AI ネットワーキング

はじめに

革新的技術として登場した人工知能（AI）は、医療から金融サービス、エンターテイメントまで、多くの産業と人々の日常生活のさまざまな側面に変革をもたらしています。リアルタイム・ゲーム、バーチャル・リアリティ、生成 AI、メタバース・アプリケーションが急速に進化し、ネットワーク、コンピューティング、メモリ、ストレージ、インターコネクト I/O の相互作用の方法も変化してきています。AI がかつてないペースで進歩し続けていることを受け、ネットワークはトラフィックの膨大な増加に適応して、何百、何千ものプロセッサ、数兆件に及ぶトランザクション、ギガビット規模のスループットに対応する必要があります。AI が研究室や研究プロジェクトの段階から本格的な導入へと急速に進むに従い、ネットワークとコンピューティングのリソースに対する需要は増大します。現在の進展は、今後 10 年間に起こる変革の基盤にすぎません。これから AI クラスタは大幅に増大すると予想されます。

こうした AI ワークロードの共通点は、データとコンピューティング・リソースの両方を大量に消費するということです。一般的な AI トレーニングのワークロードは、数十億個ものパラメーターと大規模な疎行列計算を処理する必要があり、CPU、GPU、TPU など数百から数千のプロセッサを使用します。これらのプロセッサは集中的に計算を実行した後、ピア間でデータを交換します。ピアから受け取ったデータは集約またはローカル・データとマージされ、次の処理サイクルが開始されます。この計算・交換・集約のサイクルでは、ネットワークを介する通信にジョブ時間の約 20～50%が費やされるため、ボトルネックがジョブの完了時間に大きく影響します。

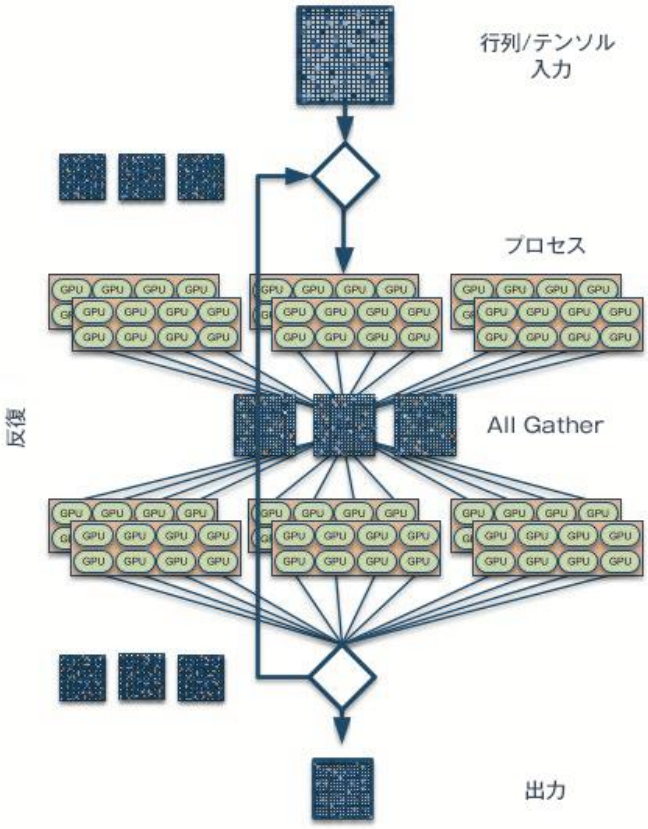

図 1：計算・交換・集約サイクル

本ホワイトペーパーでは、最適化された AI ネットワークを構築するために必要なプロトコル、プラットフォーム、トポロジなどの重要なネットワーク技術について解説します。

TCP/IP と RDMA

RDMA は、最新の AI アプリケーションに求められるスケーラブルな並列処理を実現する重要なオフロード技術です。TCP/IP ソケットでは、データはユーザー空間からカーネル空間にコピーされてから、ネットワーク・ドライバー、ネットワークへと順に到達します。AI アプリケーションに関連する大量のデータを処理するには、CPU がボトルネックになる可能性があります。

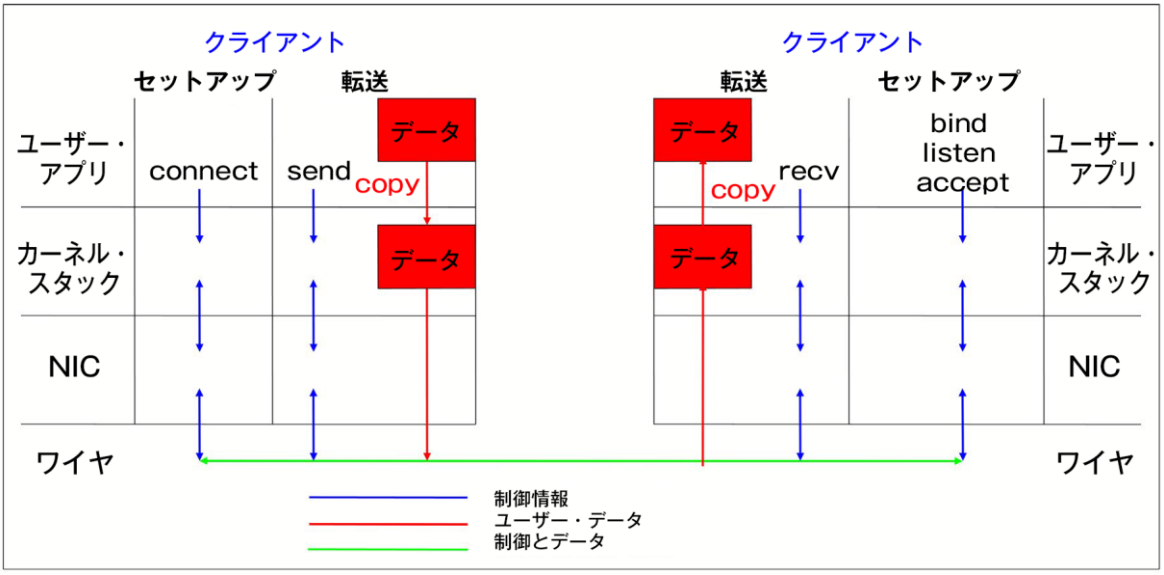

図 2：TCP/IP 転送

そこで役に立つのが、リモート・ダイレクト・メモリ・アクセス（RDMA）です。RDMA は、カーネルに依存することなくメイン・メモリ内のデータ交換を可能にするため、ハイパフォーマンス・コンピューティング・システムで広く使用されています。RDMA を使用すると CPU サイクル数が減るため、スループットとパフォーマンスが向上し、データ転送が高速化され、RDMA 対応システム間の遅延が低減されます。

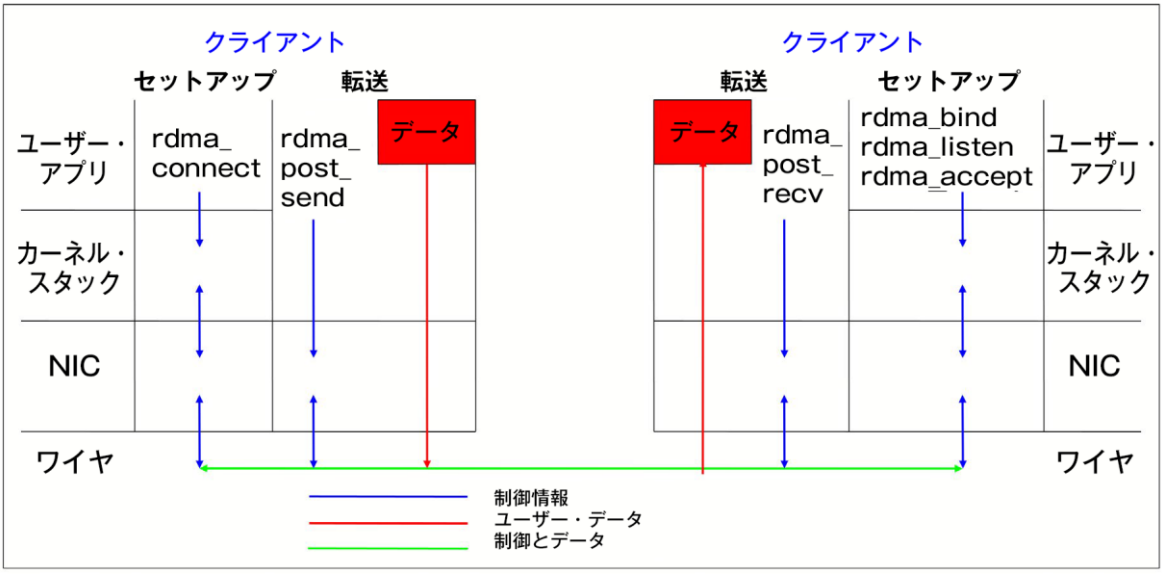

図 3：RDMA 転送

RDMA 転送のセマンティクスは、InfiniBand Verbs ソフトウェア・インターフェイスによって定義されます。これには、メモリ・ブロックの登録、記述子の交換、RDMA の読み取りおよび書き込み操作のポスティングが含まれます。このインターフェイスは、物理トランスポート層として Infiniband から独立しています。

RoCE は、イーサネット・ネットワーク経由で InfiniBand ペイロードを伝送する方法を定義します。RoCEv2 は、トラフィックのルーティングを可能にしてこのスケーラビリティと機能をさらに拡張し、イーサネット経由の RDMA のスケーリングを実現します。

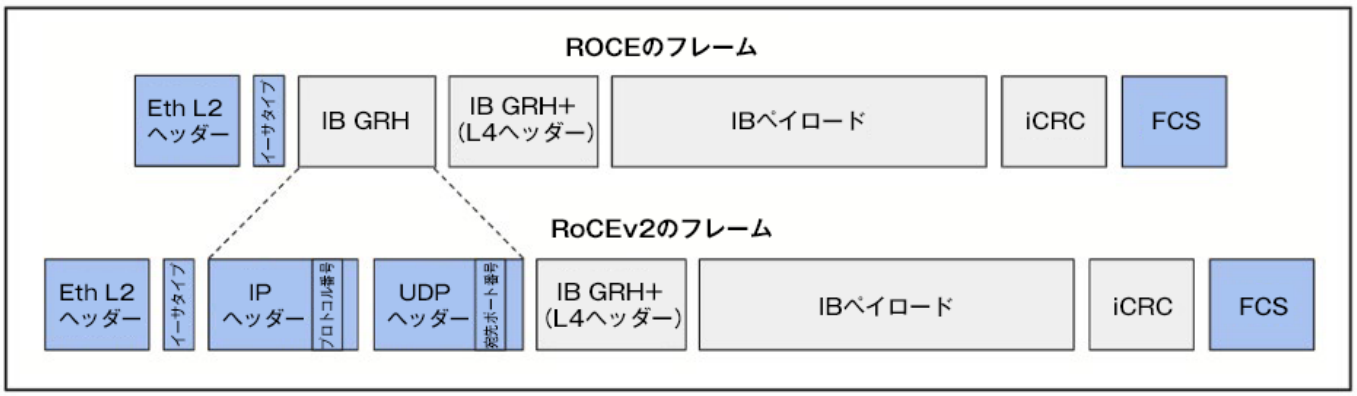

図 4：RoCE と RoCEv2 のフレーム形式

集団通信

最新の大規模言語モデルは、数十億から数兆のパラメーターを使用し、トレーニングに使われる巨大なデータセットを 1 つのホスト GPU では処理できません。そのようなデータセットとモデルを複数の GPU に分割し、トレーニングを並列して行います。これによって得られた勾配と重みは、集団通信を使用してメンバーGPU 間で集約され、同期されます。

集団通信は、同じコミュニケーターに含まれるすべてのプロセスの情報交換を可能にします。よく使用される集団通信プリミティブには、broadcast、gather、scatter、all-to-all、global reduction（allreduce）、allgather があります。最終目的は、各ステップですべてのプロセスを確実に同期することです。コミュニケーター内のプロセス間ですべてのパラメーターが同期されるまで、どのプロセスも進めることができないようにバリアが形成されます。プログラマーは、NCCL、oneCCL、RCCL、MSCCL など広く使われている集団通信ライブラリを利用して、効率的で十分に検証された通信アルゴリズムをアプリケーションに組み込むことができます。

allreduce のようにすべての GPU 間でメッセージを交換する必要がある集団通信には、リング型アルゴリズムと二分木アルゴリズムがよく使用されます。次の図は、4 つのプロセス間のメッセージ交換に使われるリング型アルゴリズムを示します。

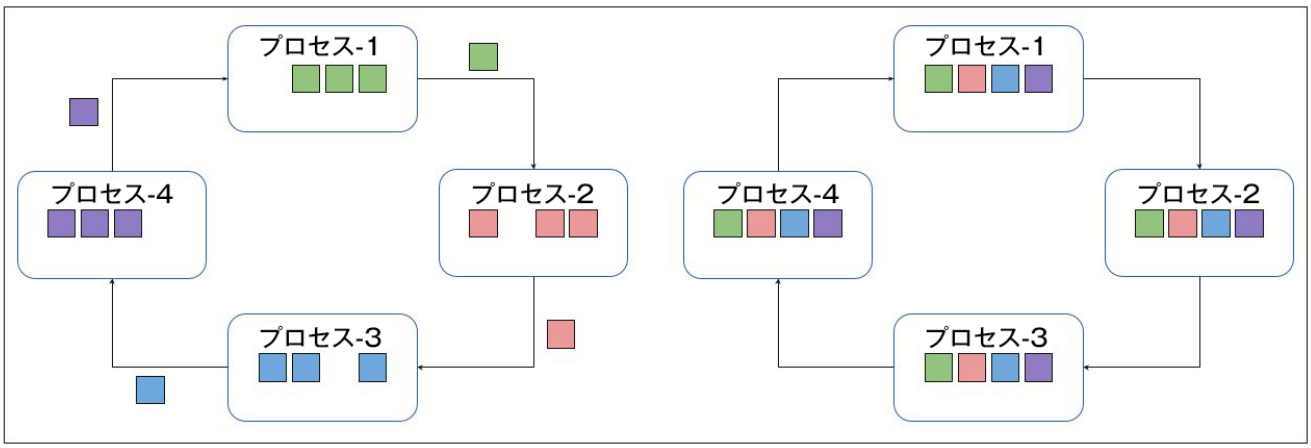

図 5：リング型アルゴリズムを使用した allreduce

リングは帯域幅に最適化され、すべてのエンド・ホスト間にラインレートの帯域幅を提供するネットワークを必要とします。リング型では帯域幅効率は高くなりますが、モデルのトレーニングに使用される GPU の数が増えると、遅延が直線的に増加します。

木アルゴリズムでは、参加するプロセスを順位付けし、重複しない二分木に分割することにより、遅延を低く維持したまま GPU をスケーリングできます。次の図では、16 個のプロセスを重複しない 2 つの二分木に分割しています。

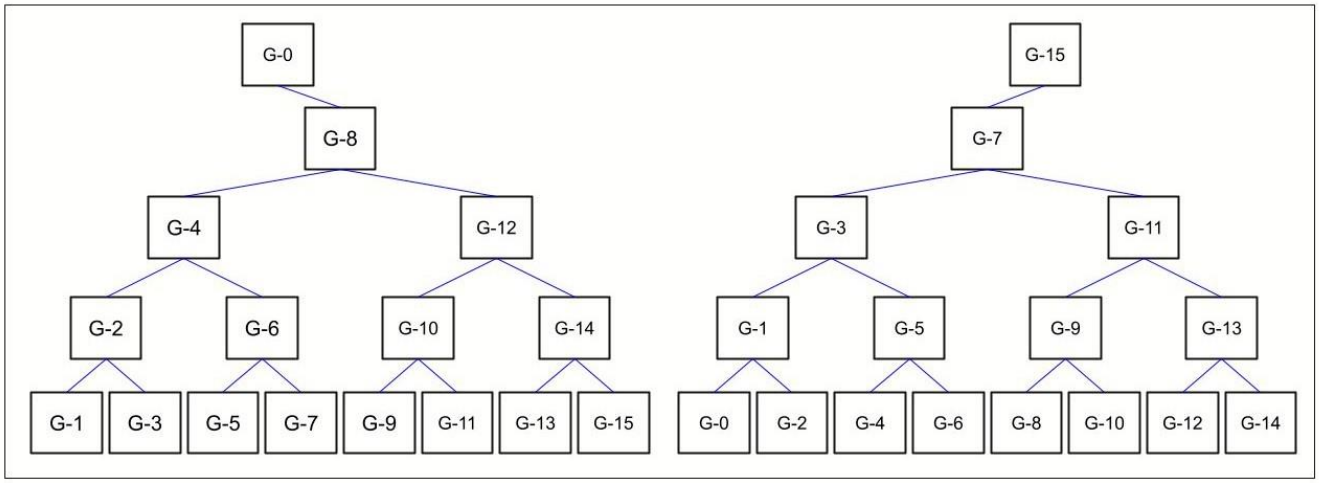

図 6：重複しない二分木

各プロセスは、2 つのピア・プロセスからメッセージを受信し、2 つのピア・プロセスにメッセージを送信します。このモデルでは、リング型のように遅延が直線的には増加しませんが、上流のプロセスが可能な限りラインレートに近い帯域幅で受信側の各プロセスにメッセージを送信できるように、ネットワークでトラフィックのインキャストを効率的に管理する必要があります。

AI ネットワークには、メッセージを効率的にやり取りでき、プロセスが各バリアを通過して次の計算段階に進むことができる、適切な相互接続を選択することが不可欠です。

AI ネットワーク向けのインターコネクト

イーサネットは、データセンター、バックボーン、エッジ、キャンパス・ネットワークにおいて、非常に低速のものから現在の 100G、200G、400G、800G、そしてロードマップで予定されている 1.6T という高速のものまで、さまざまなユースケースに広く展開されています。一方、Infiniband は、HPC クラスタで一般に使用されているネットワーク技術です。前述のとおり、AI/ML のワークロードはネットワークを大量に消費し、従来の HPC のワークロードとは異なります。

また、大規模言語モデル（LLM）の爆発的な増加に伴い、より多くの GPU とストレージ容量が絶えず求められています。最新の AI アプリケーションは、数千もの GPU とストレージ・デバイスを備えた大規模なクラスタを必要とし、これらのクラスタは、需要の拡大に応じて数万のデバイスにスケーリングする必要があります。GPU の速度は 2 年ごとに倍増しており、スケーラブルなネットワーク設計によってコンピューティングとネットワークの両方のボトルネックを回避することが重要です。アプリケーション・チームがコンピューティング・キャパシティに重点を置くのに対し、ネットワーク・チームは次のようなさまざまな要素に基づいてインターコネクトを入念に評価する必要があります。

パフォーマンス

AI クラスタのパフォーマンスを測定する重要な指標の 1 つは、ジョブ完了時間です。理想的なパフォーマンスを実現するには、ネットワークがロスレスかつノンブロッキングであり、ラインレートのリンク使用率を提供する必要があります。後述するように、RoCEv2 は、適切な輻輳制御メカニズムと効率的な負荷分散技術により、AI ワークロードに必要な最良のパフォーマンスを提供できます。

帯域幅と速度

トレーニング・ジョブが大きくなるに従い、より高速なネットワークを提供することが重要になっています。これは、高速のポート速度を備えた高密度スイッチを使用することで、より効率的に実現できます。マーチャント・シリコン・イーサネット・ソリューションでは、ネットワークの帯域幅を 2 年ごとに倍増でき、ビットあたりのコストとビットあたりの電力を削減できます。

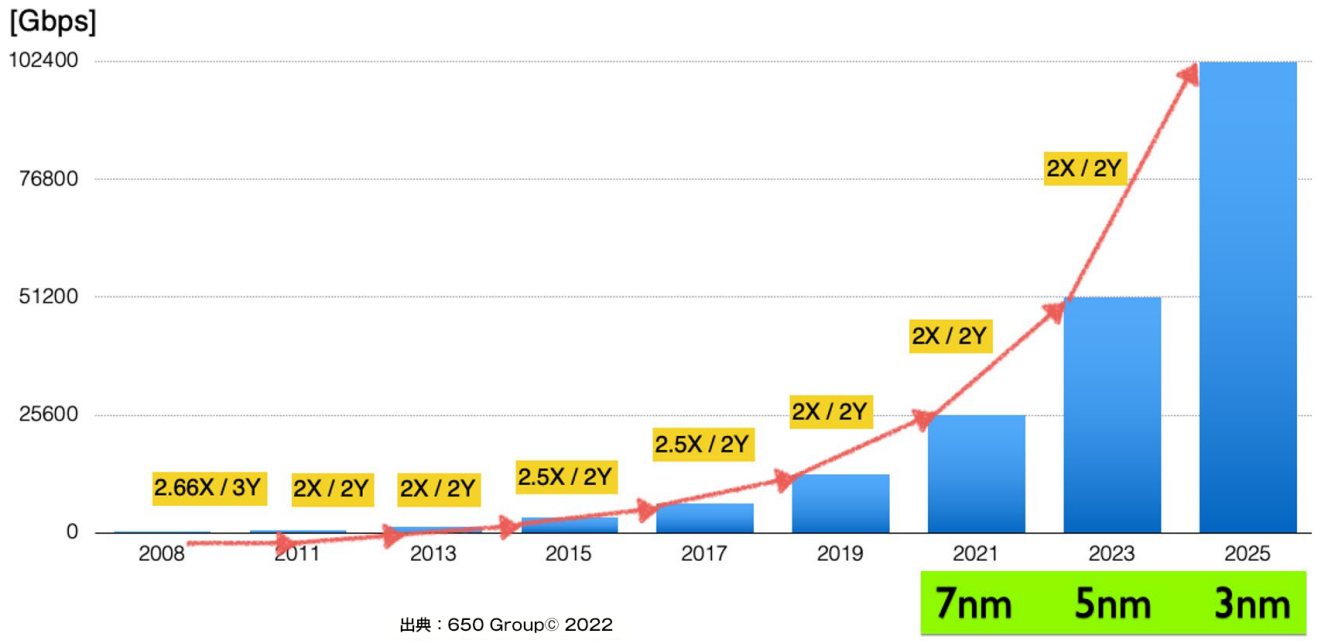

図 7：2025 年までのシングルチップ・イーサネット・スイッチ・シリコン

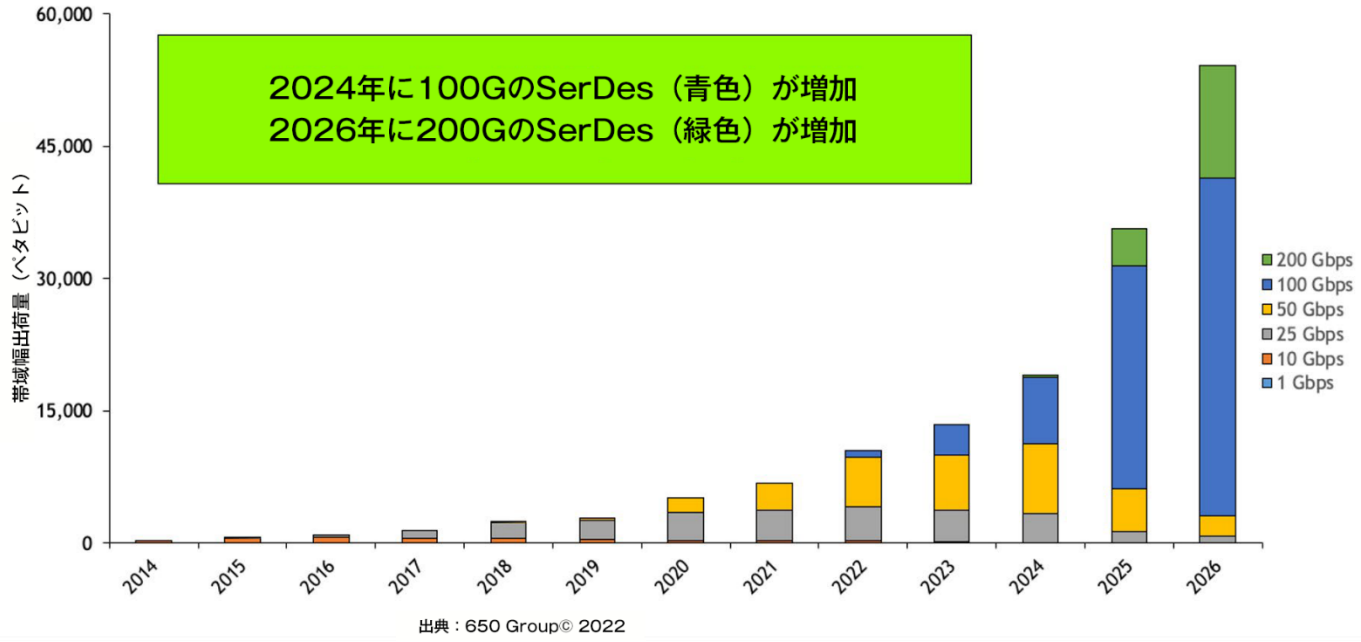

図 8：データセンター・イーサネット・スイッチング帯域幅の増大（SerDes 速度別）

ロスレス・ネットワーク

高速化は有用ですが、ジョブ完了時間にとってはロスレス・ネットワークが最も重要です。Infiniband は、クレジットベースのフロー制御を適用することによってパケット損失を回避します。送信側は、宛先ホストから利用可能なバッファを示すクレジットを受け取るまで、パケットの送信を待ちます。イーサネットも、明示的輻輳通知（ECN）と優先度ベース・フロー制御（PFC）を使用することにより、ロスレス・チャネルとして動作します。これらのメカニズムは、送信側にバックプレッシャーを適用することによって、ホストまたはスイッチのバッファのオーバーランを回避します。RDMA のパフォーマンスを最大化するには、IB フロー制御またはイーサネットと ECN/PFC による信頼性の高いトランスポートが不可欠です。

スケーラビリティ

LLM のモデル・サイズ増大は、確実かつ予測可能な能力向上をもたらしました。それはさらなる LLM のサイズ増大につながり、AI クラスタのインターコネクト拡大にもつながります。つまり、ネットワークのスケーラビリティが非常に重要になります。

イーサネットのスケーラビリティは、世界最大規模のクラウド・ネットワークで立証されています。ネットワーク・チームは、クラウド設計を採用して、Border Gateway Protocol（BGP）を実行する CLOS アーキテクチャの分散ネットワークを構築することができるようになりました。

一方、Infiniband のコントロール・プレーンは単一のサブネット・マネージャーを使用して一元化され、物理トポロジの検出と、各ノードの転送テーブルと QoS ポリシーの設定を行います。ネットワークを定期的にスイープし、トポロジの変化に応じてデバイスを再構成します。これは、小規模なクラスタでは問題ありませんが、規模が大きくなるとボトルネックになる可能性があります。パッチとして適用できる複雑な事後解決策はあります。これに対し、イーサネットの分散コントロール・プレーンは、Infiniband の最大サブネット・サイズ 48000 より大きくスケーリングでき、単一の障害点を回避して高い耐障害性を提供します。

耐障害性

Infiniband のサブネット・マネージャーに障害が発生すると、サブネット全体が停止する可能性があります。Infiniband には、特定の状況において転送を継続できる技術がありますが、それでもコントロール・プレーンは一元化されていて脆弱です。バックアップのサブネット・マネージャーへのフル・フェイルオーバーではダウンタイムが発生し、サブネットが大きくなるほどダウンタイムは長くなります（転送すべき状態が増え、ノードのスイープが大きくなります）。お客様の話によると、ダウンタイムは 30 秒から数分に及びます。ユースケースによってはこれを許容できる場合もありますが、大規模な AI/ML ワークロードの場合は、このような障害はジョブ完了時間とパフォーマンス全体に多大な影響を及ぼします。イーサネットとアリスタの SSU などの機能を使用する分散型のスケーラブルなアーキテクチャでは、リンクやノードの障害が大規模な AI ネットワークのパフォーマンス全体に及ぼす影響はごくわずかか皆無です。

ネットワーク管理

この 10 年、ネットワーク・チームはクラウド原則を採用してインフラストラクチャを 1 つのユニットとして運用し、管理してきました。AI クラスタの利用が広がりつつある中、ネットワーク・チームにとっては、AI クラスタを孤立して扱うより、共通のインフラストラクチャの一部として扱う方が望ましいでしょう。また、イーサネットは、データセンター、キャンパス、バックボーン、WAN、エッジで広く使用されており、ネットワーク・チームにはイーサネットに関する高い専門知識が蓄積されています。

可視化

ネットワークの自動化とシームレスなアクションの実行には、テレメトリと可視化が非常に重要です。ネットワーク・チームは、データセンターの汎用コンピューティングとストレージに現在使用しているツール、プロセス、ソリューションを、AI クラスタにも拡張したいと考えるでしょう。

相互運用性

多くの場合、AI ネットワークは、多様なストレージや汎用コンピューティング・インフラストラクチャとやり取りします。イーサネットベースの AI ネットワークは、こうした多様なシステムとのパイプラインのボトルネックを回避する効率的で柔軟なネットワーク設計を可能にします。IP トラフィックは物理的な Infiniband ネットワークで伝送できますが、すべてのサーバーが Infiniband HCA を搭載するか、IB ネットワークに出入りするスループットを大幅に制限する Infiniband-to-Ethernet ゲートウェイを通過する必要があります。

オープン性

イーサネットは、多数のシリコン・ベンダー、システム・ベンダー、オプティクス・ベンダーから成る非常に強力なエコシステムを擁し、ベンダー間で相互運用可能なオープンで標準ベースのソリューションの方向へ進んでいます。InfiniBand は、選択肢が限られたロックイン・ソリューションであり、大きく後れを取っています。

要するに、イーサネットは、そのスケーラビリティ、相互運用性、信頼性、費用対効果、柔軟性、馴染み深さから、AI ネットワーキングに最適なソリューションであると考えられます。イーサネットは、実績があり、広く採用され、高速ネットワーキングをサポートしているため、AI ワークロードに対応できる効率的でスケーラブルなネットワーキング・インフラストラクチャを構築しようとしている組織にとって魅力的な選択肢となっています。

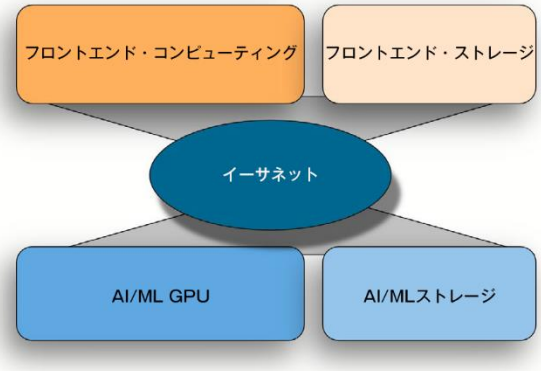

図 9：イーサネットはコンピューティング、AI/ML、ストレージに対応

AI ワークロードにイーサネットを使用する場合の重要要件を考えてみましょう。ネットワークには、RoCEv2 をサポートするロスレス・トランスポート、制御トラフィックの優先度を制御する QoS（Quality of Service）、調整可能なバッファ割り当て、効果的な負荷分散、リアルタイム監視が必要です。

Arista EOS

最新の AI アプリケーションには、100Gbps、400Gbps、800Gbps、およびそれ以上の速度で数百、数千の GPU をインターコネクトできる、高帯域幅、ロスレス、低遅延で、スケーラブルなマルチテナント・ネットワークが必要です。EOS は、データセンター量子化輻輳通知（DCQCN）、優先度制御 QoS（Quality of Service）、調整可能なバッファ割り当てスキームをサポートし、高度にロスレスの高帯域幅、低遅延ネットワークの実現に必要なツールをすべて提供します。

EOS は、Data Center Quantized Congestion Notification（DCQCN）のサポートにより、Priority-based Flow Control（PFC）と Explicit Congestion Notification（ECN）を組み合わせたエンドツーエンドの輻輳制御スキームを提供して RDMA over Ethernet をサポートします。ネットワーク・トラフィックとバッファ利用率を可視化できなければ、PFC と ECN の適切なしきい値を設定することは困難です。Arista EOS®（Extensible Operating System）は、AI アナライザ機能とレイテンシー・アナライザ機能を使用して、ワークロードのトラフィック・パターンの詳細な可視化を提供します。

AI アナライザは、マイクロ秒単位の間隔でインターフェイスのトラフィック・カウンタを監視し、レイテンシー・アナライザは、インターフェイスの輻輳とキューイング遅延を追跡し、リアルタイムでレポートします。

AI アナライザとレイテンシー・アナライザは、アプリケーションのパフォーマンスとネットワーク利用率および輻輳イベントの相関付けを支援し、アプリケーションの要件に合わせて PFC と ECN の値を最適に設定できるようにします。

GPU クラスタでは、ノード間のデータ転送に少数のキュー・ペアが使用されます。これは、各スイッチにおける高帯域幅トラフィック・フローの数は少ないということを意味します。パケット・ヘッダーのエントロピー不足により、このようなフローはコリジョンや輻輳が発生しやすく、ジョブ完了時間が長くなります。EOS は、ネットワーク・リンクのリアルタイムのトラフィック利用率を考慮し、リンク間で均一にフローのバランスを図って、ネットワーク・ホットスポットを回避します。また、送信元インターフェイスベースのハッシュを提供してオーバーサブスクリプションなしのネットワークにおけるトラフィックの減速を防ぎます。ホスト・インターフェイスに到着したトラフィック・フローを指定アップリンクに直接ハッシュでき、トラフィックのファンインとコリジョンを回避できます。さらに、パケット・ヘッダー内でユーザー定義フィールドを使用するように EOS の負荷分散を設定して、エントロピーを追加することもできます。これらの結果、ネットワークの輻輳の削減、ECN でマークされるパケットの削減、PAUSE フレームの削減、ノードの総スループットの向上、ワークロードの完了時間の短縮が実現されます。

すべての RDMA アプリケーションが同じように動作するわけではありません。遅延の要件が非常に厳しい反面、スループットの要件は厳しくないアプリケーションもあれば、可能な限り高いスループットを必要とし、遅延でのトレードオフを許容するアプリケーションもあります。大半のアプリケーションは、この 2 種類のタイプの中間に位置付けられます。EOS で QoS 分類、スケジューリング、調整可能なバッファ割り当てスキームなどのツールを使用することにより、お客様はネットワークを完全に制御して、アプリケーションの要件を満たすように調整で

きます。EOS では、VxLAN と EVPN のサポートにより、そのような複数のアプリケーションを単一のネットワークで実行でき、スケーラブルなマルチセグメンテーションのニーズに対応できます。

強力なソフトウェア機能に加え、信頼性の高い、最適に設計されたハードウェアも必要です。

プラットフォーム

AI ネットワークの帯域幅要件とスケーリング要件は、お客様やアプリケーションごとに異なります。1 つですべてに対応できる万能のものはありません。アリスタネットワークスは、クラス最高のマーチャント・シリコン・パケット・プロセッサを活用して、世界中のあらゆる AI ネットワークに最適な組み合わせのイーサネット AI リーフと AI スパインのシステムを提供します。

AI リーフ - 7060X5

7060X5 シリーズは、高密度で電力効率の高い固定構成の 800G、400GbE、200GbE、100GbE データセンター・スイッチで、高ネットワーク基数を採用し、ハイパースケール・クラウド、人工知能、機械学習環境に最適化されています。一貫した低遅延、実績のある可視化、トラフィック計測、自動化機能を提供し、大規模なハイパースケール・クラウドや IO 集約型環境に魅力的なソリューションを実現します。

7060X5 は、AI リーフとして、最大 32 の 800G ポートを 1 ラック・ユニットに収め、リーフとスパインの階層間のボトルネックを解消することで、コンピューティングとストレージの 25.6T 帯域幅パフォーマンスを最大化する効率的な高基数クラスタを実現します。また、7060X5 シリーズは、ブレークアウトを使用して最大 64 の 400G ポートまたは最大 256 の 100G もサポートし、コスト効率の高い新しい 800G オプティクスで現在の密度を倍増します。

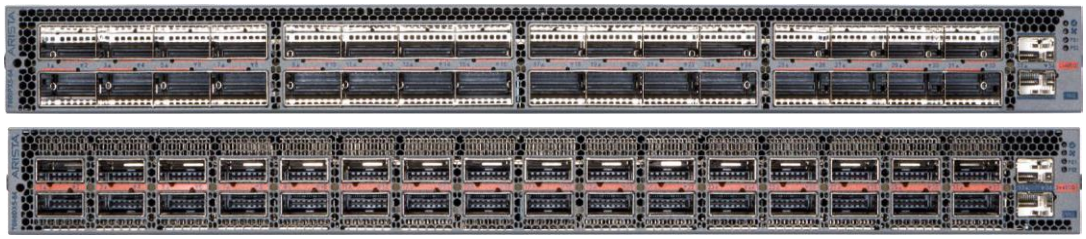

図 10：Arista 7060X5-64：32 の 800G QSFP-DD ポートまたは OSFP800 ポート、2 つの SFP+ポート

Arista 7060DX5-64S は、最大 64 の 400G ポートを収めた 2 ラック・ユニット・システムで、合計スループットは 25.6T です。QSFP-DD 400G インターフェイスを備えた 7060DX5-64S プラットフォームは、業界標準のオプティクスとケーブルをサポートしており、400G への移行を容易にします。すべてのポートは、400GbE、200GbE、または最大 256 の 100GbE インターフェイスの速度で利用できます。

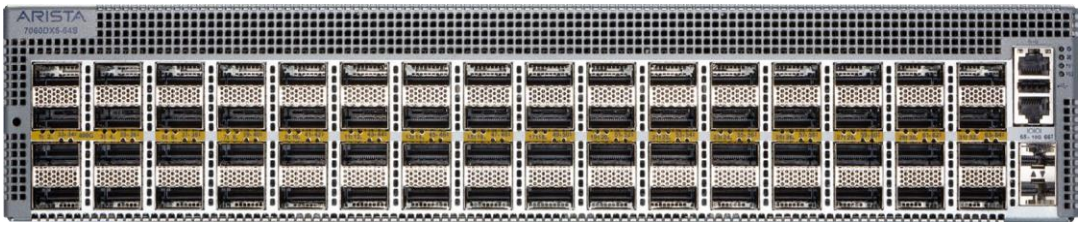

図 11：Arista 7060DX5-64S：64 の 400GbE QSFP-DD ポート、2 つの SFP+ポート

Arista 7388X5 シリーズは、高密度の 200G および 400G を提供し、高ネットワーク基数を採用して、ハイパースケール・クラウド、人工知能、機械学習環境に最適化されています。7388X5 は、1 つの 25.6Tbps 大容量パケット・プロセッサをベースに構築されたモジュール型システムで、大量のデータを消費して一貫した低遅延を必要とするワークロードに適しています。業界標準インターフェイスを柔軟に選択でき、消費電力とシステム密度の大幅な向上を実現します。

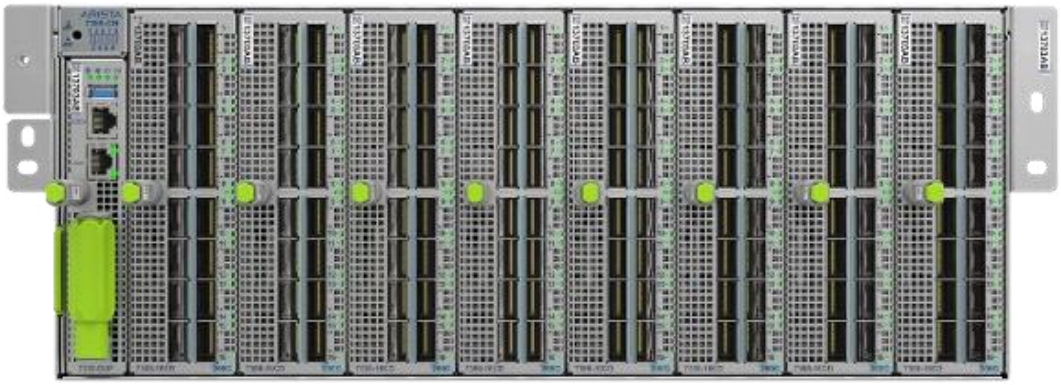

図 12：Arista 7388X5：128 の 200G ポートまたは 64 の 400G ポート

AI スパイン - 7800R3

Arista 7800R3 シリーズは、特定用途向けに設計されたモジュール型スイッチで、最大 460Tbps のシステム・スループットで業界最高のパフォーマンスを提供し、非常に大規模なデータセンターとハイパフォーマンス・コンピューティング・ネットワークの要件に対応します。Arista 7800R3 シリーズではノンブロッキングのスイッチング容量が提供され、データセンター向けの劇的に高速でシンプルなネットワーク設計が可能になるとともに、設備投資も運用コストも低減できます。

7800R3 には、AI ネットワーキングに理想的なプラットフォームとして次の重要な特長があります。

仮想出力キュー（VoQ）：スイッチ内で分散スケジューリング・メカニズムを使用して、輻輳した出力ポートへのアクセスをめぐって競合するトラフィック・フローの公平性を確保します。クレジット要求/付与ループが使用され、出力パケット・スケジューラが特定の入力パケットに対するクレジット付与を発行するまで、入力パケット・プロセッサの物理バッファで VoQ にパケットがキューイングされます。

セルベースのファブリック：セルベースのファブリックは、すべてのパケットを均一な大きさのセルに分割してから、すべてのファブリック・モジュールに均等に「スプレー（分配）」します。このスプレー動作に多くの利点があり、各転送エンジンに均等なフロー・バランスをもたらす非常に効率的な内部スイッチング・ファブリックを実現します。セルベースのファブリックは、トラフィック・パターンに関係なく 100%効率であると考えられています。

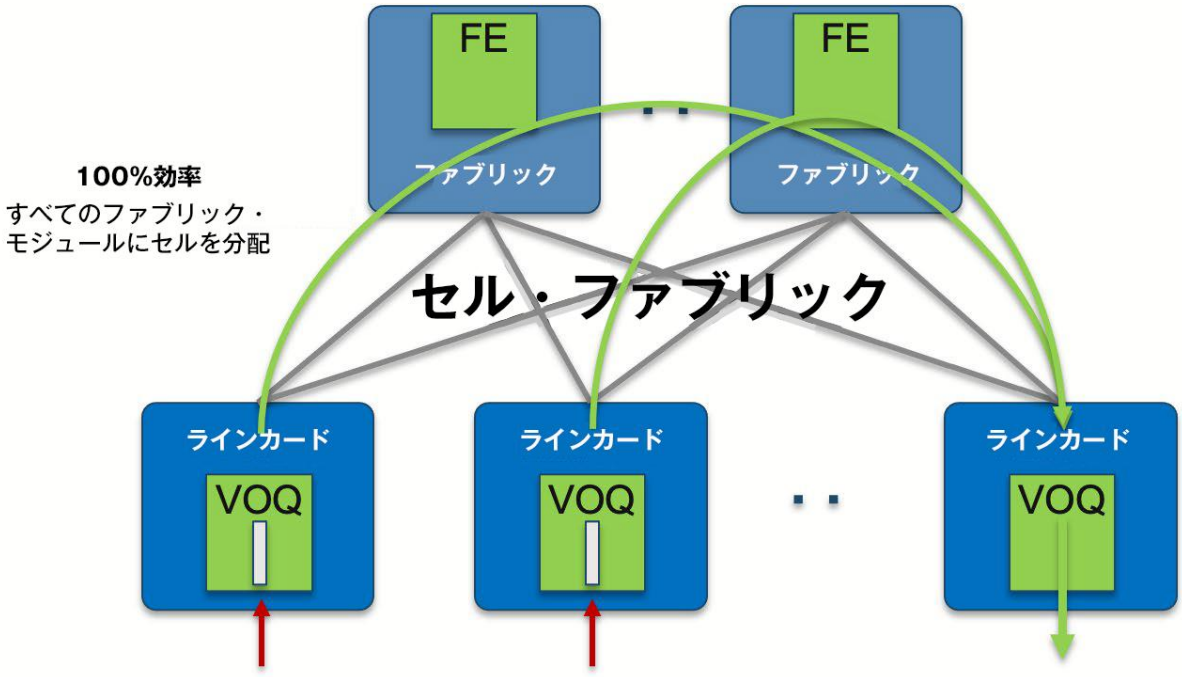

図 13：セルベースのファブリックのアーキテクチャ

セル・ファブリックは、このスプレー動作により、混合速度への対応に適しています。セルベースのファブリックは、前面パネルの接続速度に左右されないため、100G、200G、400G の混在とマッチングにもほとんど問題はありません。

さらに、セル・ファブリックは、イーサネット・ファブリックの「フロー・コリジョン」の問題の影響を受けません。フローはすべてのパスを使用して宛先に到達し、ネットワーク内に内部ホットスポットが発生しないため、セル・ファブリックは AI/ML アプリケーションで一般的な「エレファント・フロー」の大量トラフィックに特に適しています。

ディープ・パケット・バッファリング：7800R3 シリーズのラインカードは、オンチップ・バッファ（32MB）と柔軟なパケット・バッファ・メモリ（パケット・プロセッサごとに 8GB の HBM2）を組み合わせて使用します。オンチップ・バッファは非輻輳転送に使用され、瞬間的または持続的な輻輳時には HBM2 パケット・バッファをシームレスに使用します。バッファは VoQ ごとに割り当てられ、調整は不要です。また、特筆すべき点として、輻輳時にはパケットが HBM2 パケット・バッファから宛先パケット・プロセッサに直接伝送されます。

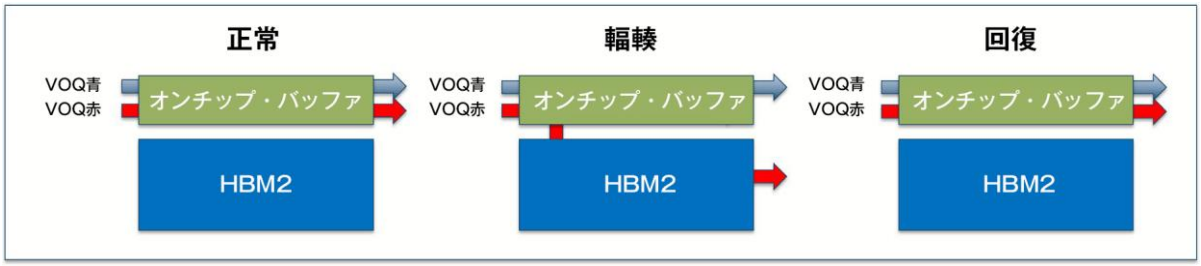

図 14：パケット・バッファ・メモリ・アクセス

HBM2 メモリは Jericho2 パケット・プロセッサに直接統合されているため、Jericho2 パケット・プロセッサに信頼性の高いインターフェイスが提供され、HMC や GDDR で必要となる高速メモリ相互接続の追加が必要ありません。その結果、同等の GDDR メモリと比較して、消費電力が 43%以上削減されます。

図 15：HBM のメモリ・パッケージング統合

予測可能なパフォーマンス：セルベースのプラットフォームで仮想出力キュー（VOQ）と高度なキューイング・クレジット・スケジューラおよびディープ・バッファ（輻輳回避用）を組み合わせることにより、7800R3 はロスレス・システムとなっています。セルベースのシステムは、どのような負荷においても予測可能なパフォーマンスを提供し、仮想出力キュー（VOQ）の追加により、輻輳時のパケット損失を防止します。これら 2 つの機能とディープ・バッファ・プラットフォームを組み合わせることにより、AI/ML ワークロードの GPU インターコネクトにおける RoCEv2 のロスレス・トランスポートが保証されます。

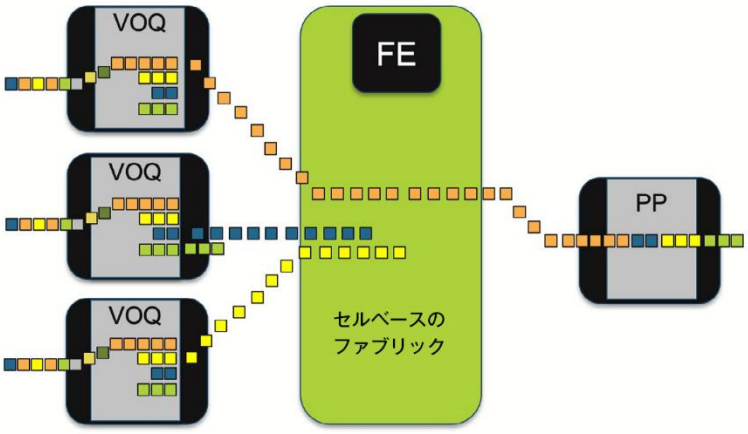

図 16：クレジットベース VoQ のアーキテクチャ

密度:7800R3シリーズは、4、8、12、16スロットのシステムで提供され、高密度100Gおよび400Gを提供する幅広いラインカードをサポートし、転送テーブルのスケールを選択できます。システム・レベルでは、16 スロットの Arista 7816R3 は、電力効率に優れた前面吸気/背面排気の 32RU の筐体で最大 460Tbps、576 の 400G ポートまでスケーリングでき、機能を損なうことなく、業界最高水準のパフォーマンスと密度を提供します。

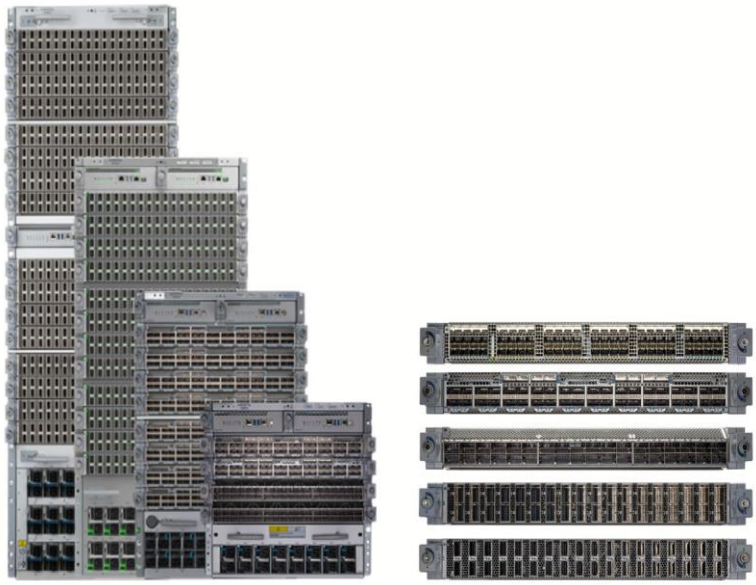

図 17:Arista 7800R3 シリーズ

柔軟性と効率:7800R3 シリーズのすべてのコンポーネントはホット・スワップ可能であり、冗長化されたスーパーバイザ、電源、ファブリック、ファンの各モジュールと、前面吸気/背面排気のエアフローを備えています。このシステムはデータセンター向けに特化して設計されており、エネルギー効率に優れ、標準消費電力は 1 つの 100G ポートあたり 25 ワット未満、1 つの 400G ポートあたり 50 ワット(フル装備の筐体の場合)です。

Arista 7800R3 のこれらのすべての特長と EOS の強力な機能セットを組み合わせることにより、7800R3 は信頼性と拡張性に優れたデータセンター・ネットワークやハイパフォーマンス・ネットワークの構築に理想的なプラットフォームとなります。

AI アプリのサイズに基づく設計

年月とともに、サーバー仮想化、アプリケーション・コンテナ化、マルチクラウド・コンピューティング、Web 2.0、ビッグデータ、ハイパフォーマンス・コンピューティング(HPC)などの新しいテクノロジーとアプリケーションによって、データセンター内の水平型と垂直型のトラフィック・パターンは大きく変わりました。これらの新しいテクノロジーのパフォーマンスを最適化し、向上させるために、分散型スケールアウトのディープ・バッファ設計 IP ファブリックが、超「水平型」のトラフィック・パターンに対応できるようにスケーリング可能な一貫したパフォーマンスを提供することが実証されています。お客様は、IP/イーサネットを使用して小規模から大規模のデータセンター・クラウド・ネットワークを構築し、最新のアプリケーションとネットワークの要件に対応することに成功しています。

AI/ML アプリケーションはこれまで、IP ファブリック内で他のアプリケーションと共存することができました。しかし、AI/ML アプリケーションが大幅に増大し、特定用途の GPU、DPU、TPU の採用に伴い、関連する複雑さも増しているため、このようなアプリケーション専用のネットワークを設計することをお勧めします。そうすることにより、オペレーターは、最新の AI/ML ワークロード特有のトラフィック・パターンをより適切に処理できるようにネットワークを調整できます。

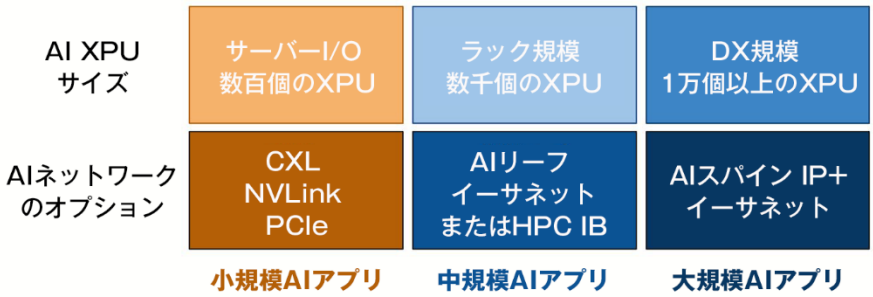

図 18:AI ネットワークの設計ガイドライン

1 ラックに収まる GPU のインターコネクトを必要とする小規模な AI アプリケーションの場合、PCIe、CXL や、NVLink のような他の独自技術などのロード/ストアインターコネクトを使って、低遅延、低消費電力で効率的にデータを移動できると考えることができます。しかし、このようなソリューションはすぐに多くのコストと電力を必要とするようになり、複数ラック間の接続には対応できなくなります。ラック間でホストをインターコネクトする必要があるアプリケーションには、イーサネットまたは InfiniBand がプロトコルの選択肢となります。

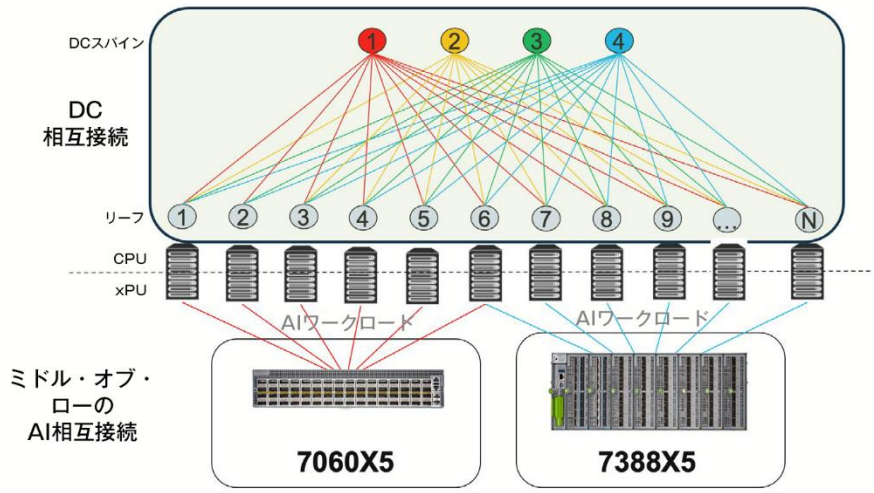

図 19:ミドル・オブ・ローの AI 相互接続

小規模な AI アプリケーション

64 の 400G ポートまたは 128 の 200G ポートを備えた Arista DCS-7060DX5-64S または DCS-7388X5 スイッチ 1 台で、数ラックの GPU を効果的にインターコネクトできます。この設計では、ノンブロッキング構成および予測可能な低遅延で、各 GPU が他のすべての GPU と通信できます。このオプションでは必要な調整が最小限で済み、運用と管理が簡素化されます。

中規模な AI アプリケーション

576 の 400G ポートをサポートする Arista 7800R3 スイッチ 1 台で、複雑な設定なしで使えるシンプルな AI スパインインターコネクトを提供して、中規模な AI アプリケーションに対応できます。この設計は、エンド・ホスト間の一貫したシングル・ホップを提供するため、遅延と電力の要件が低減されます。セルベースのノンブロッキング VOQ アーキテクチャを採用した 7800R3 では、単一の大規模なロスレス・ネットワークを実現でき、設定作業と調整は不要です。シングル・ホップ・ソリューションにより、ECN と PFC の設定が必要なのはホストに面したポートのみになるため、GPU は常にラインレートでデータを送受信できます。

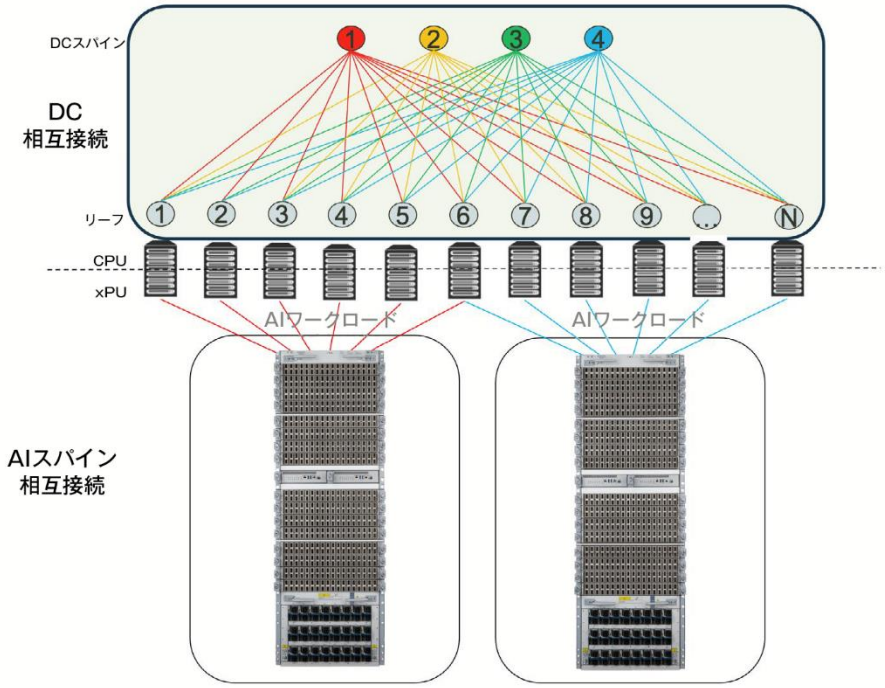

図 20:AI スパインインターコネクト

大規模な AI アプリケーション

データセンターで何万もの GPU を接続する必要がある大規模な AI アプリケーションの場合、イーサネットが最も現実的な選択肢となります。アリスタのユニバーサル・リーフ/スパイン設計は、非常にシンプルで柔軟かつスケーラブルなアーキテクチャを提供してデータセンター規模の AI ワークロードをサポートします。この設計により、予測可能で低い遅延を維持しながら、18,000 以上の 400G エンド・ホストをインターコネクトすることができます。このような設計において、Arista EOS のインテリジェントな負荷分散機能でネットワークのリアルタイムのトラフィック利用率を考慮し、トラフィック・フローを均一に分散することで、フローのコリジョンを回避することができます。Arista EOS の AI アナライザとレイテンシー・アナライザなどの高度なテレメトリ・オプションを使用すると、ネットワーク・オペレーターは、GPU がネットワーク全体でラインレート・スループットの交換ができてパケット損失を防げるように、PFC と ECN の最適な設定しきい値を簡単に決定できます。

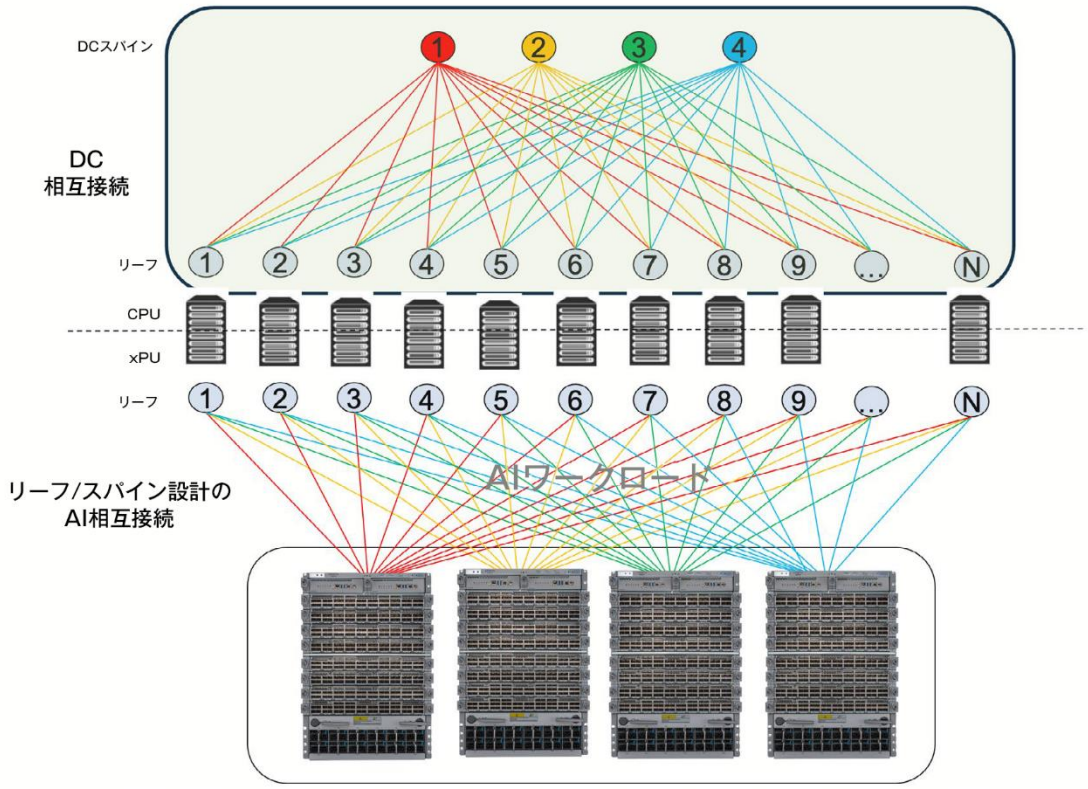

図 21:リーフ/スパイン設計の AI インターコネクト

ユニバーサル・リーフ/スパイン設計は、現在数百の GPU を必要とする AI モデルに理想的なソリューションを提供し、将来は数万の GPU まで一貫したパフォーマンスでスケールアウトできる柔軟性を提供します。

AI ネットワーク対応のストレージ

企業が AI モデルの精度向上を目指すに従い、AI で使用されるデータ量は激増しています。トレーニングの段階で AI モデルの精度を向上させるには、大規模なデータセットが必要になります。そのため、組織はペタバイト単位から始まる膨大なデータの集まりを管理する必要に迫られています。その結果、GPU とストレージ・ノード間のデータ転送を処理するネットワークに多大な負荷がかかることになります。ネットワークのボトルネックが原因となって、高価で需要の高い GPU がデータを待つアイドル状態になることを防ぐには、専用のストレージ・ネットワークが推奨されます。データを効率的に移動できるようにするため、ほとんどの GPU は、NVMe-oF を使用して、メモリとリモート・ストレージの間の直接データ・パスを可能にします。

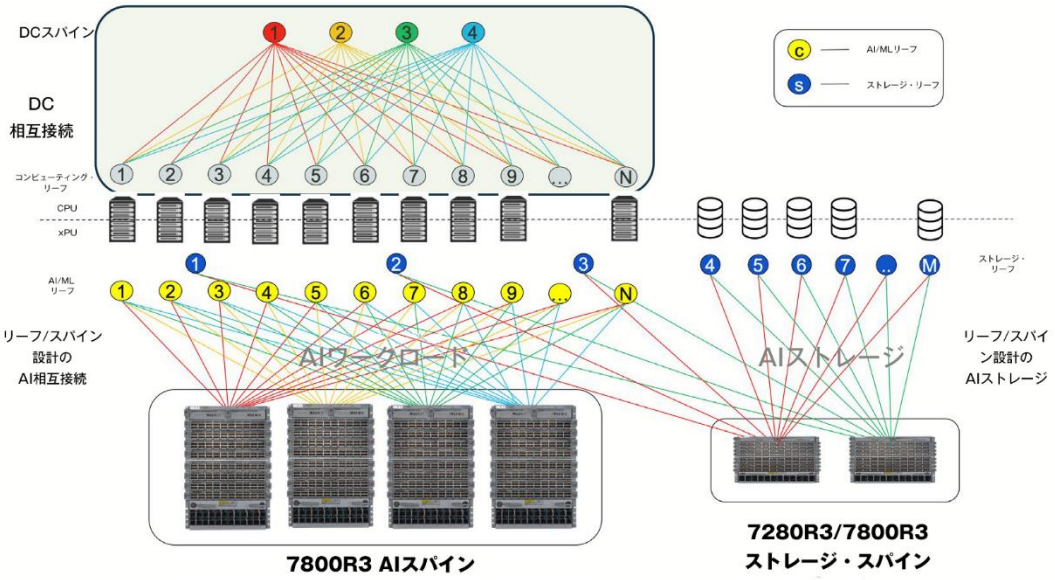

図 22：AI 対応のストレージ

7280R3 シリーズをリーフ・スイッチとして使用し、7800R3 シリーズをスパイン・スイッチとして使用するアリスタのユニバーサル・リーフ/スパイン・アーキテクチャは、NVMe/ROCE や NVMe/TCP などのプロトコルを使用するストレージ・ネットワークとして実績のあるクラス最高のソリューションです。VOQ メカニズムとディープ・パケット・バッファリング機能を統合したアリスタの 7280R3 シリーズと 7800R3 シリーズのスイッチを使用すれば、コスト効率に優れた高性能ストレージ・ソリューションを大規模に設計できます。

Ultra Ethernet Consortium

現在のイーサネットベースのソリューションは拡張性に優れていますが、ジョブ完了率をさらに向上させるには、基盤となるイーサネット・ネットワークを簡素化し、速度とスケーラビリティを高めるように再設計する必要があります。

その実現を目指し、アリスタは、2023 年 7 月に設立が公表された Ultra Ethernet Consortium（UEC）の創設メンバーになりました。他のメンバーとして、現在最大規模の AI および HPC ネットワークのサプライヤや事業者が参加しています。UEC の目標は、メンバーの長年にわたるネットワークの構築と運用の経験を活用して、イーサネットベースのオープンで相互運用可能な高性能のフル通信スタック・アーキテクチャを実現し、オンプレミスやパブリック・クラウドに展開される AI/ML および HPC のワークロードのネットワーク需要増大に対応することです。

UEC は、従来の RoCE プロトコルを Ultra Ethernet Transport に置き換えることを目指しています。これは、イーサネット/IP エコシステムの利点を維持しながら、AI アプリケーションに必要なパフォーマンスを提供するように設計された最新のトランスポート・プロトコルです。UEC は、通信スタックの変更を最小限に抑え、イーサネットの相互運用性を維持および促進しながら、要求の厳しいワークロードを総合的に向上させるために、モジュール化された互換性のある相互運用可能なレイヤとその密接な統合の体系的なアプローチを採用します。これにより、RDMA の弱点であるパケット損失、DCQCN、マルチパス機能の欠如、エンドポイントとプロセスに関するスケーリングの制約に対処します。AI モデルのビジネス資産としての機密性と価値が高まっていることを受け、UEC はネットワーク・セキュリティも設計に組み込み、将来の AI/ML および HPC ネットワークに対応できる堅牢性を目指します。

UEC プロトコルは、最新の HPC ワークロードもサポートするように設計され、MPI や PGAS など広く使用されている API を維持しながら、前述と同じトランスポート・メカニズムを採用します。UEC の詳細については、http://www.ultraethernet.org/をご覧ください。

まとめ

アリスタは、AI/ML ワークロードに対応する GPU とストレージのインターコネクトに IP/イーサネット・スイッチを使用する最適なソリューションを提供します。AI アプリケーションの急激な増大により、イーサネットのような標準化されたトランスポートを使用して、電力効率の高いインターコネクトを構築し、従来のアプローチの管理やスケールアウトの複雑さを解消することが必要とされています。アリスタのハイパフォーマンス・スイッチを利用して、IP/イーサネット・アーキテクチャを構築すれば、アプリケーションのパフォーマンスを最大化するとともに、ネットワーク・オペレーションを最適化することができます。7800R3 AI スパインと 7060 AI リーフを EOS の先進的機能と組み合わせると、最新の AI アプリケーションに理想的な選択肢となります。

アリスタネットワークスは、パケット・スプレー、柔軟な順序付け、最新の輻輳制御アルゴリズム、テレメトリによる支援、スケーラブルなセキュリティ、AI/ML に最適化された API を備えて実装された、スケーラブルで効率的なリモート・メモリ・アクセスを実現して、これからの膨大な計算量の通信ニーズに応えることを目標に、UEC をリードしていきます。

リファレンス

- RDMA - http://www.rdmaconsortium.org/
- RoCE - https://cw.infinibandta.org/document/dl/7781 - “InfiniBand Architecture Specification Release 1.2.1 Annex A17: RoCEv2” InfiniBand Trade Association.
- Arista L3LS Design Deployment Guide
- Arista 7800R3 Switch Architecture WP
- Arista UCN Deployment Guide
- Collective Communication - https://www.mpi-forum.org/docs/mpi-1.1/mpi-11-html/node64.html

アリスタネットワークスジャパン合同会社

〒100-0004　東京都千代田区大手町 1-7-2　東京サンケイビル　27F
Tel:03-3242-6401

西日本営業本部
〒530-0001　大阪市北区梅田 2-2　ヒルトンプラザウエストオフィスタワー　19F
Tel: 06-6133-5681

お問い合わせ先

Japan-sales@arista.com

www.arista.com/jp

2023 年 7 月 19 日