技術概要

クラウドネイティブスーパーコンピューティングアーキテクチャの特徴

NVIDIA

要旨

ハイパフォーマンス コンピューティング（HPC）と人工知能（AI）の導入が進むなか、研究、科学的発見、製品開発を支える主要なデータ処理エンジンとして、スーパーコンピューターの商用利用が広がっています。スーパーコンピューターのシステムは複雑なシミュレーションを実行でき、ソフトウェアが自らソフトウェアを開発する AI の新時代への扉を開きます。スーパーコンピューティングの分野で先頭に立てば、科学技術とイノベーションにおいて一歩先を行くことができるため、世界中の政府、研究機関、企業が、より高速かつ強力なスーパーコンピューティング プラットフォームの構築に投資しています。

しかし、スーパーコンピューティング システムの性能を最大限に引き出しながら効率的に利用することは、現代のクラウド コンピューティングのセキュアなマルチテナント アーキテクチャでは長らく実現できていませんでした。これは、トップクラスのパフォーマンスとクラスターの効率性に、最新のゼロトラスト モデルによるセキュリティ分離とマルチテナントを組み合わせたクラウドネイティブ スーパーコンピューティング プラットフォームによって初めて可能になります。

こうしたアーキテクチャの移行を可能にする重要な構成要素が、データ プロセシング ユニット(DPU) です。DPU は完全に一体化されたオンチップ型データセンター プラットフォームであり、スーパーコンピューティングの各ノードに 2 つの新しい機能をもたらします。1 つは、インフラストラクチャ コントロール プレーンを担うプロセッサです。コンピューティング ノードのユーザー アクセス、ストレージ アクセス、ネットワーキング、ライフ サイクル オーケストレーションを安全に行い、メインのプロセッサの処理の一部をオフロードして、ベアメタルのマルチテナントを実現します。もう 1 つは、独立したラインレートのデータ パスであり、ハードウェア アクセラレーションによってベアメタルのパフォーマンスを実現します。

はじめに

これまで、スーパーコンピューターは単一のアプリケーションの実行に合わせた設計となっており、適切に管理された少数のユーザーのみが利用していました。しかし、AI や HPC が主要なコンピューティング環境として広く商用利用されるようになった今、スーパーコンピューターにおいても、広範囲のユーザーに向けたサービスの提供や、多彩なソフトウェア エコシステムの稼働が求められており、無停止サービスを動的に提供できなくてはなりません。

新時代のスーパーコンピューターには、マルチテナント環境でベアメタルのパフォーマンスを実現できる構成が求められています。スーパーコンピューターの設計で最も重要なポイントは、パフォーマンスをできる限り高めながら、オーバーヘッドを最小限に抑えることです。ユーザーのアプリケーションが、ベアメタルのパフォーマンスと最大限の速度のネットワークを活用して、画期的な成果や新たな科学的発見をもたらすことができなくてはなりません。

クラウドネイティブ スーパーコンピューター アーキテクチャが目指すのは、スーパーコンピューターとしてのパフォーマンス特性を維持しながら、クラウド サービスの要件を満たすことです。つまり、最小権限のセキュリティ ポリシーと分離、データの保護、オンデマンドですぐに利用できる AI サービスや HPC サービスが求められます。クラウドネイティブ スーパーコンピューター アーキテクチャの開発の土台は、企業、学術

機関、政府機関などが参加するオープン コミュニティでの開発です。次世代のスーパーコンピューティングを構築するうえで、発展を続けるこうしたコミュニティの力が欠かせません。

以下のセクションでは、このような背景のもとで誕生した新しいアーキテクチャについて詳しく説明します。それは、クラウド サービスの要件を満たすインフラストラクチャ プラットフォームを基盤として、妥協のないパフォーマンスを実現する、クラウドネイティブの HPC および AI プラットフォーム アーキテクチャです。

DPU — クラウドネイティブのスーパーコンピューティング インフラストラクチャ プラットフォーム

データ プロセシング ユニット（DPU）は、インフラストラクチャ プラットフォームの種類の 1 つです。スーパーコンピューティング アプリケーション向けのインフラストラクチャ サービスを担い、アプリケーションがネイティブなパフォーマンスを発揮できるように構成および設計されています。ハードウェアのプロビジョニングと管理や、サービス（コンピューティング、ネットワーキング、ストレージ、セキュリティ）の仮想化は、すべて DPU が処理します。DPU を利用することで、アプリケーションの配置や、ネットワーク トラフィックとストレージのパフォーマンスが最適化され、マルチユーザーのスーパーコンピューターのパフォーマンスが全体として向上するとともに、サービス品質が確保されます。さらに、DPU ベースのインフラストラクチャ AI エンジンに、各種ハードウェア装置が生成したテレメトリ データを与えることで、クラスターのオペレーションを監視し、透過的な分析と最適化を可能にします。こうした AI エンジンは、スーパーコンピューター ユーザーへの課金や、スーパーコンピューターのビジネス モデルの強化に活用できます。

DPU では、インフラストラクチャ サービスをオフロードして、共有プラットフォームを構築し、サービス レベル アグリーメントに基づくセキュリティやアプリケーション分離を共有環境のなかで実現できます。また、新しい NVIDIA In-Network Computing および In-Network Storage によるアクセラレーションを可能にし、アプリケーションの総合的なパフォーマンス、処理効率、スケーラビリティを向上できます。

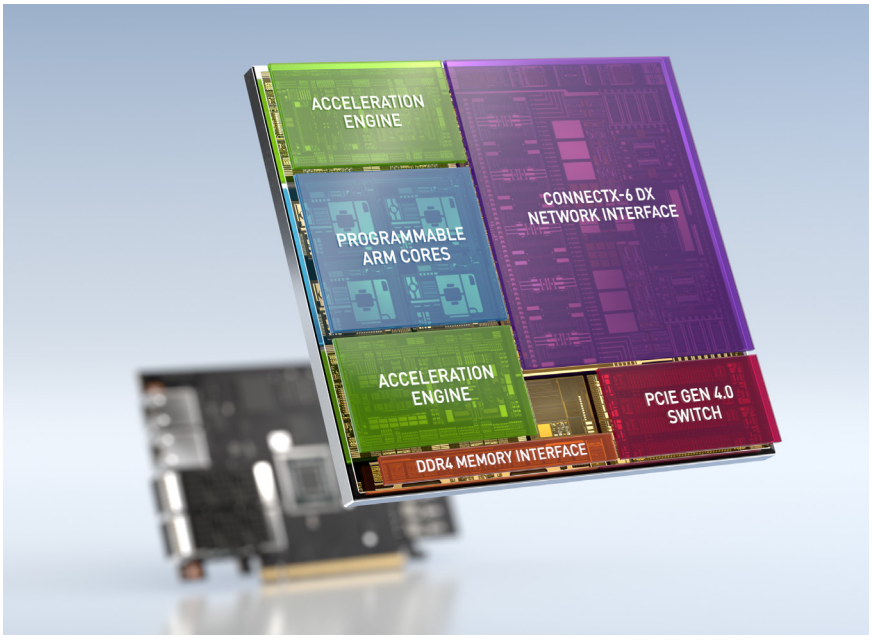

図 1：NVIDIA BlueField-2 DPU

DPU は、保護されたデータ コンピューティングもサポートしているため、機密性が高いデータの処理にスーパーコンピューティング サービスを利用できるようになります。DPU のアーキテクチャでは、クライアントのストレージとクラウドのスーパーコンピューターの間でデータがセキュアに転送され、データ暗号化がユーザーに代わって実行されます。

NVIDIA® BlueField® DPU は、業界をリードする NVIDIA ConnectX® ネットワーク アダプターに、複数の Arm® コア、個別用途向けのハイ パフォーマンス コンピューティング ハードウェア アクセラレーション エンジン、オンチップ型データセンター インフラストラクチャの完全なプログラマビリティ、PCIe サブシステムを組み合わせた構成です。

アクセラレーション エンジンと、プログラム可能な Arm コアの組み合わせにより、複雑なインフラストラクチャ管理やユーザーの分離と保護をホストから DPU に移行し、こうした処理に伴うオーバーヘッドを抑制するとともに、ハイパフォーマンスの通信やストレージのフレームワークを高速に処理できます。

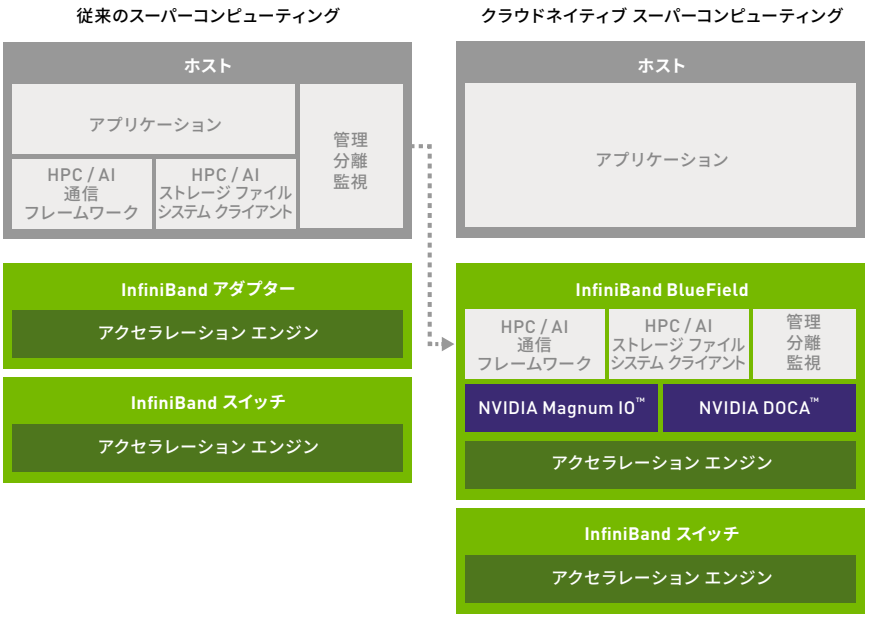

図 2：クラウドネイティブ スーパーコンピューティング アーキテクチャ

インフラストラクチャの管理、ユーザーの分離とセキュリティ、通信とストレージのフレームワークを、信頼されていないホストから信頼済みのインフラストラクチャ コントロール プレーン（DPU はこの一部）に移行することで初めて、真のクラウドネイティブ スーパーコンピューティングが可能になります。CPU や GPU の処理能力をこれまで以上にアプリケーションに振り向けることができ、動作の協調性も高まるため、全体のパフォーマンスとスケーラビリティが向上します。また、通信とストレージのフレームワークを BlueField DPU に移行することで、演算処理と通信処理のオーバーラップの度合いを高められるため、スーパーコンピューティングのパフォーマンスを最大限に引き出し、ROI を向上できます。

マルチテナントの分離：ゼロトラスト アーキテクチャに向けて

BlueField DPU は、各ノードのエッジでゼロトラストのスーパーコンピューティング ドメインを実現できます。マルチテナントのスーパーコンピューティング インフラストラクチャのなかで完全な分離と保護を行い、ベアメタル パフォーマンスを提供します。

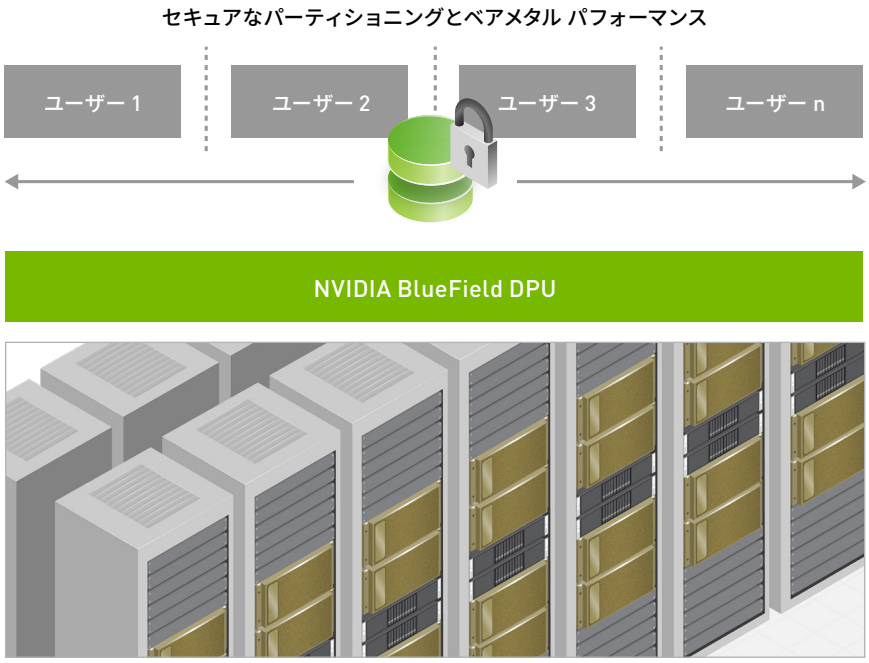

図 3：DPU が実現するベアメタルの分離

BlueField DPU では、信頼されていないマルチノード テナントをホストでき、前のテナントで使用されたスーパーコンピューティング リソースは、痕跡を一切残すことなくクリーンな状態で新しいテナントに引き渡されます。このプロセスのなかで、BlueField DPU は、ノードの完全性の保護、必要なリソースの再プロビジョニング、状態の痕跡の消去、新たにスケジュールされたテナントへのクリーンなブート イメージの提供などを行います。

HPC や AI の通信フレームワークのオフロード

HPC や AI の通信フレームワークには、Unified Communication X (UCX)、Unified Collective Communications (UCC)、Message Passing Interface (MPI)、Symmetrical Hierarchical Memory (SHMEM) などがあり、協調動作する並列処理のプロセス間でデータを交換するためのプログラミング モデルが定められています。これらのライブラリには、1 対 1 通信や集合通信のセマンティクス（データ有 / 無）があり、同期、データ収集、リダクションに対応しています。この種のライブラリはレイテンシや帯域幅の影響を受けやすく、アプリケーションのパフォーマンスを大きく左右します。

一般に、並列処理のアプリケーションの動作は、演算処理の時間と通信処理の時間を交互に繰り返す形で進んでいきます。そのため、新しい演算処理の時間を開始するには、それまでのプロセスがすべて完了していなくてはなりません。どこか 1 つのモードで遅れが生じたら、スーパーコンピューターのジョブ実行全体が遅れることになります。従来は、通信ライブラリの処理もすべてホストの CPU で行っていたため、パフォーマンスのボトルネックになっていました。

通信ライブラリの処理をホストから DPU にオフロードすることで、通信と演算を並行して（オーバーラップで）進めることができ、システムのノイズの悪影響を抑えられます。これが、エクサスケール コンピューティングの重要な戦略の 1 つであり、次世代のスーパーコンピューティング アーキテクチャを実現する鍵です。

BlueField DPU には、個別用途向けのハードウェア アクセラレーション エンジン（たとえば NVIDIA In-Network Computing エンジン）があり、通信フレームワークの特定の部分を高速化できます。たとえば、データ リダクションベースの集合通信や、タグ マッチングなどです。通信フレームワークのその他の部分は、DPU の Arm コアにオフロードでき、通信のセマンティクスを非同期で進めることができます。使用例としては、MPI のノンブロッキング All-to-All 集合通信での BlueField の活用があります。オハイオ州立大学（OSU）の MVAPICH 開発チームと、X-ScaleSolutions のチームが、OSU MVAPICH MPI でこの処理を DPU の Arm コアに移行したところ、通信処理と演算処理のオーバーラップが 100% となり、ホスト CPU で処理する場合に比べ 99% 向上しました。

Parallel Three-Dimensional Fast Fourier Transforms (P3DFFT) は、乱気流の調査や、気候学、天体物理学、材料科学など、さまざまな分野で大規模なコンピューター シミュレーションに使われているライブラリです。このライブラリは Fortran90 で記述されており、並列処理のパフォーマンスを生かせるように最適化されています。プロセッサ間の通信には MPI を使用し、MPI All-to-All のパフォーマンスに大きく依存しています。OSU と X-ScaleSolutions のチームは、BlueField を活用した OSU MVAPICH MPI により、P3DFFT で 1.4 倍のパフォーマンス高速化を達成しました。

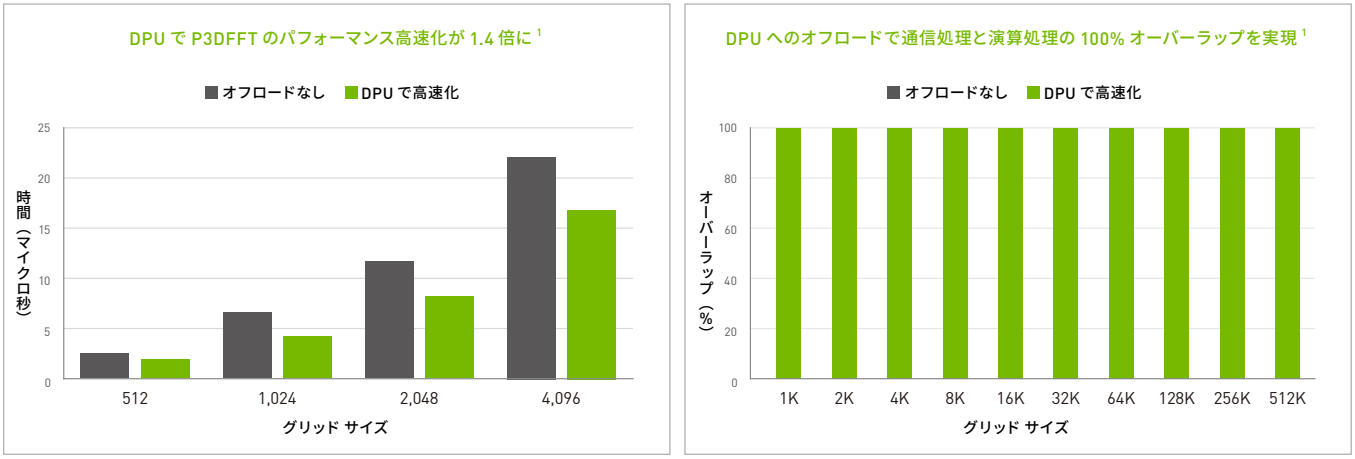

図 4：DPU で P3DFFT のパフォーマンスが向上

クラウドネイティブ スーパーコンピューティングのための BlueField HPC DOCA SDK

NVIDIA DOCA ソフトウェア開発キット（SDK）は、NVIDIA BlueField DPU を基盤とする開発を迅速に行うための SDK です。業界標準の API を活用して、ハードウェアで高速化したソフトウェアデファインドのネットワーク、ストレージ、セキュリティ、管理、および AI と HPC のアプリケーションやサービスを開発できます。ソフトウェアデファインドで高性能のクラウドネイティブな DPU 対応サービスを構築し、未来のスーパーコンピューティング インフラストラクチャをプログラミングできます。

DOCA を使用すれば、DPU 上で動作するアプリケーションやサービスを非常に柔軟な環境で開発でき、NVIDIA In-Network Computing アクセラレーション エンジンや Arm プログラマブル エンジンをシームレスに活用して、パフォーマンスとスケーラビリティを向上できます。将来的には、組み込みの GPU コアも活用してネットワーク ワークロードに AI アルゴリズムを実行し、セキュリティやパフォーマンスなどを強化できるようになります。

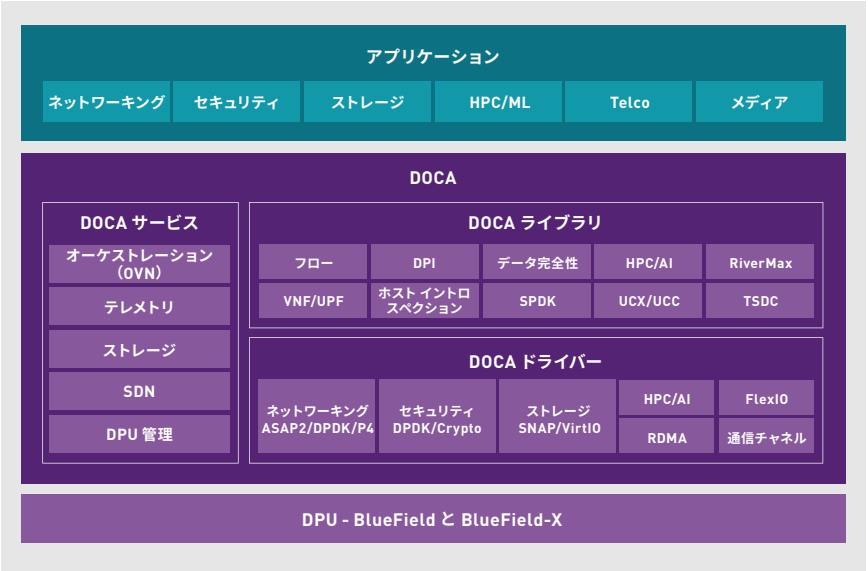

図 5：DOCA SDK

DOCA SDK は、DPU ベースのスーパーコンピューターのためのフル ソフトウェア スタックを提供し、HPC と AI のサービス デリバリ プラットフォームの開発を支えます。

DOCA のパッケージは、業界標準のオープンな API とフレームワークで構成されており、UCX による 1 対 1 通信、UCC による集合通信、NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP) ™ によるデータ リダクション、Open Smart Network API (OpenSNAPI)、ストレージ、セキュリティ、テレメトリ、管理などに対応しています。こうしたフレームワークを利用し、統合されている NVIDIA アクセラレーション パッケージを活用して、アプリケーションをシンプルにオフロードできます。DOCA ベースのサービスは、コンピュート ノードでは業界標準の入出力 (IO) インターフェイスとして公開され、インフラストラクチャの仮想化と分離を実現できます。SDK は、さまざまなオペレーティング システム、ディストリビューション、MPI および SHMEM ライブラリをサポートし、ドライバー、ライブラリ、ツール、ドキュメント、サンプル アプリケーションが含まれています。

クラウドネイティブ スーパーコンピューティングのための ソリューション — NVIDIA DGX SuperPOD

現代の AI エンタープライズに必要なのは、クラウドネイティブ スーパーコンピューティングのターンキー ソリューションです。ユーザーのニーズの変化に直ちに応え、ワークロードに合った適切なサイズのリソースをセキュア マルチテナントで提供できるソリューションが求められています。NVIDIA DGX SuperPOD は、要求の厳しいエンタープライズ環境に合わせて構築されたクラウドネイティブ スーパーコンピューターで、NVIDIA BlueField DPU と NVIDIA Base Command™ Manager を基盤としています。ユーザーとデータをセキュアに分離したマルチテナント環境でありながら、妥協のないベアメタル パフォーマンスを実現できます。

図 6：NVIDIA DGX SuperPOD

DGX SuperPOD を導入すれば、トップクラスのパフォーマンスを誇るインフラストラクチャを、クラウド同様の確実なセキュリティのもとで利用できます。

NVIDIA クラウドネイティブ スーパーコンピューティング プラットフォームの詳細については、**https://www.nvidia.com/ja-jp/networking/products/cloud-native-supercomputing/** をご覧ください。

このプラットフォームの基盤にあるテクノロジの詳細については、以下をご覧ください。

NVIDIA InfiniBand ネットワーキング https://www.nvidia.com/ja-jp/networking/products/infiniband/

NVIDIA DGX SuperPOD https://www.nvidia.com/ja-jp/data-center/dgx-superpod/

NVIDIA DOCA SDK https://developer.nvidia.com/networking/doca

NVIDIA BlueField DPU https://www.nvidia.com/ja-jp/networking/products/data-processing-unit/

NVIDIA Magnum IO https://www.nvidia.com/ja-jp/data-center/magnum-io/

1 パフォーマンス テストは、HPC-AI Advisory Council のクラスター センターで実施しました。システム構成は次のとおりです。
32 台のサーバー、各ノードはデュアルソケット Intel Xeon 16 コア CPU E5-2697A V4 @ 2.60GHz（ノード当たり計 32 プロセッサ）、256GB DDR4 2400MHz RDIMM メモリ、1TB 7.2K RPM SATA 2.5" ハード ドライブを搭載。各サーバーは、NVIDIA BlueField-2 InfiniBand HDR100 DPU および NVIDIA Quantum™ QM7800 40 ポート HDR 200Gb/s InfiniBand スイッチに接続しました。