NETAPP NOVUS：野心だけが限界を動かす世界最速のストレージで制約を超えて拡張を実現



エグゼクティブ サマリー

NetApp® Novusは、次世代AIインフラストラクチャのための革新的なスケールアウト ストレージ アーキテクチャであり、ネオクラウド、ソブリンAIプロバイダー、ハイパースケーラー、およびギガワット規模のAIファクトリーを構築する大企業に求められる速度、同時実行性、ネームスペースの規模に合わせて構築されています。GPUの性能は、モデルのサイズやネットワーク帯域幅だけでなく、データインフラストラクチャがどれだけ高速にデータを供給できるかによっても制限されます。NetApp Novusは、分離型メタデータアーキテクチャ、直接並列データアクセス、および共通ネームスペースによってこのボトルネックを解消し、GPU環境が拡張しても生産性を維持できるよう支援します。

アイドル状態のGPUは、単に利用されていないコンピューティングリソースというだけではありません。AIファクトリー規模で、AI出力ではなく、資本、電力、床面積が熱を生み出しています。NetApp Novusは、多数のONTAPベースのデータリソースを単一のデータ環境として動作させることでストレージのあり方を変革し、メタデータサービスは同時実行性と名前空間の拡張をサポートするために独立して拡張できます。ゼタバイト規模のファイルシステムとAIインフラストラクチャの効率性を目的に構築されたNetApp Novusは、オペレーターがGPUへのデータ供給を維持し、データ関連の処理停止を減らし、断片化することなく拡張し、大規模AIの経済性を向上させるのに役立ちます。

問題点

何千ものGPUがラックに設置され、稼働準備が整っている様子を想像してみてください。それらは限りある資本、電力、そして床面積を消費しています。これらは現代のデータセンターにおいて最も高価な資産の一つですが、価値を生み出すのは有用な作業：モデルのトレーニング、チューニング、推論処理、あるいはエージェント型AIワークフローの推進を行っている場合のみです。ギガワット規模のAIインフラにおいて、制限要因となるのは必ずしもモデル、GPU、ネットワークとは限りません。ますます重要になっているのは、データインフラがすべてのGPUに十分なデータ供給を維持できるかどうかです。GPUがデータを待つ1秒ごとに、生産性の低下と投資の無駄につながります。多くの大規模AI環境では、AI出力に対する需要が伸び続けているにもかかわらず、データボトルネックによって高価なGPUインフラストラクチャが著しく活用されないままになる可能性があります。問題は、チームが単に「より高速なストレージ」を必要としているというだけではありません。より根本的な問題はアーキテクチャにあります：従来のストレージシステムは、数万もの高度に並列化されたGPUクライアントに対して、メタデータの同時実行性、帯域幅、容量、およびネームスペース操作を個別に拡張するように設計されていませんでした。

あらゆるストレージシステムには、コントローラとインターコネクトによって決まる上限が存在します。独立したシステムを増やしてその上限を超えようとするとオペレーターは一つのボトルネックを別のボトルネックに置き換えることになります：ネームスペースが増え、配置決定が増え、再マウントが増え、運用上のサイロが増え、容量やパフォーマンスが活用できなくなる可能性が増えるのです。本来は単一のデータ環境として機能するはずのものが、アプリケーションチームとインフラストラクチャチームが自らつなぎ合わせなければならない、ストレージの孤立した集合体になってしまいます。

その断片化は、パフォーマンスの問題であると同時に、オペレーターの効率性の問題にもなります。AIチームは、データの保存場所、マウントするクラスター、ワークロードの再配分方法、断片化されたインフラストラクチャ全体でのチェックポイントの調整方法などを決定するために、ストレージエンジニアになる必要はありません。AIファクトリー規模になると、運用上の複雑さ自体がスケーリングの限界となります。

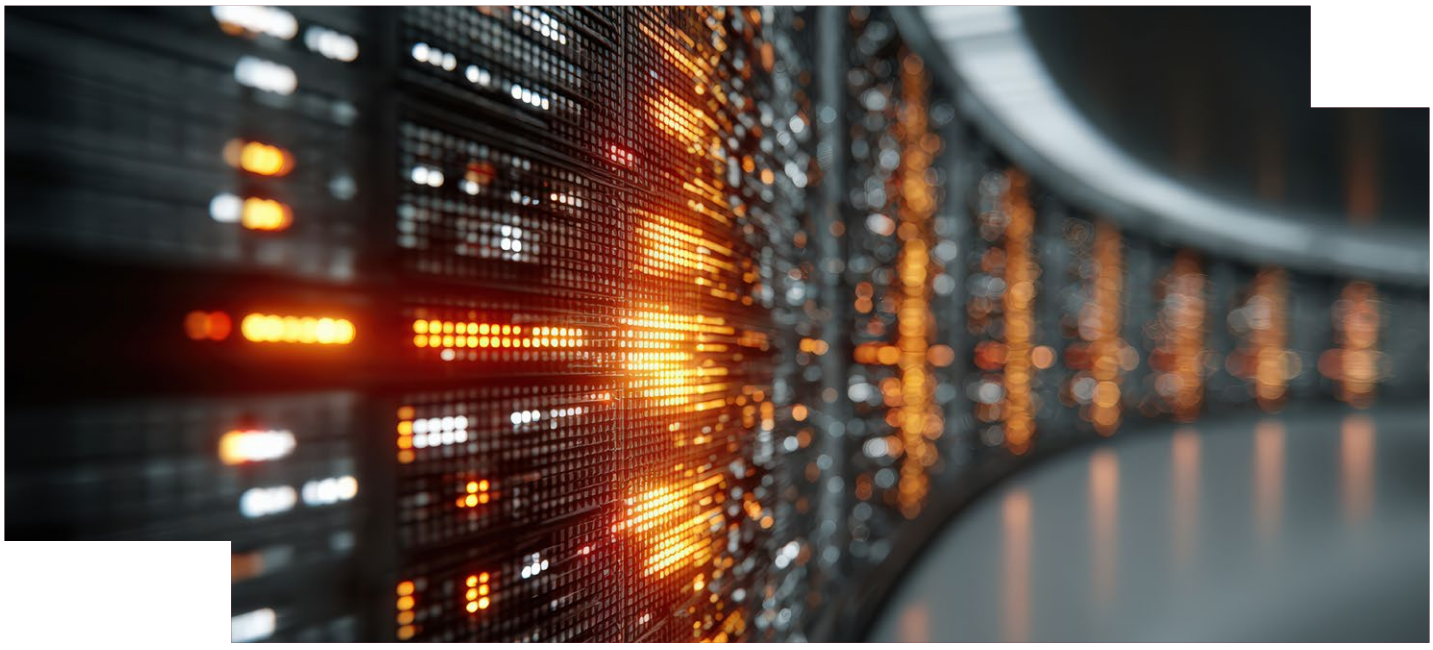
チームがたとえ1つの壁を乗り越えたとしても、別の壁に直面します：電力です。データインフラストラクチャを追従させるために費やされる1ワットは、GPUが生産的な作業を行うために利用できない1ワットとなります。アーキテクチャ上の制限を補うためにストレージを過剰に構築すると、AIオペレーターが限られた電力容量内で出力を最大化しようとしているまさにその時に、電力、冷却、設置面積、およびコストが増加する可能性があります。

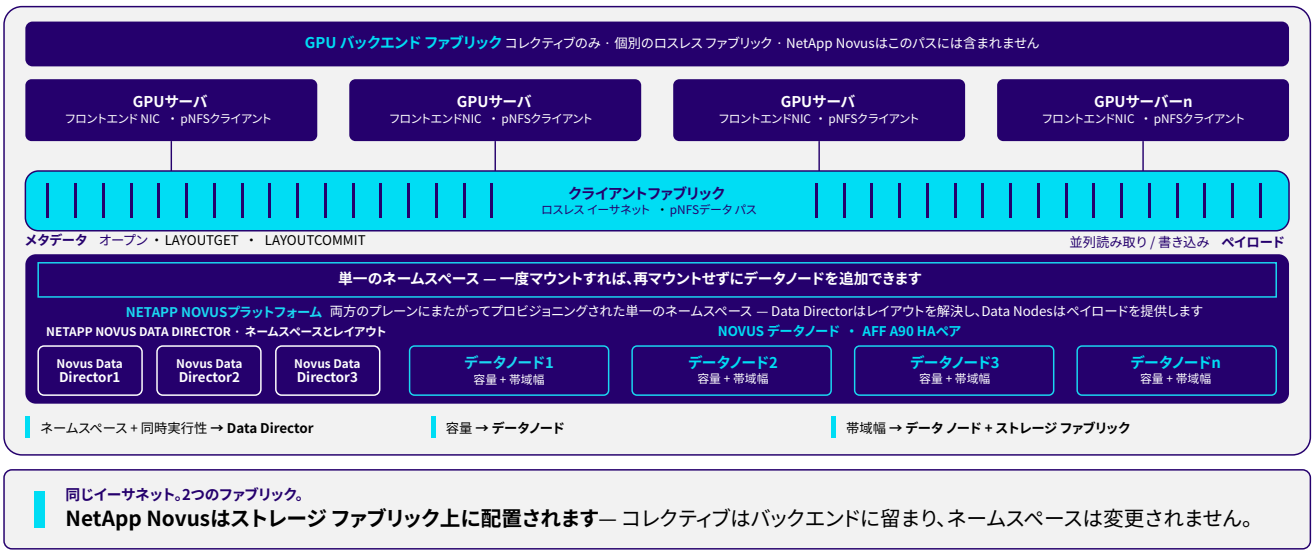
ストレージが処理速度に追いつかない場合、GPUはアイドル状態になり、電力が無駄になり、インテリジェンスを生み出すために構築されたラックは、貴重な資本、電力、床面積をAIの出力ではなく、無駄な熱に変えてしまいます。大規模AIにおいては、利用率は単なる性能指標ではありません。それは経済モデルです。

ソリューション

NetApp Novusは、メタデータ層をストレージ レイヤから分離します。メタデータは、フェデレーションされたストレージシステム全体にわたるデータレイアウトを追跡し、各GPUクライアントにそのレイアウトを渡したあとは、処理から外れます。クライアントは基盤となるストレージに直接読み書きを行い、すべてのI/Oにおいてメタデータを経由する往復通信は行いません。それが、単一のストレージシステムの限界を打ち破る鍵となるのです。ストレージレイヤ自体は実績のあるNetApp ONTAP®上で稼働し、ラインレートでデータを配信することで、そのボトルネックを解消します。このアプローチにより、パフォーマンスを段階的に向上させ、線形的にパフォーマンスを追加することが可能になります。ストレージシステムを追加することで、フェデレーションプールのパフォーマンス容量を拡張できます。メタデータも独自のパスでスケーリングされるため、ストレージ レイヤとは独立してスケールを調整できます。これは、再マウントも断片化も発生しないことを意味します。

これらはいずれも新しいクライアントを必要としません。NetApp Novusは、ほとんどのフリートですでに使用されているカーネル内NFSv4.2およびpNFS Flex Filesスタックを介してGPUノードと通信し、マルチ接続スループットを実現するnconnectと、CPUを完全にスキップしてGPUからストレージへの直接I/Oを実現するGPUDirect Storageを備えています。数千台のGPUサーバー群全体に構築、検証、展開するための独自の技術は何もありません。NetApp NovusはNFSのようにマウントされ、すべてのクライアントは既にNetApp Novusとの通信方法を知っています。





NetApp Novusアーキテクチャ：NetApp Novusはメタデータの調整とデータの移動を分離します。GPUクライアントは、NetApp Novus Data Directorメタデータ サーバーからデータ レイアウトを取得し、ONTAPベースのストレージ レイヤに対して直接I/Oを実行します。これにより、メタデータの同時実行性とストレージ スループットを独立して拡張できるようになります。

NetApp Novusの提供方法

NetApp Novusは、分離型メタデータ層とONTAPベースのデータノードを組み合わせた統合ストレージ解決策として提供されます。初期展開モデルでは、NetApp A90システムで稼働するONTAPは永続的なデータ配信基盤を提供し、pNFSメタデータサーバーサービスは認定されたサーバーインフラストラクチャ上で稼働します。メタデータ層は、ネームスペースとレイアウトをクライアントに提示します。ONTAPデータノードはデータ パスを提供します。

このアプローチにより、NetApp Novusのアーキテクチャの柔軟性を維持しながら、AIおよびGPU-as-a-Service環境向けの厳選されたハイパフォーマンスなデプロイメントモデルをお客様に提供します。現在の提供モデルは、ONTAPベースのA90データノードと、認定済みのpNFSメタデータサーバーインフラストラクチャに重点を置いています。将来の展開モデルでは、時間の経過とともにアーキテクチャが拡張されることが期待されており、これには、メタデータサービスとONTAPベースのデータサービスの両方を認定されたサードパーティ製ハードウェア上で実行できる、ソフトウェア定義型オプションが含まれます。サポート対象構成、認定の詳細、および導入要件については、最新のNetApp Novus製品ドキュメントを参照してください。

まとめ

AIファクトリーの経済性は、稼働率によって成否が決まります。競合他社も同じGPUとAIモデルを持っていますが、彼らにないのは、それらのGPUの生産性を維持するストレージ層です。NetApp Novusはボトルネックを解消し、多数のシステムをフェデレーションし、ゼタバイト規模のファイルシステム インフラを実現することで、最大のインフラ投資がアイドル状態に陥ることなく、最終的にトークンや成果を生み出せるようになります。NetApp Novusを使用すれば、すべてのGPUが生産性を維持します。

スループットとノード数を目標とするGPU規模にマッピングするには、NetApp Novusアーキテクトにご相談ください。

主なメリット

- **すべてのGPUの生産性を維持：**アイドル状態のGPUは、販売できない最も高額なコスト項目です。NetApp Novusはストレージのボトルネックを解消することで、より多くのコンピューティングリソースがフル稼働しその多くを課金対象時間へと変換します。
- **名前空間を断片化することなく、スループットを拡張：**NetApp Novusはメタデータの調整とデータ移動を分離しているため、名前空間の操作がデータ パスをブロックすることはありません。ONTAPベースのデータリソースを追加することで、共通ネームスペース下の総スループットを拡張すると同時に、メタデータ層を独立して拡張して、より多くのクライアント、テナント、およびファイル操作をサポートします。
- **AIインフラの効率性を向上：**NetApp Novusは、電力、床面積、コスト、運用といった現実的な制約の中で、オペレーターが規模を拡大できるよう支援するように設計されています。ONTAPベースのデータ レイヤは効率的なストレージ基盤を提供する一方で、ソフトウェアで定義されるメタデータ レイヤにより、チームは必要な場所にのみメタデータ容量を追加できます。将来のソフトウェアで定義されるデータサービスは、その柔軟性をさらに高め、オペレーターがインフラストラクチャに適合する構成に基づいて構築できるようにします。
- **安全なマルチテナンシーとQoS：**1つのNetApp Novus展開で多数のテナントを実行し、テナントレベルの暗号化、ポリシーベースのリソース割り当て、および人間とAIエージェントの両方に対するきめ細かなセキュリティ制御によって、互いに隔離します。
- **標準ベースのクライアント：**独自開発のエージェントを構築、検証、またはパッチ適用する必要はありません。NetApp Novusは、GPUノード上で既に稼働しているカーネル内NFSクライアント上にマウントされるため、クライアント展開プロジェクトを行うことなく、数千台規模のフリートをすぐに運用可能な状態にすることができます。

本ソリューション概要には、予定されている製品の特長および機能に関する将来予想の記述が含まれています。これらの記述は、当社の現在の計画と見通しを反映したものであり、変更される可能性があります。実際の製品の特長、機能、提供時期は記載されている内容と異なる場合があり、適用される法律で義務付けられている場合を除き、当社は特定の機能を提供する義務を負いません。

本ドキュメントは機械翻訳による参考訳です。英語版との矛盾や不一致があった場合には、英語版の内容が優先されます。

NetAppについて

NetAppはインテリジェントなデータインフラ企業として、ユニファイド データ ストレージ、統合データ サービス、CloudOpsソリューションを組み合わせることで、混沌とした世界を変革し、あらゆるお客様にビジネス チャンスをもたらしています。NetAppはデータ サイロのないインフラを構築し、オブザーバビリティーとAIを活用して業界最高のデータ管理を実現します。世界最大規模のクラウドにネイティブに組み込まれた唯一のエンタープライズクラスのストレージサービスとして、当社のデータストレージはあらゆる環境で柔軟性を提供します。当社のデータサービスは、優れたサイバーレジリエンス、ガバナンス、およびアプリケーションの俊敏性を通じてデータ優位性を生み出し、CloudOpsソリューションは、パフォーマンスと効率の継続的な最適化を提供します。データの種類、ワークロード、環境を問わず、NetAppがデータインフラを変革し、ビジネスの可能性を現実のものにします。
www.netapp.com/ja/

ネットアップ合同会社
Email:ng-sales-inquiry@netapp.com
https://www.netapp.com/ja/forms/sales-contact/

© 2026 NetApp, Inc. All rights reserved. NetApp、NetAppのロゴ、http://www.netapp.com/jp/legal/netapptmlist.aspxに記載されているマークは、NetApp, Inc.の商標です。その他の会社名と製品名は、それを所有する各社の商標である場合があります。SB-4532-0926-jaJP

4~23sep26