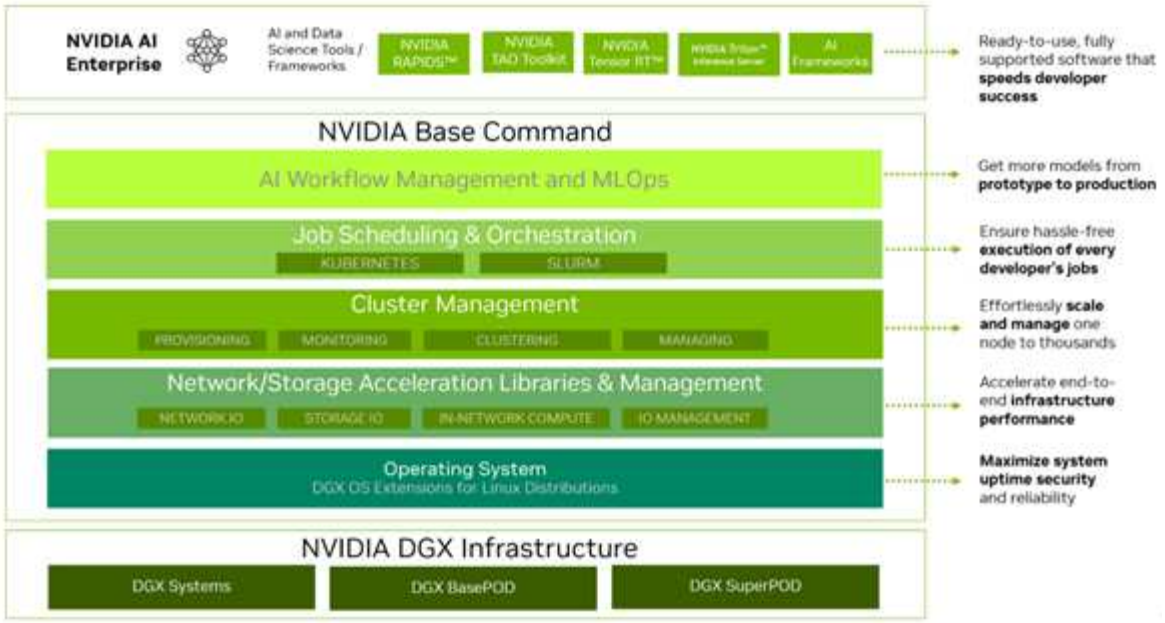
NVIDIA DGXシステムを搭載したNetApp AIPod -ソフトウェアコンポーネント

NVIDIAソフトウェア

NVIDIA Baseコマンド

NVIDIA Base Command ™はすべてのDGX BasePODを強化し、NVIDIAソフトウェアイノベーションのメリットを最大限に活用できるようにします。エンタープライズクラスのオーケストレーションとクラスタ管理、コンピューティング、ストレージ、ネットワークのインフラを高速化するライブラリ、AIワークロード向けに最適化されたオペレーティングシステム（OS）など、実績のあるプラットフォームにより、投資のポテンシャルを最大限に引き出すことができます。

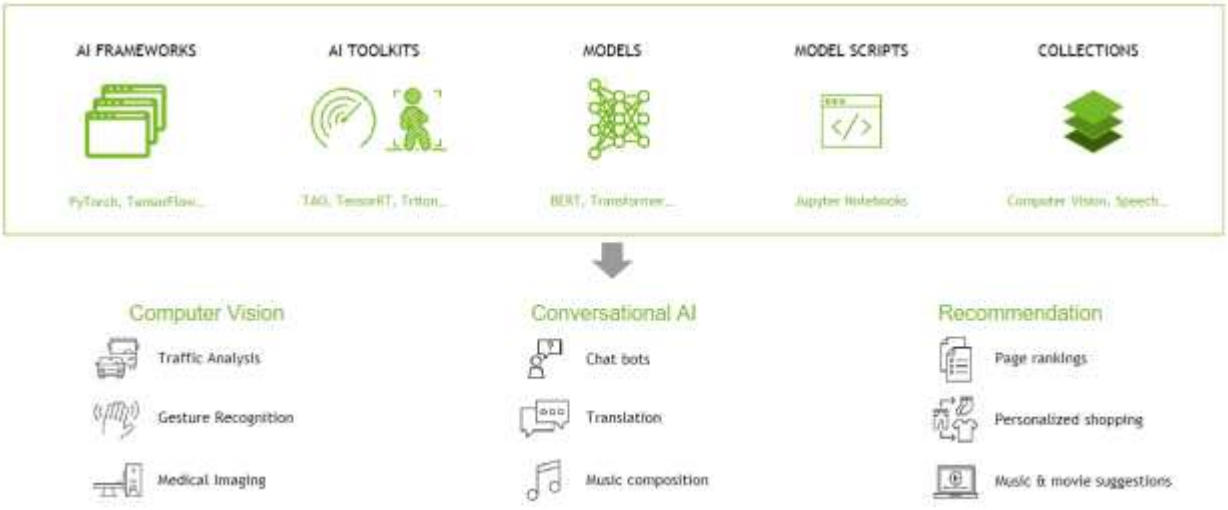
*NVIDIA*ベースコマンド解決策



NVIDIA GPU Cloud （ NGC ）

NVIDIA NGC™は、AIに関するさまざまな専門知識を持つデータサイエンティスト、開発者、研究者のニーズを満たすソフトウェアを提供します。NGCでホストされるソフトウェアは、一般的な脆弱性とエクスポージャー（CVE）、暗号鍵、秘密鍵の集合をスキャンします。複数のGPU、多くの場合はマルチノードに拡張できるようにテストおよび設計されているため、DGXシステムへの投資を最大限に活用できます。

_ NVIDIA GPUクラウド_



NVIDIA AIエンタープライズ

NVIDIA AI Enterpriseは、すべての企業にジェネレーティブAIを提供するエンドツーエンドのソフトウェアプラットフォームです。NVIDIA DGXプラットフォーム上で実行するように最適化されたジェネレーティブAI基盤モデルに対して、最速かつ最も効率的なランタイムを提供します。本番環境レベルのセキュリティ、安定性、管理性を備えているため、生成型AIソリューションの開発が合理化されます。NVIDIA AI Enterprise はDGX BasePODに含まれており、エンタープライズ開発者は事前トレーニング済みモデル、最適化されたフレームワーク、マイクロサービス、高速化されたライブラリ、エンタープライズサポートにアクセスできます。

NetAppソフトウェア

NetApp ONTAP

ネットアップが提供する最新世代のストレージ管理ソフトウェアONTAP 9を使用すれば、インフラを最新化し、クラウド対応のデータセンターに移行できます。ONTAP は、業界をリードするデータ管理機能を活用して、データの格納場所に関係なく、単一のツールセットでデータの管理と保護を実現します。エッジ、コア、クラウドなど、必要な場所に自由にデータを移動することもできます。ONTAP 9には、データ管理の簡易化、重要なデータの高速化と保護、ハイブリッドクラウドアーキテクチャ全体で次世代インフラ機能を実現する多数の機能が搭載されています。

データの高速化と保護

ONTAP は、卓越したパフォーマンスとデータ保護を実現し、以下の方法でこれらの機能を拡張します。

- パフォーマンスとレイテンシの低下：ONTAPは、NFS over RDMA、Parallel NFS（pNFS）、NFSセッショントランキングを使用したNVIDIA GPUDirect Storage（GDS）のサポートなど、可能な限り低いレイテンシで最高のスループットを提供します。
- データ保護ONTAPは、組み込みのデータ保護機能と、すべてのプラットフォームを共通で管理できる業界最高レベルのランサムウェア対策保証を提供します。
- NetApp Volume Encryption（NVE）：ONTAP は、オンボードと外部キー管理の両方をサポートし、ボリュームレベルでのネイティブな暗号化を実現します。
- ストレージのマルチテナンシーと多要素認証：ONTAP を使用すると、最高レベルのセキュリティでインフラリソースを共有できます。

データ管理を簡易化

データ管理は、AIアプリケーションの運用やAI / MLデータセットのトレーニングに適切なリソースを使用できるように、エンタープライズIT運用とデータサイエンティストにとって非常に重要です。以下に記載するネットアップテクノロジに関する追加情報 は、この検証の対象外ですが、導入環境によっては関連性がある場合もあります。

ONTAP データ管理ソフトウェアには、運用を合理化および簡易化し、総運用コストを削減するための次の機能が含まれています。

- スナップショットとクローンは、ML / DLワークフローのコラボレーション、並行した実験、強化されたデータガバナンスを可能にします。
- SnapMirrorは、ハイブリッドクラウド環境やマルチサイト環境でのシームレスなデータ移動を可能にし、必要なときに必要な場所でデータを提供します。
- インラインデータコンパクション、強化された重複排除：データコンパクションはストレージブロック内の無駄なスペースを削減し、重複排除は実効容量を大幅に増やします。この環境データはローカルに格納され、データはクラウドに階層化されます。
- 最小、最大、アダプティブのQuality of Service（AQoS）。きめ細かいサービス品質（QoS）管理機能により、高度に共有された環境で重要なアプリケーションのパフォーマンスレベルを維持できます。
- NetApp FlexGroupを使用すると、ストレージクラスタ内のすべてのノードにデータを分散できるため、非常に大規模なデータセットに対して大容量とパフォーマンスが向上します。
- NetApp FabricPool の略。Amazon Web Services（AWS）、Azure、NetApp StorageGRID ストレージ解決策 など、パブリッククラウドとプライベートクラウドのストレージオプションへコールドデータを自動的に階層化します。FabricPool の詳細については、を参照してください "TR-4598：『FabricPool best

bests』"。

- NetApp FlexCacheの略。リモートボリュームキャッシング機能を提供し、ファイル配信を簡易化し、WANレイテンシを低減し、WAN帯域幅コストを削減します。FlexCacheを使用すると、複数のサイトに分散した製品開発が可能になるだけでなく、リモートサイトから企業のデータセットにすばやくアクセスできるようになります。

将来のニーズにも対応できるインフラ

ONTAP は、次の機能を備えており、要件が厳しく、絶えず変化するビジネスニーズに対応できます。

- シームレスな拡張とノンストップオペレーションONTAPでは、既存のコントローラとスケールアウトクラスタにオンラインで容量を追加できます。NVMe や 32Gb FC などの最新テクノロジへのアップグレードも、コストのかかるデータ移行やシステム停止を行わずに実行できます。
- クラウドへの接続：ONTAP は、すべてのパブリッククラウドでSoftware-Defined Storage（ONTAP Select ）とクラウドネイティブインスタンス（NetApp Cloud Volumes Service ）のオプションを選択できる、マルチクラウドに対応した最もクラウド対応のストレージ管理ソフトウェアです。
- 新しいアプリケーションとの統合：ONTAP は、既存のエンタープライズアプリケーションをサポートするインフラを使用して、自律走行車、スマートシティ、インダストリー4.0などの次世代プラットフォームやアプリケーション向けにエンタープライズクラスのデータサービスを提供します。

NetApp DataOps ツールキット

NetApp DataOpsツールキットは、高性能なスケールアウトネットアップストレージを基盤とする開発/トレーニング用ワークスペースと推論サーバの管理を簡易化するPythonベースのツールです。DataOps Toolkitはスタンドアロンのユーティリティとして動作でき、NetApp Astra Tridentを活用してストレージの運用を自動化するKubernetes環境でさらに効果的です。主な機能は次のとおりです。

- ハイパフォーマンスでスケールアウト可能なネットアップストレージを基盤とする、大容量のJupyterLabワークスペースを迅速にプロビジョニングできます。
- エンタープライズクラスのネットアップストレージを基盤とする新しいNVIDIA Triton Inference Serverインスタンスを迅速にプロビジョニング
- 実験や迅速なイテレーションを可能にするために、大容量のJupyterLabワークスペースのクローンをほぼ瞬時に作成できます。
- バックアップ/トレーサビリティ/ベースライン化のための大容量JupyterLabワークスペースのほぼ瞬時のスナップショット。
- 大容量でハイパフォーマンスなデータボリュームのプロビジョニング、クローニング、スナップショットをほぼ瞬時に実行できます。

ネットアップアストラト **Trident**

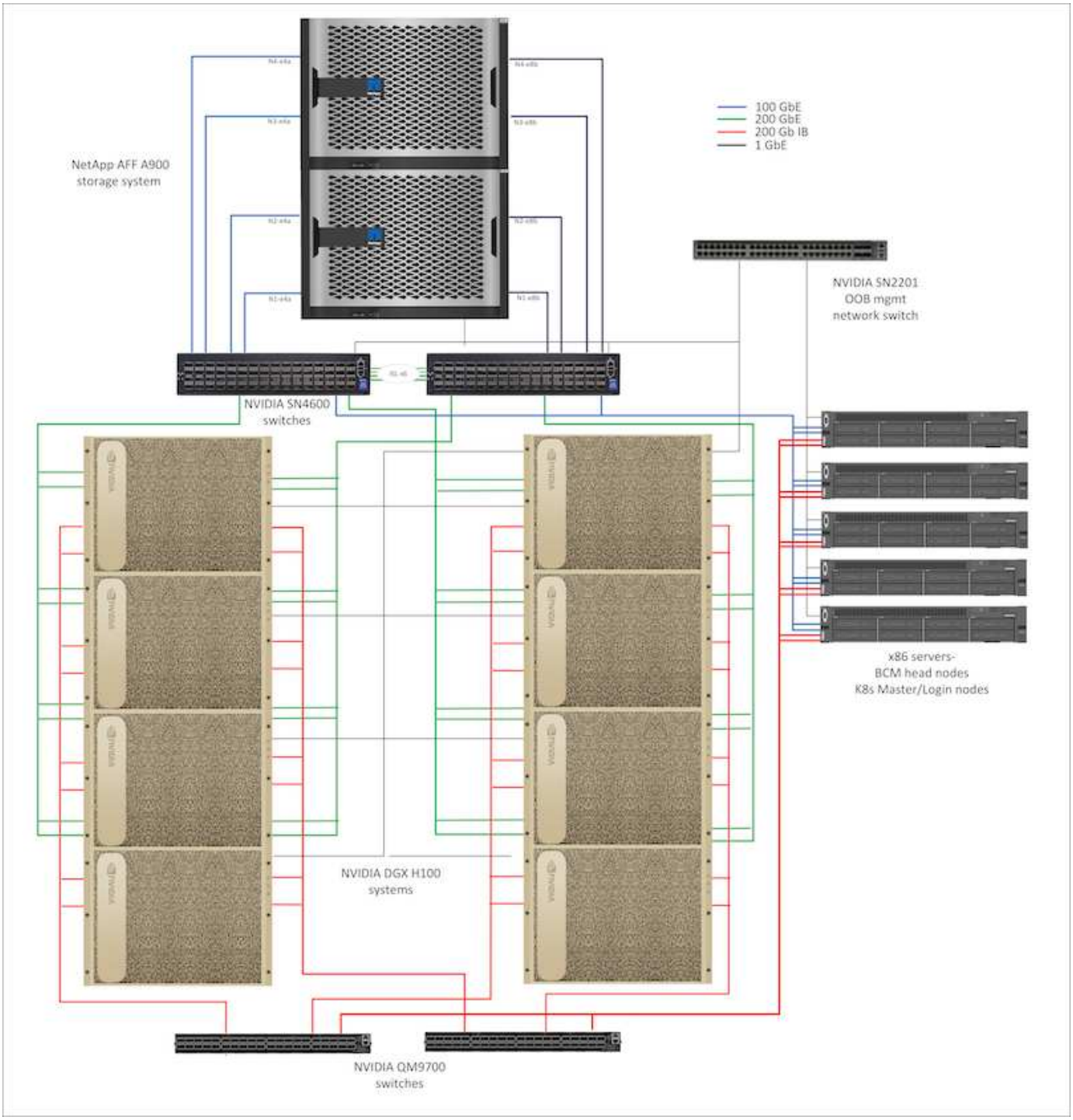
Astra Tridentは、Anthosを含むコンテナとKubernetesディストリビューション向けの、完全サポートされたオープンソースストレージオーケストレーションツールです。Tridentは、NetApp ONTAPを含むNetAppストレージポートフォリオ全体と連携し、NFS、NVMe/TCP、iSCSI接続にも対応しています。Trident を使用すると、ストレージ管理者の手を煩わせることなく、エンドユーザがネットアップストレージシステムからストレージをプロビジョニングして管理できるため、 DevOps ワークフローが高速化されます。

NetApp AIPodとNVIDIA DGXシステム-解決策アーキテクチャ

DGX H100システムを搭載したNetApp AIポッド

このリファレンスアーキテクチャでは、コンピューティングノード間の400GB/秒InfiniBand（IB）接続で、コンピューティングクラスタインターコネクトとストレージアクセスに別 々 のファブリックを利用します。次の図は、DGX H100システムを搭載したNetApp AIPodの全体的な解決策トポロジを示しています。

NetApp AIPOD解決策トポロジ_



ネットワーク構成：

この構成では、コンピューティングクラスタファブリックでQM9700 400Gb/秒IBスイッチのペアを使用します。これらのスイッチは相互に接続されて高可用性が確保されます。各DGX H100システムは、8つの接続を使用してスイッチに接続されます。一方のスイッチには偶数番のポートが接続され、もう一方のスイッチには奇数番のポートが接続されます。

ストレージシステムへのアクセス、インバンド管理、およびクライアントアクセスには、SN4600イーサネットスイッチのペアを使用します。スイッチはスイッチ間リンクで接続され、さまざまなトラフィックタイプを分離するために複数のVLANで設定されます。大規模な展開では、スパインスイッチ用にスイッチペアを追加し、必要に応じてリーフを追加することで、イーサネットネットワークをリーフスパイン構成に拡張できます。

コンピューティングインターコネクトと高速イーサネットネットワークに加えて、すべての物理デバイスを1つ以上のSN2201イーサネットスイッチに接続し、アウトオブバンド管理を行います。 DGX H100システムの接続の詳細については、 "NVIDIA BasePODドキュメント"。

ストレージアクセス用のクライアント設定

各DGX H100システムには、管理トラフィックとストレージトラフィック用に2つのデュアルポートConnectX-7アダプタがプロビジョニングされます。この解決策では、各カードの両方のポートが同じスイッチに接続されます。各カードの1つのポートがLACP MLAGボンドに構成され、各スイッチに1つのポートが接続されます。このボンドでは、インバンド管理、クライアントアクセス、およびユーザレベルのストレージアクセス用のVLANがホストされます。

各カードのもう一方のポートはAFF A900ストレージシステムへの接続に使用され、ワークロードの要件に応じて複数の構成で使用できます。NVIDIA Magnum IO GPUDirect Storageをサポートするために RDMA経由のNFSを使用する構成では、他のタイプのボンドではRDMAがサポートされないため、ポートはアクティブ/パッシブボンドとして構成されます。RDMAを必要としない環境では、ストレージインターフェイスをLACPボンディングで設定して、高可用性と追加の帯域幅を実現することもできます。クライアントは、RDMAを使用するかどうかに関係なく、NFS v4.1 pNFSおよびセッショントランキングを使用してストレージシステムをマウントし、クラスタ内のすべてのストレージノードに並列アクセスできるようにします。

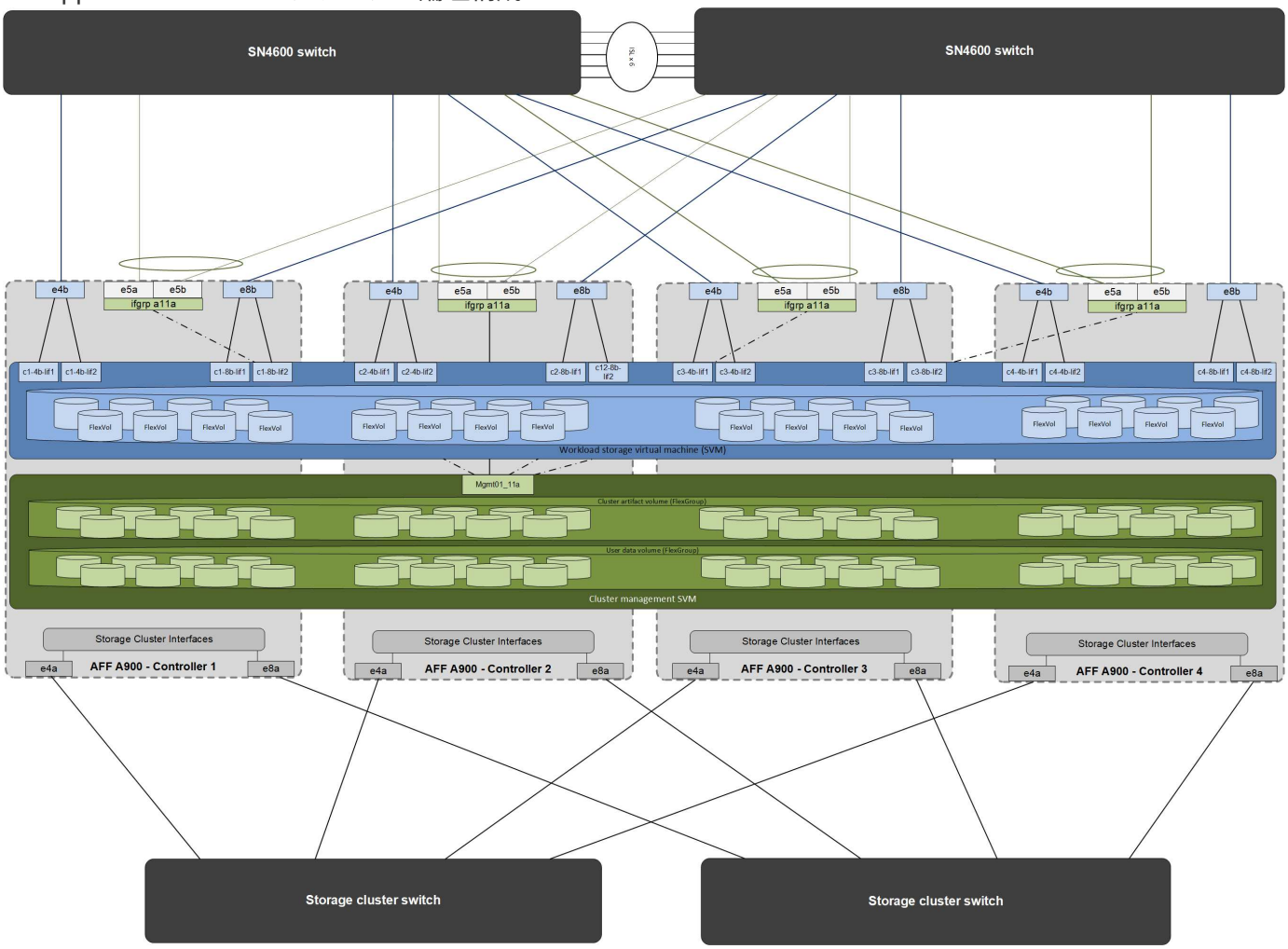
ストレージシステムの構成：

各AFF A900ストレージシステムは、各コントローラの4つの100GbEポートを使用して接続されます。各コントローラの2つのポートがDGXシステムからのワークロードデータアクセスに使用され、各コントローラの2つのポートがLACPインターフェイスグループとして構成され、クラスタ管理アーティファクトとユーザホームディレクトリ用の管理プレーンサーバからのアクセスをサポートします。ストレージシステムからのすべてのデータアクセスはNFS経由で提供されます。AIワークロードアクセス専用のStorage Virtual Machine（SVM）と、クラスタ管理専用の別のSVMがあります。

ワークロードSVMには合計8つの論理インターフェイス（LIF）が設定され、各物理ポートに2つのLIFがあります。この構成では、最大帯域幅と、各LIFが同じコントローラ上の別のポートにフェイルオーバーする手段が提供されるため、ネットワーク障害が発生した場合でも両方のコントローラがアクティブなままになります。この構成では、RDMA経由のNFSもサポートされ、GPUDirectストレージへのアクセスが可能になります。ストレージ容量は、クラスタ内のすべてのストレージコントローラにまたがる1つの大規模なFlexGroupボリューム（各コントローラに16個のコンスティチュエントボリューム）としてプロビジョニングされます。このFlexGroupにはSVM上のどのLIFからもアクセスできます。pNFSとセッショントランキングを使用したNFSv4.1を使用すると、クライアントはSVM内のすべてのLIFへの接続を確立します。これにより、各ストレージノードのローカルデータに並行してアクセスできるようになり、パフォーマンスが大幅に向上します。ワークロードSVMと各データLIFもRDMAプロトコルアクセス用に設定されます。ONTAPのRDMA設定の詳細については、 "ONTAP のドキュメント"。

管理SVMに必要なLIFは1つだけです。このLIFは、各コントローラで設定された2ポートインターフェイスグループでホストされます。他のFlexGroupは、クラスタノードのイメージ、システム監視履歴データ、エンドユーザのホームディレクトリなど、クラスタ管理アーティファクトを格納するために管理SVM上にプロビジョニングされます。次の図は、ストレージシステムの論理構成を示しています。

NetApp A900ストレージクラスタの論理構成



管理プレーンサーバ

このリファレンスアーキテクチャには、管理プレーン用に5台のCPUベースのサーバも含まれています。このうちの2つのシステムは、クラスタの導入と管理のためのNVIDIA Base Command Managerのヘッドノードとして使用されます。他の3つのシステムは、ジョブのスケジューリングにSlurmを利用する導入環境向けに、Kubernetesマスターノードやログインノードなどの追加のクラスタサービスを提供するために使用されます。Kubernetesを活用した導入では、NetApp Astra Trident CSIドライバを活用して、AFF A900ストレージシステム上の管理ワークロードとAIワークロードの両方に永続的ストレージを使用した自動プロビジョニングとデータサービスを提供できます。

各サーバは、クラスタの導入と管理を可能にするためにIBスイッチとイーサネットスイッチの両方に物理的に接続されます。また、前述したクラスタ管理アーティファクトの保存用に、管理SVMを介したストレージシステムへのNFSマウントが設定されます。

NetApp AIPodとNVIDIA DGXシステム-解決策の検証とサイジングに関するガイダンス

解決策の検証

この解決策のストレージ構成は、オープンソースツールfioを使用した一連の統合ワークロードを使用して検証されました。テストには、ディープラーニングトレーニングジョブを実行するDGXシステムで生成されるストレージワークロードをシミュレートするI/Oパターンの読み取りと書き込みが含まれます。ストレージ構成は、FIOワークロードを同時に実行する2ソケットCPUサーバのクラスタを使用して検証され、DGXシステムのクラスタをシミュレートしました。各クライアントは、前述したのと同じネットワーク構成で設定され、次の詳細が追加されました。

この検証で使用したマウントオプションは次のとおりです-
•バージョン= 4.1 # pNFSで複数のストレージノードへの並列アクセスを実現
•proto = RDMA #転送プロトコルをデフォルトのTCPではなくRDMAに設定する
•ポート=20049 # RDMA NFSサービスの正しいポートを指定してください
•max_connect = 16 #ストレージポートの帯域幅を集約するためのNFSセッショントランキングを有効にする
•write = eager #バッファ書き込みの書き込みパフォーマンスを向上
•rsize=262144、wsize=262144#はI/O転送サイズを256Kに設定

さらに、クライアントにはNFS max_session_slots値1024を設定しました。解決策では、RDMA経由のNFSを使用してテストしたため、ストレージネットワークポートにはアクティブ/パッシブボンドを設定しました。この検証で使用したボンドパラメータは次のとおりです-
•モード=アクティブバックアップ #ボンディングをアクティブ/パッシブモードに設定します。
•プライマリ=<interface name> #すべてのクライアントのプライマリインターフェイスがスイッチ全体に分散されている
•MII-MONITOR-INTERVAL = 100 #モニタリング間隔を100msに指定します。
•fail-over-mac-policy=active #は、アクティブリンクのMACアドレスがボンドのMACであることを示します。これは、ボンディングされたインターフェイス上でRDMAが適切に動作するために必要です。

2つのA900 HAペア（4台のコントローラ）と1.9TB NVMeディスクドライブを24本搭載したNS224ディスクシェルフをそれぞれのHAペアに接続して、説明のとおりにストレージシステムを構成しました。アーキテクチャのセクションで説明したように、すべてのコントローラのストレージ容量をFlexGroupボリュームを使用して結合し、すべてのクライアントのデータをクラスタ内のすべてのコントローラに分散しました。

ストレージシステムのサイジングに関するガイダンス

NetAppはDGX BasePOD認定を取得しており、テスト済みの2つのA900 HAペアは、8台のDGX H100システムのクラスタを簡単にサポートできます。ストレージパフォーマンス要件の高い大規模な環境では、AFFシステムをNetApp ONTAPクラスタに追加し、1つのクラスタに最大12のHAペア（24ノード）を追加できます。この解決策で説明するFlexGroupテクノロジを使用すると、24ノードクラスタで40PBを超える容量と最大300Gbpsのスループットを単一のネームスペースで実現できます。AFF A400、A250、C800などの他のNetAppストレージシステムは、低パフォーマンスまたは大容量のオプションを提供し、小規模な導入に低コストで対応します。ONTAP 9は混在モデルのクラスタをサポートしているため、最初は小規模な設置面積から始めて、容量やパフォーマンスの要件が増大したときに、クラスタにストレージシステムを追加したり、大容量のストレージシステムを追加したりすることができます。次の表に、各AFFモデルでサポートされるA100およびH100 GPUの概算数を示します。

_ NetAppストレージシステムのサイジングガイダンス_

		Throughput[2]	Raw capacity *(typical / max)*	Connectivity	# NVIDIA A100 GPUs supported[3]	# NVIDIA H100 GPUs supported[4]
NetApp® AFF A900	**1 HA pair[1]**	28GB/s	182TB / 14.7PB	**100 GbE**	1 - 64	1-32
	12 HA pairs	336GB/s	2.1PB / **176.4PB**		**768**	**384**
AFF A800	**1 HA pair**	25GB/s	368TB / 3.6PB	**100 GbE**	1 - 64	1-32
	12 HA pairs	300GB/s	4.4PB / **43.2PB**		**768**	**384**
AFF C800	**1 HA pair**	21GB/s	368TB / 3.6PB	**100 GbE**	1-48	1-24
	12 HA pairs	252GB/s	4.4PB / **43.2PB**		**576**	**288**
AFF A400	**1 HA pair**	11GB/s	182TB / 14.7PB	**40/100 GbE**	1 - 32	1-16
	12 HA pairs	132GB/s	2.1PB / **176.4PB**		**384**	**192**
AFF C400	**1 HA pair**	8GB/s	182TB / 14.7PB	**40/100 GbE**	1 - 16	1-8
	12 HA pairs	128GB/s	2.1PB / **176.4PB**		**192**	**96**
AFF A250	**1 HA pair**	7.4GB/s	91.2TB / 4.4PB	**25 GbE 40/100GbE**	1 - 16	1-8
	4 HA pairs	29.6GB/s	364.8TB / 17.6PB		64	32
AFF C250	**1 HA pair**	5 GB/s	91.2TB / 4.4PB	**25 GbE 40/100GbE**	1-8	1-4
	4 HA pairs	20 GB/s	364.8TB / 17.6PB		32	8

1 – 1 AFF = 1 HA pair = 2 Nodes. 12 HA pairs = 24 nodes
2 – 100% sequential read
3 – Based on workload testing in NVA-1153
4 – Based on BasePOD validation test results

NVIDIA DGXシステムを搭載したNetApp AIPod -まとめと追加情報

まとめ

DGX BasePODアーキテクチャは、同等の高度なストレージ機能とデータ管理機能を必要とする次世代のディープラーニングプラットフォームです。DGX BasePODとNetApp AFFシステムを組み合わせることで、NetApp AIPodとDGXシステムのアーキテクチャは、24ノードのAFF A900クラスタ上に最大48台のDGX H100システムまでのほぼすべてのスケールで実装できます。AFFは、NetApp ONTAPの優れたクラウド統合機能とソフトウェアで定義される機能を組み合わせることで、DLプロジェクトを成功させるために、エッジ、コア、クラウドにわたる幅広いデータパイプラインを実現します。

追加情報

このドキュメントに記載されている情報の詳細については、次のドキュメントやWebサイトを参照してください。

- NetApp ONTAP データ管理ソフトウェア—ONTAP 情報ライブラリ

  "https://docs.netapp.com/us-en/ontap-family/"

- NetApp AFF A900ストレージシステム-

  "https://www.netapp.com/data-storage/aff-a-series/aff-a900/"

- NetApp ONTAP RDMA情報-

"https://docs.netapp.com/us-en/ontap/nfs-rdma/index.html"

- NetApp DataOps ツールキット

  "https://github.com/NetApp/netapp-dataops-toolkit"

- ネットアップアストラト Trident

  "https://docs.netapp.com/us-en/netapp-solutions/containers/rh-os-n_overview_trident.html"

- NetApp GPUDirectストレージに関するブログ-

  "https://www.netapp.com/blog/ontap-reaches-171-gpudirect-storage/"

- NVIDIA DGX BasePOD

  "https://www.nvidia.com/en-us/data-center/dgx-basepod/"

- NVIDIA DGX H100システム

  "https://www.nvidia.com/en-us/data-center/dgx-h100/"

- NVIDIAネットワーク

  "https://www.nvidia.com/en-us/networking/"

- NVIDIA Magnum IO GPUDirectストレージ

  "https://docs.nvidia.com/gpudirect-storage"

- NVIDIA Baseコマンド

  "https://www.nvidia.com/en-us/data-center/base-command/"

- NVIDIA Baseコマンドマネージャ

  "https://www.nvidia.com/en-us/data-center/base-command/manager"

- NVIDIA AIエンタープライズ

  "https://www.nvidia.com/en-us/data-center/products/ai-enterprise/"

謝辞

このドキュメントは、NetAppソリューションおよびONTAPエンジニアリングチーム（David Arnette、Olga Kornievskaia、Dustin Fischer、Srikanth Kaligotla、Mohit Kumar、Rajeev Badrinath）の作業です。また、NVIDIAとNVIDIA DGX BasePODエンジニアリングチームの継続的なサポートに感謝します。

著作権に関する情報

商標に関する情報