

AI ネットワーキング

2026 年 6 月更新



人工知能(AI)の登場

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、エンターテインメントまで、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。

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、XPU、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。

AI 向けのネットワーキング

発明者のロバート・メトカーフが 1973 年にメモとして初めて発表し、1980 年に商業化されて以来、イーサネットは長い道のりを歩んできました。イーサネット技術は何度も進化し、世界中に広がっています。通信速度は、初期の 10 Mbps から現在では 1 ポートあたり 800 Gbps に高速化し、1.6 Tbps を目前に控えています。また、キャンパス・スイッチング、データセンター・スイッチング、ストレージ・ネットワーキング、高頻度取引(HFT)、ハイパフォーマンス・コンピューティング(HPC)、音声テレフォニーとストリーミング動画、クラウド・コンピューティング、そして広域通信網(WAN)さえもサポートできるようになりました。

今やイーサネットは働き方、遊び方、学び方に革命を起こす人工知能(AI)ワークロードの供給に最適な選択肢へとさらなる進化を遂げています。現在、イーサネットは大規模に展開されており、今すぐに高性能 AI ネットワーク向けに利用できます。そして、AI ネットワーキングをより強化するため、アリスタはイーサネットのさらなる進化に貢献しています。

AI センターの出現

AI ワークロードのトレーニング用に新しいバックエンド・ネットワークを構築する際、ネットワーク・オペレーターは、技術やスタッフのスキルの既存のサイロをより強固なものにするか、AI ワークロードの集合的性質を理解して、その概念を包括的な AI ソリューション全体に拡張するという最新のアプローチを AI に対して採用するかを選択できます。InfiniBand が AI に対するサイロ化アプローチの代表であるのに対し、イーサネットは AI をバックエンド・ネットワークからフロントエンド・ネットワークまで拡張し、一貫性と協調を実現する統合オプションの代表です。

ハイパフォーマンス・コンピューティング(HPC)クラスタ構築の従来型のソリューションでは InfiniBand を利用しており、それを新しいバックエンド・ネットワークでの AI トレーニングに拡張することを検討しているお客様もいます。しかし、このアプローチの根本は、これまでイーサネットを中心に構築してきた従来型のデータセンター(と AI 推論クラスタ向けのフロントエンド・ネットワーク)としてネットワーク・サイロを取り入れるというものです。お客様はこのようなネットワーク・サイロを接続するゲートウェイが必要になり、複雑さが増します。また、このモデルではバックエンドの AI ネットワークとフロントエンドの AI ネットワークで運用スキル・セットが異なるため、AI アクセラレータ処理ユニット(XPU)、汎用コンピューティング(CPU)、ネットワーキング、ストレージの運用が縦割り化されます。

AI センターはもっと優れた方法で、各レイヤでオープンスタンダードに準拠し、イーサネットでシステム全体のあらゆる要素を統合します。AI センターは、フロントエンドとバックエンドのエコシステム全体を統合し、パフォーマンスと運用を最適化して、AI 向けのスケールアウト・ネットワーキングを提供します。既存のデータセンターのワークロードやシステムだけでなく、AI ワークロード、コンピューティング、ネットワーク、ストレージの可視化、管理、制御を協調させることができます。

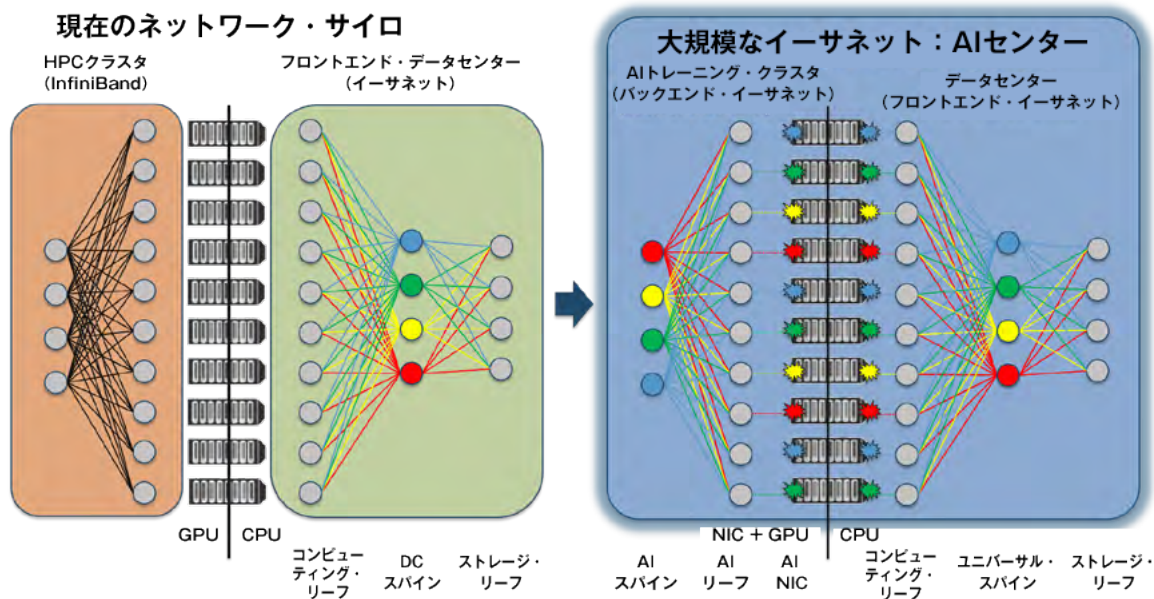


図 1: イーサネット対応の AI センターが出現し、ネットワーク・サイロを打破

アリスタのプラットフォームは、AI センターの共通のビジョンと利点を簡単に活用できるように設計されています。

アリスタの AI ファブリック: スケールアップ型、スケールアウト型、スケールアクロス型

実際、イーサネットの採用は AI センターのあらゆる領域へと広がり続けています。大規模 AI クラスターの設計に関する最近のイノベーションにより、バックエンドのトレーニング・クラスターの規模を拡大する「スケールアウト型」AI ファブリックだけでなく、個々のラック内の XPU インターコネクト（バックエンド・ネットワークのさらに奥に相当）にイーサネットを広げる新しい「スケールアップ型」AI ファブリックも出現しています。AI クラスターの大規模化が進み、必要なコンピューティング機器やネットワーク機器を 1 か所の物理的拠点には収容しきれなくなった結果、地理的に分散した AI センターや複数拠点を相互接続する新たな「スケールアクロス型」AI ファブリックへと、イーサネットの採用は広がっています。

ARISTA

AIファブリックがデータの新しい中心に

スケールアップ、スケールアウト、スケールアクロス
AI時代のための設計。

スケールアップ
性能の向上。
モデル規模の拡大。

スケールアウト
容量の拡大。
クラスター数の拡張。

スケールアクロス
拠点数の拡張。
ファブリックは1つ。

Y軸
スケールアップ

X軸
スケールアウト

Z軸
スケールアクロス

arista.com

スケールアウト型 AI ファブリック

複雑なワークロードに対応するため、多くのお客様がイーサネットベースの生成 AI クラスタを大規模に展開し、非常に大規模な言語モデル (LLM) をトレーニングして、静的および動的なテキスト、画像、音声、さらに動的な音声と画像を組み合わせたショート動画をサポートしてきました。これにはイーサネット経由の大規模な AI が必要なため、1 つのクラスタで 100,000 を超える AI アクセラレータ・ノードを持つ AI クラスタを複数展開するようになりました。この規模ならデータを取り込んでモデルをトレーニングするのに十分な計算処理能力をまかなうことができますが、このような AI クラスタに必要な前例のない物理スペースと消費電力に加え、同規模におけるネットワーク設計のオプションについて新たな課題も生じます。

16,384 台の AI コンピューティング・サーバー (約 16,000 台のサーバー) それぞれが 8 個の XPU アクセラレータで構成されている、大規模イーサネット AI クラスタの仮定のシナリオを考えてみましょう。これは、合計 131,072 個の XPU を 1 つのネットワークでインターコネクトする必要があるということです。16,000 台のサーバーを 1 つの物理ラックに収納することはできないので、多数の物理ラックに展開して、大規模なネットワーク基数を使ってすべての XPU をインターコネクトする必要があります。これを「スケールアウト」ネットワークといいます。多数のラック、サーバー、XPU ホストを取り込むため、ネットワークがデータセンターの水平方向に広がるからです。

このようなイーサネットベースの AI ネットワークに最も一般的なネットワーク設計は、ノンブロッキング、ロスレス、マルチステージの Clos ファブリックを基盤とするものです。このイーサネット AI ファブリックを構築する方法は複数ありますが、単純化するため、64 の 800 Gbps ポート (または 128 の 400 Gbps ポート) を備えた固定型で 2RU、容量 51.2 Tbps のイーサネット・スイッチをこの設計の土台としましょう。一般に、XPU 接続用の AI NIC は 400 Gbps ポートをベースとしており、アリスタの設計もこれを基盤としますが、800 Gbps の AI NIC も利用可能であり、その場合はネットワーク・インフラストラクチャ全体のポート速度がさらに向上します。

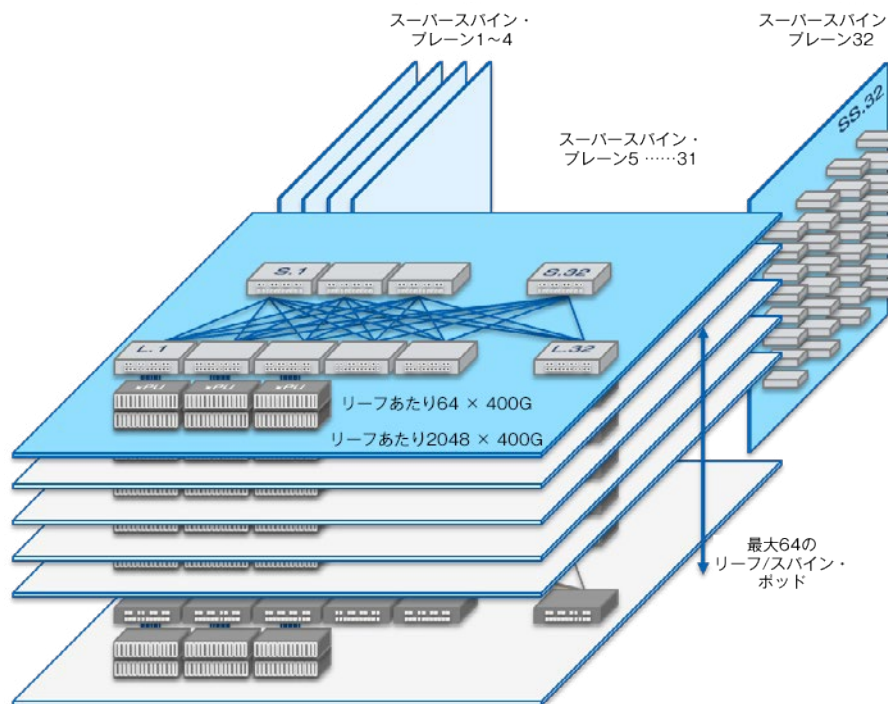


図 2: 16,000 台のサーバーと 131,000 個の 400G XPU AI アクセラレータに対応しているスケールアウト型イーサネット AI ファブリック

128 の 400GbE ポートのスイッチ・ポート基数を考えると、スケールアウト・ネットワークで 131,000 個の XPU AI ホスト・アクセラレータすべてをサポートするには、3 階層の Clos ファブリックに 5,120 台のスイッチが必要です。それには、スイッチを満載した高さ 48RU のラック 214 台に加え、1 時間あたり約 9 メガワットの消費電力と、約 328,000 本のプラグブル光トランシーバとケーブルが必要になります。これはすべてネットワーク機器のみに必要なものであり、16,000 台のサーバーと 131,000 個の XPU のためのスペースと消費電力は含まれていませんが、どちらの数字もはるかに大きくなります。

業界リーダーは、このように高い消費電力を軽減するための新技術に既に取り組んでいます。業界が今まさに重視している要素の 1 つが、オプティカル・トランスポートの消費電力の削減です。前述の仮定のシナリオのように 300,000 本を超えるプラグブル光トランシーバがあるとすると、各光トランシーバの電力の削減によってソリューション全体の消費電力削減に大きな影響が生じることはすぐにわかります。アリスタが Broadcom と共同で発明し、今では業界のマルチソース契約 (MSA) で標準となっている初期技術が、リニアドライブ・プラグブル・オプティクス (LPO) です。LPO はスイッチング・シリコンとルーティング・シリコン上で高出力・高精度の Serdes レーンを使用し、通常はプラグブル光トランシーバ上にホストされるデジタル信号処理 (DSP) 機能をホストします。DSP をプラグブル光トランシーバから取り除くことで、LPO は光トランシーバあたり 50%もの消費電力を低減できるだけでなく、光トランシーバ自体のコンポーネントを減らすことによって、遅延を短縮し、コストを削減し、信頼性を高めることもできます。

消費電力を削減するもう 1 つの方法は、ネットワーク・インフラストラクチャ製品を液冷式に変更することです。AI サーバーは、XPU あたり、サーバーあたり、ラックあたりの消費電力が非常に高いため、既に液冷式へと向かっています。液冷式は、従来の空冷式に必要なアクティブ・ファンを不要にすることで、ネットワーク・インフラストラクチャの消費電力をシステム・レベルでさらに 10%節約できる可能性があります。この節電は、大規模 AI クラスタで特に有意義です。たとえば、仮定の例では 5,120 台のスイッチが約 9MW の電力を消費していたので、10%の節電で約 900kW の消費電力を節約できます。

大規模なスケールアウト型イーサネット AI ファブリックがもたらす課題の検討に加え、高スケール AI クラスタ設計の最前線に立つ先進的な企業のお客様は、複雑で計算負荷の大きい AI ワークロード用に XPU をインターコネクトする方法を見直しています。AI の先駆者が処理しているモデルは、これまで以上に複雑になっています。また多くの場合、トレーニング用のデータセットは 1 個の XPU のメモリ内には収まらない可能性があります (新世代の XPU には、拡張された高帯域幅のメモリが引き続き追加されています)。そのため、計算の並列処理によって、複数の XPU にまたがるメモリ・バンク全体でデータセットを共有する必要があります。これにはデータ、モデル、パイプライン、またはデータ自体や演算のテンソル並列処理を含めることができます。

スケールアップ型 AI ファブリック

こうした背景から、同一データセンター・ラック内の複数の XPU 間で共有メモリにアクセスできるように設計された、「スケールアップ型」AI ファブリックという新しい概念が登場しました。スケールアップ型ファブリックは、XPU 間インターコネクトに膨大な帯域幅を必要とし、全 XPU 間に超低遅延、ロスレス・トランスポート、信頼性の高い通信も求められます。現在、1 つのラックに最大 72 個 (またはそれ以上) の XPU を収容可能であり、そのラック内の 72 個の AI アクセラレータすべての間で相互接続が必要となります。近い将来、このような「ラックスケール」設計の 1 ラックあたりの XPU 数は、72 個から 128 個、さらに 288 個にまで増える可能性も考えられます。

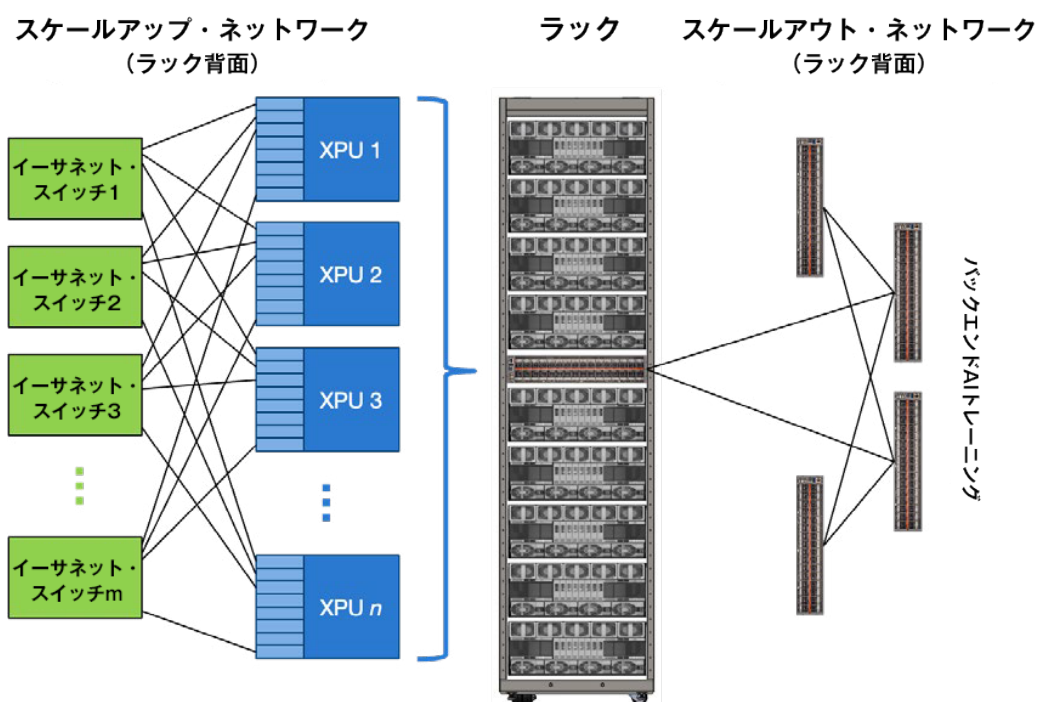


図 3: 補完的ソリューション: XPU インターコネクトを介した共有メモリ・アクセスを実現するスケールアップ型イーサネット・ファブリックと、バックエンド AI トレーニングのためのスケールアウト型イーサネット・ファブリック

現在、このような XPU 間接続は、Nvidia の NVlink など独自のインターコネクト技術で実現できます。業界内では、XPU 間接続経由で高速の共有メモリ・アクセスが可能なオープンスタンダードベースのアプローチを目指して、新たな代替アプローチが出現しています。Ultra Accelerator Link (UALink) Consortium や、Open Compute Project (OCP) の Ethernet Scale Up Networking (ESUN) などです。イーサネットがバックエンド・トレーニングでのスケールアウト AI クラスタのオープンスタンダードとして InfiniBand に取って代わりつつあるように、将来的には、ラック内の XPU 間接続用スケールアップ・ネットワークキングのオープンスタンダードとして、イーサネットが NVlink に取って代わるかもしれません。イーサネットは旧来の独自技術によって生じたサイロを打破する統合技術の役割も担い、AI センターで用いられる共通ファブリックを、スケールアップ型ファブリックによって XPU インターコネクトへと拡張します。イーサネットは、いたるところでまず InfiniBand に、次に NVlink に取って代わり、今ではスケールアップ型ファブリックへと広がるようになりました。そのため、オープンスタンダード、TCO の削減、代替性、演算の共有を特徴とする豊富なエコシステムを構築できます。

もちろん、スケールアウト型ファブリックはスケールアップ型ファブリックと共存できます。各ソリューションにはそれぞれ独自のユースケースと価値提案がありますが、互いに補い合うこともできるのです。前面と背面のあるラックを考えてみましょう。ラックの背面はラック内スケールアップ型ファブリックを担当し、高帯域幅のイーサネット・インターコネクトを使用して、XPU 間の共有メモリ・アクセスを可能にします。一方、ラックの前面では、ラックに満載されたサーバーがスケールアウト型ファブリックのリーフ・ノードとして外部と通信します。

スケールアクロス型 AI ファブリック

AI クラスタの規模が 10 万 XPU を超えて 100 万の同時ノードへと拡大を続けた結果、数十 km、数百 km、さらには数千 km にわたって地理的に分散した複数の建物を接続する「スケールアクロス型」AI ファブリックという概念が登場しました。AI クラスタがある規模に達すると、すべてのコンピューティング機器やネットワーク機器を 1 つの建物内に収容することは、物理的に不可能になります。その段階になると、WAN を介したデータセンター間相互接続 (DCI) のためにクラスタを複数の建物や拠点にスケールアクロスすることが必要となり、ネットワーク設計の前提条件や要件にも変更が必要になります。多くのお客様は、転送中のデータを保護するため、データセンターの外へ出るすべてのトラフィックを暗号化することを義務付けています。つまり、DCI 間リンクは、800 Gbps ポートごとに、MACsec または IPsec プロトコルを使用してラインレートで暗号化する必要があります。

大都市圏 (通常 40km~120km) や長距離圏 (120km 超~1,000km 以上) で利用できる光ファイバーは限られているため、WAN における AI リーフとスパインの帯域幅比を 1 対 1 に保つことは、実質的に不可能です。その結果、一般に WAN の帯域幅にはある程度のオーバーサブスクリプションが発生します。そのため、ワークロードの配置やステアリングのためのインテリジェントなルーティングが必要となり、オーバーサブスクライブされたリンクでの競合によるパケット損失を防ぐためのディープ・バッファも必要となります。GPU 側の送信をスロットリングすることで、上位層の集合通信ライブラリ (CCL) で AI ワークロードを管理し、長距離 WAN における DCI リンクのオーバーサブスクリプションを回避することは可能かもしれません。しかし、この方法は特定ベンダー固有の方式に依存することになるため、マルチベンダー、マルチ GPU の AI クラスタでは実現困難です。また、特定ベンダー独自のエコシステムに縛られることにもなるため、そのようなロックインに伴うさまざまな不利益が生じます。その結果、お客様が求めるオープンなイーサネット AI クラスタの利点の多くを損なうことになります。

そのようなシナリオでも、長距離の光の伝搬ではワークロードの送信に遅延が発生し、その遅延を処理するためのバッファ・メモリがどこかに必要となります。物理の法則と光の速度の原理からは逃れられないからです。光ファイバー・ケーブル伝送によるトラフィックの遅延は、1km あたり 5.0 マイクロ秒 (μ s) であり、これは光がファイバー・ケーブル内を 1,000km 進むのに 5 ミリ秒 (ms) かかることを意味します。ただし、これは最良の条件下での理論上のシナリオにすぎません。現実の世界でそのような長距離伝送を行うには、地理的な障害物を回避するための現実的なファイバー・ルート・マップが必要となるほか (経験則では、実際のファイバー・ケーブル長は送信元と宛先の直線距離の 1.5 倍)、アクティブ光増幅器による信号再生成も必要となり、これが追加の遅延をもたらします。通常の双方向トラフィック・フローにおけるラウンド・トリップ・タイム (RTT) の遅延は、片方向の遅延の 2 倍になります。これを踏まえると、光ファイバーで 1,000km を伝送する現実の RTT 遅延は、約 15ms になります (直線距離 1km を結ぶ実際のファイバー長は 1.5km と仮定した場合)。

必要なバッファ・メモリ量を算出するには、ビット/秒で測定されたネットワーク・トラフィック・レートをバイト数に換算し (メモリ容量はバイト単位で測定されるため)、時間単位をミリ秒から秒に統一する必要があります。次に、GBps 単位に換算したトラフィック・レートと、秒単位に換算した時間を乗算します。

- 1 バイト=8 ビットであるため、800 ギガビット(Gb)のインターフェイスは 100 ギガバイト(GB)に相当 [800 Gb * (1 バイト / 8 ビット)]
- 15 ミリ秒のトラフィックは 0.015 秒に相当 [15 ミリ秒 * (1 秒 / 1,000 ミリ秒)]
- 100 GB * 0.015 秒 = 1.5 GBps

したがって、トラフィックを 800 ギガビット/秒(Gbps)で 15 ミリ秒分処理するために必要なバッファ・メモリは、1.5 ギガバイト(GB)となります。これは、このシナリオ例の 1,000km の光伝搬に伴う遅延を処理するためだけに必要な容量です。

しかし、スケールアクロス型 AI ファブリックで異なるデータセンター間の AI ワークロードの拡張に対応するためには、ディープ・バッファ以外にも必要となるものがあります。AI ワークロード・トラフィックをローカルまたはリモートのコンピューティング・ノードへ送るために、これを適切に選択、ステアリング、監視するには、以下の機能を備えた包括的なソフトウェア・ツールキットが必要です。

- 暗号化: データセンターから送出されるすべてのトラフィックには、MACsec または IPsec プロトコルを使用して、AES-256-GCM 以上のセキュリティでワイヤースピード暗号化を行うことが必要
- スケーラブルで実績のあるクラウド規模のルーティング・スタック:
- BGP、EVPN、VXLAN、MPLS、セグメント・ルーティング(SR)、および/または SR over IPv6(SRv6) … いずれも、大規模展開に対応し、障害発生時の迅速な経路再コンバージェンスが可能な耐障害性を備えるルーティング・スタック
- 柔軟なパス選択とトラフィック・エンジニアリング: マルチテナントの顧客および XPU を分離する仮想プライベート・ネットワーク(VPN)と、トラフィック・エンジニアリング(SRv6 の SR-TE や FlexAlgo など)により、ローカルに留めるトラフィックと、DCI/WAN 経由でリモート・データセンターに送るトラフィックを識別
- 高可用性: ヒットレス・ソフトウェア・アップグレードを可能にする機能と、各デバイス内および AI ファブリック全体に組み込まれた冗長化プロトコルによるフォールト・トレランス
- 高速障害検出と経路コンバージェンス: 双方向フォワーディング検出(BFD)、トポロジ非依存ループフリー代替(TI-LFA)、高速リルーティング(FRR)、BGP プロトコル非依存コンバージェンス(PIC)
- ストリーミング・テレメトリ: ファブリック内の各ホップにおけるトラフィック・パターンと動作を詳細に把握するために必要であり、フロー、キュー、ポート、RDMA キュー・ペア(QP)、バッファ使用率、遅延(ローカルとエンドツーエンドの両方)を可視化
- 輻輳管理と負荷分散: Data Center Quantized Congestion Notification(DCQCN)、Priority-based Flow Control(PFC)、Explicit Congestion Notification(ECN)、仮想出力キュー(VoQ)、動的負荷分散(DLB)、クラスターロード・バランシング(CLB)、高速輻輳通知パケット(CNP)、パケット・トリミング、輻輳シグナリング(CSIG)
- オープン・イーサネット・スタンダード: 多様な XPU、NIC、ファブリック・スイッチに対応したマルチベンダー AI ファブリックを実現

すべてを統合: AI フロントエンドにバックエンドの AI ファブリックを接続

アリスタの Etherlink プラットフォームは、フロントエンドとバックエンドの 3 つの AI ファブリックすべてにおいて、低消費電力で大規模スケールリングを実現し、AI センター全体を支えます。

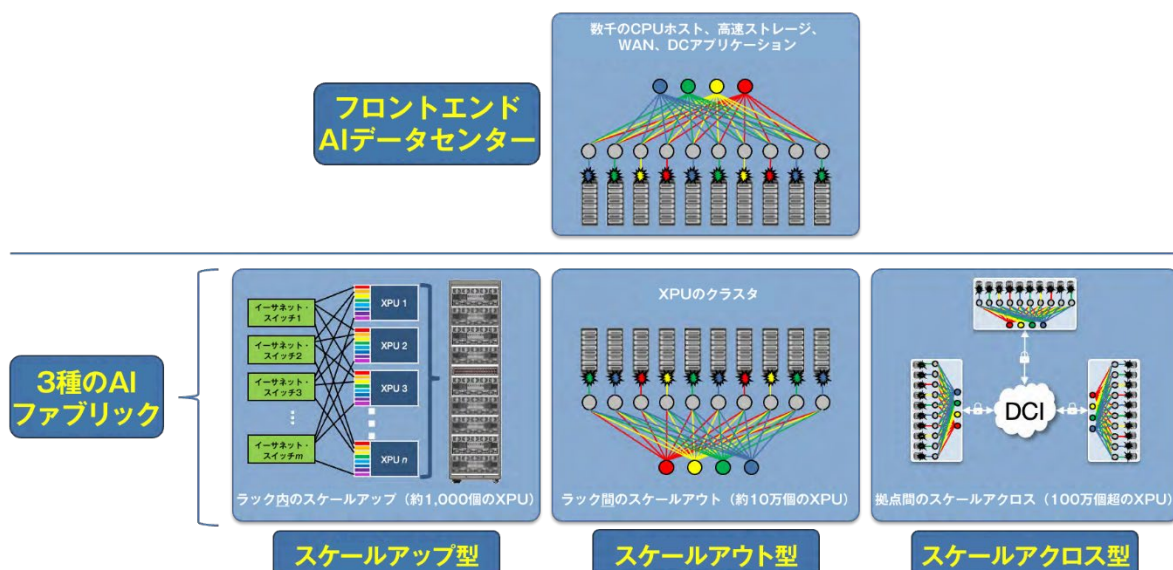


図 4: AI フロントエンドにバックエンドのスケールアップ型、スケールアウト型、スケールアクロス型 AI ファブリックを接続

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) の運営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ーや事業者が参加しています。UEC の目標は、メンバーが長年にわたり培ってきた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フル通信スタック・アーキテクチャを提供することです。UEC の標準は、オープンで相互運用可能な高性能の AI ネットワーキングを実現し、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 ワークロードのネットワーク需要増大に対応するように設計されています。

UEC は、従来の RoCE プロトコルを Ultra Ethernet Transport (UET) に置き換えることを目指して、2025 年 6 月 11 日に Ultra Ethernet 仕様のバージョン 1.0 をリリースしました。UET は、イーサネット・エコシステムの利点を維持しながら、AI アプリケーションのパフォーマンスを向上させ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、従来の InfiniBand と RDMA ベースのネットワーキングに関する数多くの重要課題に対処して、オープンなイーサネット標準を基盤とし、効率性と信頼性に優れ、高性能でセキュアな AI ネットワークのための準備を行っています。

従来の InfiniBand ネットワークと RDMA ネットワーク	最新の Ultra Ethernet ネットワーク
少数の高帯域幅のフローで厳密な順序に従ってパケット配信を行うため、利用可能なパス上での負荷分散の効率が低下	順不同でパケット配信を行うパケット・スプレーで、利用可能なすべてのパス上で非常に効率的に負荷分散
失われたパケット以降に送信されたすべてのパケットを再送信する Go-Back-N プロトコルなので、パケット損失からの復旧にコストがかかる	順不同で送信できるため、破棄されたパケットのみをシンプルに再送信
復旧に時間がかかり、障害によるテール遅延が発生	パケット損失の検出と再送信が迅速
ネットワーク・ファブリックの各階層で、運用が複雑な手作業による輻輳管理	送信者ベースの輻輳管理で、インキャスト管理には受信者ベースのクレジットを使用
ステートフル接続を開始するには時間がかかり、多くのオーバーヘッド・リソースが必要	一時的なステートレス接続により、オーバーヘッドなしでワイヤレートまで急速にスピードアップ可能
統合セキュリティなし	複数のセキュリティ・ドメインを統合して暗号化
数万ホストまで拡張	100 万の同時ホストまで拡張

図 5: 従来の RDMA ベースのネットワーキングと Ultra Ethernet ベースのネットワーキングの比較

UEC の詳細については、<http://www.ultraethernet.org/>をご覧ください。

新登場のアーキテクチャ: マルチプレーン・ファブリックと MRC

一部のお客様にとって、スケールアウト型ファブリック設計の魅力的な選択肢として浮上しているアーキテクチャの 1 つが、マルチプレーン・ファブリックです。

一部のお客様は、固定スイッチのみを使用してスケールアウト型ファブリックを構築することを選択していますが、固定スイッチはポート密度 (基数) に制限があるため、2 階層構造のリーフ/スパイン・ファブリックではスケーリングに限界があります。以前は、より大規模なシステムにスケーリングするには、3 階層 (またはそれ以上) の設計に拡張する必要がありました。3 階層 (リーフ、スパイン、スーパースパイン) にすると 1 クラスタあたりの XPU 数を増やすことはできますが、運用上の複雑さが増して非効率になり、物理的な設置面積の増大、消費電力の増加、異なる階層間を相互接続するためにファブリック内に必要な光トランシーバやケーブルの増加などの問題が生じます。

マルチプレーン・ファブリックは、ネットワークを完全に独立したファブリック「プレーン」に分割することで、水平方向のスケーラビリティを高め、3 階層アーキテクチャの非効率性を解消します。各プレーンは、それぞれ独自の並列な 2 階層の Clos ファブリックを構成しており、互いに完全に独立しています。このようなマルチプレーン・ファブリックでは、各プレーンのポート基数を (低速リンクで) 増やすことができるため、高効率の 2 階層設計においてスケーラビリティが向上し、プレーン間のフェイト・シェアリングがないため、信頼性も向上します。いずれかのプレーンで問題や障害が発生しても、残りの正常なプレーンにトラフィックを迂回させることができます。

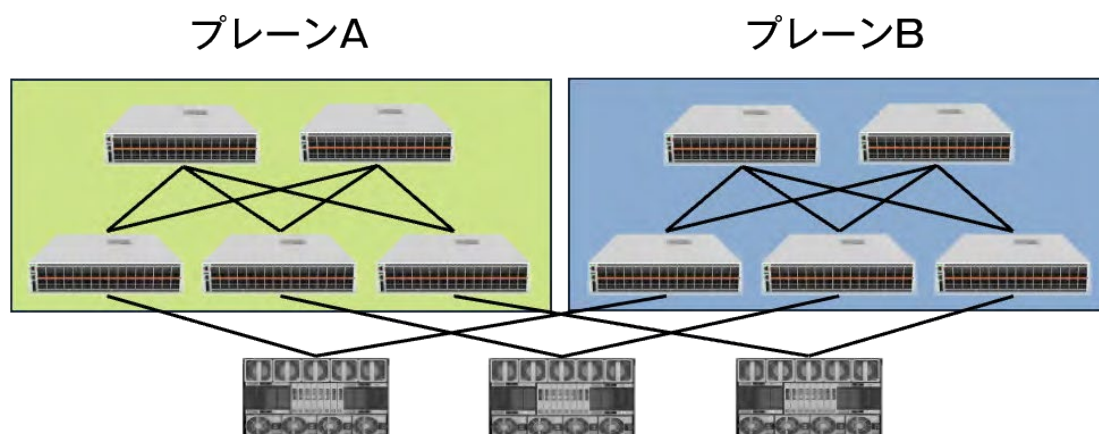


図 6: マルチプレーン・スケールアウト型 AI ファブリック

マルチプレーン・ファブリックは、各プレーンに低速の並列リンクで接続するインテリジェントな AI ネットワーク・インターフェイス・カード (NIC) によって実現されます。たとえば、800 Gbps の AI NIC は、4 本の 200G リンクを介して 4 つのプレーンに同時に接続でき、各リンクは 4 つのプレーンそれぞれの異なるスイッチに接続されます。これらの NIC は、インテリジェントなパケット・スプレーを使用して、AI ワークロードのトラフィックを各プレーン間で並列に負荷分散します。各送信元 NIC は、並列プレーン間で信頼性の高い確定的なパケット単位の負荷分散が行われるように、各プレーンの健全性と到達可能性を監視する必要があり、これにより、プレーンに障害が発生した場合にはトラフィックを再配分できるようにします。宛先 NIC は、順序が乱れたパケットを処理し、受信側の XPU に渡す前に正しい順序に再構成する役割を担います。以前は、この種のアダプティブなトラフィック・ルーティングは、各 NIC ベンダーが提供する独自ソリューションに限定されており、マルチベンダー間の相互運用性は確保できませんでした。

単一の特定ベンダー実装から脱却するため、複数のベンダーが協力し、**マルチパス信頼性接続 (MRC)** を新たなオープンスタンダードプロトコルとして OCP に提案しました。MRC は、マルチプレーン・ファブリックの高いスケール性と障害分離性を維持しながら、マルチプレーン・ファブリックを横断してパケット・スプレー対応 NIC を動作させるオープンなマルチベンダー・ソリューションを可能にします。MRC は、UEC の Ultra Ethernet Transport 仕様に含まれる、パケット・スプレー、輻輳通知、パケット・トリミングなどの技術を組み込んでいます。

MRC では、各 NIC におけるソースルーティングによるパス選択に、**Segment Routing over IPv6 (SRv6)** も利用します。これにより、MRC 対応の各 NIC は、利用可能なすべてのファブリック・プレーンにわたるすべての潜在的なパスを事前にプログラムできるようになり、ファブリック内の一時的な輻輳の通知やパス障害の発生時には、迅速なパス・リルーティングやパスの切り離しを行うことで、確定的な低遅延転送を実現します。

このように、SRv6 を採用した MRC により、マルチプレーン・ファブリックはマルチベンダーのイーサネット・ネットワークの利点を最大限に活用できるようになり、高効率の 2 階層設計において、あらゆる XPU アクセラレータ、NIC、イーサネット・スイッチを大規模に利用して、低遅延、ロスレス、高信頼のトランスポートを実現できます。

MRC の詳細については、OCP が公開しているオープン仕様 ([https://www.opencompute.org/documents/ocp-mrc-1-0-pdf\(英語\)](https://www.opencompute.org/documents/ocp-mrc-1-0-pdf(英語))) をご覧ください。

Arista Etherlink™ を搭載した AI センター

Arista Etherlink は AI センターを活用し、非常に多岐にわたる高性能なサービスを AI ネットワーキング全体に提供します。Etherlink は低消費電力で AI トレーニング・クラスタのパフォーマンスを最適化し、ネットワーク、NIC、コンピューティング、XPU リソースにまたがる AI ワークロードの協調と可視化をこれまでにないレベルで実現します。Etherlink はオープンスタンダードのイーサネットを基盤とし、すでに提供開始されていて、Ultra Ethernet Consortium (UEC) と互換性のある機能を組み込んでいます。Arista Etherlink AI プラットフォームは将来の UEC 仕様と前方互換性があり、アップグレードして準拠を維持することができます。AI センター向けの Arista Etherlink は、協調している多数の要素に展開されます。これらの要素が一つにまとまって包括的ソリューションを構成し、以下の要素を単独で利用するよりも、AI ネットワーキングの最適化を向上させます。

1. **AI プラットフォーム**: AI に最適化され、UEC と互換性があるアリスタの幅広いイーサネット・ネットワーク・システムから選択でき、わずか 32 個の XPU から 10 万個以上の並列 XPU まで、あらゆる規模の AI クラスタに対応します。これらのシステムには、Tomahawk 4 (TH4) および Tomahawk 5 (TH5) シリコンを基盤とし、400G および 800G に最適化された固定型の Arista 7060X シリーズ・スイッチや、Jericho 2C+ (J2C+) および Jericho 3-AI (J3-AI) シリコンを基盤とする仮想出力キュー (VOQ) を備え、400G および 800G に最適化されたモジュール型の Arista 7800R シリーズ・ルーターが含まれます。これらのシステムは、望ましい AI クラスタの規模に応じて、スタンドアロンまたは AI リーフ/スパイン・クラスタで展開することができます。その他に、800G に最適化された Arista 7700R シリーズ Distributed Etherlink Switch (DES) を基盤とするシングル・ホップのスケールアウト設計という展開オプションもあります。
2. **AI ソフトウェア・スイート**: 実績のある高品質な Arista EOS オペレーティング・システムで、イーサネットを経由する AI ワークロードの輸送を最適化する多様な機能を備えています。このような機能には、RDMA 対応の動的負荷分散、高度な輻輳制御、RDMA over Converged Ethernet (RoCE) をサポートするすべてのネットワーク・インターフェイス・カード (NIC) への信頼性の高いパケット配信などが含まれます。
3. **AI エージェント**: ネットワークや NIC をまたいだ協調、制御、可視化を行い、コンピューティング環境とネットワーク環境のパフォーマンス最適化や一元管理を実現します。将来的には、SmartNIC 上やサーバー上に EOS ベースの AI エージェントを展開できるようになります。
4. **AI 可観測性**: ストリーミング・テレメトリを取り込むためのアリスタのネットワーク・データレイク (NetDL)、アリスタの CloudVision ネットワーク自動化ソフトウェア、各システム上の EOS に組み込まれたアリスタのレイテンシ、コレクティブ、AI アナライザ機能を組み合わせ、AI クラスタ全体の AI ワークロードのパフォーマンスについて、詳細なインサイトと可視化を提供することを目的としています。

Arista Etherlink: かつてないパフォーマンス、拡張性、耐障害性

Arista の Etherlink AI ソリューションは、システム、ソフトウェア、AI エージェントを協調させて、これまでに類を見ないインサイトと制御を提供し、エンドツーエンドのパフォーマンスを最適化します。主な利点は次のとおりです。

- **パフォーマンスの最適化**: Etherlink が提供する AI ソリューションは高性能で、最適化されていない従来のイーサネット・システムのパフォーマンスに比べ、最大 65% パフォーマンスが向上します。このようなパフォーマンスの向上は、ロスレスで低遅延のプラットフォームと、AI ジョブ完了時間 (JCT) を最小限に抑える革新的な RDMA 対応負荷分散機能と輻輳回避機能を組み合わせることによって実現されます。

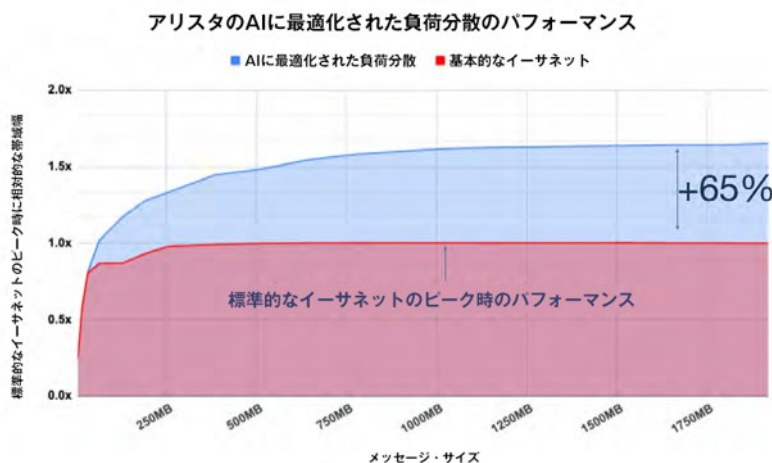


図 7: Arista Etherlink AI プラットフォームのパフォーマンスと従来のイーサネット・システムの比較

パフォーマンスはいくつかの方法で特徴付けることができます。一般的に最も重要なのは、システムのあらゆる要素が正常に動作している最良の条件でピーク時のパフォーマンスを発揮できるかどうかですが、実際には故障が発生します。何千ものシステム、NIC、XPU がすべて脆弱な光ケーブルでインターコネクトされている大規模 AI クラスタでは、特にその可能性が高まります。どのような AI クラスタも、システム全体の中のいずれかのコンポーネントが故障すればパフォーマンスが直接影響を受けるため、ピーク時のパフォーマンスの測定には、故障の発見からクラスタが最適な動作に戻るまでの効率も使用する必要があります。アリスタの Etherlink はこの点でも優れており、障害復旧は InfiniBand の最大 30 倍の高速コンバージェンスです。

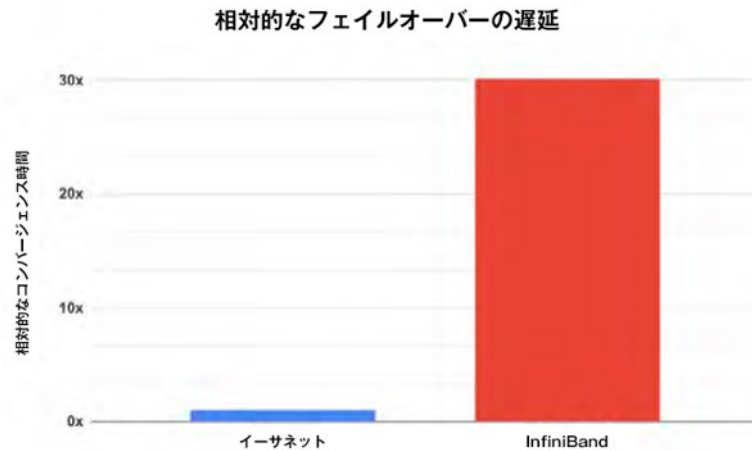


図 8: Arista Etherlink AI プラットフォームの障害コンバージェンスの効率と InfiniBand の比較

- **最大限の柔軟性と選択肢:** Etherlink はあらゆる規模の AI ワークロードに対応し、オープンスタンダードに完全に準拠していて、どの XPU アクセラレータ、NIC、ワークロードとも相互運用可能です。すべての Etherlink システムは、Ultra Ethernet Consortium (UEC) の新しい機能強化と完全な前方互換性があります。さらに、Etherlink は、1 台の固定型システムまたはモジュール型シャーシを使用する単一階層設計、リーフまたはスパインの役割の固定型およびモジュール型システムを使用して柔軟に設計できる多階層のリーフ/スパインまたは平面設計、革新的な新しいシングル・ホップの Distributed Etherlink Switch (DES) システムで展開することができます。
- **大規模に展開されているイーサネット:** Etherlink が提供する選択肢は柔軟で非常に幅広く、固定型とモジュール型のスタンドアロン・システム、シングル・ホップの Distributed Etherlink Switch、固定型とモジュール型のシステムを選択できる大規模なリーフ/スパインおよび平面設計など、市場で最も拡張性が高くなっています。
 - » 単一システム: 1 台のモジュール型 7800R4 シリーズ・システムに最大 576 の 800G ポートまたは 1,152 の 400G ポート、または 1 台の固定型 7060X6 シリーズ・システムに最大 64 の 800G ポートまたは 128 の 400G ポートを搭載。
 - » シングル・ホップの Distributed Etherlink Switch: 7700R4 シリーズ・システムに最大 16,000 の 800G ポートまたは 32,000 の 400G ポートを搭載。
 - » 2 階層および 3 階層のリーフ/スパインまたは平面トポロジ: リーフおよびスパインの役割のモジュール型 7800R4 シリーズ・システムに 10 万個を超える XPU を搭載。小規模な場合は、固定型 7060X6 シリーズ・システムをリーフやスパインとして並列で使用することもできる。
- **最適な消費電力:** Etherlink は、システム・レベルと光インターコネクト・レベルの両方で、AI クラスタ全体の消費電力を大幅に削減することができます。Arista の Etherlink システムでは、業界最高水準の 5nm シリコンを利用して、7nm またはそれ以前の世代のプラットフォームに比べ、Gb あたりの消費電力を 25% 以上削減しています。また、リニアドライブ・プラガブル・オプティクス (LPO) を使用すると、従来のプラガブル光トランシーバとアクティブ光ケーブル (AOC) に比べ、正味消費電力をさらに 50% 削減できます。個別に見ると、これらのソリューションそれぞれが消費電力を削減できますが、Etherlink はそれらを組み合わせて AI ワークロード全体の消費電力を最小化し、貴重なリソースを解放してコンピューティング容量を増やしています。
- **持続可能な品質:** Etherlink が基盤とする EOS オペレーティング・システムは高品質で、新しい AI ネットワークを非常に迅速に取り入れて、トレーニング時間を短縮します。EOS の共通脆弱性識別子 (CVE) と不具合は、市場の代替オペレーティング・システム、特にオープン・ソースのオペレーティング・システムに比べ、明らかに少なくなっています。
- **制御や可視化の協調:** Etherlink は重要なエンドツーエンドのインサイトと多岐にわたる制御をネットワーク・チームに提供し、ネットワークからコンピューティング・ノード内の NIC まで、パフォーマンスを最適化できるようにしています。多くの場合、ネットワーク・チームとサーバー・チームが確認できるのは AI ソリューション全体のパフォーマンスの一部のみで、各チームはそれぞれの分野の設定を別々に管理しています。このため、NIC とネットワークの QoS 設定で不注意による不一致が発生し、AI クラスタ全体にパフォーマンスの問題が生じる可能性があります。Arista の EOS ベースの AI エージェントは包括的な Etherlink ソリューションの重要な要素であり、ネットワーク分野とコンピューティング分野で設定を協調させ、一元的に可視化することによって、そのような問題を回避してエンドツーエンドの AI パフォーマンスを最適化できるようにします。

- **UEC 対応:** Etherlink は実績ある標準ベースのイーサネットを以前から利用していますが、Ultra Ethernet Consortium(UEC)の 1.0 Ultra Ethernet 仕様にも準拠しています。イーサネットでは、InfiniBand などの代替技術に関連する数多くの重要な利点を活用して、豊富なサプライヤのエコシステムから大量展開を行うことで、光トランシーバとケーブルを含めてコスト優位性をもたせたり、長年にわたる技術革新で新しいユースケースに適した機能を導入したりできます。現在、Etherlink の AI ワークロードのパフォーマンスは、InfiniBand と比べて最大 10% 高くなっています。また、イーサネットのパフォーマンスの強みをさらに伸ばすため、新しい UEC 仕様へのアップグレードに対応中です。

NCCLのAll-to-Allのパフォーマンス

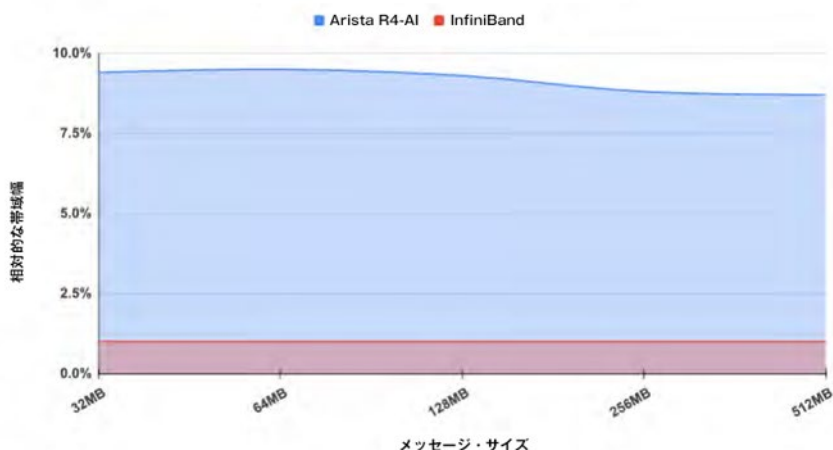


図 9: AI に関する Arista Etherlink AI プラットフォームのパフォーマンスと InfiniBand のパフォーマンスの比較

Arista Etherlink AI 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。アリスタのネットワークは、業界最高水準のシリコン・パケット・プロセッサを利用して、あらゆる規模の AI ネットワークに最適化された幅広いハードウェア・システムを提供します。アリスタの固定型とモジュール型のシステムは、スタンドアロンおよびスパイン/リーフの AI トポロジでも、シングル・ホップの AI ネットワーク向けの Distributed Etherlink Switch でも使用できます。

Arista Etherlink AI ポートフォリオ

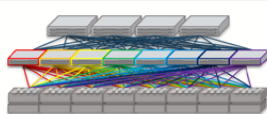
単一スイッチ	リーフ/スパイン	Distributed Etherlink Switch
 固定型とモジュール型 51.2T ➡ 460.8T 64~576の800Gノード 128~1152の400Gノード 低コスト、低消費電力、低複雑性	 2階層または3階層のリーフ/スパインまたは平面構造 マルチペタビットの二分帯域幅 DLB、PFC、ECN、数万ノード	 シングル・ホップ・インターコネクト 完全にスケジューリングされたロスレス 100%効率、セル・スプレー マルチペタビットの二分帯域幅 数万ノード

図 10: Arista Etherlink AI ポートフォリオ

アリスタは、Etherlink ポートフォリオの一部として、7060X シリーズ固定型プラットフォーム、7800R モジュール型シャーシ、7700R Distributed Etherlink Switch という AI に最適化された幅広いプラットフォームを提供しています。

7060X シリーズ: 固定型 AI リーフ

7060X シリーズは、固定型のフォーム・ファクタで大容量かつ低遅延のイーサネット・スイッチングを提供し、大規模な AI クラスタにおけるリーフとしての利用に最適であり、固定構成のみの設計のスパンインとしても使用可能です。最新の 7060XE7 シリーズは、お客様の導入ニーズに応じて、100% 液冷および/または空冷のフォーム・ファクタで、102.4T のスイッチング容量を提供します。

7060XE7 シリーズには大きく 2 つのバリエーションがあり、それぞれ異なるポート速度に最適化された異なるシリコン基盤を採用しています。1 つ目のオプションは、200 Gbps の Serdes レーンを備えた 3nm Tomahawk 6 シリコン (TH6-200) を基盤としており、システムあたり 64 ポートの 1.6TbE ポートを提供します。もう 1 つのオプションは、100 Gbps の Serdes レーンを備えた 3nm Tomahawk 6 (TH6-100) を基盤としており、システムあたり 128 ポートの 800GbE ポートを提供します。すべてのシステムは、低速リンクで高密度化を実現する高ポート基数構成をサポートしており、高密度 800G (1.6T ポートのみ)、または 400G、200G、100G のポート・オプションにブレークアウト可能です。



図 11: Arista Etherlink AI プラットフォーム: 7060XE7 固定型 102.4T

7060XE7 シリーズは、AI ネットワーキングの消費電力の削減を飛躍的に進歩させています。前世代に比べ、容量 1Gbps あたりの消費電力を 25% 削減するだけでなく、リニアドライブ・プラグブル・オプティクス (LPO) もサポートし、従来のフルリタイミング光トランシーバに比べてさらに 50% 正味消費電力を削減することができます。100% 液冷モデルは、低消費電力と Open Rack Version 3 (ORv3) ラック向けの小型フォーム・ファクタ (2OU) を兼ね備えており、大規模なラックスケール展開を可能にするだけでなく、液冷式 XPU コンピューティング・サーバーと同じラック内に設置することもできます。

補完的な選択肢である 7060X6 シリーズは、2RU または 1RU の固定型フォーム・ファクタに最大 64 の 800G ポートまたは 128 の 400G ポートを備え、最大 51.2T の容量を提供します。

7800R シリーズ: モジュール型 AI スパンイン

Arista 7800R シリーズのモジュール型システムは最大容量 460 Tbps を提供し、最も要求の厳しい AI ワークロードのニーズや、最新のマルチサービス・ルーティングおよびスイッチングのユースケースに対応します。7800R システムは、4 スロット、8 スロット、12 スロット、16 スロットのモジュール型シャーシ・オプションで利用できます。16 スロットのシャーシは最大 576 の 800G ポートまたは 1,152 の 400G ポートをサポートし、中小規模の AI ワークロード向けとして、または大規模 AI スパンイン/リーフ・クラスタのスパンインとして、単一システムの密度を最大化します。

7800R シリーズは 100% 公平でセルベースのスイッチング・ファブリックを基盤とし、効率と組み込みのファブリック・スピードアップにより、インGRESS・ネットワーク・ポートとノンブロッキング・アーキテクチャ全体のオーバープロビジョニングを統合します。AI ワークロードのパフォーマンスは、フォワーディング・パイプラインとファブリックに組み込まれたロスレスで完全にスケジューリングされた仮想出力キュー (VOQ) の設計によって最適化されます。ディープ・パケット・バッファと組み合わせることにより、輻輳とパケット損失を回避できます。

7800R4 の AI に最適化されたラインカードは、Jericho 3 AI シリコン・パケット・プロセッサのペアを基盤とし、36 の 800G OSFP ポートを備えて容量 28.8 Tbps を提供します。このラインカードは、AI ワークロードのニーズに特化した低遅延パケット・パイプラインを提供し、VOQ と 32G のディープ・パケット・バッファリングを組み合わせ、AI 向けに最適化された機能セットとスケーラビリティを提供しています。その結果、7800R4 の AI ラインカードは極めて低い消費電力を実現しており、低消費電力のリニアドライブ・プラグブル・オプティクス (LPO) によって、さらなる削減が可能です。

7800R は、400G に最適化された 7800R3A ラインカードもサポートしています。このラインカードは Jericho 2C+シリコン・パケット・プロセッサのペアを基盤とし、36 の 400G OSFP ポートまたは QSFP-DD ポートを備えて容量 14.4 Tbps を提供します。



図 11: Arista Etherlink AI プラットフォーム: 7060X シリーズ、7800R シリーズ、7700R シリーズ

7700R4: Distributed Etherlink Switch

Arista 7700R Distributed Etherlink Switch (DES) は AI 中心のネットワーキングにおける新しいパラダイムの代表的なもので、リーフとスパインの拡張性を実現しつつ、単一の論理システムとして動作し、コンピューティング・クラスタへのシングル・ネットワーク・ホップとなります。7800R4 モジュール型シャーシは中間スイッチ・ファブリックでラインカード同士を接続しますが、7700R4 DES は分散型と固定型のリーフ・プラットフォームを提供して、専用のスイッチ・ファブリック・システム経由で互いにトラフィックを転送します。7800R4 も 7700R4 も、VOQ、ディープ・パケット・バッファリング、ロスレスでセルベースのファブリックで共通のアーキテクチャを共有しています。

7700R4 DES システムは、従来のリーフ/スパイン・トポロジと同じように見えるかもしれませんが、実際、ケーブルのインターコネクトは同じです。しかし、イングレスとイグレス両方のリーフ・ノードとスパインのすべてが独自のフォワーディング・プロセッサを持つために 3 ホップ・フォワーディングを必要とするリーフ/スパイン設計と異なり、7700R4 はシングル・ホップ・フォワーディングのパラダイムに対応しています。その結果、7700R4 DES システム全体が完全にスケジューリングされた単一の論理クラスタとして管理され、システム内のすべてのノード間で 100% 公平なロスレス・トランスポートを提供します。DES は導入直後から 100% 効率に自動調整されるため、PFC/ECN やその他の RDMA 対応負荷分散機能など、通常は DES ファブリック内部の AI ネットワークに必要な QoS 機能が不要です。

DES システム内の固定型リーフ・ノードそれぞれが、AI アクセラレータ接続用の 18 の 800G ポートに加え、一元化されたスイッチ・ファブリックへのアップリング専用 20 の 800G ポートを提供します。各ファブリックが 128 の 800G ポート接続をホストして、複数のリーフ・ノードを集約します。したがって、DES システムは 100ms での自動検出とクラスタ内のリンクやシステムの障害を回避するリルーティングを利用して、ファブリックのオーバープロビジョニングによる冗長性とオーバースピードによるトラフィック管理を自動的に提供します。各 DES リーフ・ルーターは Jericho 3-AI シリコン・パケット・プロセッサを基盤とし、専用の VOQ とディープ・パケット・バッファを備え、7800R4 モジュール型シャーシの J3-AI ラインカードと同様の低遅延、低消費電力という特徴があります。また、J3-AI ラインカードと同様に、7700R4 DES システムはリニアドライブ・ブラガブル・オプティクス (LPO) をサポートし、従来のブラガブル光トランシーバに比べ 50% の消費電力を節約します。

Arista EOS の AI ソフトウェア・スイート

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。Arista EOS® (Extensible Operating System) は、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。

AI ワークロードに求められる第 1 の要件は、無停止可用性です。AI ファブリック自体が不安定だったり、停止したりするようでは、ジョブ完了時間(JCT)の最適化に注力しても意味がありません。そのため、Arista EOS は、転送中のワークロード・トラフィックに影響を与えることなく、Etherlink プラットフォーム上のソフトウェアをプロアクティブにアップグレードできるようにするスマート・ソフトウェア・アップグレード(SSU)を提供します。AI ファブリックが常に稼働し続け、最新のセキュリティ・パッチや運用パッチが常に適用されることを確保するヒットレス SSU は、高速 JCT を実現する重要な要素の 1 つです。

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管理ソリューションを提供して RDMA over Converged Ethernet (RoCEv2)をサポートします。また、動的負荷分散(DLB)を提供して等価コスト・マルチパス(ECMP)リンク経由での AI フローの負荷分散をサポートし、リンクあたりの効率が最大で約 90%の帯域幅使用率になります。

けれども、DLB では各リーフから複数のスパイン・ノードへの複数の並列リンクを考慮してリーフからスパインの負荷分散を効果的に行うことができますが、通常は特定のスパインから特定のリーフへの 1 つのリンクしかないため、リバース・パスでうまく機能しません。1 つのパスだけでは、スパイン・レイヤからリーフへの異なる AI フローを効果的に負荷分散することができないのです。AI クラスタ・ファブリックの双方向(リーフからスパインとスパインからリーフ)で AI フローの負荷分散を行うには、帯域幅の効率も重要です。AI クラスタは集散的処理を基盤としており、クラスタに含まれるすべての XPU に結果を分散させる必要があるからです。

アリストは EOS でこれを解決します。アリストが発明した新機能のクラスタ・ロード・バランシング(CLB)は、グローバルな視点を提供して両方のパスで同時に負荷分散を行えるようにするとともに、各パケットを詳細に調べ、RDMA キューのペアを負荷分散アルゴリズムへの追加入力として検査します。この新機能の効果によって AI ワークロードのパフォーマンスが高まるため、帯域幅の効率が 98%以上に向上し(DLB では 90%)、クラスタ内のすべての XPU のテール遅延が均一に低く抑えられます。

通常、AI クラスタには、少量の超高帯域幅フローが存在します。

そのため、従来の負荷分散オプションは多くの場合非効率的で、フロー間でパフォーマンスに一貫性がなく、テール遅延が高くなっていました。

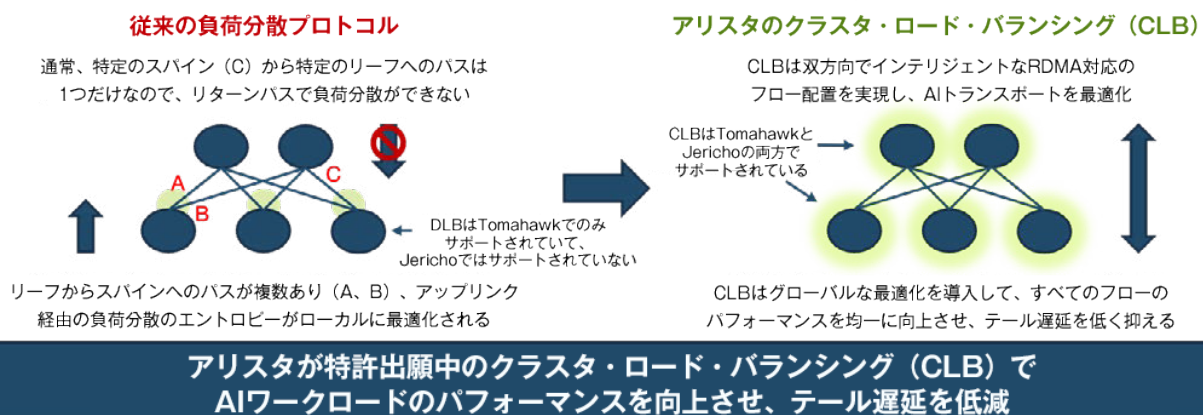


図 12: アリストのクラスタ・ロード・バランシング(CLB)機能の概要

アリストは、AI フローの可観測性のための強力なソフトウェア・スイートも提供しています。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。EOS は、**AI アナライザ機能**と**レイテンシ・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・パターンを詳細に可視化します。

AI アナライザは、パフォーマンスの正確な最適化とトラブルシューティングのために、100 マイクロ秒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・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・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。

XPU 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。

Arista EOS は、スケールアウト型、スケールアップ型、スケールアクロス型の各種 AI ネットワーク・モデルにおいてトラフィックを分析できる強力なソフトウェア・スイートを提供します。スケールアクロス型アーキテクチャの採用が進むにつれ、相互接続された AI センターを管理する課題を克服し、XPU as a Service (XPUaaS) を提供するためには、可視化と監視が極めて重要になります。EOS は、フロー、パス、ワークロードのトラフィック・パターンをエンドツーエンドで分析するための基盤として、Secure Traffic Analyzer (STANZ) という強力な機能を提供します。これには、EOS AI エージェントが提供する XPU および CPU の統計情報を使用してホストから行うテナントまたはワークロードのトラフィック分析や、ネットワーク側が提供する、パケット、フロー、インターフェイス、パス (ECMP、トンネルなど)、プロトコル、そしてテナント単位 (VLAN、VRF、L3VPN、EVPN VNI、VTEP など) の各種テレメトリ情報の解析が含まれます。これにより、お客様は不要な情報を排除して重大な問題を解決し、ユーザー・エクスペリエンスへの影響を最小限に抑えることができます。CloudVision AI にエクスポートされたテレメトリ・データにより、オペレーターは障害データを迅速に収集し、障害の発生箇所を追跡し、障害の修復や復旧までの平均時間を短縮できる重要な指標を提供することができます。

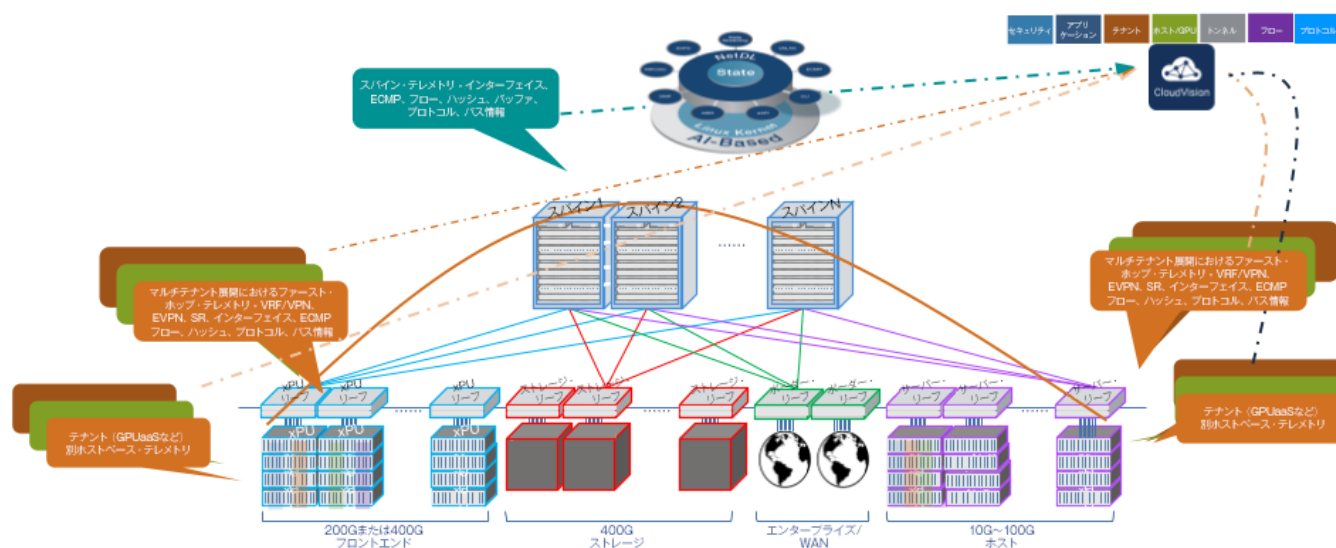


図 13: Arista EOS ベースの AI エージェントと一元的な可視化

Arista EOS ベースの AI エージェント

ジョブの完了時間を最小化し、それによって高価な XPU アクセラレータの利用率を最大化するため、AI ワークロードのパフォーマンスと可用性を常に最適化する必要があります。問題が発生した場合、ネットワーク運用チームとサーバー運用チーム間の誤設定や責任の押し付け合いは決して許されません。

アリスタはネットワーキング分野とコンピューティング分野の間でパフォーマンスの最適化を協調させるとともに、エコシステム全体を一元的に可視化して、最適化または改善する領域を特定します。EOS ベースの AI エージェントを SmartNIC 上に直接、またはサーバー CPU 上に配置して、NIC のローカルな構成管理を提供したり、NIC のパフォーマンスのストリーミング・テレメトリを直接接続の Arista EOS ベース・スイッチに送ったりすることができます。これにより、AI 最適化のための QoS パラメータを NIC からネットワークまで不一致なく適用して、根本原因を簡単に診断できないパフォーマンス・ボトルネックを発生させる可能性がある誤設定を回避できます。そして、AI NIC から AI ネットワーキング・プラットフォームまでのテレメトリ・データを使用して、ネットワーク運用チームがトラフィック・パス全体を包括的に可視化し、パフォーマンスと問題に関するインサイトを迅速に獲得することができます。

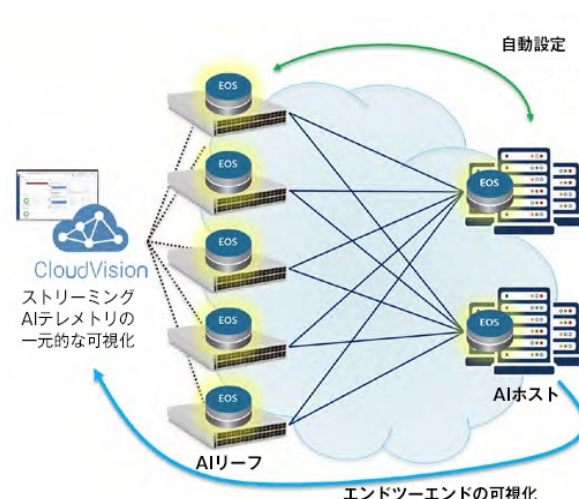


図 14: Arista EOS ベースの AI エージェントと一元的な可視化

アリスタの AI 可観測性

はじめに AI エコシステムのあらゆる要素からストリーミング・テレメトリによって収集されたリアルタイム・データを照会して洞察を得て、AI ネットワークの状態を把握しなければ、その AI ネットワークを最適化することは困難です。わかっていることが増えれば増えるほど、より適切な判断を下し、ワークロードとインフラストラクチャ全体を連動させて、さらに最適化することができます。

アリスタの AI 可観測性はオープンスタンダードに準拠して設計され、ネットワーク運用に幅広い情報を提供することを目的として、適切なインサイトを獲得するために最適なデータをリアルタイムで共有します。このソリューションには次のものが含まれます。

- ボックス型、モジュール型、DES プラットフォームなど Arista EOS ベースの AI プラットフォーム、および SmartNIC 上やサーバー CPU 上にホストされている EOS ベースの AI エージェント経由のコンピューティング・ノードからのストリーミング・データ。
- 全管理対象デバイスからのすべてのストリーミング・データを一元化するリポジトリを提供する Arista NetDL ネットワーク・データレイク。NetDL は、特にパフォーマンス・モニタリングの分析やフォレンジック・テストを行うための信頼できる唯一のデータ・ソースとなります。
- EOS に搭載された Arista の AI アナライザ機能とレイテンシ・アナライザ機能により、トラフィック統計、インターフェイスの輻輳、キューイングのレイテンシ・メトリクスをマイクロ秒間隔でリアルタイムに詳細に把握し、AI ワークロードの動作とネットワーク特性を相関させることができます。
- Arista Cloudvision ネットワーク自動化プラットフォームで、ネットワークとコンピューティングのトポロジ、健全性、設定を可視化します。
- アリスタの Autonomous Virtual Assistant (AVA) を Cloudvision や NetDL に保存されているデータと連携させて、ネットワーク・オペレーターに会話型のアシスタントを提供します。これにより、自然言語で質問して、エンドツーエンドの AI エコシステムの状態を問い合わせたり、潜在的な問題の原因に関するインサイトを収集したりすることができます。

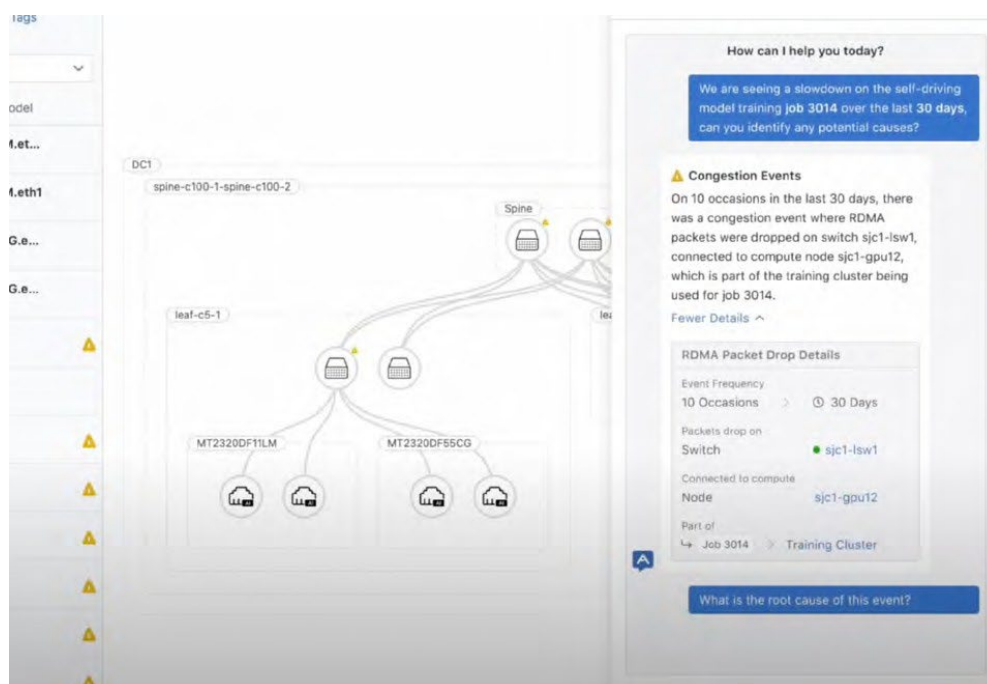


図 15: Arista AVA による NIC からスイッチまでの AI 輻輳の問題の診断の表示

大規模な AI センターの設計オプション

近年では、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 East-West および North-South 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、極端な「East-West」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 XPU、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。

AI XPU サイズ	サーバー I/O 数十個の XPU	ラック規模 数百個の XPU	データセンター規模 1 万個以上の XPU
	AI リーフ 固定型 イーサネット	AI スパイン モジュール型 イーサネット	AI リーフ + スパイン 固定型、 モジュール型、 DES
AI ネットワークの オプション	小規模 AI アプリ	中規模 AI アプリ	大規模 AI アプリ

図 16: AI ネットワークの設計ガイドライン

小規模 AI センター: 固定型 AI プラットフォーム 1 台で数十個の XPU 向け

64 の 800G ポートまたは 128 の 400G ポートを備えた Arista 7060X6-64PE スイッチ 1 台で、数ラックの XPU 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 XPU が他のすべての XPU 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。

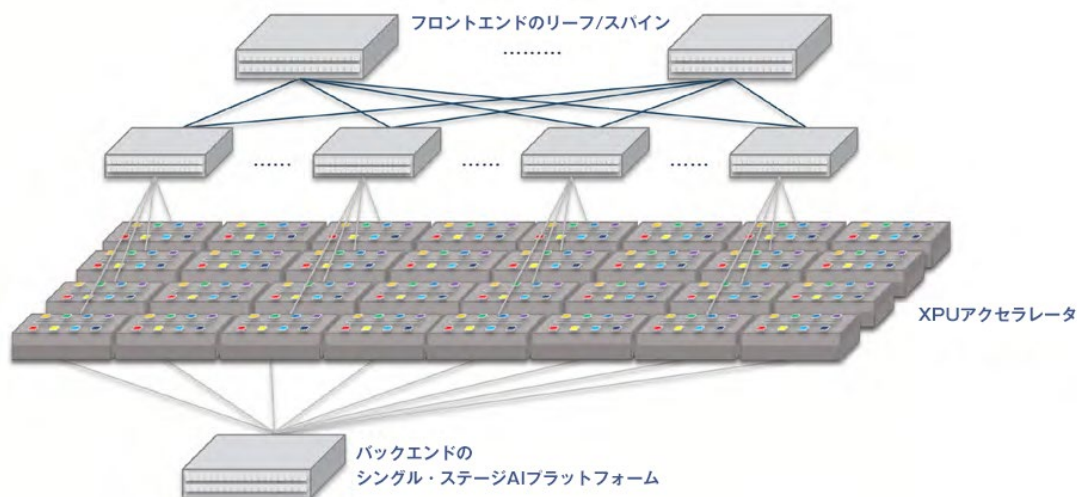


図 17: 小規模 AI センター、Arista 7060X6 固定型 AI プラットフォーム 1 台を基盤とするシングル・ステージ

中規模 AI センター: モジュール型 AI プラットフォーム 1 台で数百個の XPU 向け

576 の 800G ポートまたは 1,152 の 400G ポートをサポートする Arista 7800R4 シャーシ 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、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。7800R4 はセルベースのノンブロッキング VOQ アーキテクチャを基盤としているため、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、XPU は常にラインレートでデータを送受信できます。

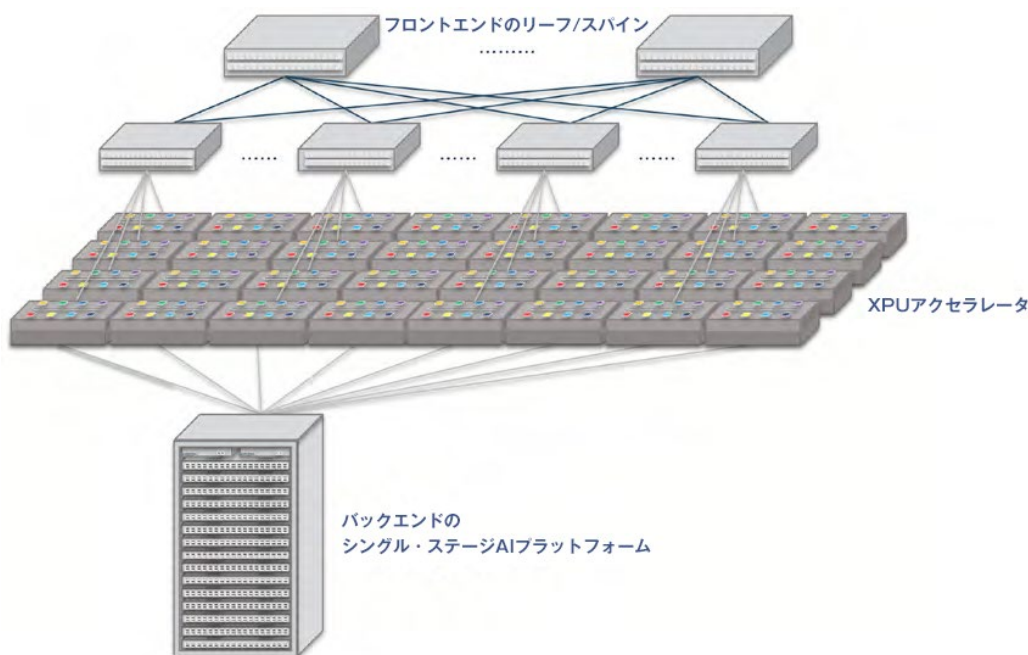


図 18: 中規模 AI センター、Arista 7800R モジュール型シャーシ 1 台を基盤とするシングル・ステージ

大規模 AI センター: シングル・ホップの Distributed Etherlink Switch AI プラットフォームで数千個の XPU 向け

パフォーマンスを犠牲にすることなく、32,000 個の XPU アクセラレータで構成される大規模アグリゲーションとシンプルな運用を組み合わせる新しい AI 設計オプションが登場しました。それが Distributed Etherlink Switch (DES) です。この革新的な設計は、従来のリーフ/スパイン・ネットワークの拡張性を提供しますが、単一の論理クラスタとして管理され、接続しているホストへのシングル・ホップ・パスとなります。DES はユニバーサル 800G リンクに最適化されています。これには、一時的な帯域幅のバーストを管理する固有のファブリック・オーバープロビジョニングなど、さまざまな光トランシーバを使用したポートあたり 2 個の 400G ホスト接続のサポートオプションが含まれます。DES は単一のクラスタとして管理されるため、フォワーディングの判断はすべてリーフ・レイヤで行われ、セルベースのファブリックが自動的に 100% 公平なロスレス・トランスポートや 100ms での障害検出および復旧を提供します。

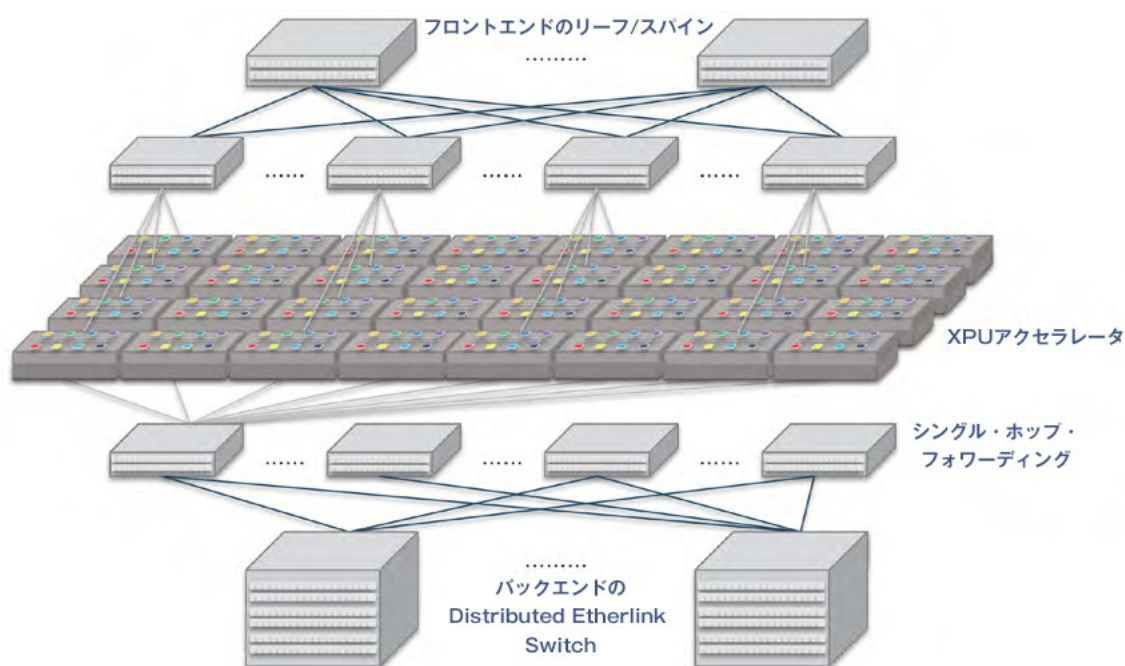


図 19: 大規模 AI センター、7700R4 Distributed Etherlink Switch を使用したシングル・ホップ設計

超大規模 AI センター: マルチステージのリーフ/スパイン AI プラットフォームで 10 万個以上の XPU 向け

データセンターの何万個もの XPU を接続したり、10 万個もの並列 XPU を接続したりする必要がある超大規模 AI アプリケーションの場合、スケールアウト型リーフ/スパイン設計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、100,000 以上のエンド・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・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、XPU 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。

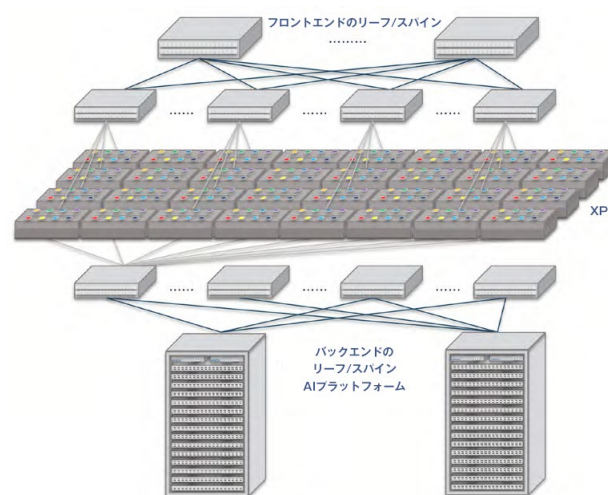


図 20: 超大規模 AI センター、Arista 7800R AI スパイン + 7060X AI リーフのマルチステージ

任意の AI クラスタに必要な XPU の数に応じて、AI リーフの選択肢は固定型 AI プラットフォームから大容量のモジュール型シャーシまで多岐にわたります。AI リーフとスパインの役割のどちらでも、モジュール型プラットフォームを使用すると XPU 密度が最大になります。

ユニバーサル・リーフ/スパイン設計は、現在数百の XPU 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 XPU 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。

まとめ

アリスタの Etherlink AI プラットフォームが提供する包括的ソリューションは、個々の要素を単独で利用するよりも、協調して AI ネットワーキングの最適化を向上させます。Etherlink は、AI に最適化されたイーサネット・ハードウェア・システム、業界をリードする AI 輻輳回避機能を備えた EOS ベースのソフトウェア、ネットワーキングと NIC を協調させる EOS ベースの AI エージェント、AI インフラストラクチャのエンドツーエンドの可観測性など最高水準の製品で構成されています。Etherlink は、現時点で高性能 AI クラスターの構築に最適な選択肢であるイーサネットを基盤としており、UEC の新しい機能強化を取り入れて進化し続けます。イーサネットの AI ワークロードのパフォーマンスは、次第に InfiniBand よりも高くなります。アリスタの Etherlink AI プラットフォームは優れた障害復旧機能を備え、基本的なイーサネット・ソリューションに比べてパフォーマンスを最大 65%向上させます。

AI センターは、ネットワーキング・エコシステム全体を統合し、新しい AI バックエンド・ネットワークと既存のデータセンター・システムやインフラストラクチャを結び付けることができます。イーサネット技術との一貫性を維持しているため、組織全体でスキル・セットを協調させ、運用を一致させることができます。そして、ネットワークと運用を一元管理することで、既存データセンターの他のシステムとの連携、協調的な可観測性を使用したネットワークからコンピューティング・ノードまでの新しいインサイトの収集、AI のパフォーマンス最適化が簡単になります。

リファレンス

ブログ: [オープンな AI ネットワーキングの到来](#)、Jayshree Ullal

ブログ: [AI の新時代: ネットワーキングのための AI、AI のためのネットワーキング](#)、Jayshree Ullal

ブログ: [AI センターの新時代](#)、Jayshree Ullal

ブログ: [すべてのイーサネット AI ネットワーキングの強化](#)、Vijay Vusirikala、John Peach

プレスリリース: [Ultra Ethernet Consortium \(UEC\) Launches Specification 1.0 Transforming Ethernet for AI and HPC at Scale](#)、2025 年 6 月 11 日、Ultra Ethernet Consortium (UEC)

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

Copyright © 2026 Arista Networks, Inc. All rights reserved.
CloudVision、EOS は、Arista Networks, Inc. の登録商標です。Arista Networks は Arista Networks, Inc. の商標です。その他の企業名はすべて、それぞれの所有者の商標です。

本書に記載されている情報は、予告なく変更される場合があります。

一部の機能は、まだ利用できない場合があります。

Arista Networks, Inc. は、本書に記載された誤りについて一切の責任を負いません。

www.arista.com/jp

ARISTA

2026 年 6 月 9 日 02-0097-05