Oriole

PRISM：世界初の AI 用フルフォトニックネットワーク

コンピューティングは高速でもネットワークが足を引っ張っている可能性がある

今日のハイパフォーマンスコンピューティング（HPC）システムと分散ディープラーニング（DDL）システムでは、パフォーマンスを低下させているのはプロセッサではなく、メモリとネットワークです。ハードウェアのコンピューティング能力が飛躍的に向上しているにもかかわらず、ネットワークのボトルネックのためにピークパフォーマンスのごく一部しか達成できず、現実世界のパフォーマンスはしばしば壁にぶつかります。

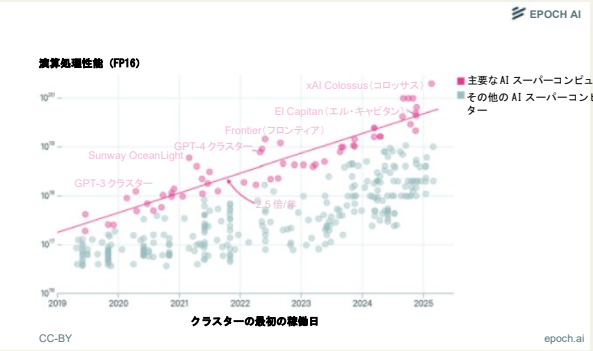


図 1：主要な AI スーパーコンピューターの性能は 9 カ月ごとに倍増

ネットワーク関連の遅延は、128 ノード程度の比較的小規模なセットアップでも、ディープニューラルネットワーク（DNN）の総トレーニング時間の最大 40～60%を占める可能性があります。また、ニューラルネットワークのパラメータ数は数カ月ごとに倍増しているため、相互接続への圧力は高まる一方です。そして、こうしたネットワークオーバーヘッドの大きさは、スケーリングの方法によって大きく異なってきます。

弱スケーリングでは、ワーカーを追加することで、通常はデータの並列性によってスループットを向上させ、規模拡大時のコンピューティング時間と通信時間の安定も保たれます。そのため、現状のネットワークの制約にうまく対応でき人気があります。また、バッチサイズが大きい限り、適度な帯域幅（Gbps レンジ）で負荷を処理できます。しかし、非常に大規模なモデルでは、トレーニングの効率が常に向上するわけではなく、メモリに対する要求が低下するわけでもないため、壁にぶつかります。

一方、強スケーリングでは、モデルの並列性を利用してモデル自体をスライスすることで、各トレーニングの反復を高速化します。このアプローチはメモリを節約できますが、はるかに頻繁で重い通信を必要とし、超高速（ノード当たり数 Tbps）で低レイテンシのネットワークが必要です。

現在の電子パケット交換（EPS）システムは、最大 72 台のデバイスをホストするスケールアップ領域内でのみ、顕著なネットワークオーバーヘッドを招くことなく強スケーリングを効率的にサポートできます。

パケット交換ネットワークから回線交換ネットワークに戻る

強スケーリングを達成する一つの方法は、電子パケット交換（EPS）システムの容量を増やすことですが、言うは易く行うは難しです。I/O 帯域幅とトランジスタ密度が物理的な限界に達すると、電力とコストが急速に上昇し、パフォーマンスの向上を維持することが難しくなります。より大きなスイッチを構築したり、EPS ネットワーク全体を複数用意したりすると帯域幅を増やすことができますが、エネルギー、コスト、複雑さという重大なトレードオフが発生します。

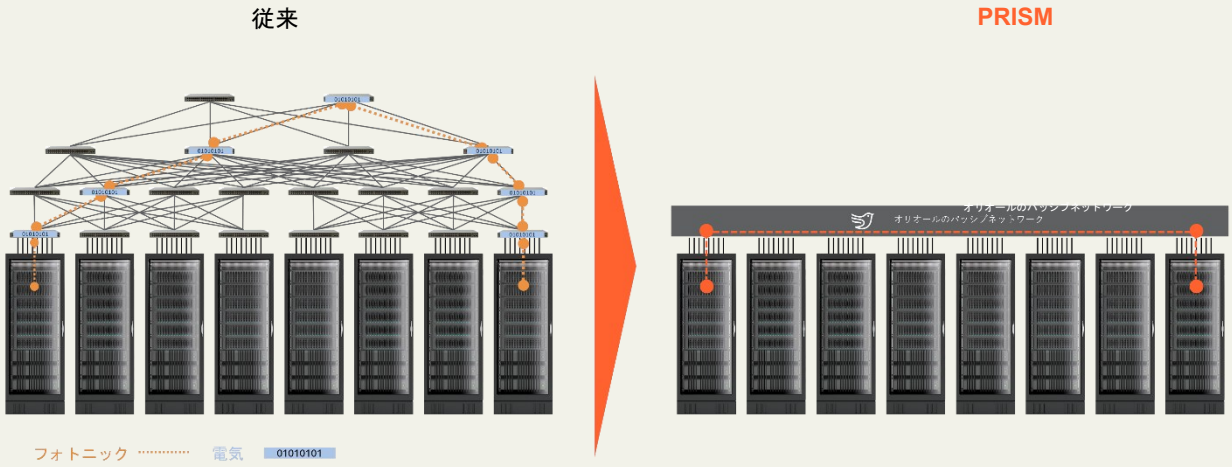
Dragonfly や Torus のような接続性の限られたトポロジーは、理論上は優れたスケーラビリティを提供しますが、データが宛先に到達するために多くの中間ノードを通過しなければならないため、しばしば非効率的な帯域幅の使用と高レイテンシにつながります。

有望な代替案は、パケット交換網における電子相互接続を光相互接続に置き換えることです。EPS は柔軟性と拡張性に優れ、混合トラフィックをうまく処理できるため広く使用されていますが、その柔軟性は高くつきます。

EPS を使用したネットワークは、輻輳制御、フォワーディングおよびルーティングテーブル、キューイングなどの複雑な制御ロジックに依存していますが、これらの制御ロジックは、データ処理とストレージに対する要求が高く、そのまま光学的に実装することはできません。その結果、非同期光パケットスイッチング（OPS）では高パフォーマンスワークロードの要求を満たすことができません。

そこで登場するのが、光ハードウェアの強みを活かした、よりシンプルで決定論的なアプローチである光回路スイッチング（OCS）です。OCS ではコントロールプレーンとデータプレーンを分け、ルーティングとスケジューリングは論理コントローラ（集中型または分散型）によって処理され、光回路はデータの高速移動に集中します。これにより、電力消費が低減され、複雑さが軽減され、拡張性が向上します。また、通信は論理コントローラによってポイントツーポイント要求に変換される「論理回路」で事前に調整されます。短時間のハードウェア再構成の後、データはシームレスに流れ、パイプライン化されたスケジューリングによって遅延なく事が進みます。

OCS ネットワークには大きな可能性がありますが、大規模なデータセンターや HPC システムに広く採用されるには、依然としていくつかの大きなハードルがあります。この技術は、高速で大規模なトラフィックスケジューリング、ハードウェア（トランシーバーとスイッチ）のセットアップと再構成にかかる時間、数千台のデバイス間での厳密な同期などの問題にしばしば直面してきました。何よりも、OCS ネットワークの導入は、ネットワークアーキテクチャ全体の見直しを意味します。オリオールネットワークスの新しいアプローチは、まさにこれらの課題を克服するために開発されました。



PRISM の特長

ここでは、オリオールネットワークスの PRISM（Photonic Routing Infrastructure for Scalable Models）について紹介します。PRISM は、AI データセンターネットワーキング（DCN）、ハイパフォーマンスコンピューティング（HPC）、分散ディープラーニング（DDL）のワークロードのためにゼロから構築された、初の大規模、大容量、全帯域幅アーキテクチャです。PRISM は、今日の相互接続の限界を打ち破り、将来のニーズに対応できるように設計されています。PRISM の特長は以下のとおりです。

1. **すべてのトラフィックを処理：**波長と空間のスイッチングを用いたナノ秒レベルのスイッチングにより実現します。時間、波長、空間領域でのスイッチング（TDM、WDM、SDM）の組み合わせが、高速回路構成による大規模ネットワーキングを可能にします。これは、大規模および小規模のデータ転送、いわゆるエレファントフローとマウスフローで効率的に動作する光データ転送の小規模バーストをサポートします。そのスイッチング速度のため、システムは決定論的な集合通信と非決定論的な動的トラフィックの両方をサポートします。
2. **真の All-to-All 接続性：**任意のエンドポイントが任意のエンドポイントに到達できるポートレベルの All-to-All 通信を提供します。
3. **電力消費と温度の低減：**完全にパッシブな相互接続とスイッチにより、ネットワークコアはクリーンで効率的な状態を維持します（冷却が不要で電力散逸がありません）。すべてのコントロールがネットワークエッジに移ることで複雑さが大幅に軽減されます。
4. **耐障害性と信頼性の高い設計：**単一障害点がありません。各ノードは互いに複数のパスを持つため、何かが壊れても通信は継続されます。
5. **規模拡大を前提に構築：**最大 100 万のエンドポイントを持つシステムを 1 ホップで処理するため、ますます複雑化する分散ワークロードをサポートできます。
6. **AI ワークロードを考慮した設計：**スケジュールレスかつコンテンションフリーのデータ伝送を可能にする、光回路スイッチング（OCS）ネットワークに合わせた集合通信戦略のための専用アルゴリズムです。最大規模のネットワークであっても動作はわずか数ステップで完了するため、レイテンシが大幅に短縮され、通信に

よって止められることなくデータが GPU に供給され続けます。

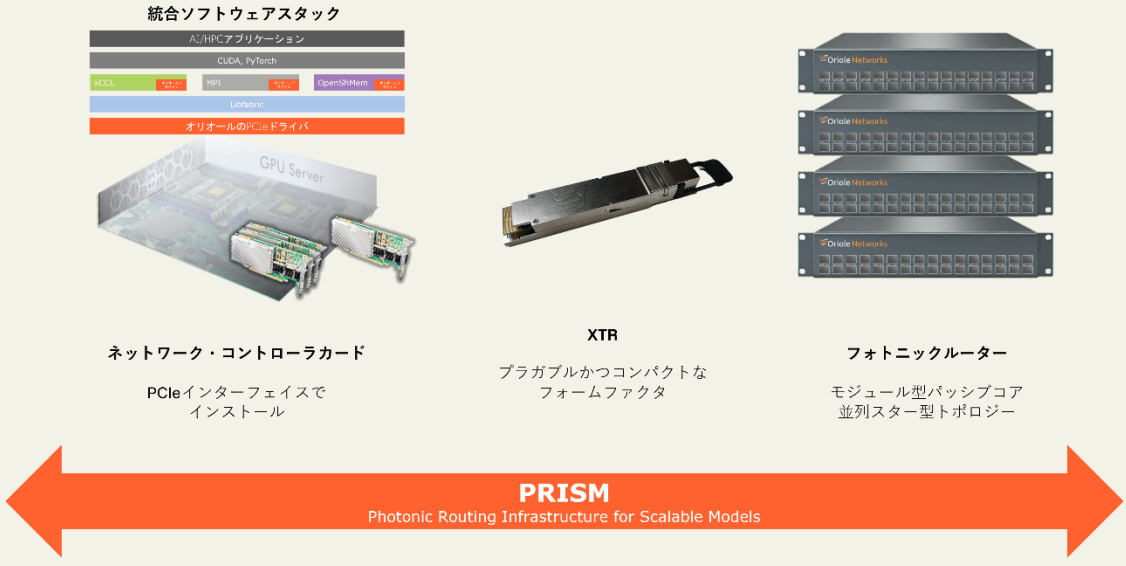
オリオールネットワークスの PRISM は、こうしたパフォーマンス、スケーラビリティ、耐障害性の組み合わせにより、次世代の高性能データセンターに強固な基盤を提供します。

ネットワークとインフラストラクチャーの概要

PRISM は、GPU を高速かつシンプルに、大規模に接続することに特化したフルスタックの全光学ファブリックを採用しています。このアーキテクチャは、現在の HPC/AI ネットワークの運用方法を大胆に変えるものですが、各コンポーネントは既存の AI データセンターにぴったり収まるように設計され、高いパフォーマンスと複雑さの軽減を両立できます。

ソフトウェアスタック：NCCL その他の集合通信ライブラリと互換性があり、PRISM のネットワークスタックに直接統合できるカスタム xCCL プラグインを開発しています。このプラグインは、これらのライブラリで使用されるデフォルトのトランスポート層（通常は Ethernet または InfiniBand）を、PRISM の光回路スケジューラおよび集合通信ロジックへの直接的なフックで置き換えます。集合演算を PRISM のアーキテクチャに合わせることで、データフローがフォトニックネットワークのトポロジーに最適化され、大規模な分散トレーニングのパフォーマンスが最大限に発揮できるようになります。

高性能ネットワークインターフェイスカード：このカードは PCIe インターフェイスを採用し、コンパクトなフォームファクタで 800 Gbps の帯域幅を提供します。ほとんどの標準的なサーバーに適合するように構築され、クラスター内の完全な二分帯域幅を提供するとともに、プラガブルなインターフェイスを介して接続性を管理します。一組の専用のコントローラカードセットが、転送要求を許可し、マイクロ秒未満の更新レートで回路を再構成し、ネットワークヘルスを監視してネットワークの転送を処理します。



ネットワークインターフェイスカードは、これらの完全な冗長性とパイプライン処理により、実時間応答性を実現します。

XTR－統合フォトニックスイッチ・トランシーバー：このプラガブルなモジュールは、1 台で送信機、スイッチ、受信機のすべての役割を担います。モジュールはネットワークカードに直接挿入され、波長と経路を臨機応変に選択します。光エンジンと統合 ASIC を組み合わせ、最大 100 万ノードへの接続性を可能にします。

パッシブルーター：ネットワークの中心は、電源が不要なモジュール型光ルーターです。トップオブラックと集中配置の両方をサポートし、デュアルパスの耐障害性を提供します。データプレーンとコントロールプレーンを分け、すべてを光とパッシブに保つことで、コアの電力消費をゼロにして超低レイテンシを実現します。

これらがもたらすメリット

電力に限りのある業界における低電力と低コスト

NVIDIA のジェンスン・フアン CEO は、GTC 2025 で「AI からの収益は電力により制限を受けている」と強調しました。今日の大規模な AI クラスターでは、価値を生み出す能力は、どれだけ計算をさせることができるかだけでなく、どれだけ効率的に電力を供給できるかにも関係しています。無駄に消費できる電力はないのです。

オリオールネットワークスは、ネットワークアーキテクチャ、通信モデル、スケジューリングを一つにまとめた、緊密に協調設計されたシステムで GPU ボトルネックの問題を解決するための、新しいアプローチを採用しています。オリオールのアーキテクチャは、クラスターのネットワーク部分と計算部分の両方の電力消費を大幅に低減し、GPU の通信が原因のアイドル時間を事実上ゼロ（1%未満）にします。高価で複雑で電力消費の大きいスイッチをより少ない台数の調整可能なトランシーバーと手頃な価格のルーターに置き換えることで、システムの複雑さを軽減し、信頼性を高めます。

GPU 使用率の低さは、HPC や分散ディープラーニング（DDL）システムの効率を低下させる隠れた要因の一つでもあります。信じられないほど強力な（そして高価な）GPU がアイドル状態になっているのは、データや計算を待っているからではなく、ネットワークが原因で止まってしまっているのです。大規模なトレーニングや推論、シミュレーションのワークロードでは、GPU は他のノードからのデータの到着を待つだけで驚くほどの時間を費やすことがよくあります。これは、システムの規模が拡大するにつれて、より大きな問題になります。こうしたオーバーヘッドを低減することは、パフォーマンスだけでなく、収益性にとっても不可欠です。

推論、特に最新の Mixture of Experts（MoE）モデルでは、モデルは各トークンをエキスパートネットワークの小さなサブセットに動的にルーティングするため、推論はコンピューティングノード間の低レイテンシ通信に大きく依存します。従来の光相互接続では、エキスパートの選択とデータ移動の間にミリ秒単位の遅延が発生し、リアルタイム推論のボトルネックが発生します。オリオールのコンテンションフリーのソリューションは、これらの問題をことごとく回避してパフォーマンスを向上させます。MoE アーキテクチャは特殊な All-to-All 通信に大きく依存していますが、EPS システムの（輻輳による）非効率性は、DeepSeek によって実証されたように複雑なパイプライン処理によって対処できます。しかし、オリオールのハードウェアはこれらの問題を本質的に回避します。オリオールのネットワークはナノ秒スケールで再構成される

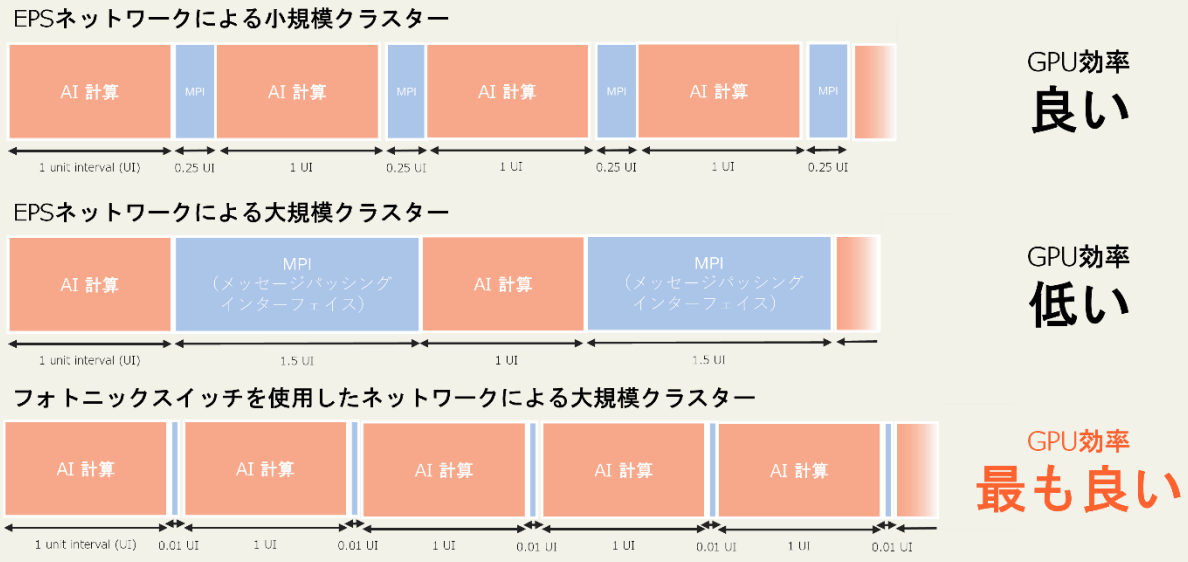
ため、事前に割り当てたルートや静的トポロジーなしに必要の都度エキスパートを起動できます。これにより、エキスパートの使用率が大幅に向上し、テールレイテンシが短縮され、推論をより多くのエキスパートに効率的にスケーリングできます。オリオールのアーキテクチャは、マイクロ秒未満の粒度で通信遅延を最小化することにより、生産環境でスパースアクティベーションされる MoE モデルのパフォーマンスのポテンシャルを最大限に発揮させます。

これこそが、PRISM のフルフォトニック、ナノ秒レベルのスイッチングが大きなインパクトを持つところです。オリオールのネットワークは、光回路を迅速に再構成することで GPU のアイドル時間を最小限に抑え、GPU のコンピューティング能力を最大限に使用することができます。ネットワークが追いつくのを待つ必要はなく、ピーク時の負荷に対応するためだけにリンクを過剰にプロビジョニングする必要もありません。その結果、EPS システムと比較して、システム全体で GPU の使用率が向上し、エネルギーの無駄が減り、ワットあたりのパフォーマンスが向上します。

パフォーマンスを低下させず AI クラスターをスケーリング

AI インフラストラクチャーのスケーリングにおける最大の課題の一つは、ノードを追加した際に一貫してパフォーマンスを維持させることです。従来のネットワークでは、システムの規模が大きくなるのに伴い、輻輳、パケット損失、および変動する平均とテールのレイテンシのすべてがアプリケーションのスループットを低下させる可能性があります。これに対し、PRISM のようなスケジュールされた同期決定論的な完全フォトニックネットワークは、大きな帯域幅とナノ秒レベルの再構成により GPU のアイドル時間を最小限に抑え、スケーリングに伴うパフォーマンスの問題を軽減することができます。

PRISM は、従来のソリューションから**ネットワークパフォーマンスを不連続的に向上させ、AI のトレーニングと推論のパフォーマンスを最大限に発揮できるようにする**と同時に、電力消費と構築コストを削減します。つまり、速度が向上するだけでなく、予測可能なパフォーマンスを大規模に得ることができ、これはまさに HPC と DDL のワークロードが必要とするものです。

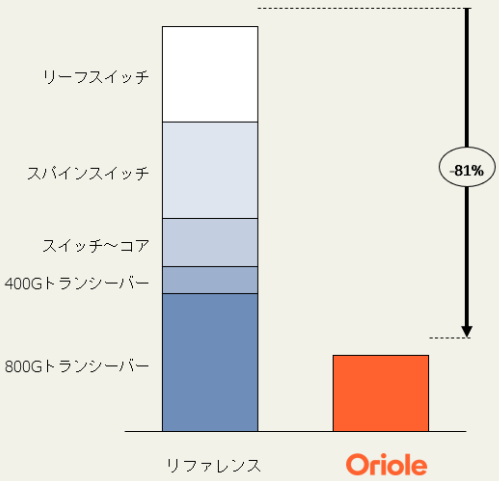


従来の AI クラスターと光 AI クラスターの比較（32k ノード時）

アーキテクチャの概要

	リファレンス アーキテクチャ		オリオールネットワークスの アーキテクチャ
ネットワーク層	3		1
リーフスイッチ	1024		0
スパインスイッチ	1024		0
コアスイッチ	512		
スイッチ合計	2,560	完全に削減！	0
エンドポイントトランシーバー	32,768		32,768
スイッチトランシーバー	163,840		0
トランシーバー合計	196,608	83%削減	32,768

電力消費［MW］



複雑さの軽減と信頼性の向上

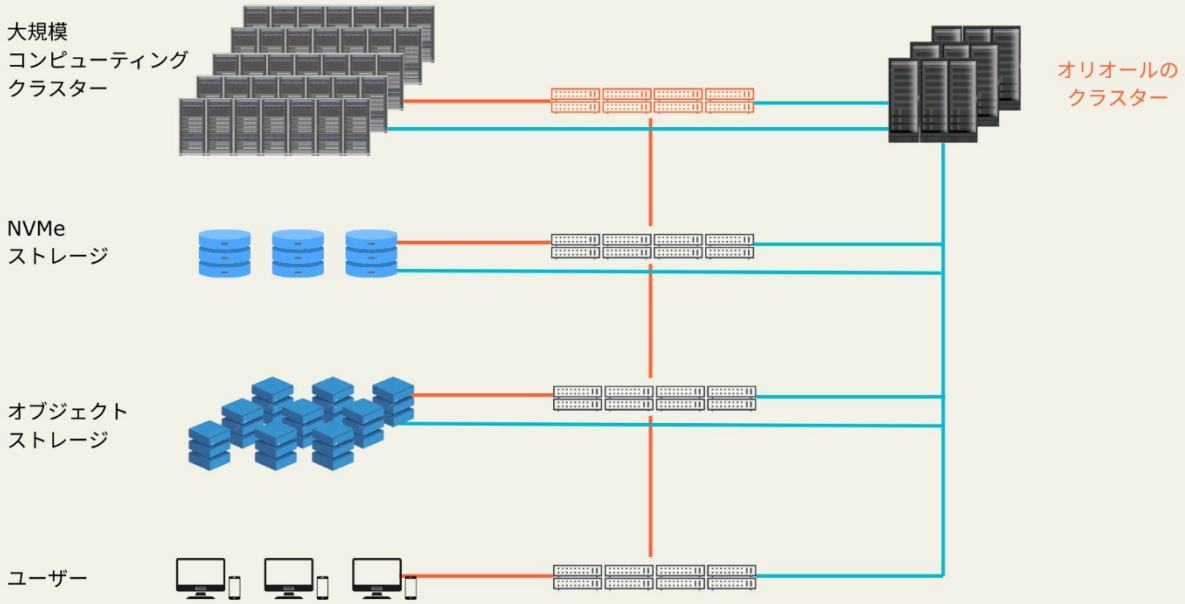
オリオールは、ネットワークコアからスイッチを完全になくすことで、従来の EPS ネットワークが依存していた大量の電子機器を取り除きます。つまり、ASIC の数が減り、電力消費の大きいチップの数が減り、管理すべき熱も大幅に減ります。さらに、EPS スイッチとそれに関連するトランシーバーを完全に削減することにより、潜在的な障害点の数が減少し、システムの信頼性が大幅に向上します。その結果、よりシンプルでクリーンなアーキテクチャとなり、規模の拡大、冷却、管理が容易になります。上の図は、オリオールネットワークスのアーキテクチャとリファレンスアーキテクチャを比較したものです。スイッチ（および関連するトランシーバー）を完全に削減することにより、ネットワークコアの電力消費をリファレンスアーキテクチャと比較して、EPS スイッチ基準で 81%削減できることを示しています。

この変化は持続可能性にも大きなメリットをもたらします。PRISM は、ループ内のハードウェアを減らすことで、データセンターのインフラストラクチャーの規模が大きくなるのに伴いますます緊急の問題となっている、電子機器廃棄物となる電子機器の量を削減します。さらに、ネットワークコアに高密度のカスタムシリコンを使用しないようにすることで、調達が難しくなっている（そして地政学的にも影響を受けやすくなっている）レアアース金属や高純度銅などの重要鉱物への依存度を下げることができます。さらに、冷却要件の低減は、データセンターの給水ニーズが減ることを意味します。

PRISM は既存のエコシステムにどのように適合するか

PRISM の目的はパフォーマンスだけではありません。実用的な統合を目的として構築されているため、すべてを取り壊すことなく、既存のデータセンターに容易に導入できます。既存のセットアップのほとんどは、従来の EPS スイッチングを使用したスパインアンドリーフ・トポロジーを使用しており、オリオールのネットワークはそれと並行する高性能クラスターとして導入できます。PRISM と既存のインフラストラクチャーは、軽いイーサネットゲートウェイで橋渡しされるため、高帯域幅でレイテンシセンシティブなワークロードを光側に移動しても、レガシー側のコンピューティングリソースとストレージリソースは働き続けることができます。

これにより、ネットワークを時間とともに容易に進化させることができます。具体的には、最も必要とされる場所に PRISM を導入し、需要の増加に応じて光側の割合を大きくできます。PRISM は、一度に完全に置き換えるのではなく、アップグレードのパスに沿ってプラグイン式に導入します。つまり、働いているものを中断することなく、迅速な導入とスムーズなスケーリングが可能です。



まとめ

GPU の高速化とモデルの大規模化が続く世界では、ネットワークが速度低下の原因であってはなりません。PRISM は本当のボトルネックに正面から取り組み、スループットの向上、レイテンシの短縮、電力消費の大きいハードウェアの大幅な削減を実現します。スイッチのファブリックが無秩序に広がることはもうありません。ケーブルの山も改善されます。クリーンなフォトニックコアで、データセンターを過酷な場所にすることなく規模を拡大できます。

PRISM は、パフォーマンスを向上させながら複雑さと電力消費を低減することで、最新の AI ワークロードに対応し続け、さらなる規模拡大を支援することができます。

このように、PRISM は、DDL および HPC のインフラストラクチャーを、大規模なディープラーニングネットワーク用に特別に設計されたネットワークアーキテクチャに接続する、飛躍的にスマートな方法なのです。

オリオールネットワークスの PRISM は、世界初の高速スイッチングとエネルギー効率に優れたフルフォトニックネットワークにより、次世代の分散 AI トレーニングと推論のパフォーマンスを最大限に発揮させるための鍵となるものです。

Oriole Networks Ltd

ロンドン
4 City Road,
London, EC1Y 2AA
United Kingdom

電話：+44 (0)20 814 81981

info@oriolenetworks.com

ペイントン
EPIC, White Rock Park, Waddeton Close,
Paignton, TQ4 7RZ
United Kingdom

電話：+44 (0)18 037 14763

oriolenetworks.com

パロアルト
380 Portage Ave
Palo Alto, CA 94306
United States