ARISTA

UECの謎解きと~~UALink~~ESUNの誘い

-Demystifying Ultra Ethernet and temptation of ~~Ultra Accelerator Link~~Ethernet for Scale-Up Networking-

Shishio Tsuhciya
shtsuchi@arista.com

UECの目的と結成メンバー

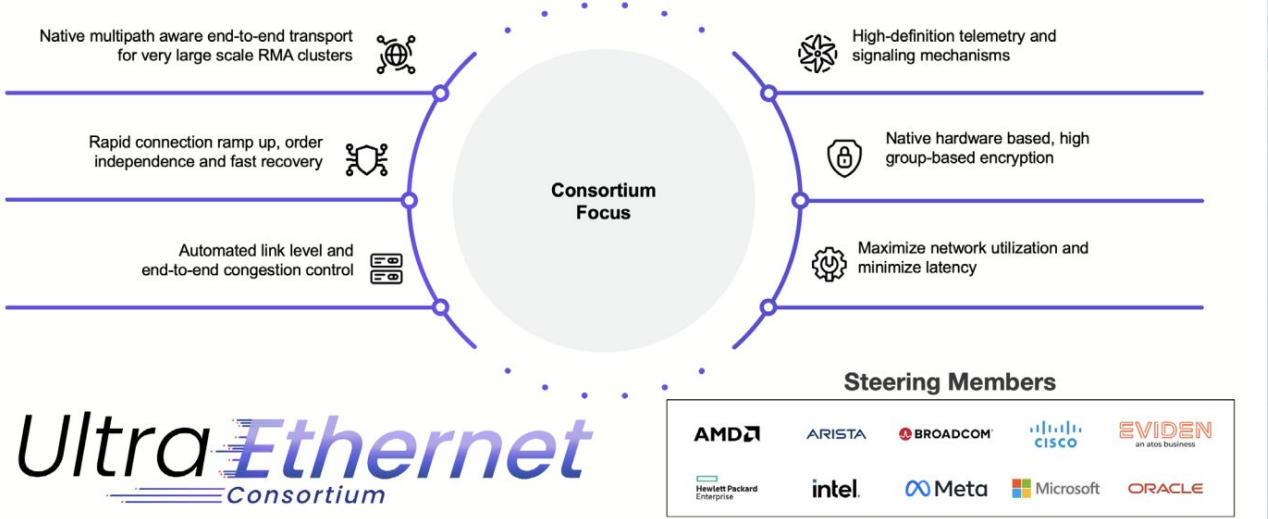
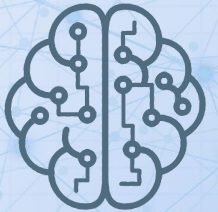


Figure 1: UEC Goals and Founding Members

- AI/MLは初期段階では従来のネットワークとは独立して開発/運用されていた
- 現在はあらゆるビジネスで不可欠となっていて、共通の技術パラダイムが必要
- ウルトラ・イーサネット・コンソーシアムは、AIとHPCのニーズに応えるためイーサネットを強化することを目的として設立された標準化団体
  - 100社以上の加盟企業と1000人以上

誰がAI/MLをどこで要求してるか



産業	使用用途
企業	カスタマーサポート 文章作成 自動要約
製造業	品質管理 メンテナンス予測
医療	診断サポート 文章解析
金融	予測分析 クレジットスコアリング
ゲーム	コンテンツ作成 カスタマイズされたユーザ体験

産業	使用用途
小売業	商品レビュー分析 需要予測
エネルギー	エネルギー効率向上 予測メンテナンス
交通・運輸	輸送の最適化(2024年問題) 交通予測
農業	作物管理 病害虫予測
教育	個別指導 教材作成

ARISTA

Ultra Ethernet Transport (UET)

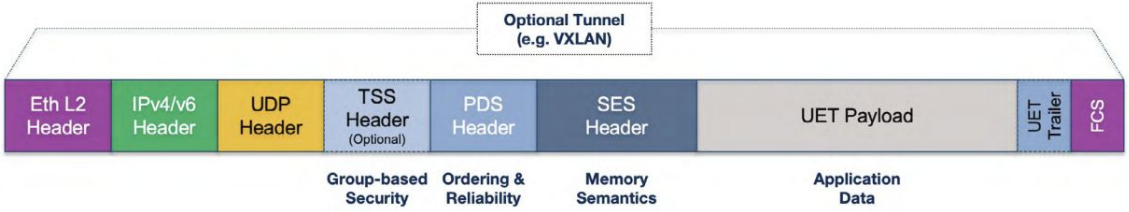
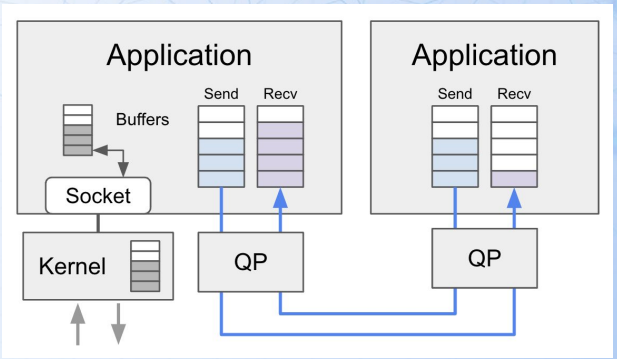
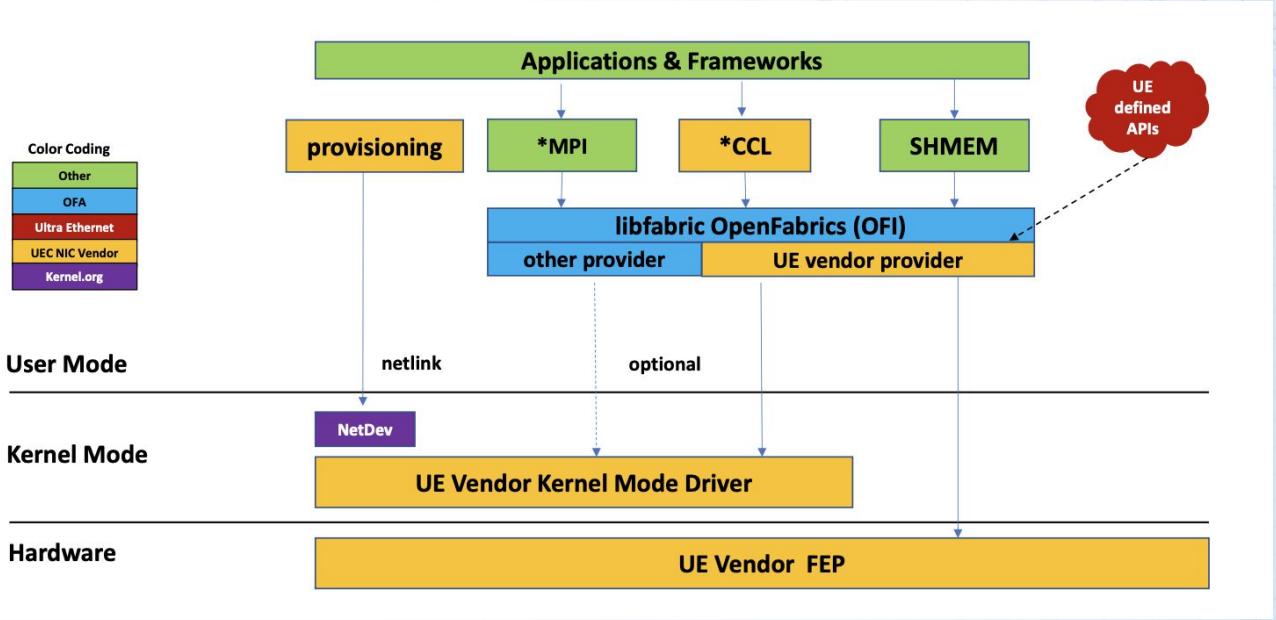


Figure 2: UET Packet Format



- UECの中核はRDMAの再構築
- 現在400Gbps/800Gbpsでネットワーク上でバースト通信する
  - ワークロードを分散出来る
  - 複数のアクセラレーターにまたがる並列計算を実現
- UET(Ultra Ethernet Transport)では標準のイーサネットおよびIPにトラフィック分散セマンティクス(論理定義）と最新の輻輳制御を追加

Native Libraries



- Open Fabrics Allianceによって標準化され、HPC環境で広く実装されている成熟した汎用APIであるlibfabric 2.0を使用
- 多様なRDMAセマンティクスを単一の中央インターフェースに集約する事が出来、異なるシステムやアクセラレータアーキテクチャ間でのアプリケーション移植が容易になる

Traffic Forwarding

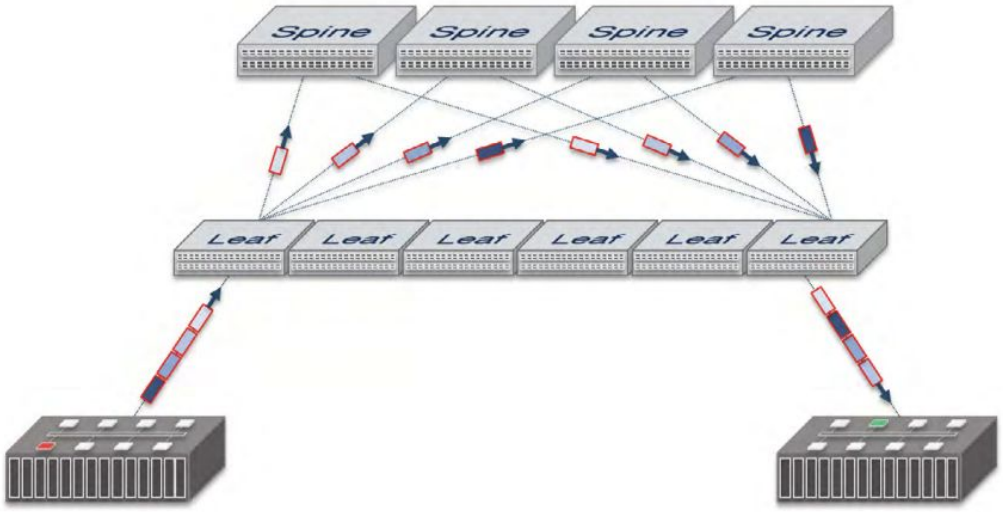


Figure 3: Packet Spraying Overview

- フローベースのトラフィック分散からNIC毎のパケットスプレイに
- NICは順不同で来たパケットを再構築してアプリケーションに渡す
- 個々の損失パケットの再送信が可能になる

ARISTA

Congestion Management

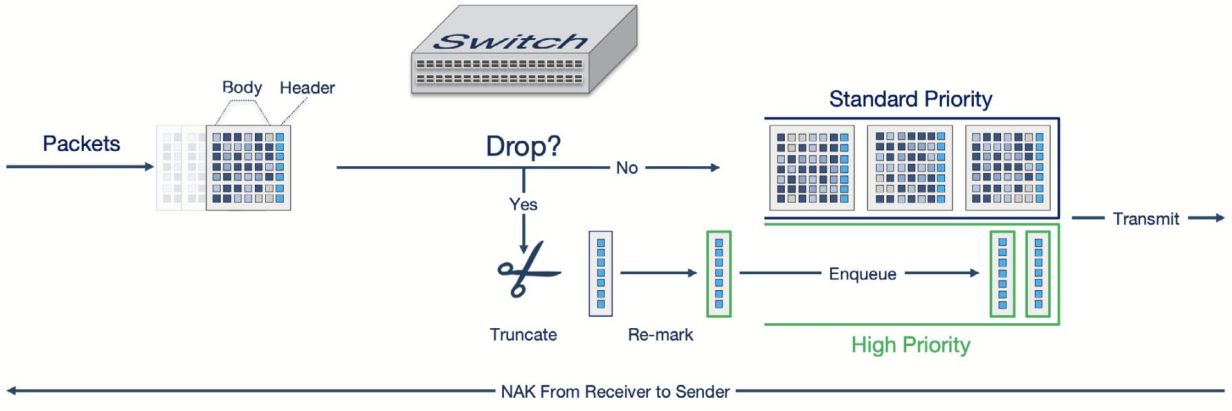


Figure 4: Packet Trimming Mechanism

- 従来はシーケンス番号でパケットの取りこぼしを検出 go-back N
  - 1つのパケット損失の為に多くのパケットが再送され非効率
- UECではパケットが輻輳状態のスイッチに到達するとパケットを廃棄する代わりにパケットを最小に切り込む。
- 切り込まれたパケットは優先キューに入れられ、宛先はこれを見て NAKの形で送信元に戻す
- これは2つの大きな役割を果たす
  - ドロップした事を明示的に通知
  - 輻輳通知となり、転送レートを落とし、輻輳パスを通らない様に迂回する

Advanced Connection Setup and Host-based Flow Control

- そもそも輻輳を回避する事がより良い方法になる。UECには“Ephemeral Connections” (儚い接続)と2つの輻輳制御スキームを導入してる
- エフェメラル・コネクションは、データが流れ始めるまでのラウンドトリップ・ハンドシェイクの遅延をなくすことで、高速な接続開始を可能にする。明示的な終了は必要無い

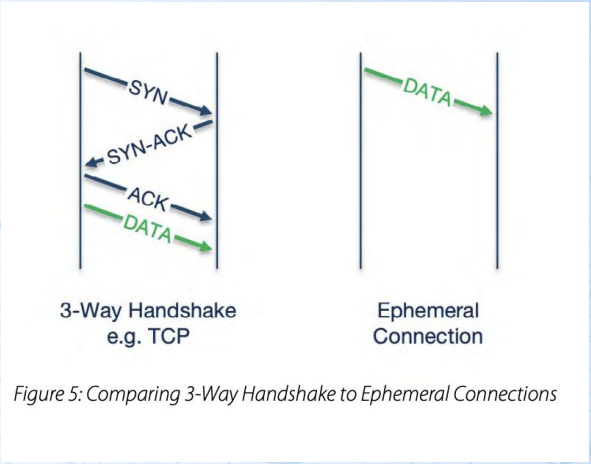


Figure 5: Comparing 3-Way Handshake to Ephemeral Connections

Advanced Connection Setup and Host-based Flow Control

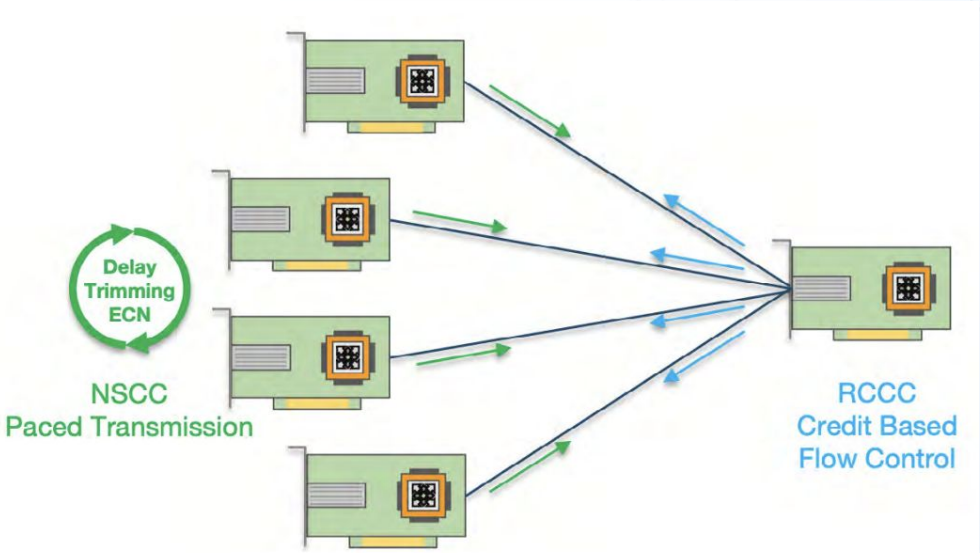


Figure 6: NSCC and RCCC Operation

- Control (NSCC)は送信者ベースの輻輳回避手法で下記を利用し送信ペースを配分する
  - ネットワーク遅延
  - トリミングパケット
  - ECN通知
- Receiver Credit Congestion Control: RCCC)は受信者ベースの輻輳回避。各受信者がクレジットを作成し、全ての送信者に公平に割り当てる。これによりTCPインキャストの様な事を排除出来る。
- NSCCとRCCCは独立して動作する事が出来る

Security

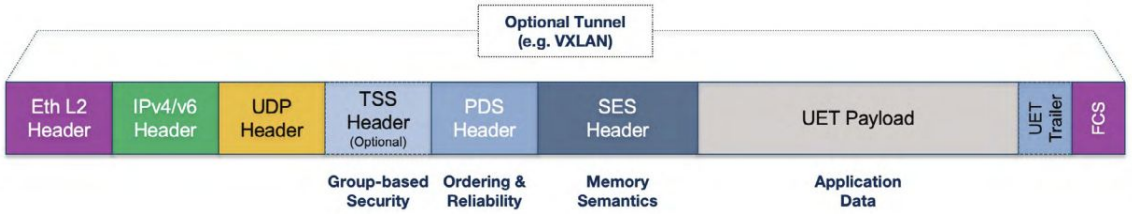


Figure 2: UET Packet Format

- UETではセキュリティは第一の目的
- AES-GCM、量子計算機暗号(PQC),Key Derivation Functions (KDF)、リプレイ防止などの実績のある技術を活用したオプションのエンドツーエンドの暗号化と認証が、UETホスト間で動作
- 1つのグループ・キーは、例えば1つのテナントが運用するすべてのXPUなど、1つのjobの全メンバーで共有される。

  Public.

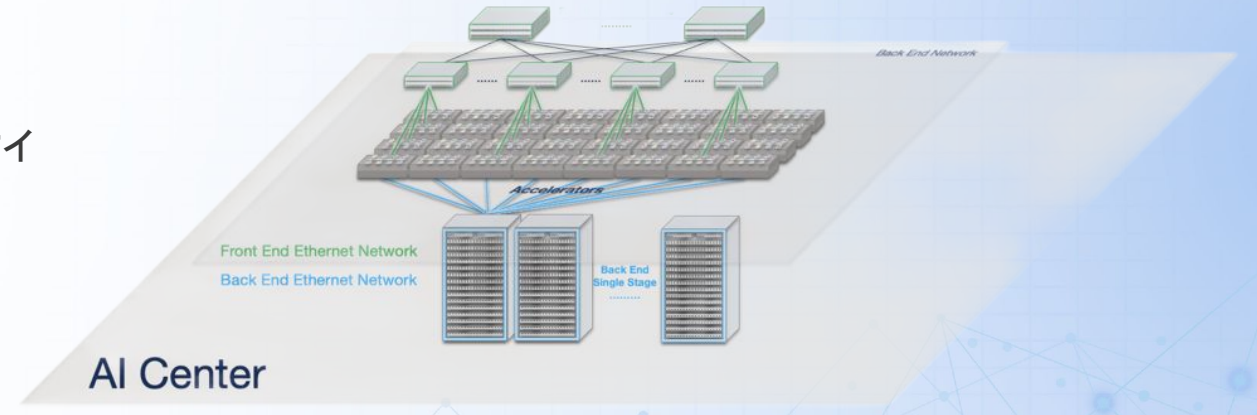
Additional Future Capabilities

- LLR(Link Level Retry)
  - 各スイッチ・ポートに小さなバッファを実装することで、個々のリンク単位で再送信を行うメカニズム。
- Credit-Based Flow Control (CBFC)
  - 受信スイッチが予約できるスペースと同数のパケットを正確に要求することができる。
- これらの新機能は、スイッチング・シリコン内で新たな論理設計を必要とし、将来の次世代システムで利用可能になる。

UALinkの誘い

AIセンター

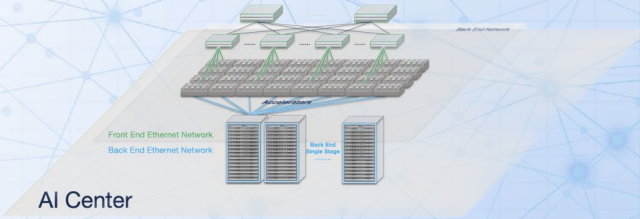
- End to End の単一技術パラダイム
  - キャンパス、WANからデータセンター
  - フロント側、ストレージのネットワーク
  - AI学習と推論のバックエンドネットワーク
- 単一のツールと運用
  - 構築、運用
  - ツールやセキュリティ
- 投資保護
  - 拡張への経済性
  - オープン、標準
  - 再配備



Ethernet – コストとパフォーマンスを最適化したAIインフラを実現する鍵

ARISTA

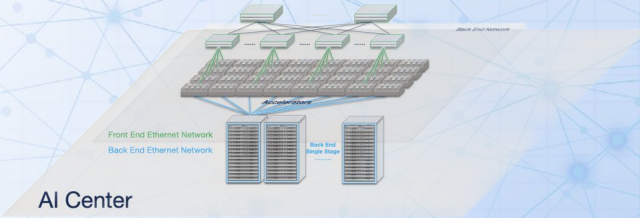
AI導入への解析
- フロントエンド・ネットワーク



- フロントエンド・ネットワーク（学習）
  - 学習をサポートするサービスを接続
  - 高帯域幅ストレージ
  - オーケストレーションとモニタリング
  - クラスタ・ニーズに基づく仕様
- フロントエンド・ネットワーク（推論）
  - 顧客とクライアントをクラスタに接続
  - 受信クエリと応答
  - APIやマイクロサービス
  - アプリケーション／クライアントのニーズによって駆動される仕様

フロントエンド・ネットワークは最新のデータセンターのパラダイムに従う

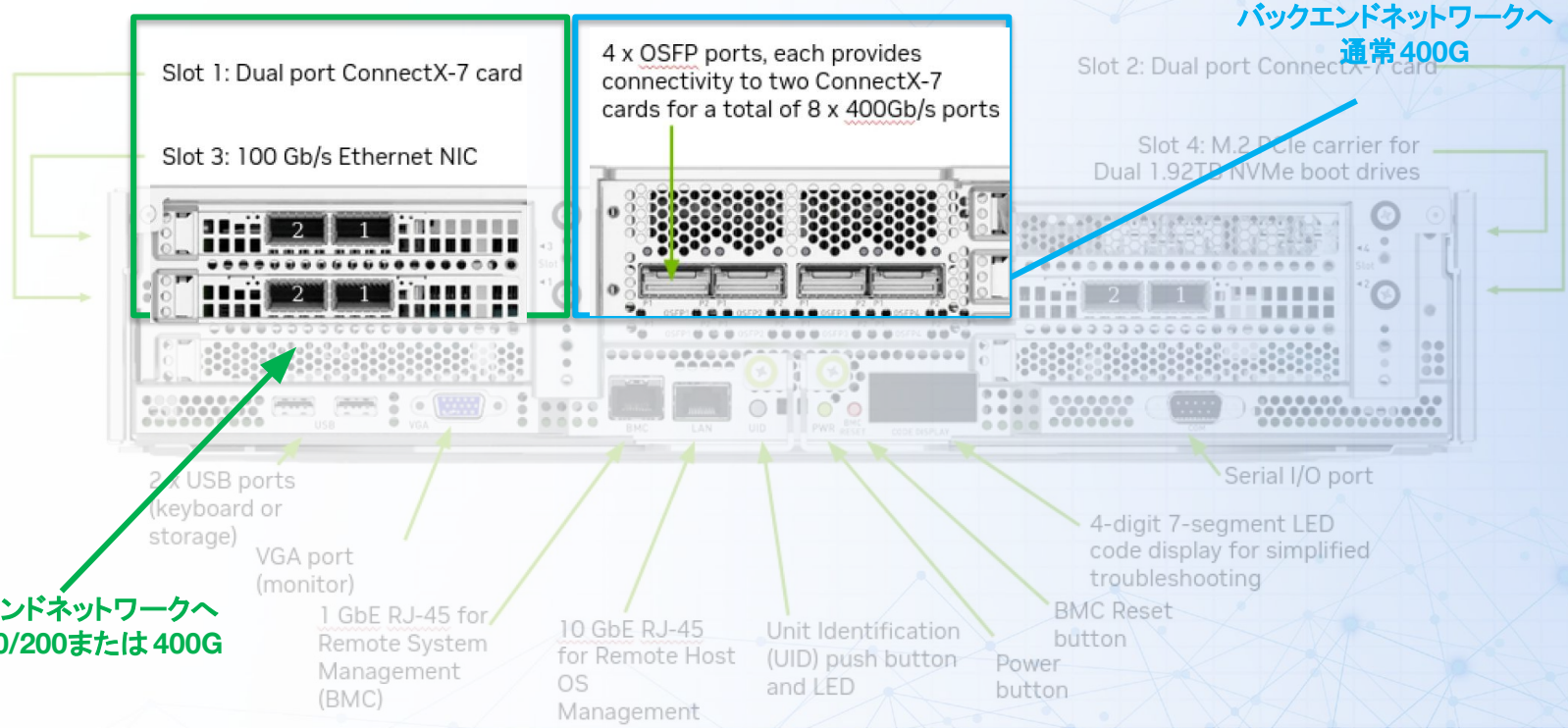
AI導入への解析
- バックエンドネットワーク



- バックエンドネットワーク（スケールアウト）
  - 高性能なXPU間バックボーン
  - 学習および推論タスク用
- 高いノンブロッキングスケールアウト帯域幅
  - 数百 → 数千 → 10万以上のアクセラレータ
  - 400G → コンピュートノードへの800G
  - オーバーサブスクリプションしないと回復力のあるフェイルオーバー
  - ロスレスフォワーディング
- トラフィック・プロファイルへのチェレンジ
  - 大きなメッセージサイズ
  - 少ないエントロピー
  - 長寿命のフロー
  - 高度に同期されたバーストトラフィック

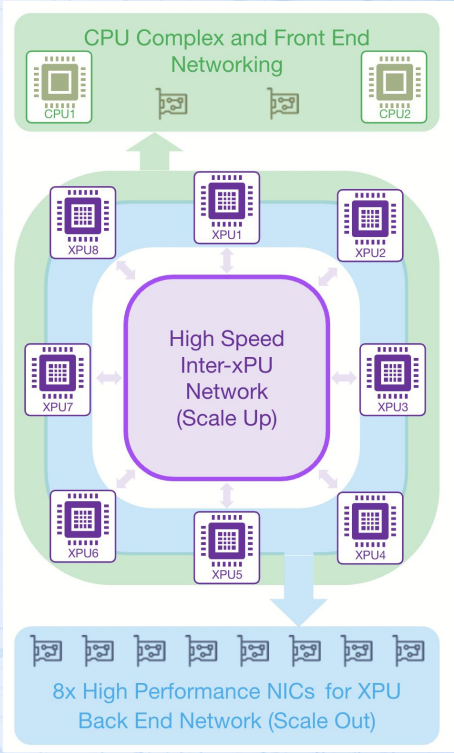
AI学習のワークロードは相関性が高い　- 低下や遅延の影響を受けやすい

接続するアクセラレーター (GPU)の例



スケールアップ - 第3のネットワーク

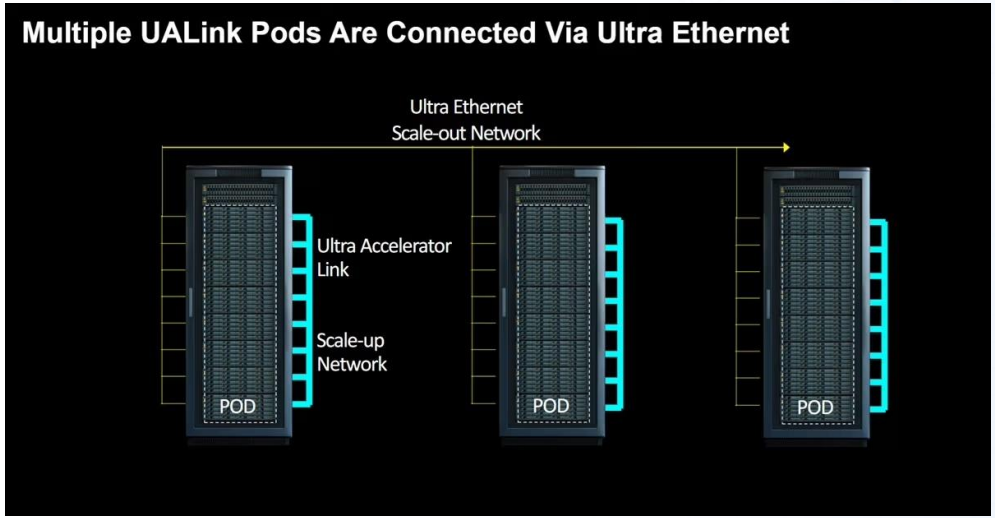
- Fast Path Local Interconnect
  - 同じホスト（またはラック）内のxPU
  - スイッチまたはメッシュのいずれか
  - **独自規格（NVLinkなど）またはオープン（イーサネット）のいずれか**
- xPU 間の直接通信に使用
  - 高速データ交換
  - メモリ共有
  - レールホッピング
- 用途は
  - ジョブの最適化
  - 同一タスクへのプロセッサのローカル割り当てに依存

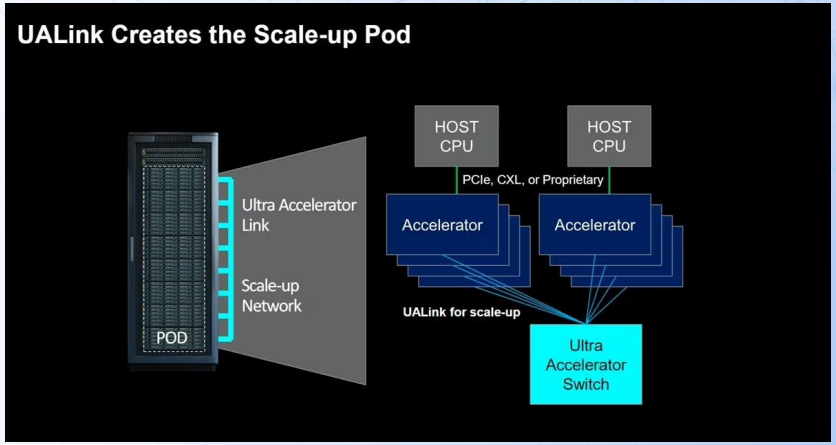


ARISTA

Data center leaders create Ultra Accelerator Link group for AI connectivity

https://venturebeat.com/ai/data-center-leaders-create-ultra-accelerator-link-group-for-ai-connectivity





- ウルトラアクセラレータリンク（UALink）は、次世代AI/MLクラスタの性能を向上させる高速アクセラレータ相互接続技術
- AMD、Broadcom、Cisco、Google、HPE、Intel、Meta、Microsoftなどがオープンな業界標準化団体を設立
- 1.0仕様では、AIコンピューティングポッド内で最大1,024個のアクセラレータを接続可能

Ultra Accelerator Link Consortium, Inc. Specification UALink_200 Rev 1.0

https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-Specification-Overview_FINAL-1.pdf

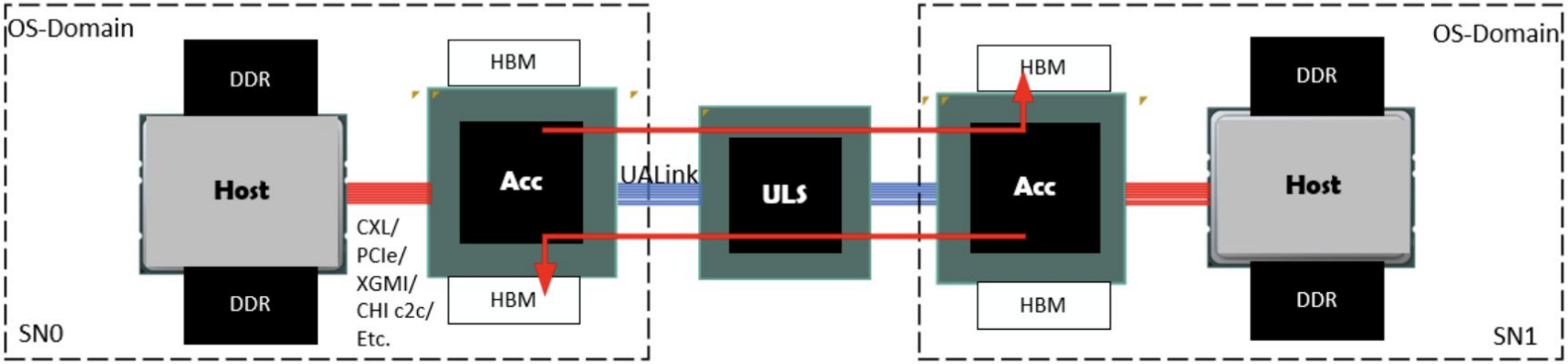


Figure 0-1 UALink Connectivity Overview

- UALinkはアクセラレーター間通信の為のもの
- アクセラレーター間の直接ロード/ストア/アトミック操作
- 1.0仕様では**Infinity Fabric Protocol**を使用

Routing a Transaction from End-to-End

https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-Specification-Overview_FINAL-1.pdf

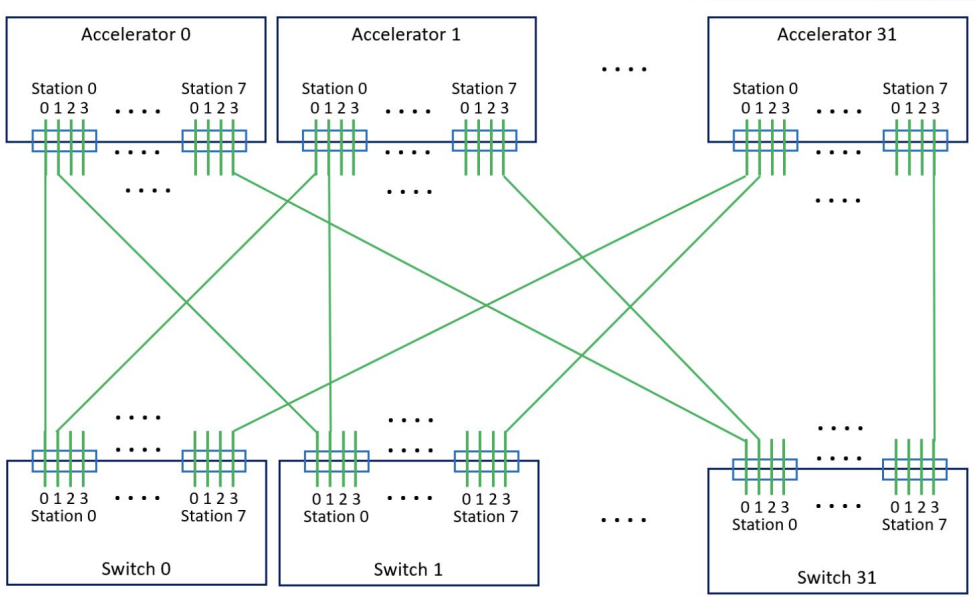


Figure 2-8 Example system with 32 Accelerators with 32 x1 UALink Links

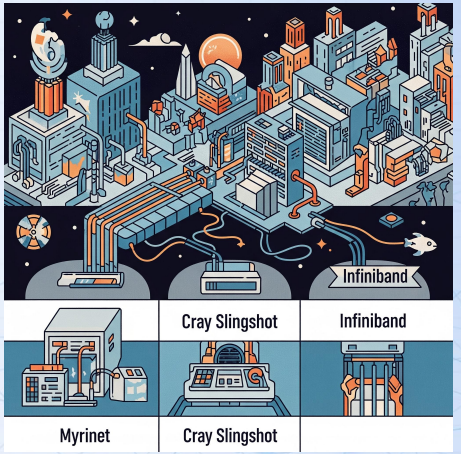
- 32アクセレーターと32x1 UALinkリンクの例
- UALinkステーションは4つのUALinkレーンを有す
- ポッド内の全UALinkステーション（スイッチ上およびアクセラレータ上の両方）は同一方式で分岐されなければならない。
- 各アクセラレータ上のUALinkポート数は等しくなければならない。
- ポッド内では、物理スイッチのポート数はポッド内のアクセラレータ数以上を有し、物理スイッチのポート数はポッド内のアクセラレータ数と等しくなければならない

  Public.

Tomahawk Ultra: Ethernet to the rescue for HPC and AI Scale-up

https://www.broadcom.com/blog/tomahawk-ultra-ethernet-to-the-rescue-for-hpc-and-ai-scale-up

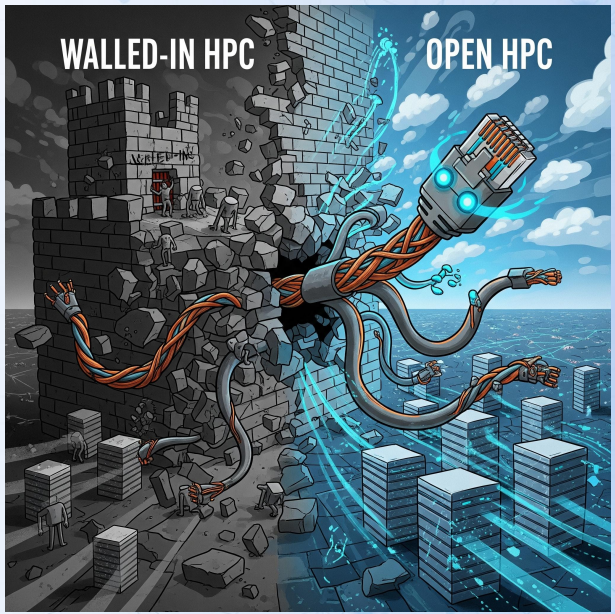
- HPCの世界ではどうして独自技術に閉じ込められたのか？
  - ファーストイーサネットでは遅すぎ、Myrinet/Cray Slingshot/Infinibandと独自機構に移行
  - HPCシステムは高価になったが、導入数が少なく政府機関などの導入によりコストやオープン性はここでは問題なかった
  - HPCシステムでは
    - 最速のポートスピード
    - 最小遅延
    - 高い信頼性
    - 非常に小さいサイズのパケットのサポート



Tomahawk Ultra: Ethernet to the rescue for HPC and AI Scale-up

https://www.broadcom.com/blog/tomahawk-ultra-ethernet-to-the-rescue-for-hpc-and-ai-scale-up

- どうすれば救い出せるか？
  - 250nsの超低遅延
  - 超高信頼性 -LLR(Link Layer Retry)やCBFC(クレジットベースフローコントロール)
  - 小さいパケットサイズでフル帯域のサポート(ppsは通常のものの2倍)
  - 少ないオーバーヘッドのフレームサポート



Tomahawk Ultra / BCM78920 Series

https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm78920-series

- 64 × 800GbE 106.25G PAM4 256 x 200GbE
- シングルチップで51.2Tbpsをサポート
- 250ns 超低遅延
- ハードウェアにてLLR(Link Layer Recovery)を実施して信頼性向上
- CBFC(Credit-based Flow Control)にてシステムパフォーマンス向上
- In-Network Collectives (INC)でネットワークベースでCCLのAllReduceをサポート
- 12Bに圧縮されたネットワークヘッダー (AI Fabric Headers, AFH)をサポート

SUE Requirements

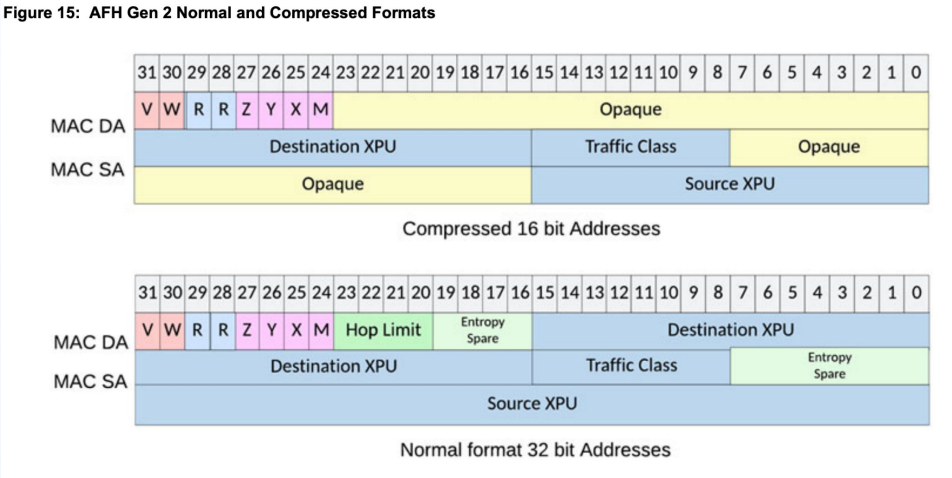
https://docs.broadcom.com/doc/scale-up-ethernet-framework

	要件	注意事項
XPUの数	1024まで	シングルホップ/低遅延
トランザクションの種類	片方向演算 メモリーロード/ストアアトミック演算 小規模転送	
メモリアーキテクチャー	シェアドメモリ	
エンドツーエンド遅延	<2µs RTT	
ケーブル長	10m以内 SUEからスイッチ	
SUE毎の帯域	800Gbps	1.6Tbpsまでスケール
SerDes	200Gbps	100Gと50Gをサポート
ポート設定	1,2または4	
バーチャネルチャネル	4まで	
イーサネットコンプライアンス	標準イーサネットをサポート	圧縮ヘッダーとLR.CBFCサポート

Scale-Up Ethernet Framework

https://docs.broadcom.com/doc/scale-up-ethernet-framework

- AFH Gen 2は、既存のイーサネット規格に準拠し変更を最小限に抑えつつ、はるかに小さなネットワークヘッダーを提供する最適化ヘッダー
- IEEE 802.c-2017に基づき12Bヘッダーと6Bヘッダーの2つのオプションが定義
- XPU識別子は32ビットまたは16ビット値にマッピングされ、宛先アドレス領域および送信元アドレス領域に格納：
  - M = 0/1（マルチキャスト）
  - V = 0（現行バージョン）
  - – W = 0 (ホップカウントとエントロピーを含む通常形式)
  - – W = 1 (圧縮形式、ホップカウントやエントロピーなし)
  - V = 1 (将来)
  - X = 1 (ローカル割り当て)
  - Y = Z = 0 (SLAPに基づくAAIエンコーディング)



まとめ

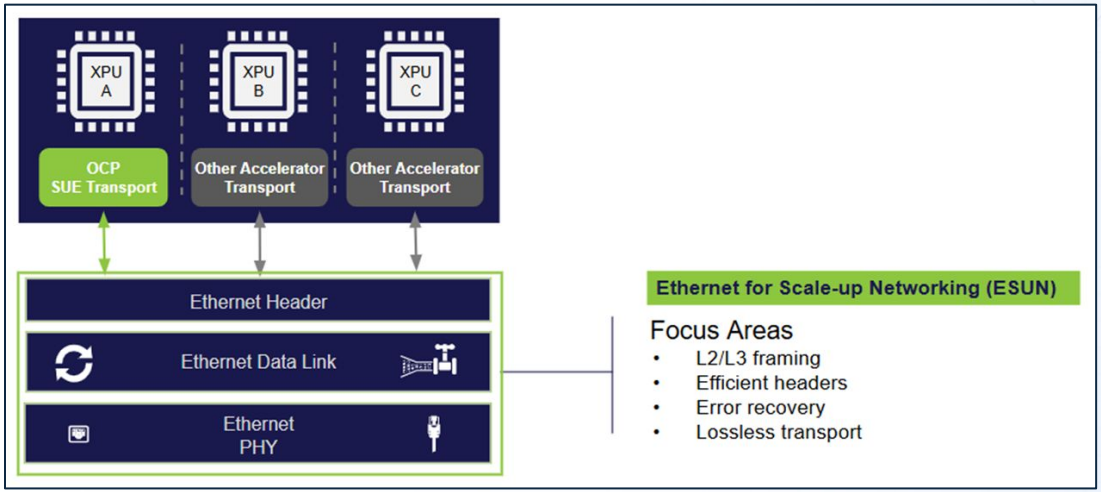
- UEC1.0の概要を説明した
- 更にAIファブリックにおける第三のネットワークスケールアップネットワークおよび、そこで使われるのを目的としてUALinkやSUEの概要を説明した

ESUNの誘い

ここまでの資料が2025年9月30日に作成/応募

Introducing ESUN: Advancing Ethernet for Scale-Up AI Infrastructure at OCP

https://www.opencompute.org/blog/introducing-esun-advancing-ethernet-for-scale-up-ai-infrastructure-at-ocp

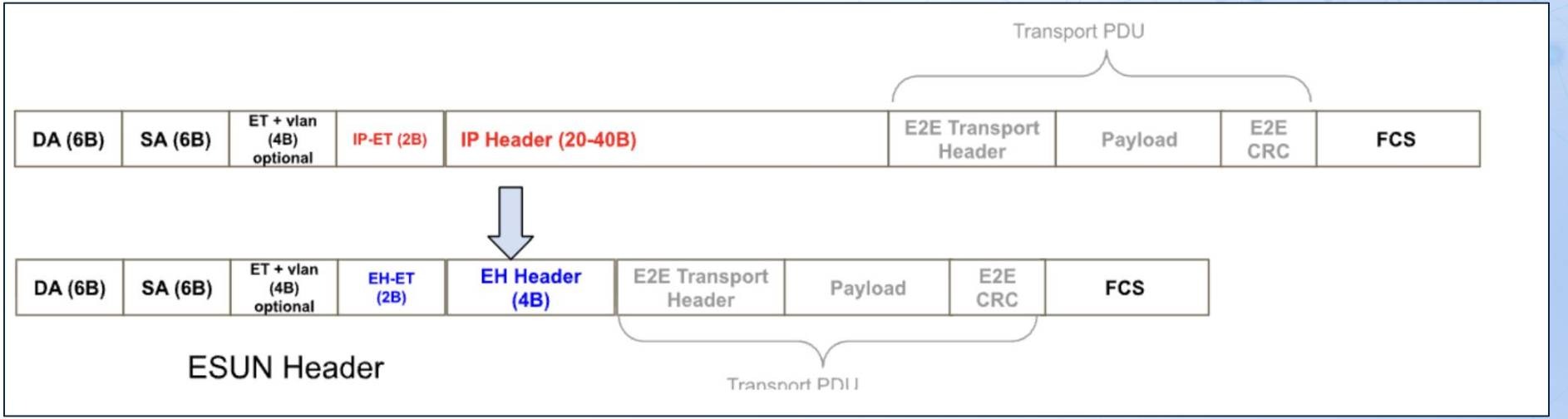


主な重点分野:
技術コラボレーション:
相互運用性:
技術的な焦点:L2/L3 イーサネット フレーミングとスイッチング
標準の整合: UEC (Ultra-Ethernet Consortium) や IEEE 802.3 (Ethernet) などの組織と積極的に連携
エコシステムの有効化: ESUN は、イーサネットの成熟したハードウェアおよびソフトウェアエコシステムを活用することで、多様な実装を促進し、業界全体での急速な導入を促進

- 2025年10月13日 OCPにて新たな取り組み ESUN(Ethernet for Scale-Up Networking)を発表
- 初期メンバー
  - AMD/Arista/ARM/Broadcom/Cisco/HPE Networking/Marvell/Meta/Microsoft/Nvidia/OpenAI/Oracle

OCP ESUN - Network Operator Requirements (Rev 1.0)

https://drive.google.com/file/d/1y72w4PwaCrOimmqNlH59DOR9dz8jnwl7/view



- 標準のIP/イーサネットと比較すると小規模なノード間通信に必要な新しい仕様の定義
- IPヘッダーは使わずに新たなESUN Header(EH)を定義、わずか4Bのヘッダーを使って転送効率を大幅に向上

ESUNヘッダー

Rev (2b)	F	EH-CoS (3b)	EH-ECN (2b)	Flow label (16b)	TTL (4b)	UD (2b)	RSVD (2b)

ESUN Header Definition

- EH-ECN (2bit):
  - ネットワークの輻輳を通知し、パケットロスを回避
- EH-CoS (3bit):
  - トラフィックの優先順位を識別
- EH-Flow Label (16bit):
  - ロードバランシングのためのエントロピーを提供
- EH-TTL (4bit):
  - 意図しないネットワークループを防ぐための生存時間を管理

スイッチの要件

- 転送方式の変更（L2ベースの静的転送）
  - 宛先ベースの転送
  - 静的設定
  - ブロードキャストは制限する
  - ローカル管理アドレスのサポート
- トラフィック管理と優先制御（CoS）
  - EH-COSを内部トラフィッククラスにマッピング
  - イーサネットCOSやCBFSの優先順位よりもEH-COSを値を優先
- ロードバランスの高度化
  - Flow labelを活用し、ロードバランスを実現
- 輻輳管理とループ防止
  - ECNのマーキング：混雑状態に基づいてEH-ECNに通知
  - TTL管理:パケット転送時にTTLを減算する
  - FCS：ECNやTTLを変更した際にFCSを再計算する
- リンクレベルの信頼性向上
  - PFCサポート:優先度ベースのフローコントロールは必須
  - LLRとCBFC:サポートされている場合はポート毎に出来る必要がある。

UALink/SUE/ESUN

UALink	OCP/ESUN	OCP/ESUN
UALink TL / Protocol	UALink TL / Protocol	SUE
UALink Link Layer Support	UALink Link Layer Support	
UALink Logical Link Control	UALink Logical Link Control	
Link Layer Retry (LLR; incl. CRC)	ESUN/Ethernet	ESUN/Ethernet
Ethernet		

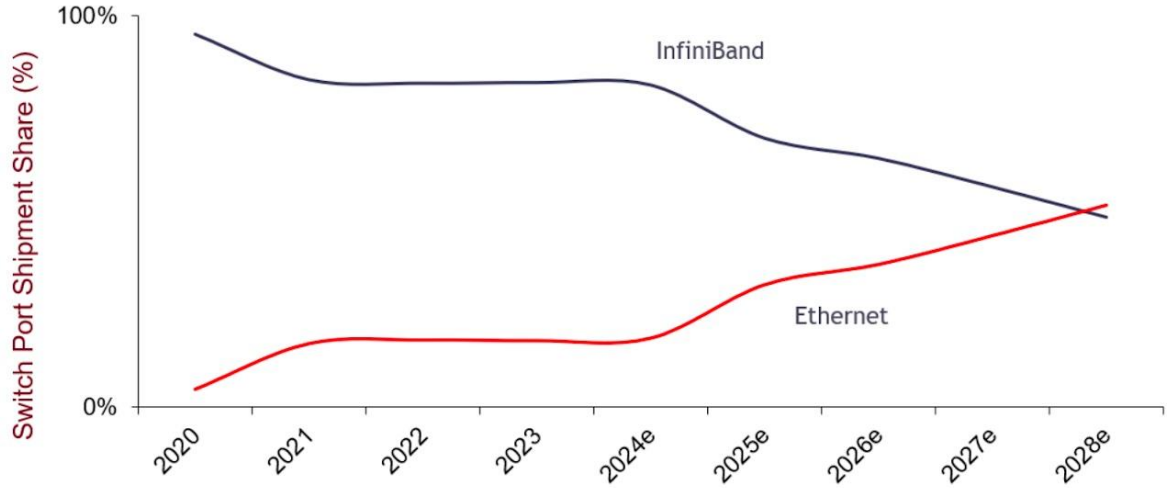
Special Thanks:Ebisawa-san

ARISTA

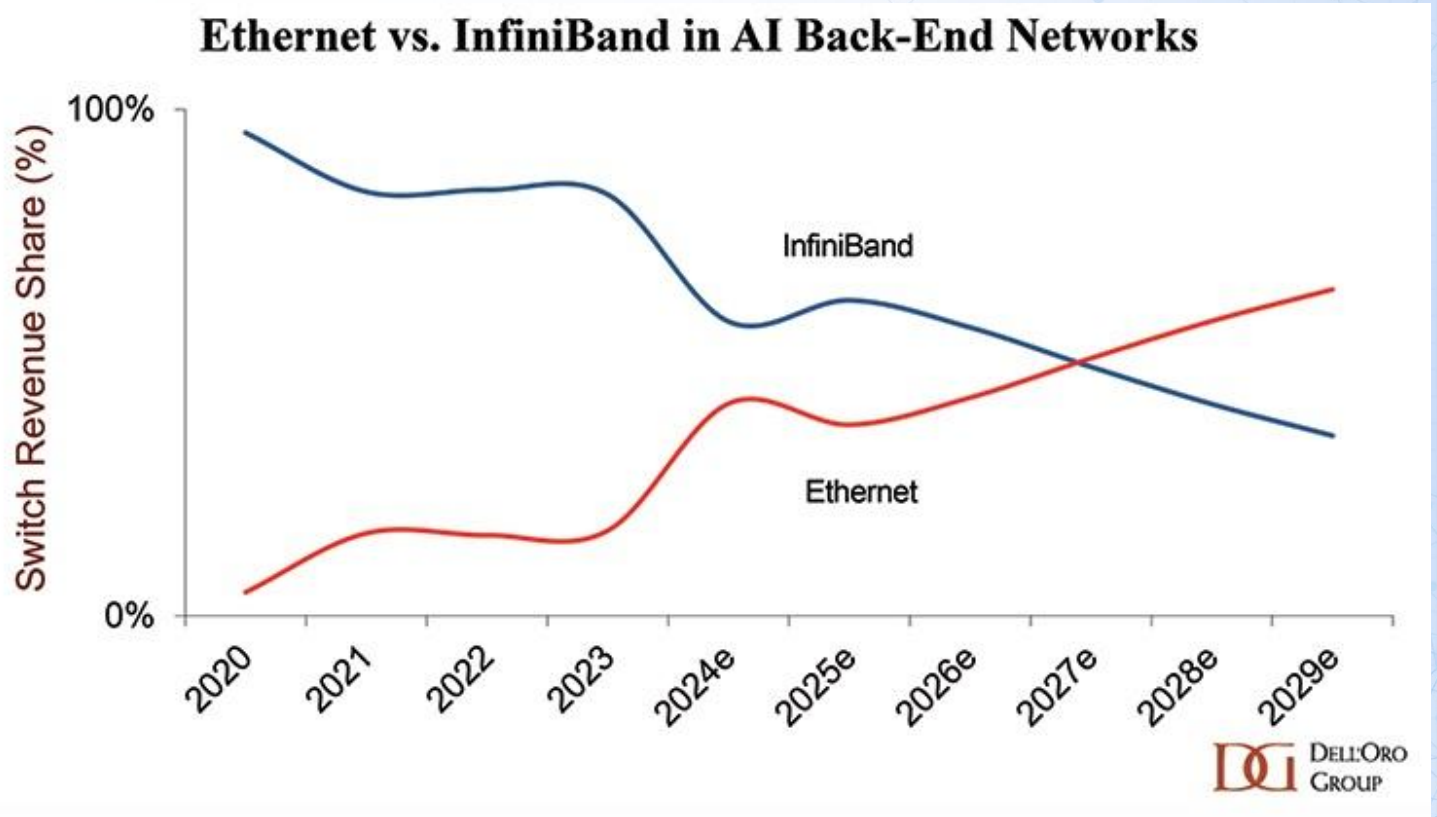
議論

- そのAIファブリックオープンですか？？
- バックエンドネットワークの構築はイーサネット? InfiniBand?
- スケールアップネットワークのリンクは何を使ってますか？

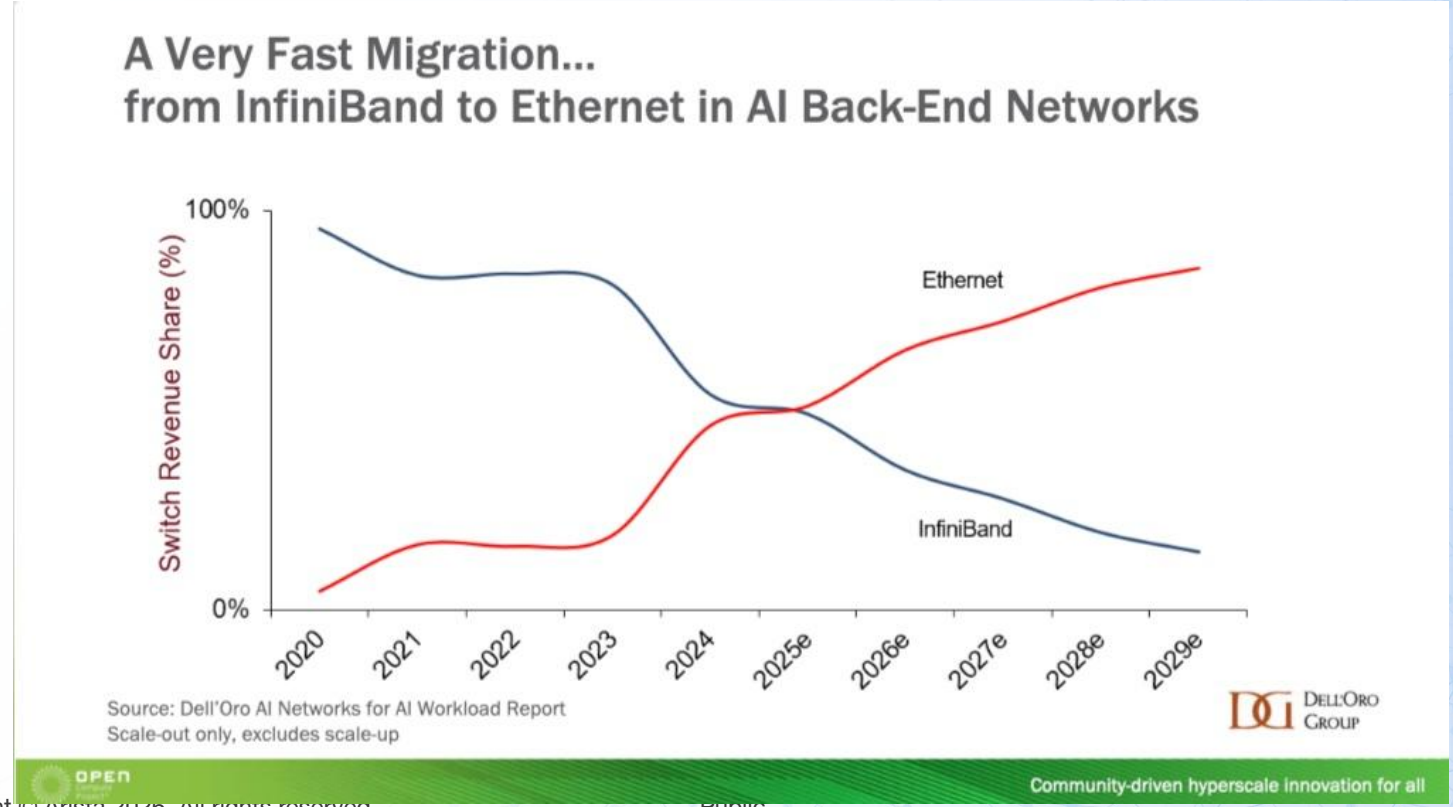
少なくともスケールアウトネットワークのイーサネット化は進んでる 2024年10月の予想



少なくともスケールアウトネットワークのイーサネット化は進んでる 2025年2月の予想



少なくともスケールアウトネットワークのイーサネット化は進んでる 2025年8月の予想



参考

- Demystifying Ultra Ethernet
  - https://www.arista.com/assets/data/pdf/Whitepapers/Demystifying-Ultra-Ethernet-WP.pdf
- Ultra EthernetTM Specification v1.0
  - https://ultraethernet.org/wp-content/uploads/sites/20/2025/06/UE-Specification-6.11.25.pdf
- The Ultra Ethernet Consortium Launches Specification 1.0
  - https://www.youtube.com/watch?v=jfC-1u8BR4Y
- Scale-Up Ethernet Framework
  - https://docs.broadcom.com/doc/scale-up-ethernet-framework
- #TechUpdate: Broadcom's Tomahawk Ultra — Reimagining Ethernet Switching
  - https://www.youtube.com/watch?v=h4RQbh4931M

ARISTA

Thank You

www.arista.com