

AIインフラ解説資料（AI/HPC & UEC技術概要）

Table of Contents

AIインフラ解説資料の読み進め方	2
既存のAI/HPC関連ネットワーク技術	3
参考文献	3
過去のJANOGセッション	3
AI/HPCネットワーク関連（日本語）	4
AI/ML関連：その他（日本語）	5
UEC概要	5
設立趣旨	5
メンバー企業のタイプ	5
組織と運営方法	6
Working Groups	7
普段の活動	7
UECの活動を知る方法	8
UEC BLOG	8
UECが参加したカンファレンスの一覧（スライドやビデオのアーカイブ）	9
UEC関連団体	10
UEC取り組みの背景	11
"UEC取り組みの背景" に含まれる情報について	11
イーサネットのアドバンテージ	11
マルチパスとパケットスプレー	12
Flexible Ordering（柔軟な配信順序）	12
AIやHPCに最適化された輻輳制御メカニズム	13
エンドツーエンドのテレメトリー	14
大規模化、安定性、信頼性	14
ロスレスファブリックの課題	14
UEC Technology Overview（全体像やトピック毎の詳細解説）	15
UEC技術の全体像	15
UEC Profile	16
将来的に検討が進められている技術（公開情報のみ記載）	16
レイヤー毎のUEC技術一覧	16
UET：AIやHPC向けトランスポートプロトコル	17
Packet Spray (Posted Buffer)	18
Packet Trimming (optional)	19
UETの輻輳制御アルゴリズム	19

Ephemeral Connection (エフェメラルコネクション)	20
イーサネットの拡張機能 (LLR, CBFC, LLDP)	21
Link-Layer Retransmission (optional)	21
In Network Collectives (INC)	21
UECを利用するための検討ポイント	23
まとめ	23
UECを利用することによる変化や注意点	23
技術スタックの変化	24
デバイス構成による利用可能な技術の比較	24
UEC技術と関連するコンポーネント (まとめ)	25
UECを利用する際の検討ポイント	26
Infiniband に対する Ethernet を利用する理由	26
NVIDIA Suites を利用しない理由	27
UECとRoCEの比較	28
既存環境でのチューニングの必要性について	28
UEC仕様や製品のロードマップ	28
UEC対応製品（出荷前の製品を含む）	29
AMD	29
Broadcom	30
Marvell Technology	30
Marvel NIC	30
Arista	30
Asterfusion	30
Cisco Nexus 9000 Series Switches	31
Mercury AI-SuperNIC	31

AIインフラ解説資料の読み進め方

JANOG55 AIインフラ解説資料 (PDF) は [janog55-ai-infra.pdf](#) からダウンロード可能です。

AIワーカロードやインフラに関する理解度に応じて、以下のように読み進めていただくと良いでしょう。

- 既存のAI/HPC関連ネットワーク技術をまず学習したい人は、最初の "既存のAI/HPC関連ネットワーク技術" から、"UEC概要"、"UEC Technology Overview" と順に読み進めてください。
- 既存技術について十分理解している人は、"UEC概要" から読み始めてください。
- UECの背景や組織ではなく、UECの技術にのみ興味ある方は "UEC Technology Overview" から読み始めてください。
- UEC技術や製品ロードマップ、および検討のポイントをざっと把握したい人は "UECを利用するための検討ポイント" を参照してください。

注意

本資料に含まれるUEC関連情報は（2025年1月）時点での公開情報をもとに執筆しています。 UEC

技術の利用検討や評価を実施する際には、公式サイトの情報や、2025年前半に一般公開予定の "UEC Specification (UEC仕様書)" を参照してください。

既存のAI/HPC関連ネットワーク技術

既存のAI/HPC関連ネットワーク技術は、大きく分類するとロスレスイーサネット、輻輳制御、ロードバランシング、の3つを実現する技術から構成されます。これらの既存技術に関してはJANOGなどで解説されていますので、UECの技術を理解するための前提知識として "参考文献" に記載した資料を一読しておくと良いでしょう。

ロードバランシング技術は様々な技術がありますが、ベンダを問わずサポートされているのはDLBです。

Table 1. 既存のAI/HPC関連ネットワーク技術

機能	代表的な技術
ロスレスイーサネット (Lossless Ethernet)	<ul style="list-style-type: none">PFC (802.1Qbb, Priority Flow Control)
輻輹制御 (Congestion Control)	<ul style="list-style-type: none">DCQCN (ECN, Sender-driven)
ロードバランシング (Load Balancing, Fabric utilization)	<ul style="list-style-type: none">Dynamic Load Balancing (flowlet)NVIDIA Adaptive Routing (flowlet?)Packet Spray (Packet based)Scheduled Fabric (VoQ) (Cell based)

参考文献

過去のJANOGセッション

- JANOG54 生成AI向けパブリッククラウドサービスをつくってみた話
 - <https://www.janog.gr.jp/meeting/janog54/sakura/>
 - 井上 喬視, 高峯 誠, 平田 大祐 さくらインターネット株式会社
- JANOG54 ネットワークオペレーションにおける生成AI技術の活用検討について
 - <https://www.janog.gr.jp/meeting/janog54/genai/>
 - 佐藤 亮介, 白井 嵩士, 田口 順史, 近藤 優吉 株式会社NTTフィールドテクノ
 - 宮坂 拓也, 株式会社KDDI総合研究所
 - 仲松 匠, KDDI株式会社
 - AI/MLネットワークではなくAIをネットワーク運用に活用する内容。GPTの概要を解説している。
- JANOG53 AI(人工知能)の為のネットワーク
 - <https://www.janog.gr.jp/meeting/janog53/ainw/>
 - 土屋 師子生, アリスタネットワークスジャパン合同会社
- JANOG52 AI/ML基盤の400G DCネットワークを構築した話

- <https://www.janog.gr.jp/meeting/janog52/aiml400/>
 - 内田 泰広, 小障子 尚太朗, 株式会社サイバーエージェント
- JANOG50+ パケットロスと遅延
 - <https://www.janog.gr.jp/meeting/janog50plus/docs/janog50plus-maz-losslatency.pdf>
 - 松崎 吉伸(株式会社インターネットイニシアティブ)
- JANOG43 LINEのネットワークをゼロから設計した話
 - <https://www.janog.gr.jp/meeting/janog43/application/files/7915/4823/1858/janog43-line-kobayashi.pdf>
 - Masayuki Kobayashi, LINE Corporation

AI/HPCネットワーク関連（日本語）

- GPUクラスタネットワークとその設計思想 (Rethinking AI Infrastructure Part 2)
 - <https://techblog.lycorp.co.jp/ja/20250115a>
 - LINEヤフー株式会社 小林、深澤
- "GPUネットワーク設計・運用 基礎勉強会 Lossless Ethernet - PFC/ECN編"
 - <https://speakerdeck.com/markunet/ecnbian>
 - LINEヤフー株式会社 小林正幸
- AI時代のデータセンターネットワーク
 - https://speakerdeck.com/lycorptech_jp/dcnw_in_the_ai_era
 - 第40回 情報ネットワーク・ネットワークシステム研究ワークショップ
 - LINEヤフー株式会社 小林正幸
- EthernetベースのGPUクラスタ導入による学びと展望
 - https://speakerdeck.com/lycorptech_jp/20241202
 - NVIDIA AI Summit Japan 2024
 - LINEヤフー株式会社 小林正幸、道下幹也
- Podcast: fukabori.fm "124. AI時代のGPUクラスタ、DCネットワーク"
 - <https://fukabori.fm/episode/124>
 - LINEヤフー株式会社 小林正幸、道下幹也
- PFNにおけるアクセラレータ間通信の実際 / MPLS Japan 2024
 - <https://speakerdeck.com/pfn/mpls-japan-2024>
 - Yuichiro Ueno / Preferred Networks, Inc.
- "イーサネットが高信頼ネットに進化する, [3] 中核はフロー制御をつかさどる「PFC」と「ETS」"
 - <https://xtech.nikkei.com/it/article/COLUMN/20091006/338383/>
 - 2009-10-15 日経クロステック

AI/ML関連：その他（日本語）

- Distributed Cache Empowers AI/ML Workloads on Kubernetes Cluster / KubeCon + CloudNativeCon North America 2024
 - <https://speakerdeck.com/pfn/kubecon-plus-cloudnativecon-north-america-2024>
- 生成AI向け機械学習クラスタ 構築のレシピ - 北海道石狩編
 - <https://speakerdeck.com/pfn/20240615-cloudnatedayssummer-pfn>
 - Cloud Native Days Summer 2024
- 実際に運用してわかった！多種GPU混載Kubernetesクラスタの使われ方と運用省力化
 - <https://speakerdeck.com/pfn/cloudnatedaystokyo23-shiota>
 - Tetsuya Shiota 株式会社PreferredNetworks
 - CNDT2023: CloudNative Days Tokyo 2023
- (2023年末) LLMの現在 - 今のLLMを取り巻く状況について紹介します。
 - <https://speakerdeck.com/pfn/llmnoxian-zai>
 - Preferred Networks, いもす (imos@preferred.jp)

UEC概要

設立趣旨

Ultra Ethernet Consortium (UEC) は、AIやHPCの要求に応える高性能なイーサネット通信を実現するために設立された業界団体です。

[top500.org](#) の TOP500 LIST - NOVEMBER 2004 ^[1] にランクインされているHPCシステムでもGPUが利用されているなど、今後AIとHPCのワークロードおよびネットワーク要件が重複していくことが予想されています。そのため、UECではAIに限定せず、AIとHPCワークロード (RoCEv2) 両方の課題を解決するプロトコルや技術の "オープンスタンダード" を "マルチベンダー" で開発することを目的としています。

UECは、2023年に AMD, Arista, Broadcom, Cisco, Eviden (an Atos Business), HPE, Intel, Meta, Microsoft により設立され ^[2]、2024年12月 現在では、合計90社のメンバー企業で構成されています。

NVIDIAによるアプリケーション、アクセラレータ (GPU) 、ネットワークの垂直統合による独占に対抗するために設立されたと言われることもありますが、2024年からNVIDIAもメンバー企業として活動に参加しています。

UECは Linux Foundation 傘下の JDF (Joint Development Foundation) ^[3] の一部として組織されています。 JDFは、企業や団体が共同で技術標準やオープンソースプロジェクトを迅速に立ち上げるための法的および運営的な基盤を提供する団体で、共同開発を効率的に推進するための仕組みを提供しています。

メンバー企業のタイプ

メンバー企業や組織は Steering Members, General Members, Contributor Members の3種類に分類されます。新規加入企業が選択できるのは General と Contributor のいずれかです。

2025年1月1日現在、10社の Steering Members、26社の General Members、54社の Contributor Members、合計90社のメンバー企業で構成されています。

Steering Members

メンバーは AMD, Arista, Broadcom, Cisco, Eviden, HPE, Intel, Meta, Microsoft, Oracle であり、共同設立企業 + Oracleで構成されています。

General Members

General Members はTACへの参加が可能で、新しいワーキンググループの立ち上げやワークアイテムの作成など、UECの活動内容について影響を与えることが可能です。 また、過半数よりも多い（2/3 や3/4など）賛成が必要な議題への投票（Supermajority）権利も持ります。

Contributor Members

Contributor Members は General Members と異なり新しい活動の提案はできませんが、アクセス可能な情報や技術貢献も可能で、Supermajorityを除く投票にも参加する権利を持ちます。 そのため、通常は General Members と変わらない活動が可能です。

日本企業も数社メンバー企業として参加しています。

UECメンバーの日本企業

- General Member: Preferred Networks
- Contributor Member: Fujitsu, IIJ

メンバー企業の一覧は、UEC Webサイトのトップページ <https://ultraethernet.org/>^[4] で参照可能です。

組織と運営方法

UECは以下のような組織から構成されます。 技術的には、各 Working Groups の活動状況や、その結果として広報されるホワイトペーパーやBLOG、仕様書などを参照すると良いでしょう。（仕様策定プロセスの詳細は、内部情報のため省略しています）

Steering Committee (SC)

組織運営を統括し、UEC全体の方向性を決定します。

Marketing Committee

外部とのコミュニケーションに責任を持ち、イベントなどのコーディネートを行います。

Technical Advisory Committee (TAC)

SCの元で、UECの技術活動の範囲や優先順位を定め、監督します。 また、ハイレベルな技術ロードマップや、ユースケースのスコープや技術目標を定めます。 各 Working Group が提案する技術仕様を承認し、UECの仕様全体の整合性を保ちます。 TACに参加するには General Members の会員タイプである必要があります。

Working Groups

技術分野毎に設立され、UEC技術仕様のうち、担当する技術分野の仕様検討や仕様書を作成します。

Working Groups

2024年12月 現在の Working Group と活動内容は以下の通りです。 [5]

Table 2. UEC Working Group と活動内容 (2024年12月 現在)

Working Group	活動内容
Physical Layer (PH)	物理レイヤーでの、遅延の低減、管理改善、などに取り組んでいます。イーサネット物理層、電気および光信号特性、APIやデータ構造の仕様策定を行います。
Link Layer (LL)	リンクレイヤーでの、遅延の低減、管理改善、などに取り組んでいます。イーサネットの効率、セキュリティ、スケーラビリティを最適化する仕様策定を行います。
Transport Layer (TR)	エンドツーエンドのデータ配信に不可欠な、トランSPORTレイヤーでのスループットの向上、遅延の低減、スケーラビリティの向上、管理の改善を実現するトランSPORTの仕様策定を行います。UEC技術の中核となる Ultra Ethernet Transport (UET) の仕様策定を行っています。
Software Layer (SL)	AI/HPCの幅広いユースケースとアプリケーションをサポートするため、ソフトウェアAPIを含む技術仕様の策定や、オープンソース実装の開発を行います。技術分野としては、リモートメモリアクセスの最適化、INC (In Network Collective) の実現、セキュリティ、管理改善、ストレージを含みます。
Storage Working Group (ST)	UECベースのAI/HPCワークロードで利用可能なストレージサービスを、他 Working Groups と協力しながら取り組みます。新たな仕様策定だけでなく、既存ストレージ管理のベストプラクティスの取り込みを含みます。
Compliance Working Group (CT)	サービスとデバイスがUECの定義した技術に適合していることを確認することに取り組みます。UEC実装を評価するためのテストを作成し、UEC標準への厳格な準拠を保証します。また、UECが定義したAI/HPCプロファイルに従って、UEC準拠のネットワークデバイス (NIC、PCIe NIC、スイッチなど) 間の相互運用性目標を定義します。
Management Working Group (MG)	UEC Fabricの管理性を強化するための仕様策定を行います。UEC準拠のTransport Fabric End Point (FEP) のモデル、管理エレメント、RPC (Remote Procedure Call)などを定義します。トポロジー・ディスカバリー、ケイパビリティ・ディスカバリー、モニタリング、インターフェラビリティ・クエリーなども含みます。
Performance and Debug Working Group (PD)	進化するUECの仕様に合わせ、AI/HPCワークロードの性能指標、ベンチマーク、デバッグ機能、ツールを定義します。また、UEC準拠の実装における可視性とデバッグ可能性を強化することで、開発者、DevOps、ネットワーク運用チームを支援します。

普段の活動

普段の活動は各 Working Group 毎に仕様検討や仕様書を作成し、定期的にTACがレビュー承認することで全体の整合性を保ちながら仕様策定を進めています。

普段の議論はメーリングリストとオンライン会議など、オンラインで技術検討や仕様策定は進んでいくためリモートでの参加が可能です。但し、オンライン会議は日本時間では深夜早朝にあたる北米の日中に行われています。

Working Group よっては、年数度のオフライン会議で集中的に課題を議論する場合もあります。ま

た、2025年はメンバー会合が開催される予定です。

2025年前半に最初の仕様が公開された後は活動頻度や形態も変わっていく可能性がありますので、メンバーとして加入を検討する際には最新の活動状況を確認すると良いでしょう。

UECの活動を知る方法

メンバー企業に所属していない人でも、以下方法でUECの活動を知ることが可能です。

- カンファレンス
 - カンファレンスのセッションを聴講することで、活動状況や技術に関する最も詳細な情報に触れることが可能です。
 - また、展示がある場合は展示会場でUECメンバーと会話することが可能です。
 - 詳細は、表 [UECが参加したカンファレンスの一覧（スライドやビデオのアーカイブ）](#) を参照してください。
- Webサイト：<https://ultraethernet.org/>
 - Webサイトには、UECの組織構成、メンバー企業の一覧、Working Groups、などの情報が記載されています。
 - UEC設立のモチベーションと、策定中の仕様（概要）が記載されたホワイトペーパー^[6]をダウンロード可能です。
 - Blog: Latest News^[7]には最新情報が掲載されますので、定期的にチェックすると良いでしょう。
- メンバー企業のプレスリリースやWebサイト
 - メンバー企業がUECでの活動について発信する場合があります。
 - 特にUECに対応した製品の最新情報については、各メンバー企業からの発信される情報を参照すると良いでしょう。
- LinkedIn: <https://www.linkedin.com/company/ultraethernet/posts/>
 - LinkedInのUECアカウントのポストを追うことで、UECが参加するカンファレンスや発信した情報を知ることができます。
 - また、メンバー企業や他組織の投稿をリポストする場合もあります。全てのメンバー企業の情報をチェックするのは大変ですので、まずはUECの LinkedIn アカウントをフォローするのも良い方法です。

UEC BLOG

UECの活動や技術内容を理解するために参考になりそうなBLOG記事を抜粋しました。

- March 18, 2024: UEC Progresses Towards v1.0 Set of Specifications
 - <https://ultraethernet.org/uec-progresses-towards-v1-0-set-of-specifications/>
- August 29, 2024: Ultra Ethernet Specification Update
 - <https://ultraethernet.org/ultra-ethernet-specification-update/>
- November 14, 2024, Interview of UEC Chair, J Metz (By Next Platform, but also on UEC blog)

- The Collaboration That Will Drive Ethernet Into The HPC And AI Future
- Part#1: <https://www.nextplatform.com/2024/11/12/the-collaboration-that-will-drive-ethernet-into-the-hpc-and-ai-future/>
- Part#2: <https://www.nextplatform.com/2024/11/14/uec-doesnt-want-to-kill-infiniband-but-it-wants-ethernet-to-beat-it/>

UECが参加したカンファレンスの一覧（スライドやビデオのアーカイブ）

Table 3. UECが参加したカンファレンスの一覧（スライドやビデオのアーカイブ）

カンファレンス	リンク
OCP2024 (Open Compute Project)	<p>https://www.opencompute.org/events/past-events/2024-ocp-global-summit</p> <ul style="list-style-type: none"> • Leveraging UEC for Next Generation AI Networks (video 37min) <ul style="list-style-type: none"> ◦ Presented by UEC: Uri Elzur, Intel (Architect, Computer Networks, GPU Networks) • Overview of Ultra Ethernet (video 15min) <ul style="list-style-type: none"> ◦ Presented by UEC, J Metz, AMD (Chair Steering Committee, UEC) • Accelerating AI/HPC: OCP and UEC's Collaborative Vision for High-Performance Networking (video 22min, slides) <ul style="list-style-type: none"> ◦ J Metz, AMD (Chair Steering Committee, UEC) ◦ Uri Elzur, Intel (Architect, Computer Networks, GPU Networks) • Future of AI Networks: UAI and Ultra Ethernet (video 26min, slides) <ul style="list-style-type: none"> ◦ J Metz, AMD (Chair Steering Committee, UEC) ◦ Kurtis Bowman (AMD, Director, Architecture and Strategy)
SC24 (Super Computing)	<p>https://sc24.supercomputing.org/program/proceedings-archives/</p> <p>注：アーカイブへのアクセスは有料の登録者のみ</p> <ul style="list-style-type: none"> • Industry Standards Working Together to Accelerate Innovation in AI and HPC <ul style="list-style-type: none"> ◦ https://sc24.supercomputing.org/proceedings/panel/panel_pages/pan114.html
NANOG 92 (2024年10月)	<p>https://nanog.org/events/nanog-92/</p> <ul style="list-style-type: none"> • Keynote: Networking for AI and HPC, and Ultra Ethernet <ul style="list-style-type: none"> ◦ Hugh Holbrook, Arisa Networks (VP Software Engineering) ◦ Video: https://youtu.be/0roIi1pscts?si=XmZAjfBFM3CibWBb ◦ Slide: 20241021_Holbrook_Keynote_Networking_For_v1.pdf

カンファレンス	リンク
HOTI31（2024年） (Hot Interconnects)	<p>https://ieeexplore.ieee.org/xpl/conhome/10664198/proceeding</p> <ul style="list-style-type: none"> • Day 2: Invited Talk: Ultra Ethernet Consortium (UEC) overview <ul style="list-style-type: none"> ◦ Uri Elzur, Intel ◦ Video: https://www.youtube.com/watch?v=LtifmYChRTo

UEC関連団体

UECはLinux Foundation以外にも、以下の団体と連携をとりながら仕様策定を進めています。連携の密度は団体毎に濃淡がありますが、UECの技術を深く理解したい場合には、これら団体の動向や技術も参考にすると良いでしょう。

Open Compute Project (OCP)

OCPはコンピュート・インフラへの要求を効率的にサポート可能なハードウェアの仕様や設計のオープンソース化を進める非営利団体です。2011年にFacebook(元Meta)が自社設計のデータセンター機器の設計仕様書を公開したことに始まります。公開された仕様書を元にハードウェアベンダが機器を提供可能にしたことにより、コスト削減と調達のしやすさが向上しました。サーバーやストレージ、ネットワークなどの製品分野ごとにプロジェクトが存在し、参加企業による新規開発や設計仕様書の公開が進んでいます。ネットワークの分野の例としては、ホワイトボックススイッチやそれを制御するスイッチOS (SONiC等) の開発が行われています。（SONiCは2022年にLinux Foundationに移籍しました）年1回の OCP Global Summit を始め、OCPのカンファレンスでは多くのUEC関連セッションが開催され、UECの動向理解の参考になります。

- OCPホームページ：<https://www.opencompute.org/>
- OCP Global Summit (スライドや動画アーカイブ)：<https://www.opencompute.org/summit/global-summit>

OpenFabrics Alliance (OFA)

The OpenFabrics Alliance (OFA) は、高性能ネットワーキング技術を推進する業界団体です。主にデータセンター、ハイパフォーマンスコンピューティング (HPC)、クラウドインフラストラクチャ向けに、RDMA (Remote Direct Memory Access) やInfiniBand、iWARP、RoCEなどのネットワーキング技術の普及を支援しています。また、オープンソースソフトウェア (OFED: OpenFabrics Enterprise Distribution) を提供しています。OpenFabrics Interfaces Working Group では、Open Fabrics Interfaces (OFI) とも呼ばれる **libfabric**^[8] ライブラリを開発しています。

- OFAホームページ：<https://www.openfabrics.org/>

ULTRA ACCELERATOR LINK (UALink)

Ultra Accelerator Link™ (UALink™) は、アクセラレーター間を接続するオープンな業界標準の内部インターフェクトです。2024年の時点では、UECはイーサネットを外部インターフェクトとして利用したスケールアウト技術を策定しており、スケールアップに必要な内部インターフェクト技術はUALinkなど、UECではなく、他団体の技術を活用する前提となっています。

- UALink Consortiumホームページ：<https://www.ualinkconsortium.org/>

Storage Networking Industry Association (SNIA)

SNIAは、ストレージを中心とした情報管理に関する技術標準を開発や教育プログラムを提供する業

界団体です。 UEC Storage Working Groupが存在するように、UECの技術はストレージにも活用されることを想定し、技術策定を進めています。

- SNIAホームページ : <https://www.snia.org/>
- SNIA日本支部 : <https://www.snia-j.org/>

IEEE

IEEE（アイ・トリプル・イー）にはイーサネットに関連する技術を扱うワーキンググループが存在します。 例えば IEEE 802.3 は、有線イーサネットの物理層とデータリンク層のメディアアクセス制御（MAC）を定義するワーキンググループです。 また、IEEE 802.1 では、UECで拡張が検討されている LLDP (IEEE 802.1AB) の標準を策定しています。

- IEEEホームページ : <https://www.ieee.org/>

UEC取り組みの背景

本セクションでは、UECホワイトペーパー^[9]の内容を元にUECの目的や取り組んでいる技術について紹介します。

"UEC取り組みの背景" に含まれる情報について

"UEC取り組みの背景" で紹介している内容は、2023年にUECから発表された "UECホワイトペーパー"^[10]を元に執筆しています。 また、その他の内容も、UECホームページやカンファレンスでの発表資料など、公開情報を元にまとめています。

UECでは2025年Q1のリリースに向けて仕様策定が進められており、様々な議論が行われている状況です。そのため、仕様がリリースされるまでの間に、記載の内容から変更がある可能性があります。 UEC技術の利用検討の際には、UEC技術仕様を含め最新の情報に当たるようにしてください。

イーサネットのアドバンテージ

イーサネットのアドバンテージとして、以下のような項目が挙げられています。

- 複数のベンダーが参加するエコシステムにより、互換性のあるイーサネットスイッチ、NIC、ケーブル、トランシーバー、光学デバイス、管理ツール、ソフトウェアが広く提供されている。
- IPネットワーク（ルーティング）の拡張性が実証されており、ラック規模、建物規模、データセンター規模のネットワークを実現可能。
- イーサネットネットワークをテスト、測定、展開、効率的に運用するための幅広いツールが存在する。
- 競争のあるエコシステムと規模の経済により、コスト削減が実現してきた実績がある。
- IEEEイーサネット標準が、多くの物理層および光学層において迅速かつ定期的に進化を遂げる能力が実証されている。

このようなイーサネットのアドバンテージを活用し、

- スループットの最大化
- Tail Latency の最小化

といった目標を達成するために必要なネットワークの要件として以下を挙げ、実現する技術を策定しています。

- マルチパスとパケットスプレー
- Flexible Ordering（柔軟な配信順序）
- AIやHPCに最適化された輻輳制御メカニズム
- エンドツーエンドのテレメトリ
- 大規模化、安定性、信頼性

マルチパスとパケットスプレー

従来のイーサネットネットワークでは、スパンニングツリープロトコルにより1つの経路しか利用できない状況が続いてました。その後、データセンターファブリックでは、ノード（スイッチ）間を Layer 3 (IP) で接続する事により、Equal-Cost Multipath (ECMP) のような技術を利用し複数のパスを利用できるようになりました。ECMPは通常ハッシュを用いてトラフィックの振り分けをするため、トラフィックの偏りが生じるデメリットがあります。また、パケットのリオーダーを防ぐために、同じフロー (Layer 4: TCP, UDP セッション) のトラフィックは同じパスを通す必要があります。そのため、フロー数が少ないとハッシュ計算の元となるヘッダ情報が同じトラフィックが増える事により、偏りが顕著になります。

近年では Dynamic Load Balancing (DLB) のように各ポートの使用率などから動的に利用するパスの振り分けを行う機能もありますが、ノードが多段になった場合、ネットワーク全体で最も負荷の低いパスを確実に選択する事は困難です。

そのため、UECでは全てのパスにパケットを分散させる "packet spraying"（パケットスプレー）という技術を採用しています。パケットスプレーの難点はパケットのリオーダーが発生する事ですが、これにはトランスポート層 (UET) で対応しています。

Flexible Ordering（柔軟な配信順序）

RoCEv2など、従来の技術ではパケットが送信された順に受信する必要があります。具体的には、パケットのリオーダーやパケットロスによりアウトオブオーダーパケットが発生すると、Go-Back-N により順序通り届かなかったパケット以降の全てのパケットを再送する必要があります。これにより、利用可能なスイッチ間リンクの利用率低下やテールレイテンシの増加が発生し、ジョブ完了までにより多くの時間がかかります。

しかし、理想的にはすべてのリンクが使用され、AIのワークロードがそれを必要とする場合にのみ順序が強制されるべきです。

AIワークロードにおける集団通信の操作は、All-Reduce や All-to-All が多くを占めます。これらの操作完了を早くする鍵は高速なバルク通信（データ転送）であり、AIアプリケーションにとって "メッセージの全てのデータが届いたか？" が重要であり、データの到着順序は重要ではありません。

Flexible Ordering は、パケット到着順序の制約を緩和する事で、データ転送を効率化します。例えば、パケットスプレーを行った時に発生するアウトオブオーダーパケットの理オーダリングが不要になります。

AIやHPCに最適化された輻輳制御メカニズム

輻輳が発生する場所は、以下3か所に分類できます。

1. 送信元サーバーとスイッチ間のリンク
2. スイッチファブリックの内部（送信元と送信先サーバーが接続されているスイッチ間の経路）
3. 送信先サーバーとスイッチ間のリンク（Incast）

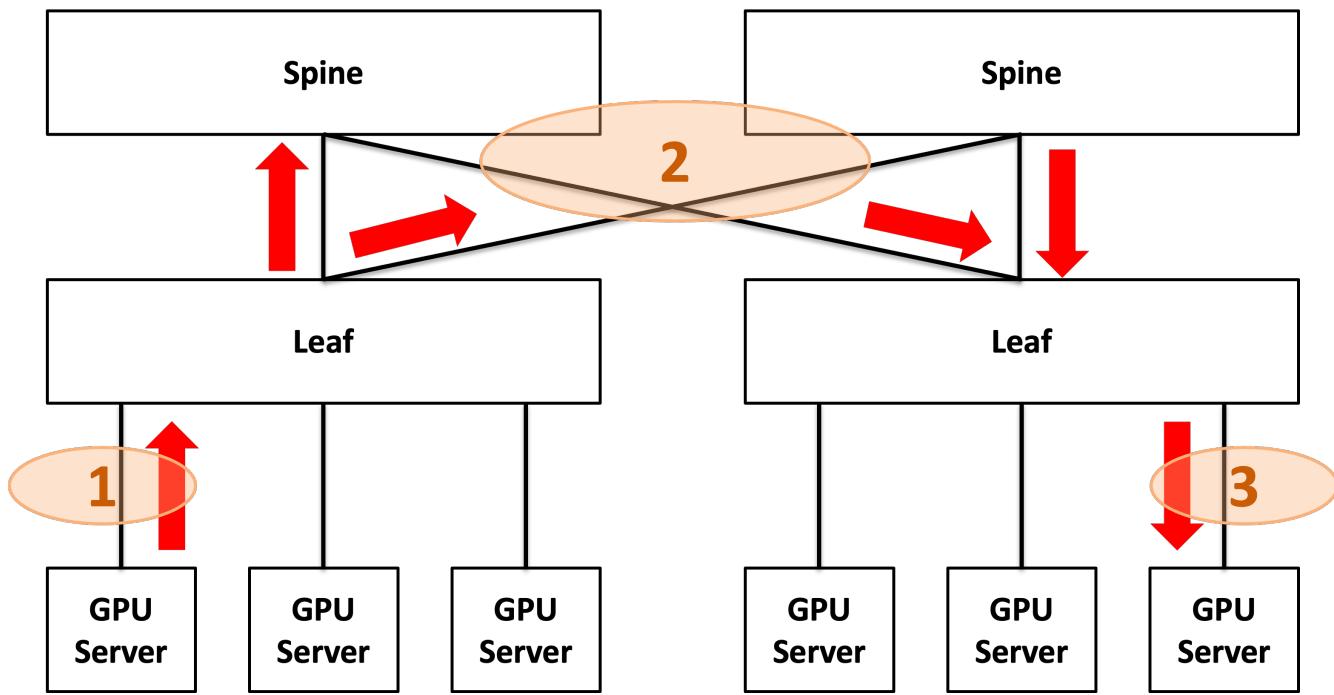


Figure 1. 輻輳発生場所の分類

"1. 送信元サーバーとスイッチ間のリンク" の輻輳は、送信元サーバーが全ての送信トラフィックを把握可能なため、制御が可能です。

"2. スイッチファブリックの内部" の輻輳はマルチパスとパケットスプレーにより最小化されます。

"3. 送信先サーバーとスイッチ間のリンク（Incast）" は All-to-All 操作など、複数の送信元から同じ送信先にデータ転送される際に発生します。この Incast による輻輳制御のために DCQCN 等のアルゴリズムが用いられてきましたが十分ではなく、以下のようないくつかの要件を満たす輻輳制御メカニズムが必要と考えられます。

転送レートの立ち上がりの速さ

広帯域かつ低遅延のネットワークで、既存のトラフィックのパフォーマンスを低下させることなく、転送開始時に速やかにワイヤ・レートまで送信速度を上げること。

最終リンクの公平な共有

パケットロスや再送信、テールレイテンシの増加を招くことなく、最終リンクを公平に共有することで Incast を制御可能のこと。

設定やチューニングの最小化

トラフィックミックスの変化、コンピュートノードの進化、リンク速度の向上、ハードウェアの進化に応じて、(DCQCNでは必要であった) チューニングや設定を必要としない（最小化する）こと。

UECでは、これらの要件をサポートし、マルチパスやパケットスプレーと連動するような将来のAIワークロードのための輻輳制御アルゴリズム（UET）を設計しています。

注

将来的には、マルチテナンシーサービスを実現するために、送信元サーバーの仮想化やネットワークのパーティションが必要になった場合など、輻輳発生場所が変わってくる可能性があります。

エンドツーエンドのテレメトリ

輻輳制御アルゴリズムの最適化は、エンドツーエンドのテレメトリによって実現されます。送信元もしくは送信先が転送スケジュールを管理する際に、ネットワークノード（スイッチ）からパケット送信スケジューラーやペーサーへの輻輳通知を迅速に行うことにより、輻輳制御アルゴリズムの反応性や正確性が向上します。

これにより、輻輳が緩和され、パケットの取りこぼしが減り、キューが小さくなるなど、テールレイテンシの改善が可能になります。

大規模化、安定性、信頼性

UECでは、100万のエンドポイントまで対応可能なスケーラビリティを目指しています。

口レスファブリックの課題

口レスイーサネットを前提としたRoCEは広くデプロイされてきましたが、最大限の性能を引き出すためには専門家によるチューニングや運用監視が必要であり、総所有コスト（TCO）の削減が困難です。

また、口レス実現のためPFCを用いたパケット転送の一時停止には、以下のような課題があります。

HOLブロッキング (Head-Of-Line Blocking)

- HOLブロッキングとは、キュー内の先頭パケットが転送されるポートが輻輳している場合、他のパケットが転送可能であってもキュー全体がブロックされる現象です。
- PFCは特定の優先度のトラフィックを一時停止させるため、これがキュー内での先頭データの滞留を引き起こし、後続のトラフィックもブロックされます。
- これにより、スループットの低下、遅延の増加、といった影響があります。

輻輳伝播 (congestion spreading)

- あるリンクで輻輳（例：バッファが満杯になる）が発生すると、そのリンクはデータを受信できなくなり、上流デバイスに PAUSE フレームを送信します。
- その結果、上流デバイスも同様に輻輳が生じ、さらにその上流に PAUSE フレームを送信します。
- このように輻輳の伝播が発生し、輻輳の影響が最初に発生した一部のリンクに留まらず、ネットワーク全体に波及することができます。

デッドロックの発生

- 複数のデバイスが互いに PAUSE フレームを送り合う状況に陥ると、ネットワーク全体が停止する "デッドロック" が発生することがあります。

UECは、これらのパケットロスの防止やリカバリ、輻輳制御といった課題はトランスポート層で解決されるべきと考え、RoCEv2を置き換える UET (Ultra Ethernet Transport) 、UETと協調して利用されるネッ

トワーク層の技術、イーサネットの拡張技術（リンク層）、In Network Collectives (INC)、等の開発を進めています。

UEC Technology Overview （全体像やトピック毎の詳細解説）

本章では、最初にUECが取り組む技術の全体像を俯瞰し、各レイヤーでどのような技術の開発が進められているかを紹介します。

その後、更に技術を深掘りした人向けに、UECで仕様策定が進められている各技術の詳細を解説します。

UEC技術の全体像

UECでは、イーサネットをAI/HPCアクセラレータのノード間インターフェクトとして利用する際に重要なRMAを高性能かつ効率的に行うために、トランスポートレイヤーからリンクレイヤーにかけ、幅広いレイヤーで技術開発が行われています。その中には、既存の RMA over Ethernet 技術であるRoCEv2の運用経験からくる改善点を含みます。

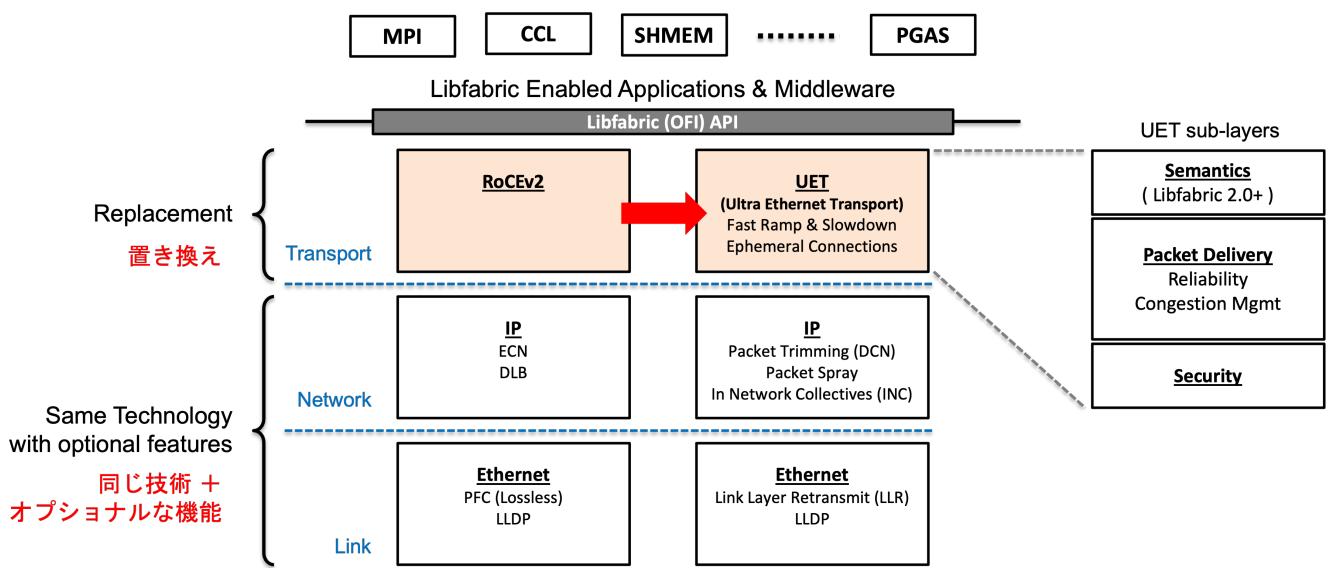


Figure 2. UECの技術スタックとRoCEv2との比較

最も大きな特徴は、図 : [\[uec_tech_stack\]](#) のようにトランスポートレイヤーをRoCEv2から、新しく策定した UET (Ultra Ethernet Transport) に置き換えたことです。UETでは、モダンなHPC向け RMA API でありHPCアプリケーションで幅広く利用されている libfabric API を採用し、拡張しています。

NOTE 図 : [UECの技術スタックとRoCEv2との比較](#) は Libfabric API を前提としたスタックとなっていることに注意してください。Libfabricを利用しているアプリケーションやミドルウェアはAPI拡張への対応でUEC技術を利用可能な場合もありますが、Libfabricを利用していないアプリケーションやミドルウェアの場合、UETをサポートするために大きな変更（プログラムの改修）が必要となる可能性があります。

UETは名前の通り輻輳制御などを担うトランスポートレイヤーの技術ですが、効率的なトランスポートプロトコルやアルゴリズムを実現するために、ネットワークレイヤーでも新しい技術が開発され、密に連携しながら動作しています。

そのため、ネットワークレイヤーの技術であっても、UETの利用が前提となっています。

但し、UEC技術のインクリメンタルな採用が可能なように、ネットワークレイヤー以下の機能はオプショナルとなっており、既存のスイッチでもUET自体は利用可能です。

その他、In Network Collectives (INC) と呼ばれる、スイッチファブリック内で集合計算の一部を実行する技術も開発されています。また、AI/HPCにより生成されたモデルや、モデル作成に利用するデータのビジネス的な重要性が増していることから、セキュリティ機能も追加されています。（セキュリティ機能に関しては解説は省略しています）

NOTE AI/HPCワークロードでのアプリケーションは、データをパケットでは無くメッセージ単位で送受信（指示）するため、効率的なデータ送受信のためにはメッセージ境界を意識した方式が必要になります。メッセージは1つまたは複数のパケットから構成されるため、メッセージまたはパケットどちら（もしくは両方）の順序が意味を持つのか、順序に依存しないのか、を意識することが重要です。

UEC Profile

UECでは、AIやHPCなど異なるワークロードに対応可能な技術仕様を策定しています。また、前述通り、UEC技術の全てが必須サポートではなく、オプショナルとして定義された技術や機能があります。

UECではプロファイルを定義することで、ワークロードの種類毎に必要な技術や機能を選択的に利用可能になっています。

将来的に検討が進められている技術（公開情報のみ記載）

UECでは2025年Q1に公開予定のversion 1.0以降も開発は続けられ、将来的には以下のようなユースケースや機能の拡張が段階的に行われていく予定です。

- ストレージへの適用
 - Storage APIs on UET
- 管理機能
 - OpenConfig や RedFish による設定管理
- テレメトリー機能
 - Congestion Signaling (CSIG)^[11] や Back to Sender (BTS)^[12]
- その他
 - 性能測定やデバッグ手法の標準化
 - プロファイル毎のコンプライアンステスト（オプショナルな機能を含む）

レイヤー毎のUEC技術一覧

表：[レイヤー毎のUEC技術一覧](#) に、現在検討が進められていることが公表されているUEC技術をレイヤー毎にまとめました。（UETを中心に他レイヤーの技術に依存している機能もありますので、各技術毎に中心となるレイヤーに分類してあります。）

以降、それぞれの技術について概要を解説していきます。

Table 4. レイヤー毎のUEC技術一覧

Layer	Features
Application	(Libfabric と UEC拡張に対応した様々なAI/HPCアプリケーション)
Software APIs	<ul style="list-style-type: none"> • Libfabric 2.0 の拡張
Transport	<ul style="list-style-type: none"> • UET (Ultra Ethernet Transport) <ul style="list-style-type: none"> ◦ Posted Buffer (out-of-order delivery) ◦ Fast Ramp and Slowdown ◦ Ephemeral Connections ◦ Receiver-generated Credit (manage incast) ◦ Optimistic Transmission (before credits received) • Security (encryption, host-level security and authorization)
Network Layer (IP)	<ul style="list-style-type: none"> • Packet Trimming • Packet Spray
Link Layer (Ethernet)	<ul style="list-style-type: none"> • Link Layer Retransmit (LLR) • LLDP (capability negotiation)

UET：AIやHPC向けトランスポートプロトコル

ロスレスファブリックの課題を解決し、ロスレスを必要としないトランスポートプロトコルとして、Ultra Ethernet Transport (UET) が開発されました。

UETは以下の要件を満たすよう設計されています。

- IPおよびイーサネット上で動作することを最初から設計された、オープンなプロトコル仕様
- ネットワークやワーカロード固有の輻輳アルゴリズムのチューニング（パラメータ調整）が不要
- 集中型の負荷分散アルゴリズムやコントローラーが不要
- マルチパスおよびパケットスプレーの採用（輻輳やヘッドオブラインブロッキングの防止）
- インキャスト管理メカニズム（ファンインを最小限のドロップで制御）
- 効率的なレート制御アルゴリズム（迅速にワイヤレートに到達可能）
- 順序の異なるパケット配信を可能とするAPI（オプションで順序通りの配信もサポート）
 - ネットワークとアプリケーションの並行性の最大化
 - メッセージ遅延の最小化
- 将来のネットワークに対応し、100万のエンドポイントをサポート可能なスケーラビリティ
- 800G、1.6T、さらには将来のより高速なイーサネットで、市販のハードウェアを使用したワイヤレート性能が達成可能な設計

これら要件を実現するため、UETはトランスポートレイヤーに留まらず、セマンティックレイヤー (Semantic Layer) も拡張しています。そのため、Libfabric などのライブラリの対応も必要となります。UECでは Libfabric Provider の参照実装を公開予定です。

このように、UETはUECの中核となる技術でありRoCEv2を置き換えることから影響範囲が大きいため、利用する際にはどのコンポーネントの変更が必要か、十分な理解が必要です。

NOTE

- パケットスプレーやそのために必要なFlexible Orderingの他、後述するトリミング(Packet Trimming) やエフェメラルコネクション(Ephemeral connections)といった技術は、従来のDCQCN(+RoCEv2)では利用できず、トランスポートプロトコルとしてUETが利用されている必要があります。
- Libfabricでは "エンドポイント(endpoint)" は Socket API (TCP/UDP) のソケットに該当します。そのため、スケーラビリティで "100万エンドポイント" は必ずしも100万ノードを意味しません。

Packet Spray (Posted Buffer)

スイッチにおけるハッシュの偏りを防ぐために Dynamic Load Balancing (DLB) などの技術が利用されていますが、パケットのリオーダリングを防ぐため flowlet を認識するためのidle閾値、パスの品質判定のための使用帯域の閾値の粒度など、様々なパラメーターに依存し、ワークロード毎に応じたチューニングが必要と言われています。[NOTE]

UETではこれらを不要にするために、利用可能な全てのパスを利用してパケットを送信する "Packet Spray" 方式を採用しています。Packet Spray は、multipathing や out-of-order delivery (OOO) とも呼ばれます。

本方式により発生するパケットのリオーダリングには、(スイッチなどネットワークファブリックではなく) サーバーサイド(トランスポートレイヤー)で対応しています。効率的にリオーダーに対応するため、UETでは "Posted Buffer" を採用しています。(プロダクション環境では、トランスポートレイヤーの処理はNICにオフロードされることが想定されます)

Posted Buffer とは、パケット毎に書き込むバッファのIDが振られており、どの順番で受信しても正しいバッファにデータが書き込まれる方式です。これにより、パケット受信後のリオーダー処理が不要となり、途中のパケット受信を待つことなく、直接受信プロセスのメモリへの書き込み(RMA)が可能となります。

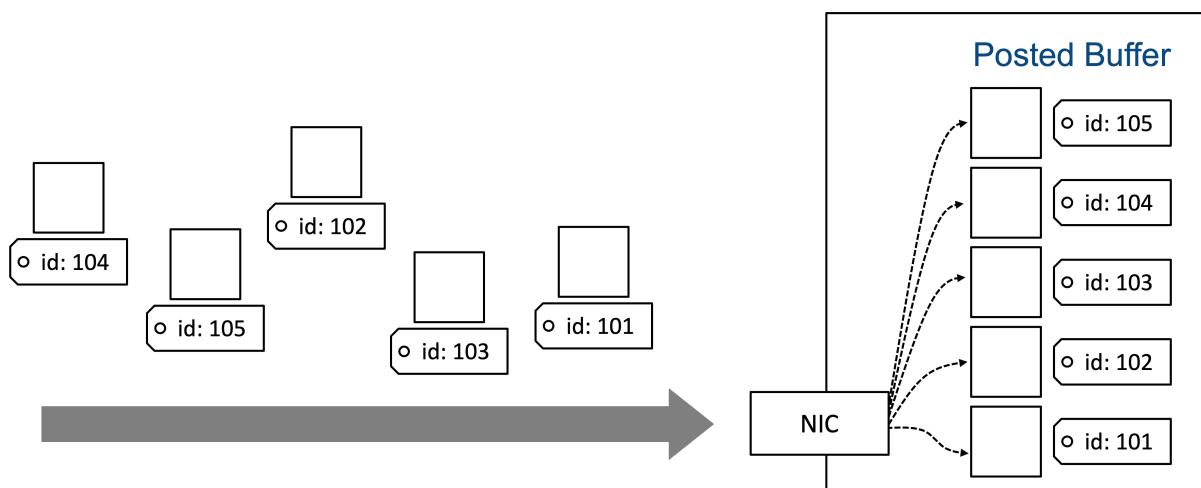


Figure 3. Posted Buffer

NOTE

既存環境でのチューニングの必要性については、[UECを利用するための検討ポイント](#)で考察しています。

Packet Trimming (optional)

トランスポートレイヤーにおけるパケットロスの検知には、以下のような方法があります。

- out-of-order (OOO) パケットの検知
- timeout (タイムアウト)

しかし、Packet Spray ではOOOを許容しているため、OOOを用いたパケットロスの検知ができません。また、タイムアウトを利用する場合、Packet Sprayでは様々なパスを経由するため、遅延の揺らぎを考慮しある程度余裕を持ったタイムアウト値にする必要があります、パケットロス検知までに時間が長くなり性能に悪影響を及ぼします。

そのため、UECではパケットロスを検知する新しい方法として "Packet Trimming" を採用しました。

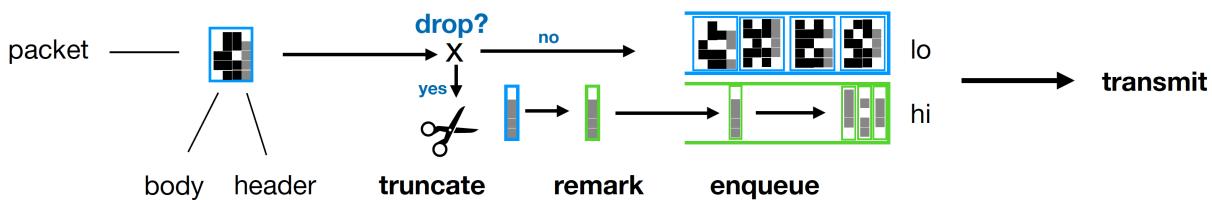


Figure 4. Packet Trimming の動作 : NANOG92^[1] の図を引用。解説は著者により日本語訳

- ドロップする代わりに 64byte に切り捨て (Trim)
- DSCPを "trimmed" としてマーク
- 高優先キューを用いて送信 (enqueue)

Packet Trimming はスイッチの機能で、図 : [Packet Trimming の動作] のように輻輳が発生した場合にパケットをドロップするのではなく、パケットをトランスポートレイヤーで一意に認識可能な必要最小限のヘッダ情報を残し切り捨て (Trim, Truncate)、高優先キューを用いて転送します。これにより、"輻輳が発生し" "パケットがドロップされた" という情報を素早く受信側に伝え、再送要求や送信ペースの調整といった輻輳回避を含む対応をタイムアウト方式に比べ低遅延で実行可能になります。

本機能は、特にLLM等のAIワークロードではメッセージサイズ (パケットサイズ) が大きいため効果的です。 (例えば4096バイトパケットを64バイトにTrimした場合、64倍のデータ削減になります)

- NOTE**
- Packet Trimming は、Broadcom Tomahawk 5 では Drop Congestion Notification (DCN) という機能名でサポート済みです。^[13]
 - SAIへの追加は "PR#2077 Add packet trimming API"^[14] で議論されており、デザイン文書 "SAI-Proposal-Packet-Trimming.md"^[15] が参考になります。

UETの輻輳制御アルゴリズム

AI/HPCワークロードの特性として、データ転送と計算を何度も繰り返す事に加え、Packet Spray により広帯域を利用可能なことから、"大きなデータを短時間で送受信" するという特徴があります。 例えば、800Gbpsの帯域を利用可能な場合、1MBのデータ送信は10usecで完了します。

このため、TCPのスロースタートといったアルゴリズムではなく、AI/HPCワーカーロードに適した以下特徴を持つ輻輳制御アルゴリズムを採用しています。

UETの輻輳制御アルゴリズムの特徴

- Fast Ramp: 送信開始時、最速でワイヤーレートまで転送レートを上げる
- Fast Slowdown: 輻輳検知時、素早い送信レートの削減（back off）

また、オプションとして受信側での輻輳制御機能も存在します。

受信側での輻輳制御機能

- receiver-generated credit manages incast
- optimistic transmission before credits received

Ephemeral Connection （エフェメラルコネクション）

UETではセッション開始時の遅延を削減するために、Ephemeral Connection （エフェメラルコネクション）を採用しています。直訳すると "儵いコネクション" であり、以下ののような特徴を持ちます。

Ephemeral Connection の特徴

- ハンドシェイクを不要とし、転送開始までの遅延を排除
- コネクション毎のステートの削減

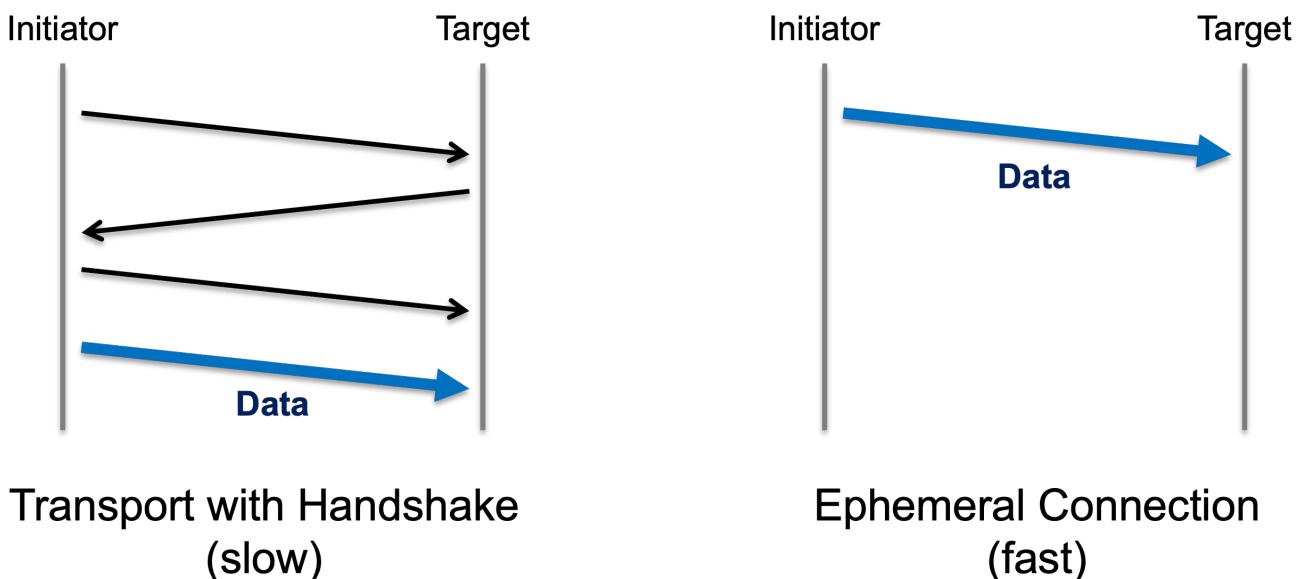


Figure 5. Ephemeral Connection

AI/HPCワーカーロードでは、短い時間(ms~us)のコネクションが大量に発生するという特徴を持ちます。そのため、コネクションの開始終了に必要な時間（遅延）を削減することが、性能向上に貢献します。

また、高速通信におけるトランスポートレイヤーの処理はNICへのオフロードが必須となっており、限られたリソースで多くのセッションを効率的に処理するために、コネクション毎のステートの削減は貴重なNICリソースの削減に大きな効果があります。

イーサネットの拡張機能 (LLR, CBFC, LLDP)

UECではリンクレイヤーであるイーサネットの拡張機能として、Link-Layer Retransmission (LLR) を開発しました。また、LLRへの対応有無を含めたキャパビリティネゴシエーション手段として、LLDPを拡張しています。

これらはオプショナルな機能であるため、利用しない場合は既存のスイッチでUEC技術（UET）を利用可能です。

NOTE UECのBlog^[16]などでは、Retransmissionではなく Retry を用いる場合もあります。

Link-Layer Retransmission (optional)

Link-Layer Retransmission (LLR) は以下の目的として開発された、CBFC: Credit Based Flow Control を利用しリンクレベルでのパケットロスを防ぐ技術です。

- ・ローカル再送 (local retransmission) を行い end-to-end再送を不要にすることで tail latency を削減
- ・リンクやトランシーバーの障害に対応

AI/HPCでは大量のトランシーバーを利用するため、リンクレイヤーの障害にどう対応するかも、性能や可用性向上に影響があると考えられます。

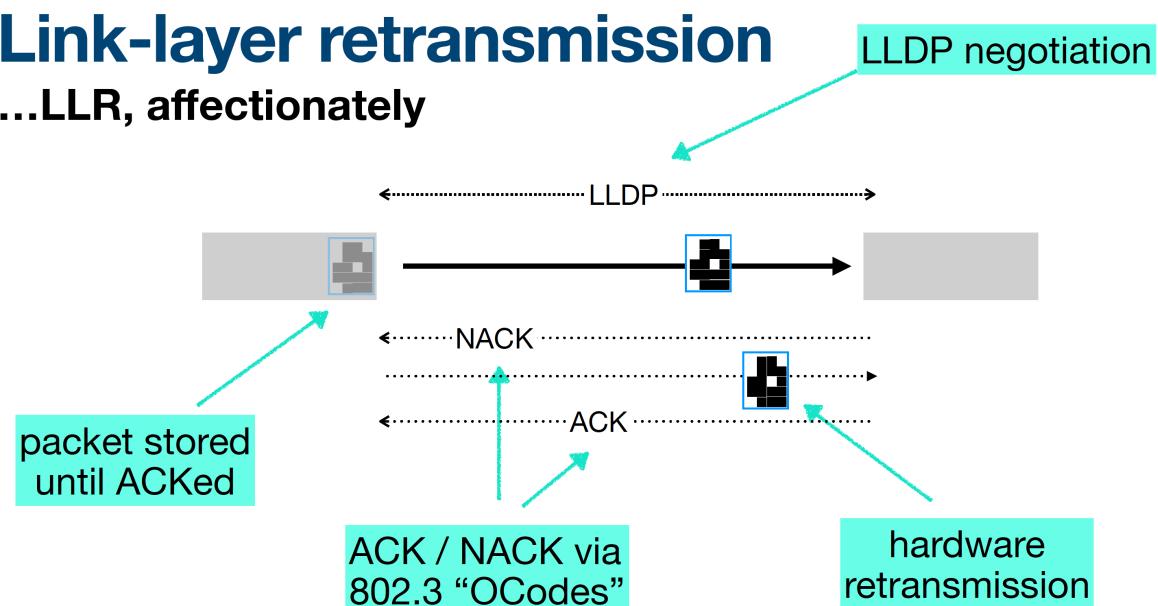


Figure 6. Link-Layer Retransmission (LLR): NANOG92^[2] の図を引用

In Network Collectives (INC)

In Network Collectives (INC) とは、Collective Communication Operation（集団通信操作）をネットワーク（スイッチファブリック）上のノードで実行することにより、遅延の削減やネットワーク帯域の消費を抑え、処理の高速化を実現する技術です。

一般的に INC は In-Network Computing (Computation) の略語として利用されており、集団通信操作に

INCの表記について

現状、UEC SW Working Group 内の議論では、In Network Collectives (INC) の In と Network の間にハイフン (-) を入れない表記を利用しているため、本文書でもあえて入れてません。但し、UEC仕様検討の過程でハイフン有りになる可能性もあります。2025年Q1にリリース予定のUEC INC仕様書が公開された以降は、そちらの文書で定義された Terminology に従いましょう。

逆に、In-Network Computing など、UEC外ではハイフンを入れることが多い（と著者は感じている）ため、ハイフンを入れています。例えば NVIDIA の公式文書では In-Network Computing と表記しています。

先行技術としては Mellanox (元NVIDIA) が開発した Scalable Hierarchical Aggregation and Reduction Protocol (SHARP™) がありますが、SHARP は InfiniBand でのみ動作し、NVIDIA製のNIC (ConnectX, BlueField) が必要となります。

これに対し、UECのINCはオープンな仕様で、マルチベンダのイーサネットスイッチ (ASIC) やNICで動作可能になる予定です。

UEC INC の詳細は公開されていませんが、SHARPとの違いは以下の通りで、概念的には非常に似ています。

- イーサーネット上で動作
- データ転送にUETを用いる

そのため、仕様公開まではSHARPについて調査することで、UEC INCについてもある程度の動作をイメージすることが可能です。

- SHARPポータルページ："NVIDIA Docs Hub > NVIDIA Networking > Accelerator Software" [\[17\]](#)
- SHARPのオリジナル論文：IEEE Xplore "Scalable Hierarchical Aggregation Protocol (SHArP): A Hardware Architecture for Efficient Data Reduction" [\[18\]](#) [\[19\]](#)
 - NVIDIAのWEBサイトからダウンロードできます [\[20\]](#)

上記のような公式ドキュメント以外にも、YouTubeにチュートリアルや研究発表のカンファレンス講演が掲載されていますので検索してみてください。

In Network Collectives に関する、著者の個人的な感想メモです。十分な理解ができるおらず考えがまとまってないので、一緒に調査や考察をする人を募集中です。

- NOTE
- UECはマルチベンダのため、INCをサポートするスイッチASICやサーバーサイドのNIC、ミドルウェア、などが色々出てくると思われる。
 - そのため、導入検討や評価を行う際には、INCの実装間の比較や、SHARPとの比較をするポイントを整理する必要がある。
 - 評価視点の例：
 - ワークロードの特徴を確認もしくは仮定した上で評価する必要がある。

- HPCに比較し、AI(LLM)はメッセージサイズが大きい。
- Switch ASIC のメモリは限られているため、AIワークロードのようにメッセージサイズが大きいと分割して転送、計算、結合、などする必要があるのでは？（仮説）
- 分割して計算するアルゴリズムは様々なものが考えられるので、同じINCでも、INC Switchとライブラリの組み合わせによってアルゴリズムが異なり、性能にも大きく影響することがあるのでは？
- 得意なワークロード、データサイズ、それによる性能の高低、などを判断し最適な実装を選択するには、INC Switchの実装と、MPI CC操作をINCで実行するアルゴリズムの、両方を深く理解する必要がありそう。この部分がUECで標準化されるかはまだ不明。
- これ参考になる？ Session 13 "Designing In-Network Computing Aware Reduction Collectives in MPI"
 - <https://www.openfabrics.org/2024-ofa-virtual-workshop-agenda/>

UECを利用するための検討ポイント

従来のAI/HPCネットワーク技術からの大きな変更点として、UECではトランスポートをRoCEv2からUETに置き換えています。そのため、UECの技術の利用検討にあたっては、ハードウェアとソフトウェアの両面で "変更（開発）が必要なものと既存のまま利用が可能なもの" を整理し、システム、サービス、構築や運用の人員、などへの影響と、変化によるメリットやデメリットを理解しながら進める必要があります。

本章では、UEC技術の利用検討を進める際のポイントを以下視点から整理しました。

- UECを利用することによる変化や注意点
- UECを利用する際の検討ポイント
- UEC仕様や製品のロードマップ、UEC対応製品（対応製品はどのベンダーからいつ提供されるか？）

まとめ

- UECはRoCEv2をUETと置き換える "スタック全体の改善"
 - AIアプリケーションやミドルウェアを含め、影響範囲を理解することが重要
 - GPU, NIC, Switch ASIC, Switch OS, 等、様々なベンダと連携した技術検証が必要
- UECの各要素技術がリリースされるタイミングは様々
 - 段階を追った技術検証が必要
- リンクやネットワークレイヤーの機能はオプショナルなものが多い
 - 既存のイーサネットスイッチを利用した検証から開始可能

UECを利用することによる変化や注意点

技術スタックの変化

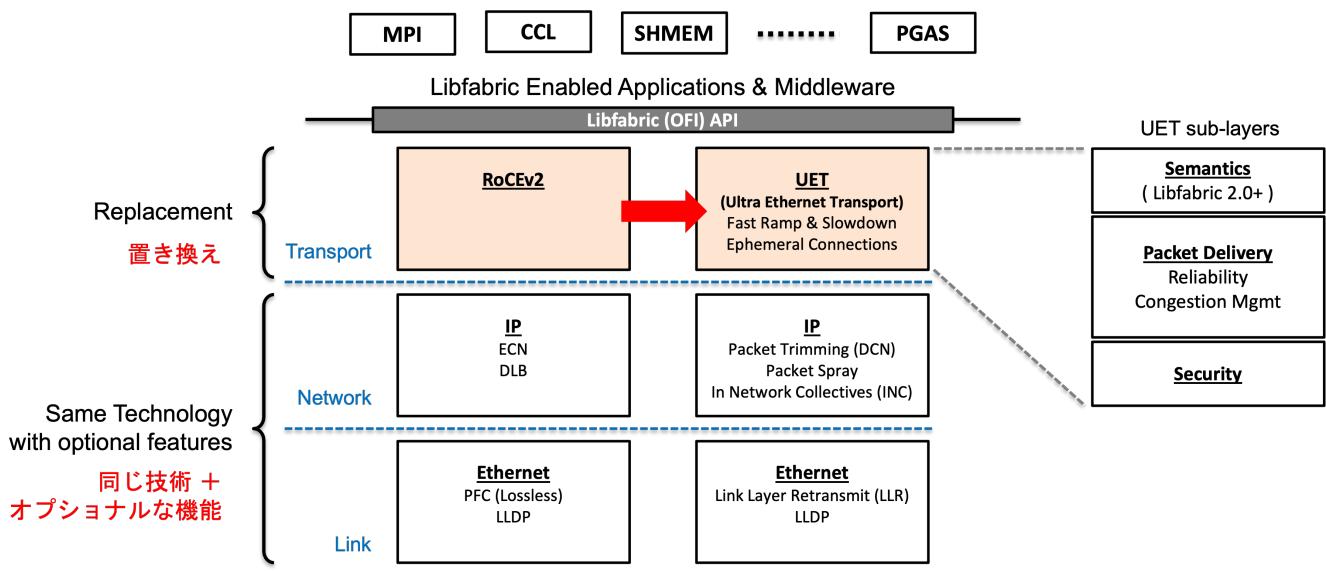


Figure 7. UECの技術スタックとRoCEv2との比較

RoCEv2を用いたAI/HPCネットワーク技術とUECの最も大きな違いは、図：[UECの技術スタックとRoCEv2との比較](#) のようにトランスポートレイヤーをRoCEv2から、新しく策定した UET (Ultra Ethernet Transport) に置き換えたことです。 UETでは、モダンなHPC向け RMA API でありHPCアプリケーションで幅広く利用されている libfabric API を採用し、拡張しています。

すなわち、libfabric を利用しているアプリケーションやミドルウェアはAPI拡張への対応でUEC技術を利用可能な場合もありますが、Libfabricを利用してないアプリケーションやミドルウェアの場合、UECをサポートするために大きな変更（プログラムの改修）が必要となる可能性があります。

NOTE

libfabric 以外のAPIへの対応も時々議論に上がりますが、現状では明確なロードマップはありません。もし libfabric 以外のAPIへの対応が必要なユーザーが多い場合は議論が進む可能性がありますので、ユースケースや対応すると嬉しいライブラリなどUECメンバー企業に対してインプットすると良いでしょう。

逆にIPやイーサネットに対する拡張はオプショナルな機能のみであるため、既存のスイッチでもUETは利用可能です。但し、UETを活かした性能を得るために、最低限 Packet Trimming などIPレイヤーの機能は対応が必要と考えられます。

また、Lossless Ethernet が不要となることにより、PFC/バッファ設定のチューニングなどが不要となるため、ネットワーク層の設定は従来よりシンプルになると考えられます。 逆に、UETの実装である Libfabric Provider の安定性や性能はドライバを含めてNIC毎に異なる可能性があります。そのため、サーバーサイドのチューニングの必要性は未知数であり、少なくとも初期段階では、NIC毎に検証が必要と予想されます。

デバイス構成による利用可能な技術の比較

UECを利用する場合はもちろん、従来のRoCEv2の場合でもデバイス構成により利用可能な技術が異なります。

図：[AI/HPC Hardware Matrix](#) に、GPU/NIC/Switch の組み合わせに対し利用可能な技術を整理しました。利用したい技術と必要な機材について検討する際に参考にしてください。

例えば、Adaptive Routing を利用するには NVIDIA で機材を揃える必要があります。また、2025年1月時点でINCを利用するためには InfiniBand が必要となりますが、将来的にUEC対応のNICやSwitchを選択することでイーサネットでもINC利用可能になります。

	NVIDIA Infiniband	NVIDIA Ethernet Suite	NVIDIA GPU with non-NVIDIA switch	non-NVIDIA GPU & Switch	non-NVIDIA Lossless	non-NVIDIA UEC
GPU	NVIDIA	NVIDIA	NVIDIA	non-NVIDIA	non-NVIDIA	non-NVIDIA
NIC	NVIDIA (BF/CX)	NVIDIA (BF/CX)	NVIDIA (BF/CX)	NVIDIA (BF/CX)	non-NVIDIA	non-NVIDIA (UEC)
Switch	InfiniBand (NVIDIA)	Ethernet (NVIDIA)	Ethernet (non-NVIDIA)	Ethernet (non-NVIDIA)	Ethernet (non-NVIDIA)	Ethernet (UEC)
Lossless?	Lossless	Lossless (PFC)	Lossless (PFC)	Lossless (PFC)	Lossless (PFC)	Best Effort
Load Balancing	Adaptive Routing?	Adaptive Routing (Packet Spray)	Dynamic LB (flowlet)	Dynamic LB (flowlet)	Dynamic LB (flowlet)	Packet Spray
Congestion Notification		ECN	ECN	ECN	ECN	Packet Trim (DCN)
INC In Network Computing Collectives	SHARP	no	no	no	no	UEC INC

Figure 8. AI/HPC Hardware Matrix

UEC技術と関連するコンポーネント（まとめ）

UECの各技術を利用する際に関連するコンポーネントをまとめました。

基本的にUECを利用する際には、NIC, Switch ASIC, Switch OS (NOS), AIアプリケーション、全てを考慮する必要がありますが、各技術のサポート有無について確認する製品やベンダーを絞り込みたい時に参照してください。

なお、"AIアプリケーション" については、アプリそのもの、ミドルウェア、など、対応が必要な範囲が異なるため、更に詳細に分類して影響を確認する必要があると思われます。

Technology	Description	Switch ASIC	Switch OS	NIC (driver)	AIアプリ
UEC Transport (UET)	RoCEv2の置き換えとなるトランスポートプロトコル。Out-of-order パケットの受信、ベストエフォート（非ロスレス）、等に対応。	optional	optional	YES	YES
Packet Trimming	Drop Congestion Notification (DCN) とも呼ばれる。輻輳発生時に、パケットを一定のサイズにトリムし、高優先度で転送することにより受信ノードに輻輳を伝える。これにより、ドロップ検知Timerによる遅延を無くす。	YES	YES	YES	NO
Link Level Retry (LLR)	イーサネットのリンクレイヤでパケットロスが発生しないように送信管理や再送を行う。 Credit Based Flow Control (CBFC)	YES	YES	YES	NO
LLDP Negotiation	LLRのサポートなど、キャパビリティをノード間でネゴシエーションする。従来のLLDPにUEC拡張が行われる。	YES	YES	YES	NO

Technology	Description	Switch ASIC	Switch OS	NIC (driver)	AIアプリ
Packet Spray, Ordered(ROD) and un-ordered(RUD)	利用可能なパス全てにパケットをSprayすることで、パケットレベルのロードバランシングを実施し、イーサネットファブリックの利用率を向上させる。 UETの機能であり、パケット（メッセージ）にメモリのどの場所に保存すべきか識別可能なIDを埋め込むことにより、受信ノードでパケットのリオーダー（バッファリング）を不要にする。	YES?	YES?	YES	NO
Ephemeral connections (短命コネクション)	最初のパケットにセッション情報入れてハンドシェイクを不要にする。（PDC,PDS）バースト的なデータ転送を繰り返すワーカロードでハンドシェイクのオーバーヘッドを排除しバースト毎の転送時間の短縮によるスループット向上を実現する。	NO	NO	YES	NO
In Network Collectives (INC)	Collective Communication Operation を Switch Fabric 内のスイッチにオフロードする。恩恵として性能向上（輻輳やデータ転送量の削減、遅延の削減）、リソース利用効率の向上（プロセッサ、アクセラレータ、メモリ）、消費電力の削減、が期待される。	YES	YES	YES	??

UECを利用する際の検討ポイント

UECを利用する際の検討ポイントとして思いついたものを列挙しました。

組織（ユーザー、メーカー（ベンダー）、システムインテグレーター）、役割（ネットワークエンジニア、サーバーエンジニア、アプリケーション開発者）、ユースケース、予算、等によってそれぞれ重視すべき検討ポイントは異なると考えられます。

ここには2025年1月時点で著者が思い浮かんだもののみを記載していますので、「こんな観点もある」「この観点から考えるとメリット・デメリットが異なる」、などありましたら是非フィードバックお願いします。

Infinibandに対するEthernetを利用する理由

UECの前に検討が必要なのが、InfiniBandとイーサネットどちらを選択するか、であり、例えば以下のような観点が考えられます。

- 安定性：遅延やスループットが安定しているか？（Job完了時間の短縮や安定性が最終指標）
- 柔軟性：構成変更が容易か？
- 価格
- 人材確保の容易さや教育コスト

- ・マルチテナントの必要性
- ・AI/HPC以外のトラフィックを利用するか？

立場や状況により重視する視点は様々ですが、サービス事業者の観点からは2025年1月に公開された LINE ヤフーのBLOG "GPUクラスタネットワークとその設計思想 (Rethinking AI Infrastructure Part 2)"^[21] でわかりやすく整理されていました。

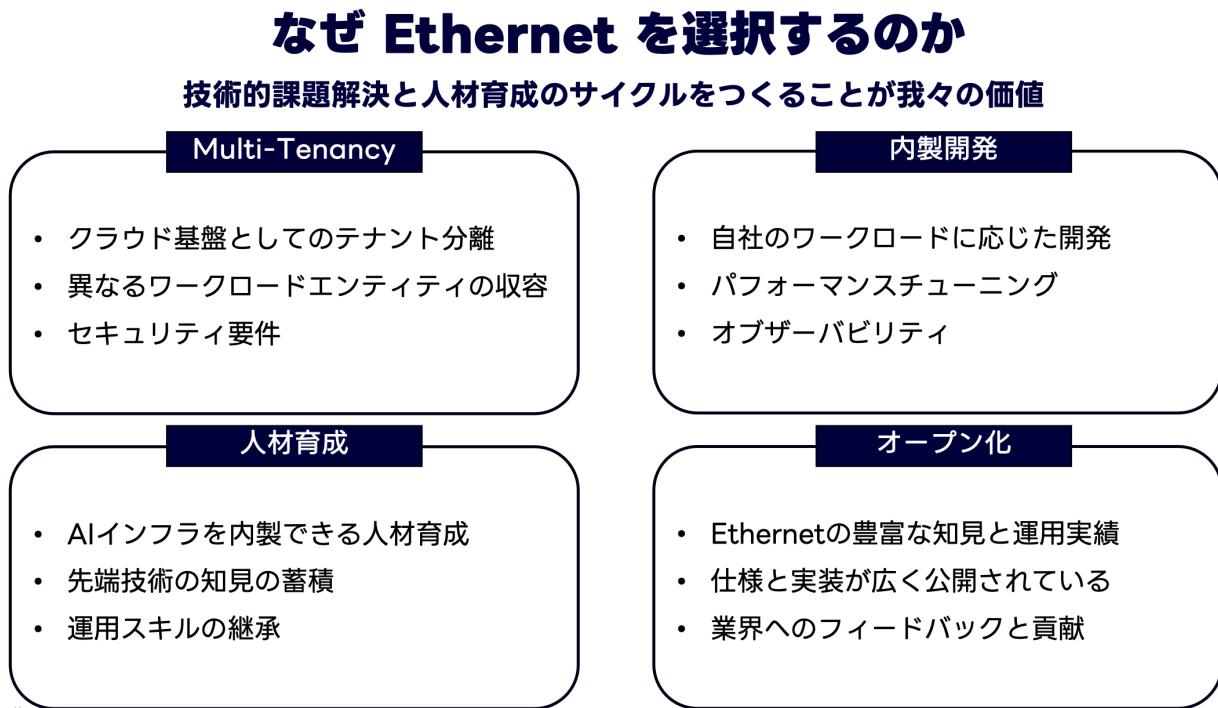


Figure 9. なぜEthernetを選択するのか (LINEヤフーBLOG^[2]から引用)

逆に大規模なトレーニングで利用する場合は、METAなどハイパースケーラー企業が公開しているBLOG や論文を参照すると、また異なる視点について学ぶことができます。

NVIDIA Suites を利用しない理由

周辺ライブラリやソフトウェアエコシステムの成熟度から、AI分野では GPU, NIC, Switch 全てを NVIDIA の機材を利用する、というのが最初に選択肢と上がることが多いでしょう。

- ・NVIDIA Suites: BlueField/ConnectX + Spectrum + Cumulus, Spectrum X, etc.

NVIDIAもUECのメンバー企業であるため、将来的にNCCLといったNVIDIAライブラリへの対応が進んで行く可能性はありますが、最初のリリースでは対応していないと予想されます。そのため、「なぜNVIDIAスイートを選択しないのか？」ということをクリアにすることが、UECに限らずNVIDIAではない機材や技術を選択する際に重要であり、例えば以下のようないわゆる視点が考えられます。

- ・価格
- ・納期
- ・AI専用チップなど、NVIDIA GPU以外のAIアクセラレータの利用
- ・ベンダーロックインの回避

UECとRoCEの比較

主な点を列挙してみました。

ここに記載した内容が正解というわけでもなく他の視点も沢山あるかと思いますので、引き続きJANOG55などでの議論を通じて、整理・追記していきたいと考えています。

- メリット
 - オープンな技術と実装（ライブラリのソースコードへのアクセス？）
 - Lossless Ethernet が不要 ⇒ チューニング不要なシンプルなスイッチファブリック
 - アクセラレータの選択肢拡大
- デメリット
 - NVIDIAエコシステムからの移行の手間や動作しないリスク
 - NVIDIA任せにできない（UECシステムパックを提供するベンダが出現するかも？DELL,HPE？）
 - システムインテグレーションに伴う検証や相性問題のリスク

既存環境でのチューニングの必要性について

RoCEv2が前提とするロスレスイーサネットでの性能限界を達成するためには、バッファサイズや閾値など、ワークロードに応じたチューニングが必要と言われています。しかし、達成したい性能とチューニングにかけられるコスト（時間と機材）のバランスや優先度は、各サービス（システム）や組織により異なります。

そのため、例えば以下のような観点を考慮し、「チューニングを行わない方が良い」、という判断になる可能性もあると考えられるため、"チューニングが不要" というのがUEC導入理由になるかは、利用するアクセラレータに応じて、各種ベンチマークや性能以外の指標や利便性を考慮しながらの判断になると考えています。

- デフォルト設定での性能
- サービスの規模（巨大サービスでは小さな改善でも大きな効果がある）
- チューニングが可能な人材のコスト
- チューニングに必要な期間（その間の機会損失）
- ネットワーク、サーバー、などのコストバランス
- 利用するアクセラレータ（GPU）の技術スタックの成熟度

UEC仕様や製品のロードマップ

UECの仕様や製品がいつリリースされるかは、利用検討の重要な要素です。

図："UEC Target Timelines, 2024-10-15" は2024年10月に開催されたOCP Global Summitで発表されたロードマップです。これによると、UECの仕様や製品がリリースされるのはそれぞれ以下のタイミングになりそうです。

- UEC仕様(v1.0) : 2025年3月頃（第一四半期）
- UEC対応製品 : 2025年（おそらく中旬以降？）

但し、2023年時点では2024年中に仕様公開を目指すと発表されていましたので、今後ロードマップが変更される可能性もあります。 UECのプレスリリース、BLOG、ベンダー各社からの発表などを通じて、最新の動向を確認しながら検討を進めましょう。

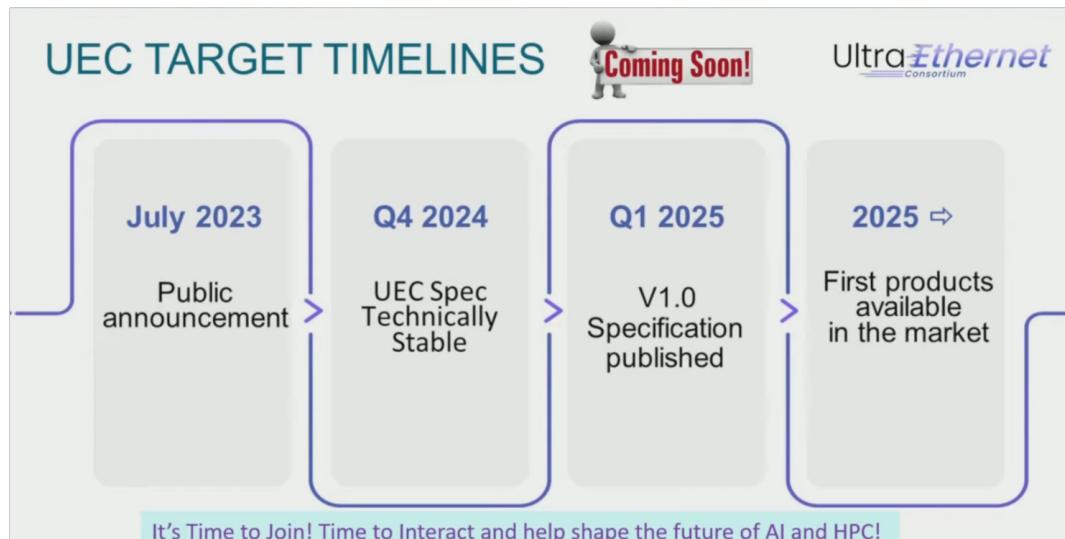


Figure 10. UEC Target Timelines, 2024-10-15, "Accelerating AI/HPC: OCP and UEC's Collaborative Vision for High-Performance Networking, Uri Elzur, OCP2024" ^[4] から引用

UEC対応製品（出荷前の製品を含む）

2024年後半には各社からUEC対応製品についてのアナウンスが相次ぎました。ここではUEC対応製品を列挙します。アナウンスがあった製品はそのリンクを、アナウンスはまだ無いがおそらくUEC対応製品をリリースすると思われるベンダーに関してはプレースホルダとして企業や製品名だけを記載しています。

なお、UECに限らずデータセンター向け製品は、最初はハイパースケーラーなど大量購入する企業に割り当てられて、その後に大規模～中規模の事業者で購入が可能になる傾向があります。そのため、製品リリースと実際に入手可能になるタイミングにはズレがあることが多いことに注意が必要です。

AMD

AMD Pensando Pollara 400 (SmartNIC)

- AMD Pensando Pollara 400, the first Ultra Ethernet Consortium ready NIC, reduces the complexity of performance tuning and helps improve time to production.
 - <https://ir.amd.com/news-events/press-releases/detail/1218/amd-unveils-leadership-ai-solutions-at-advancing-ai-2024>
- AMD unveils industry's first Ultra Ethernet ready network card for AI and HPC News, October 11, 2024
 - <https://www.tomshardware.com/networking/amd-unveils-industrys-first-ultra-ethernet-ready-network-card-for-ai-and-hpc>

Broadcom

Thor 2 NIC Chip

- 2024-07-02, "Word on the Street: Broadcom high-performance 400G RoCE / RDMA NICs"
 - <https://www.broadcom.com/blog/400g-roce-rdma-nics>

Switch ASIC

- Tomahawk 5: Packet Trim (DCN) に対応
- Jericho 3-AI: ???

Marvell Technology

Marvel NIC

- SmartNIC をリリースする可能性あり？

Marvell Teralynx Ethernet Switch

- Marvell Technology: As a member of the UEC, Marvell is committed to advancing Ethernet technology for AI and accelerated computing. Their Teralynx® Ethernet switches are optimized for low-latency fabrics between compute nodes, aligning with UEC's objectives.
 - <https://multiplatform.ai/marvell-teralynx-10-switch-enters-production-to-meet-surge-in-ai-cloud-demands/>

Arista

Arista Etherlink AIプラットフォーム

- Arista Etherlink AIプラットフォーム
 - <https://www.arista.com/jp/solutions/ai-networking>
- <https://blogs.arista.com/blog/new-ai-era>
 - Arista Etherlink is standards-based Ethernet with UEC-compatible features. These include dynamic load balancing, congestion control, and reliable packet delivery to all NICs supporting RoCE. Arista Etherlink will be supported across a broad range of 800G systems and line cards based on Arista EOS. As the UEC specification is finalized, Arista AI platforms will be upgradeable to be compliant.
- <https://www.arista.com/jp/company/news/press-release/19841-jp-pr-20240605>
 - すべてのEtherlinkスイッチが新たに設立されたUltra Ethernet Consortium (UEC) の標準をサポートしています。この標準によって、近い将来UECのNICが利用可能になったとき、パフォーマンスのメリットがさらに大きくなると期待されています。

Asterfusion

UEC仕様が公開されたのち、将来的にUEC対応するとアナウンスしている。

- 2024-09-03, The Ultimate Switches for Artificial Intelligence
 - <https://medium.com/@Asterfusion/the-ultimate-switches-for-artificial-intelligence-80fb8033ce86>
 - Asterfusion AI Switches Offers Forward-compatible Products with UEC Standard.
 - As the Ultra Ethernet Consortium (UEC) completes its expansion to improve Ethernet for AI workloads, Asterfusion is building products that will be ready for the future. The Asterfusion CX-N AI data centre switch portfolio is the definitive choice for AI networks, leveraging standards-based Ethernet systems to provide a comprehensive range of intelligent features. These features include dynamic load balancing, congestion control, and reliable packet delivery to all ROCE-enabled network adapters. **As soon as the UEC specification is finalised, the Asterfusion AI platform will be upgradeable to comply with it.**
- CX864E Data Sheet
 - <https://cloudswit.ch/wp-content/uploads/2024/06/Datasheet-CX864E-N-Ultra-Ethnet-Switch.pdf>
 - "Line-rate programmability to support evolving UEC (Ultra Ethernet Consortium) standards" と記載されているため、将来的にUECに対応可能（だがまだサポートしていない）とも解釈できる
- 2024-10-24, "The Ultimate In-Depth Exploration of Ultra Ethernet Consortium (UEC) Technology"
 - <https://cloudswit.ch/blogs/exploration-of-ultra-ethernet-consortium-uec/>

Cisco Nexus 9000 Series Switches

- 2024-12-12, "Nexus Improves Load Balancing and Brings UEC Closer to Adoption"
 - <https://blogs.cisco.com/datacenter/nexus-improves-load-balancing-and-brings-uec-closer-to-adoption>
 - Cisco Nexus 9000 is Ultra Ethernet ready

Mercury AI-SuperNIC

- 2025-01-06, "DreamBig Announces world leading 800G AI-SuperNIC chip (Mercury) with Fully HW Offloaded RoCE v2 + UEC RDMA Engine"
 - DreamBig Mercury chip features a hardware accelerated RDMA engine that supports existing RoCE (RDMA over Converged Ethernet) v2 and new UEC (Ultra Ethernet Consortium) standards, delivering best-in-class bandwidth (800Gbps) and throughput (800Mpps) with lowest power, ultra low latency, and smallest area.
 - https://www.prnewswire.com/news-releases/dreambig-announces-world-leading-800g-ai-supernic-chip-mercury-with-fully-hw-offloaded-roce-v2—​uec-rdma-engine-302342748.html

Mercury is designed with fully programmable Congestion Control to adapt to any data center and provides the following critical functions for AI applications

- Multi-pathing and packet spraying
- Out-of-order packet placement with in-order message delivery
- Programmable congestion control for RoCE v2 and UEC algorithms
- Advanced packet trimming and telemetry congestion notifications
- Support for selective retransmission

[1] "Keynote: Networking for AI and HPC, and Ultra Ethernet", NANOG 92, Hugh Holbrook, Arisa Networks

[2] "Keynote: Networking for AI and HPC, and Ultra Ethernet", NANOG 92, Hugh Holbrook, Arisa Networks

[3] <https://techblog.lycorp.co.jp/ja/20250115a>

[4] <https://www.opencompute.org/events/past-events/2024-ocp-global-summit>

[1] <https://www.top500.org/lists/top500/list/2024/11/>

[2] <https://ultraethernet.org/leading-cloud-service-semiconductor-and-system-providers-unite-to-form-ultra-ethernet-consortium/>

[3] <https://jointdevelopment.org/>

[4] <https://ultraethernet.org/>

[5] <https://ultraethernet.org/working-groups/>

[6] UECホームページの "Download White Paper" をクリック

[7] <https://ultraethernet.org/blog/>

[8] <https://ofiwg.github.io/libfabric/>

[9] UECホームページの "Download White Paper" をクリック

[10] UECホームページの "Download White Paper" をクリック

[11] <https://datatracker.ietf.org/doc/draft-ravi-ippm-csig/>

[12] BTS is a sub-RTT backward congestion signaling from the switch back to the sending Network Adapter.

[13] <https://www.broadcom.com/blog/cognitive-routing-in-the-tomahawk-5-data-center-switch>

[14] <https://github.com/opencomputeproject/SAI/pull/2077>

[15] <https://github.com/marian-pritsak/SAC/blob/master/doc/SAC-Proposal-Packet-Trimming.md>

[16] <https://ultraethernet.org/ultra-ethernet-specification-update>

[17] <https://docs.nvidia.com/networking/software/accelerator-software/index.html#nvidia-sharp>

[18] <https://ieeexplore.ieee.org/document/7830486>

[19] R. L. Graham et al., "Scalable Hierarchical Aggregation Protocol (SHArP): A Hardware Architecture for Efficient Data Reduction," 2016 First International Workshop on Communication Optimizations in HPC (COMHPC), Salt Lake City, UT, USA, 2016, pp. 1-10, doi: 10.1109/COMHPC.2016.006.

[20] https://network.nvidia.com/pdf/solutions/hpc/paperieee_copyright.pdf

[21] <https://techblog.lycorp.co.jp/ja/20250115a>