AI infrastructure explanatory materials (AI/HPC & UEC technology overview)

Table of Contents

How to read AI infrastructure explanatory materials

The JANOG55 AI infrastructure explanatory document (PDF) can be downloaded from janog55-ai-infra.pdf . (https://github.com/ebiken/janog/blob/main/JANOG55/janog55-ai-infra.pdf)

Depending on your level of understanding of AI workloads and infrastructure, you might want to read through the following:

- If you would like to learn about existing AI/HPC-related network technologies first, please read the first "Existing AI/HPC-related Network Technologies", then "UEC Overview", "UEC Technology Overview", etc.
- If you are familiar with existing technologies, please start reading with "UEC Overview".
- If you are interested only in UEC's technology and not its background or organization, please start reading from "UEC Technology Overview".
- For those who would like to get a general understanding of UEC technology, product roadmap, and points to consider, please refer to "Points to consider when using UEC".
- UET uses the Libfabric API. For more information about Libfabric, see the attached document "Libfabric (OFI)" hpc-libfabric.pdf . (https://github.com/ebiken/janog/blob/main/JANOG55/hpc-libfabric.pdf)

  Note

  The UEC-related information contained in this document was written based on publicly available information as of January 2025. When considering the use or evaluating UEC technology, please refer to the information on the official website and the "UEC Specification," which is scheduled to be publicly released in the first half of 2025.

Existing AI/HPC related network technologies

Existing AI/HPC-related network technologies can be broadly categorized into three categories: lossless Ethernet, congestion control, and load balancing. These existing technologies are explained in JANOG and other publications, so it is a good idea to read the materials listed in "References" as prerequisite knowledge for understanding UEC's technology.

There are various load balancing technologies, but DLB is the only one that is supported regardless of vendor.

Table 1. Existing AI/HPC-related network technologies

function	Representative technologies
Lossless Ethernet	• PFC (802.1Qbb, Priority Flow Control)
Congestion Control	• DCQCN (ECN, Sender-driven)
Load Balancing, Fabric utilization	• Dynamic Load Balancing (flowlets) • NVIDIA Adaptive Routing (flowlet?) • Packet Spray (Packet based) • Scheduled Fabric (VoQ) (Cell based)

References

Past JANOG Sessions

- JANOG54: Creating a public cloud service for generative AI
  - https://www.janog.gr.jp/meeting/janog54/sakura/
  - Takashi Inoue, Makoto Takamine, Daisuke Hirata Sakura Internet Co., Ltd.
- JANOG54 Consideration of the use of generative AI technology in network operations
  - https://www.janog.gr.jp/meeting/janog54/genai/
  - Ryosuke Sato, Takashi Shirai, Junji Taguchi, Yukichi Kondo NTT Field Techno Corporation
  - Takuya Miyasaka, KDDI Research, Inc.
  - Takumi Nakamatsu, KDDI Corporation
  - The content is about utilizing AI for network operations instead of AI/ML networks. It explains the overview of GPT.
- JANOG53 Network for AI (Artificial Intelligence)
  - https://www.janog.gr.jp/meeting/janog53/ainw/
  - Shishio Tsuchiya, Arista Networks Japan LLC
- A story about building a 400G DC network based on JANOG52 AI/ML
  - https://www.janog.gr.jp/meeting/janog52/aiml400/
  - Yasuhiro Uchida, Shotaro Koshoji, CyberAgent, Inc.
- JANOG50+ Packet Loss and Latency
  - https://www.janog.gr.jp/meeting/janog50plus/docs/janog50plus-maz-losslatency.pdf
  - Yoshinobu Matsuzaki (Internet Initiative Japan Inc.)
- JANOG43 The story of how we designed the LINE network from scratch
  - https://www.janog.gr.jp/meeting/janog43/application/files/7915/4823/1858/janog43-line-kobayashi.pdf
  - Masayuki Kobayashi, LINE Corporation

AI/HPC network related (Japanese)

- GPU Cluster Network and Its Design Philosophy (Rethinking AI Infrastructure Part 2)
  - https://techblog.lycorp.co.jp/ja/20250115a
  - LINE Yahoo Japan Corporation Kobayashi, Fukasawa
- "GPU Network Design and Operation Basics Seminar: Lossless Ethernet - PFC/ECN Edition"
  - https://speakerdeck.com/markunet/ecnbian
  - LINE Yahoo! Corporation Masayuki Kobayashi
- Data Center Networks in the AI Era
  - https://speakerdeck.com/lycorptech_jp/dcnw_in_the_ai_era
  - 40th Information Network and Network System Research Workshop
  - LINE Yahoo! Corporation Masayuki Kobayashi
- Learnings and prospects from the introduction of Ethernet-based GPU clusters
  - https://speakerdeck.com/lycorptech_jp/20241202
  - NVIDIA AI Summit Japan 2024
  - LINE Yahoo! Corporation Masayuki Kobayashi, Mikiya Michishita
- Podcast: fukabori.fm "124. GPU Clusters and DC Networks in the AI Era"
  - https://fukabori.fm/episode/124
  - LINE Yahoo! Corporation Masayuki Kobayashi, Mikiya Michishita
- Actual state of accelerator communication at PFN / MPLS Japan 2024
  - https://speakerdeck.com/pfn/mpls-japan-2024
  - Yuichiro Ueno / Preferred Networks, Inc.
- "Ethernet evolves into a highly reliable network, [3] The core is "PFC" and "ETS" which are responsible for flow control"
  - https://xtech.nikkei.com/it/article/COLUMN/20091006/338383/
  - 2009-10-15 Nikkei Crosstech

AI/ML related: Other (Japanese)

- Distributed Cache Empowers AI/ML Workloads on Kubernetes Cluster / KubeCon + CloudNativeCon North America 2024
  - https://speakerdeck.com/pfn/kubecon-plus-cloudnativecon-north-america-2024
- Recipe for building a machine learning cluster for generative AI - Ishikari, Hokkaido
  - https://speakerdeck.com/pfn/20240615-cloudnativedayssummer-pfn
  - Cloud Native Days Summer 2024
- What I learned from actually using it! How to use a Kubernetes cluster with multiple GPUs and how to reduce the workload
  - https://speakerdeck.com/pfn/cloudnativedaystokyo23-shiota
  - Tetsuya Shiota PreferredNetworks, Inc.
  - CNDT2023: CloudNative Days Tokyo 2023
- (End of 2023) Current State of LLM - We will introduce the current situation surrounding LLM.

- https://speakerdeck.com/pfn/llmnoxian-zai
- Preferred Networks, Imos (imos@preferred.jp)

Overview of UEC

Purpose of Establishment

The Ultra Ethernet Consolrtium (UEC) is an industry consortium established to realize high-performance Ethernet communications that meet the demands of AI and HPC.

GPUs are also used in HPC systems ranked in top500.org (https://www.top500.org) 's TOP500 LIST - NOVEMBER 2004[1], and it is expected that the workloads and network requirements of AI and HPC will overlap in the future. Therefore, UEC aims to develop "open standards" for protocols and technologies that solve the issues of both AI and HPC workloads (RoCEv2) in a "multi-vendor" manner, not limited to AI.

The UEC was founded in 2023 by AMD, Arista, Broadcom, Cisco, Eviden (an Atos Business), HPE, Intel, Meta, and Microsoft[2] and as of December 2024, consists of a total of 90 member companies.

It is said that the group was established to counter NVIDIA's monopoly on vertical integration of applications, accelerators (GPUs), and networks, and NVIDIA has been participating in the group as a member company since 2024.

UEC is organized as part of the Linux Foundation's Joint Development Foundation (JDF)[3]. JDF provides a legal and operational foundation for companies and organizations to quickly launch technical standards and open source projects together, and provides mechanisms to efficiently promote joint development.

Member Company Type

Member companies and organizations are classified into three categories: Steering Members, General Members, and Contributor Members. New companies can choose between General and Contributor.

As of January 1, 2025, the organization consists of 90 member companies, including 10 Steering Members, 26 General Members, and 54 Contributor Members.

Steering Members

Members are AMD, Arista, Broadcom, Cisco, Eviden, HPE, Intel, Meta, Microsoft, and Oracle, and consist of the founding companies + Oracle.

General Members

General Members can participate in the TAC and can influence the activities of the UEC, such as launching new working groups and creating work items. They also have the right to vote (Supermajority) on agenda items that require more than a majority (such as 2/3 or 3/4).

Contributor Members

Unlike General Members, Contributor Members cannot propose new activities, but they can provide accessible information and contribute technology, and have the right to vote except for Supermajority. Therefore, they can usually do the same activities as General Members.

Several Japanese companies also participate as member companies.

UEC member Japanese companies

- General Member: Preferred Networks

- Contributor Member: Fujitsu, IIJ

A list of member companies can be found on the UEC website homepage at https://ultraethernet.org/ [4].

Organization and operation method

UEC is composed of the following organizations. From a technical standpoint, it is a good idea to refer to the activities of each Working Group and the resulting white papers, blogs, specifications, etc. (Details of the specification development process are omitted as they are internal information.)

Steering Committee (SC)

It oversees organizational operations and determines the overall direction of UEC.

Marketing Committee

Responsible for external communications and coordinating events etc.

Technical Advisory Committee (TAC)

Under the SC, we define and oversee the scope and priorities of UEC's technical activities. We also define high-level technology roadmaps, scopes of use cases, and technical goals. We approve technical specifications proposed by each Working Group and ensure the integrity of UEC's specifications. To join the TAC, you must be a General Member.

Working Groups

Each technical field is established, and they review and prepare specifications for the UEC technical specifications in their respective technical fields.

Working Groups

As of December 2024, the current Working Groups and their activities are as follows. [5]

Table 2. UEC Working Groups and Activities (as of December 2024)

Working Group	Activities
Physical Layer (PH)	It works on reducing latency and improving management at the physical layer. It defines the Ethernet physical layer, electrical and optical signal characteristics, APIs and data structures.
Link Layer (LL)	We are working on reducing latency and improving management at the link layer, and are formulating specifications to optimize the efficiency, security, and scalability of Ethernet.
Transport Layer (TR)	It develops transport specifications that will improve throughput, reduce latency, increase scalability, and improve management at the transport layer, which are essential for end-to-end data delivery. It is currently developing specifications for Ultra Ethernet Transport (UET), which is the core of UEC technology.
Software Layer (SL)	To support a wide range of AI/HPC use cases and applications, we will develop technical specifications including software APIs and open source implementations. Technical areas include optimizing remote memory access, enabling In Network Collectives (INC), security, management improvements, and storage.
Storage Working Group (ST)	We will work with other Working Groups to develop storage services that can be used with UEC-based AI/HPC workloads, including not only developing new specifications but also incorporating existing storage management best practices.

Working Group	Activities
Compliance Working Group (CT)	Work to ensure that services and devices comply with UEC defined technologies. Create tests to evaluate UEC implementations and ensure strict compliance with UEC standards. Also define interoperability targets between UEC compliant network devices (NICs, PCIe NICs, switches, etc.) according to UEC defined AI/HPC profiles.
Management Working Group (MG)	This standard will develop specifications to enhance the manageability of the UEC Fabric. It will define the model of UEC-compliant Transport Fabric End Point (FEP), management elements, RPC (Remote Procedure Call), etc. It will also include topology discovery, capability discovery, monitoring, and interoperability queries.
Performance and Debug Working Group (PD)	As the UEC specification evolves, it will define performance metrics, benchmarks, debug capabilities and tools for AI/HPC workloads, and will empower developers, DevOps and network operations teams by enhancing visibility and debuggability in UEC-compliant implementations.

Usual activities

On a day-to-day basis, each Working Group reviews specifications and prepares specification documents, which are then periodically reviewed and approved by the TAC to ensure overall consistency while developing specifications.

Regular discussions are conducted via mailing lists and online meetings, and technical discussions and specification formulation are carried out online, so remote participation is possible. However, online meetings are held during the day in North America, which is late at night and early in the morning in Japan.

Some Working Groups will meet offline several times a year to intensively discuss issues, and a member meeting is planned for 2025.

The frequency and form of activities may change after the first specifications are released in the first half of 2025, so if you are considering joining as a member, it is a good idea to check the latest activity status.

How to find out about UEC activities

Even if you are not affiliated with a member company, you can learn about UEC's activities in the following ways.

- Conference
  - Attending conference sessions will give you the most in-depth information on activities and technologies.
  - Also, if there is an exhibition, you will be able to talk to UEC members at the exhibition venue.
  - For more details, see Table List of conferences attended by UEC (archive of slides and videos) .
- Website: https://ultraethernet.org/
  - The website contains information about UEC's organizational structure, a list of member companies, Working Groups, and more.
  - A white paper [6] describing the motivation for establishing the UEC and the specifications being developed is available for download.
  - Blog: Latest News [7] will contain the latest information, so be sure to check it regularly.
- Member company press releases and websites
  - Member companies may publish information about their activities at UEC.

  - For the latest information on UEC-compatible products, it is a good idea to refer to the information provided by each member company.
- LinkedIn: https://www.linkedin.com/company/ultraethernet/posts/
  - By following the posts on the UEC LinkedIn account, you can find out about conferences that UEC is participating in and information that it is publishing.
  - We may also repost posts from member companies and other organizations. It can be a hassle to check the information of all member companies, so a good way to get started is to follow the UEC LinkedIn account.

UEC BLOG

We have excerpted some BLOG articles that may be helpful in understanding UEC's activities and technical content.

- 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/

List of conferences that UEC has attended (archive of slides and videos)

Table 3. List of conferences attended by UEC (archive of slides and videos)

Conference	link

OCP2024 (Open Compute Project)	https://www.opencompute.org/events/past-events/2024-ocp-global-summit
	• Leveraging UEC for Next Generation AI Networks (video 37min)
	  ◦ Presented by UEC: Uri Elzur, Intel (Architect, Computer Networks, GPU Networks)
	• Overview of Ultra Ethernet (video 15min)
	  ◦ 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)
	  ◦ 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)
	  ◦ J Metz, AMD (Chair Steering Committee, UEC)
	  ◦ Kurtis Bowman (AMD, Director, Architecture and Strategy)

Conference	link
SC24 (Super Computing)	https://sc24.supercomputing.org/program/proceedings-archives/ Note: Archive access is available to paid subscribers only • Industry Standards Working Together to Accelerate Innovation in AI and HPC ◦ https://sc24.supercomputing.org/proceedings/panel/panel_pages/pan114.html

NANOG 92 (October 2024)	https://nanog.org/events/nanog-92/
	• Keynote: Networking for AI and HPC, and Ultra Ethernet
	  ◦ Hugh Holbrook, Arisa Networks (VP Software Engineering)
	  ◦ Video: https://youtu.be/0roIi1pscts?si=XmZAjfBFM3CibWBb
	  ◦ Slide: 20241021_Holbrook_Keynote_Networking_For_v1.pdf (https://storage.googleapis.com/site-media-prod/meetings/NANOG92/5182/20241021_Holbrook_Keynote_Networking_For_v1.pdf)

HOTI31 (2024) (Hot Interconnects)	https://ieeexplore.ieee.org/xpl/conhome/10664198/proceeding
	• Day 2: Invited Talk: Ultra Ethernet Consortium (UEC) overview
	  ◦ Uri Elzur, Intel
	  ◦ Video: https://www.youtube.com/watch?v=LtifmYChRTo

UEC-related organizations

In addition to the Linux Foundation, UEC is working with the following organizations to develop specifications. The level of collaboration varies from organization to organization, but if you want to gain a deeper understanding of UEC's technology, it is a good idea to refer to the trends and technologies of these organizations.

Open Compute Project (OCP)

OCP is a non-profit organization that promotes open sourcing of hardware specifications and designs that can efficiently support the demands of compute infrastructure. It began in 2011 when Facebook (formerly Meta) published the design specifications for its own data center equipment. By allowing hardware vendors to provide equipment based on the published specifications, costs were reduced and procurement became easier. There are projects for each product field, such as servers, storage, and networks, and participating companies are developing new products and publishing design specifications. As an example in the field of networks, white box switches and the switch OS (SONiC, etc.) that controls them are being developed. (SONiC moved to the Linux Foundation in 2022) Many UEC-related sessions are held at OCP conferences, including the annual OCP Global Summit, which can help you understand the trends of UEC.

- OCP home page: https://www.opencompute.org/
- OCP Global Summit (slides and video archive): https://www.opencompute.org/summit/global-summit

OpenFabrics Alliance (OFA)

The OpenFabrics Alliance (OFA) is an industry group that promotes high-performance networking technologies. It supports the adoption of networking technologies such as RDMA (Remote Direct Memory Access), InfiniBand, iWARP, and RoCE for datacenters, high-performance computing (HPC), and cloud infrastructure. OFA also provides open

source software (OFED: OpenFabrics Enterprise Distribution). The OpenFabrics Interfaces Working Group develops a library of interfaces also known as Open Fabrics Interfaces (OFI) `libfabric` [8].

- OFA homepage: https://www.openfabrics.org/

ULTRA ACCELERATOR LINK (UAlink)

Ultra Accelerator Link™ (UALink™) is an open industry standard internal interconnect that connects accelerators. As of 2024, UEC will be formulating scale-out technology that uses Ethernet as an external interconnect, and the internal interconnect technology required for scale-up will be based on the assumption that it will use technology from other organizations, such as UALink, rather than UEC.

- UALink Consortium homepage: https://www.ualinkconsortium.org/

Storage Networking Industry Association (SNIA)

SNIA is an industry organization that develops technical standards and provides educational programs related to information management, with a focus on storage. As the UEC Storage Working Group exists, we are formulating technology with the assumption that UEC technology will also be used in storage.

- SNIA Homepage: https://www.snia.org/
- SNIA Japan Branch: https://www.snia-j.org/

IEEE

The IEEE (Institute of Electrical and Electronics Engineers) has working groups that handle technologies related to Ethernet. For example, IEEE 802.3 is a working group that defines the media access control (MAC) for the physical layer and data link layer of wired Ethernet. IEEE 802.1 is also developing the standard for LLDP (IEEE 802.1AB), which is being considered for extension by UEC.

- IEEE Home Page: https://www.ieee.org/

Background of UEC Initiatives

In this section, we will introduce UEC's objectives and the technologies we are working on based on the contents of the UEC white paper [9].

About the information contained in "UEC Initiative Background"

The content introduced in "Background of UEC's Initiatives" is based on the "UEC White Paper" [10] published by UEC in 2023. In addition, other contents are also compiled based on publicly available information such as the UEC website and presentation materials at conferences.

The UEC specification is currently being formulated for release in Q1 2025, and various discussions are currently underway. Therefore, there may be changes to the contents described here before the specification is released. When considering the use of UEC technology, please be sure to check the latest information, including the UEC technical specifications.

The Ethernet Advantage

The advantages of Ethernet are as follows:

- A multi-vendor ecosystem offers a wide range of compatible Ethernet switches, NICs, cables, transceivers, optical devices, management tools and software.
- The scalability of IP networks (routing) has been proven, making it possible to realize rack-scale, building-scale, and data center-scale networks.
- A wide range of tools exist to test, measure, deploy and efficiently operate Ethernet networks.
- There is a proven track record of reducing costs through competitive ecosystems and economies of scale.
- The IEEE Ethernet standard has demonstrated its ability to evolve rapidly and regularly across many of its physical and optical layers.

Taking advantage of these advantages of Ethernet,

- Maximize Throughput
- Minimizing Tail Latency

We have identified the following network requirements necessary to achieve these goals and are formulating the technologies to achieve them.

- Multipass and packet spray
- Flexible Ordering
- Congestion control mechanism optimized for AI and HPC
- End-to-end telemetry
- Large scale, stable, reliable

Multipass and packet spray

In conventional Ethernet networks, only one route could be used due to the spanning tree protocol. Later, in data center fabrics, by connecting nodes (switches) with Layer 3 (IP), it became possible to use multiple paths using technologies such as Equal-Cost Multipath (ECMP). ECMP usually uses hashing to distribute traffic, which has the disadvantage of causing traffic bias. In addition, to prevent packet reordering, traffic of the same flow (Layer 4: TCP, UDP session) must pass through the same path. Therefore, if the number of flows is small, the bias becomes more noticeable as the number of traffic with the same header information, which is the basis of the hash calculation, increases.

In recent years, there are functions such as Dynamic Load Balancing (DLB) that dynamically allocate paths to be used based on the utilization rate of each port, but when there are multiple nodes, it is difficult to reliably select the path with the lowest load across the entire network.

Therefore, UEC uses a technique called "packet spraying" that distributes packets to all paths. The drawback of packet spraying is that it causes packet reordering, but this is handled by the transport layer (UET).

Flexible Ordering

Traditional technologies, such as RoCEv2, require packets to be received in the order they were sent. Specifically, when an out-of-order packet occurs due to packet reordering or packet loss, Go-Back-N must resend all packets after the out-of-order packet. This results in underutilization of available inter-switch links and increased tail latency, requiring more time to complete the job.

But ideally all links should be used and ordering should only be enforced when the AI workload requires it.

In AI workloads, collective communication operations such as All-Reduce and All-to-All are common. The key to completing these operations quickly is high-speed bulk communication (data transfer), and for AI applications, what matters is "Did all the data in the message arrive?", not the order in which the data arrives.

Flexible Ordering improves the efficiency of data transfer by relaxing the constraints on the order in which packets arrive. For example, it eliminates the need for reordering of out-of-order packets that occurs when performing packet spraying.

Congestion control mechanism optimized for AI and HPC

Congestion can occur in three places:

1. The link between the source server and the switch
2. Inside the switch fabric (the path between the switches that connect the source and destination servers)
3. Link between destination server and switch (Incast)

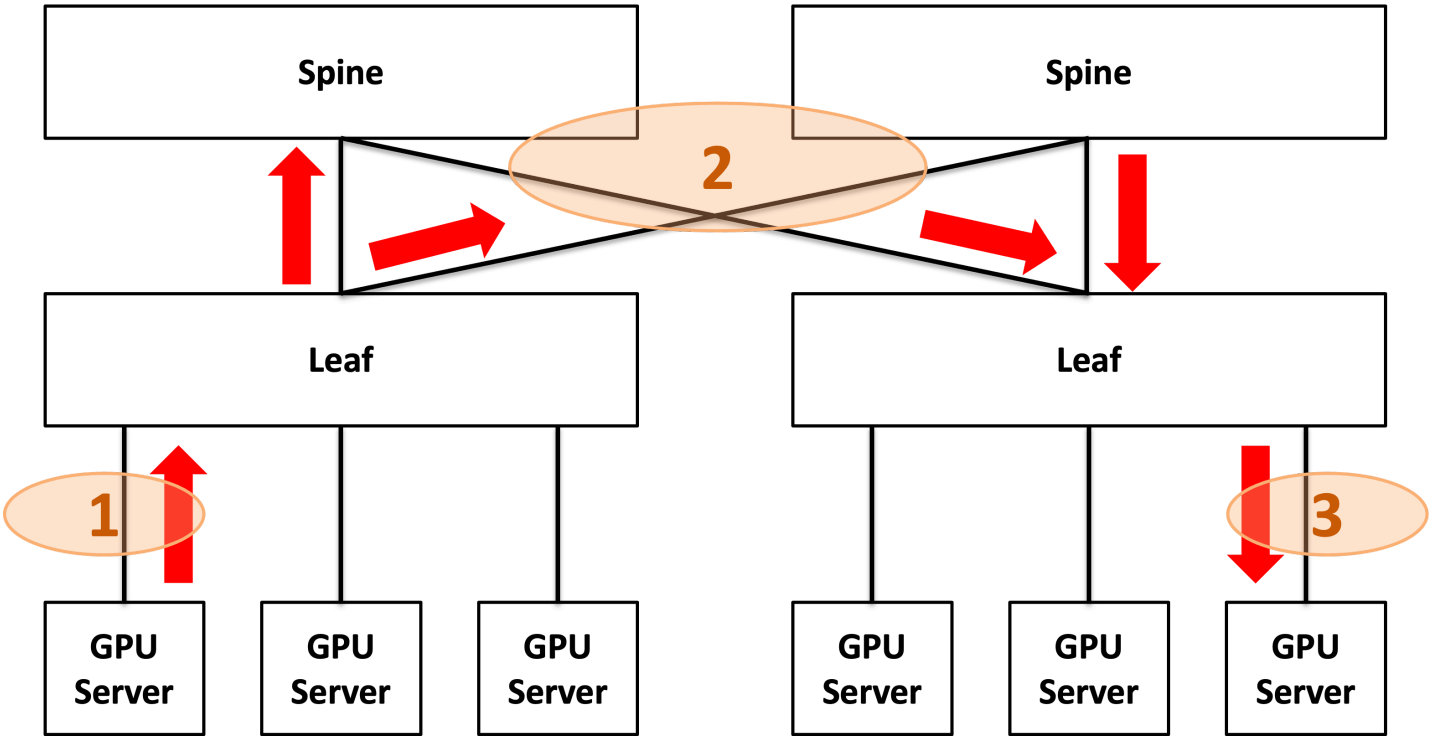

Figure 1. Classification of congestion locations

Congestion on the link between the source server and the switch can be controlled because the source server can see all of the traffic being sent.

"2. Congestion inside the switch fabric" is minimized by multipathing and packet spraying.

"3. Link between destination server and switch (Incast)" occurs when data is transferred from multiple sources to the same destination, such as in an All-to-All operation. Algorithms such as DCQCN have been used to control congestion caused by Incast, but they are not sufficient, and a congestion control mechanism that meets the following requirements is considered necessary.

How quickly the transfer rate starts

To quickly increase the transmission speed up to wire rate at the start of transmission in a wide-bandwidth, low-latency network without degrading the performance of existing traffic.

Fair sharing of final links

Incast can be controlled by sharing final links fairly without incurring packet loss, retransmissions, or increased tail latency.

Minimal configuration and tuning

No (minimized) tuning or configuration is required (as was required with DCQCN) as traffic mix changes, compute nodes evolve, link speeds increase, and hardware evolves.

At UEC, we are designing congestion control algorithms (UETs) for future AI workloads that support these requirements and work in conjunction with multipath and packet spraying.

Note

In the future, the location of congestion may change if it becomes necessary to virtualize source servers or partition networks in order to achieve multi-tenancy services.

End-to-end telemetry

Optimization of congestion control algorithms is achieved through end-to-end telemetry: rapid congestion notification from network nodes (switches) to packet transmission schedulers and pacers when the source or destination manages the transmission schedule, making congestion control algorithms more responsive and accurate.

This reduces congestion, reduces dropped packets, and reduces queues, resulting in improved tail latency.

Large scale, stable, reliable

UEC aims to be scalable to support up to 1 million endpoints.

The Challenges of Lossless Fabric

RoCE, which is based on lossless Ethernet, has been widely deployed, but it requires expert tuning and operational monitoring to achieve maximum performance, making it difficult to reduce total cost of ownership (TCO).

In addition, temporarily suspending packet forwarding using PFC to achieve lossless transmission has the following challenges:

Head-Of-Line Blocking

- HOL blocking is a phenomenon in which when the port through which the first packet in the queue is forwarded is congested, the entire queue is blocked even if other packets can be forwarded.
- PFC pauses traffic of a certain priority, which causes the head of the queue to back up, blocking subsequent traffic.
- This has the effect of reducing throughput and increasing latency.

Congestion spreading

- When a link experiences congestion (e.g., its buffers become full), it can no longer receive data and will send a PAUSE frame to the upstream device.
- As a result, the upstream device becomes congested as well and sends a PAUSE frame further upstream.
- In this way, congestion propagation occurs, and the effects of congestion can spread beyond the few links where it first occurs, and can affect the entire network.

Deadlock occurs

- If multiple devices end up sending PAUSE frames to each other, a "deadlock" can occur that brings the entire network to a halt.

UEC believes that issues such as preventing and recovering packet loss and controlling congestion should be resolved at the transport layer, and is therefore developing UET (Ultra Ethernet Transport) to replace RoCEv2, network layer technologies to be used in cooperation with UET, Ethernet extension technologies (link layer), In Network Collectives (INC), etc.

UEC Technology Overview (Overall overview and detailed explanations of each topic)

In this chapter, we will first provide an overview of the technologies that UEC is working on and introduce the technologies that are being developed at each layer.

After that, for those who want to dig deeper into the technology, we will provide detailed explanations of each technology for which specifications are being developed at UEC.

Overview of UEC technology

At UEC, we are developing technologies at a wide range of layers, from the transport layer to the link layer, to achieve high performance and efficiency in RMA, which is important when using Ethernet as an interconnect between AI/HPC accelerator nodes. These include improvements based on operational experience with RoCEv2, an existing RMA over Ethernet technology.

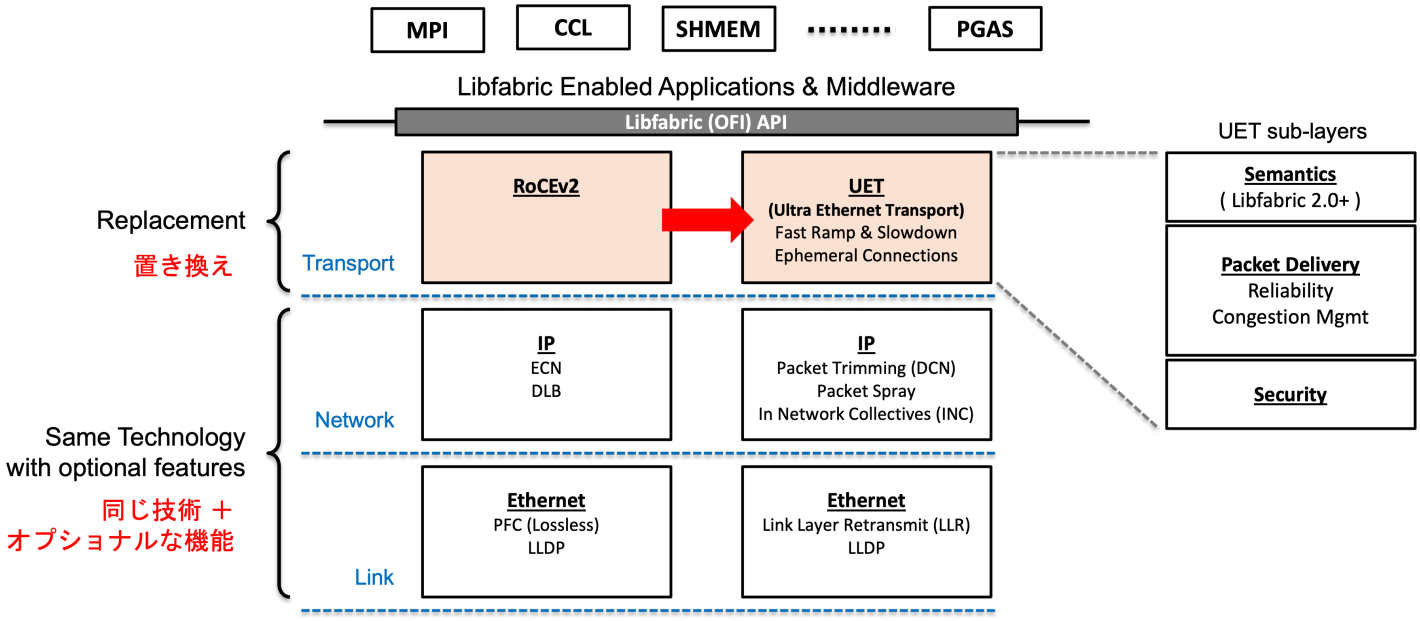

Figure 2. UEC technology stack and comparison with RoCEv2

The most notable feature is that the transport layer has been replaced from RoCEv2 to the newly defined UET (Ultra Ethernet Transport), as shown in Figure [uec_tech_stack] . UET adopts and extends the libfabric API, a modern RMA API for HPC that is widely used in HPC applications.

> NOTE
>
> Figure: Please note that the comparison of the UEC technology stack with RoCEv2 is based on the Libfabric API. Applications and middleware that use Libfabric may be able to use UEC technology by responding to API extensions, but applications and middleware that do not use Libfabric may require major changes (program modifications) to support UET.

As its name suggests, UET is a transport layer technology responsible for congestion control, etc., but in order to realize efficient transport protocols and algorithms, new technologies have also been developed at the network layer, and they work in close cooperation with each other.

Therefore, **even for network layer technologies, the use of UET is a prerequisite.**

However, to enable incremental adoption of UEC technology, functions below the network layer are optional, and the UET itself can be used with existing switches.

In addition, a technology called In Network Collectives (INC) is being developed to execute part of collective calculations within the switch fabric. In addition, security features are being added because the business importance of models generated by AI/HPC and the data used to create the models is increasing. (An explanation of security features is omitted.)

> NOTE
>
> Applications in AI/HPC workloads send and receive (instruct) data in messages rather than packets, so efficient data transmission and reception requires a method that is aware of message boundaries. Because a message is made up of one or more packets, it is important to be aware of whether the order of messages and/or packets is significant or whether they are order-independent.

UEC Profile

UEC is formulating technical specifications that can support different workloads such as AI and HPC. Also, as mentioned above, not all UEC technologies are mandatory, and some technologies and functions are defined as optional.

UEC allows you to define profiles, allowing you to selectively use the technologies and functions required for each type of workload.

Technologies currently under consideration for the future (only publicly available information is listed)

UEC will continue development after version 1.0, which is scheduled to be released in Q1 2025, and plans to gradually expand use cases and functions in the future, such as the following:

- Storage Applications
  - Storage APIs on UET
- Management Features
  - Configuration management with OpenConfig and RedFish
- Telemetry Function
  - Congestion Signaling (CSIG)[11] and Back to Sender (BTS)[12]
- others
  - Standardization of performance measurement and debugging methods
  - Compliance testing for each profile (including optional features)

UEC technology list by layer

Table: List of UEC technologies by layer shows the UEC technologies that have been announced as being currently under consideration, grouped by layer. (Some functions are dependent on technologies in other layers, centered on UET, so each technology is classified into its core layer.)

Below, we will provide an overview of each technology.

Table 4. UEC technology list by layer

Layer	Features

Layer	Features
Application	(Various AI/HPC applications that support Libfabric and UEC extensions)
Software APIs	• Libfabric 2.0 Extensions

Transport	• UET (Ultra Ethernet Transport)
	  ◦ 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)	• Packet Trimming • Packet Spray
Link Layer (Ethernet)	• Link Layer Retransmit (LLR) • LLDP (capability negotiation)

UET: Transport Protocol for AI and HPC

To address the challenges of lossless fabrics, Ultra Ethernet Transport (UET) was developed as a transport protocol that does not require losslessness.

UET is designed to meet the following requirements:

- An open protocol specification designed from the beginning to run over IP and Ethernet
- No need to tune network or workload specific congestion algorithms
- No need for a centralized load balancing algorithm or controller
- Employs multipathing and packet spraying (to prevent congestion and head-of-line blocking)
- Incast management mechanism (controlling fan-in with minimal drops)
- Efficient rate control algorithm (rapidly reaches wire rate)
- API allows out-of-order packet delivery (with optional support for in-order delivery)
  - Maximizing network and application concurrency
  - Minimizing Message Delay
- Future-proof your network with scalability to support 1 million endpoints
- Designed to achieve wire-rate performance using off-the-shelf hardware for 800G, 1.6T and future faster Ethernet speeds

To achieve these requirements, UET extends not only the transport layer but also the semantic layer. This requires support for libraries such as Libfabric, and UEC plans to release a reference implementation of Libfabric Provider.

As such, since UET is a core technology of UEC and replaces RoCEv2, the scope of impact is large, so when using it, it is necessary to fully understand which components need to be changed.

NOTE

- In addition to packet spraying and the flexible ordering required for it, technologies such as packet trimming and ephemeral connections, which will be described later, cannot be used in conventional DCQCN (+RoCEv2) and require the use of UET as a transport protocol.
- In Libfabric, an "endpoint" corresponds to a socket in the Socket API (TCP/UDP), so in terms of scalability, "1 million endpoints" does not necessarily mean 1 million nodes.

Packet Spray (Posted Buffer)

Technologies such as Dynamic Load Balancing (DLB) are used to prevent hash bias in switches, but they depend on various parameters, such as the idle threshold for recognizing flowlets to prevent packet reordering, and the granularity of the bandwidth threshold for determining path quality, and are said to require tuning according to the workload. *[NOTE]*

To make these unnecessary, UET uses a "Packet Spray" method that sends packets using all available paths. Packet Spray is also known as multipathing or out-of-order delivery (OOO).

Packet reordering caused by this method is handled on the server side (transport layer) (not in the network fabric such as a switch). In order to handle reordering efficiently, UET uses "Posted Buffer". (In a production environment, it is expected that transport layer processing will be offloaded to the NIC.)

Posted Buffer is a method in which a buffer ID is assigned to each packet, and data is written to the correct buffer regardless of the order in which it is received. This eliminates the need for reordering after packet reception, and allows data to be written directly to the memory of the receiving process (RMA) without waiting for intermediate packets to be received.

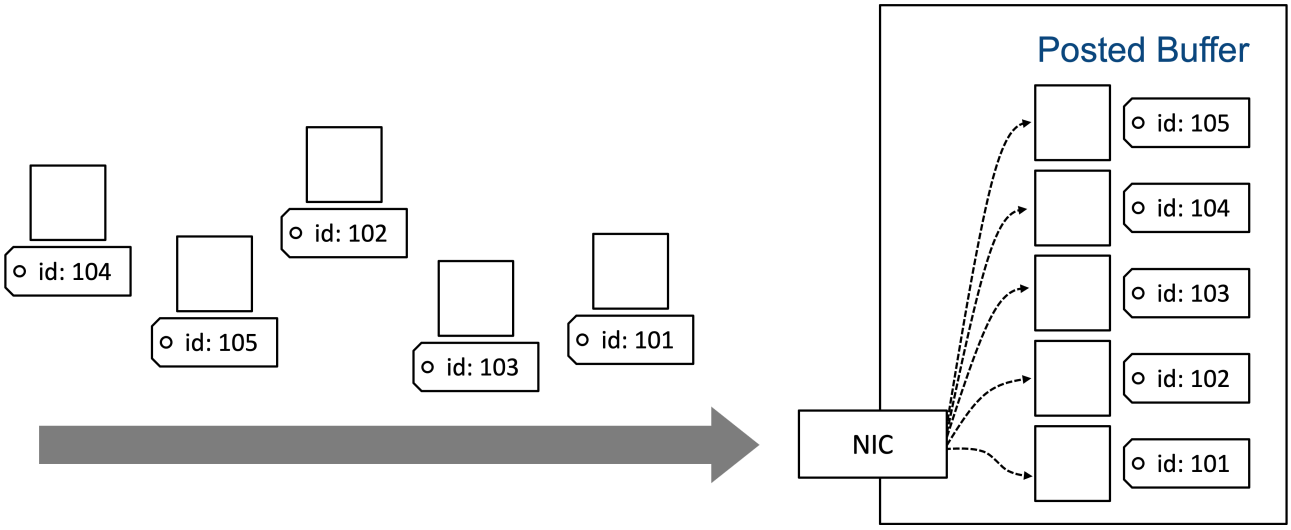

Figure 3. Posted Buffer

NOTE

The need for tuning in an existing environment is discussed in the section on points to consider when using UEC .

Packet Trimming (optional)

There are several ways to detect packet loss at the transport layer:

- Out-of-order (OOO) packet detection
- timeout

However, because Packet Spray allows OOO, it cannot detect packet loss using OOO. Also, when using a timeout, because Packet Spray uses various paths, it is necessary to set a timeout value with a certain margin considering delay fluctuations, which increases the time until packet loss is detected, adversely affecting performance.

Therefore, UEC has adopted "Packet Trimming" as a new method to detect packet loss.

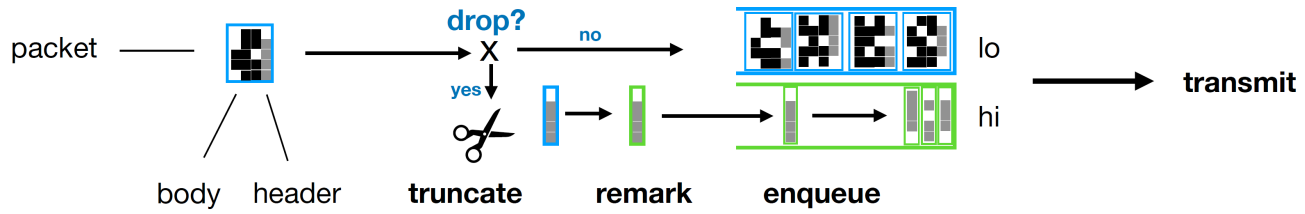

Figure 4. Packet Trimming in action: Figure taken from NANOG92[1]. Commentary translated by the author.

- Truncate to 64 bytes instead of dropping
- Mark DSCP as "trimmed"
- Send using high priority queue (enqueue)

Packet Trimming is a switch function that, instead of dropping packets when congestion occurs as shown in the figure: [Packet Trimming Operation] , trims and truncates the packet leaving only the minimum necessary header information that can be uniquely recognized at the transport layer, and forwards it using a high-priority queue. This allows the receiving side to be quickly notified that "congestion has occurred" and "a packet has been dropped," and makes it possible to take measures, including congestion avoidance, such as requesting retransmission or adjusting the transmission pace, with less delay than with the timeout method.

This feature is particularly effective in AI workloads such as LLM, where message sizes (packet sizes) are large. (For example, trimming a 4096-byte packet to 64 bytes results in a 64-fold data reduction.)

NOTE

- Packet Trimming is already supported on the Broadcom Tomahawk 5 with a feature called Drop Congestion Notification (DCN).[13]
- Additions to SAI are discussed in "PR#2077 Add packet trimming API"[14] and the design document "SAI-Proposal-Packet-Trimming.md"[15] is a useful reference.

Congestion control algorithm for UET

AI/HPC workloads are characterized by the fact that data transfer and calculations are repeated many times, and that large amounts of data can be sent and received in a short period of time because of the availability of a wide bandwidth via Packet Spray. For example, if an 800Gbps bandwidth is available, sending 1MB of data takes 10usec.

For this reason, rather than using algorithms such as TCP slow start, we use a congestion control algorithm with the following characteristics that are suitable for AI/HPC workloads.

Characteristics of UET Congestion Control Algorithms

- Fast Ramp: When transmission starts, the transfer rate is increased to the wire rate as quickly as possible.
- Fast Slowdown: Quickly reduce (back off) the sending rate when congestion is detected.

There is also an optional congestion control feature at the receiver.

Congestion control function on the receiving side

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

Ephemeral Connection

UET uses Ephemeral Connection to reduce the delay when starting a session. It is a short-lived connection and has the following characteristics:

Features of Ephemeral Connection

- No handshake required, eliminating delays before transfers start
- Reduced per-connection state

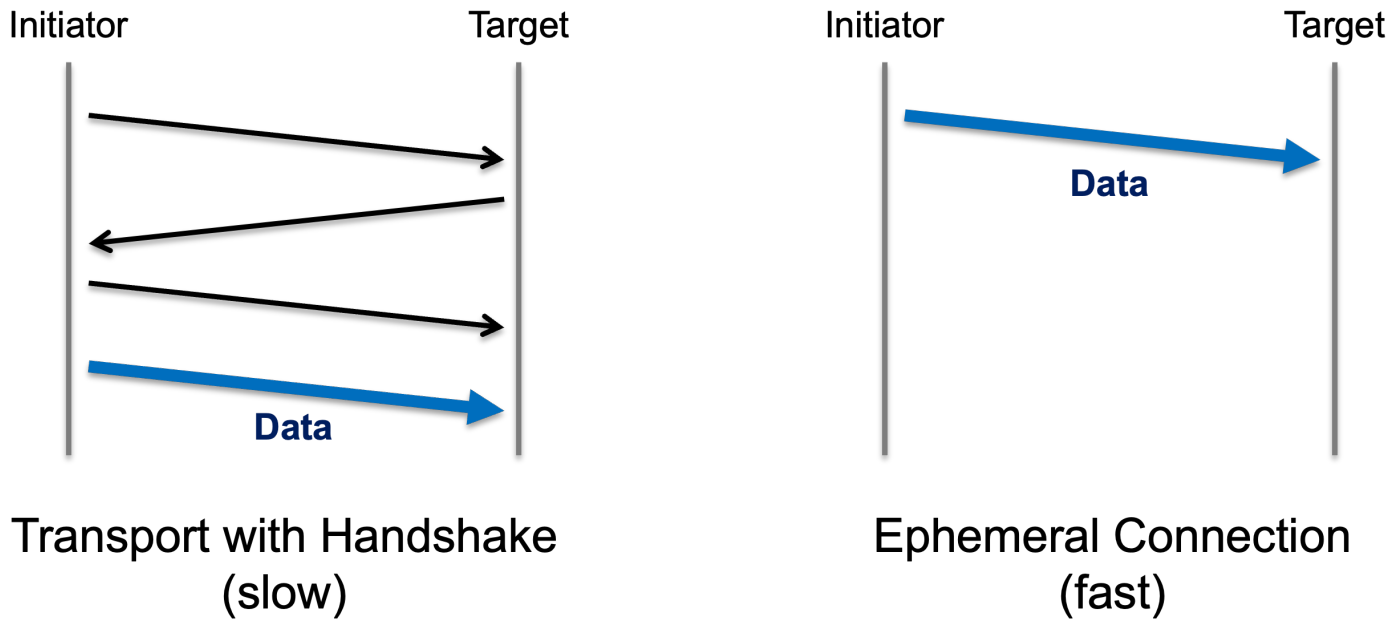

Figure 5. Ephemeral Connection

AI/HPC workloads are characterized by a large number of short-duration (ms~us) connections. Therefore, reducing the time (latency) required to start and end a connection contributes to improving performance.

In addition, in high-speed communications, offloading transport layer processing to the NIC is essential, and in order to efficiently process many sessions with limited resources, reducing the state per connection has a significant effect on reducing valuable NIC resources.

Ethernet extensions (LLR, CBFC, LLDP)

UEC has developed Link-Layer Retransmission (LLR) as an extension of the Ethernet link layer, and has also extended LLDP as a means of capability negotiation, including support for LLR.

These are optional features, so if you don't want to use them you can use UEC Technology (UET) with your existing switch.

> NOTE | In some cases, such as the UEC Blog [16], Retry is used instead of Retransmission.

Link-Layer Retransmission (optional)

Link-Layer Retransmission (LLR) is a technology developed for the following purposes to prevent packet loss at the link level by using CBFC: Credit Based Flow Control.

- Reduce tail latency by using local retransmission and eliminating the need for end-to-end retransmission
- Handling link and transceiver failures

Since AI/HPC uses a large number of transceivers, how link layer failures are handled is also likely to have an impact on performance and availability.

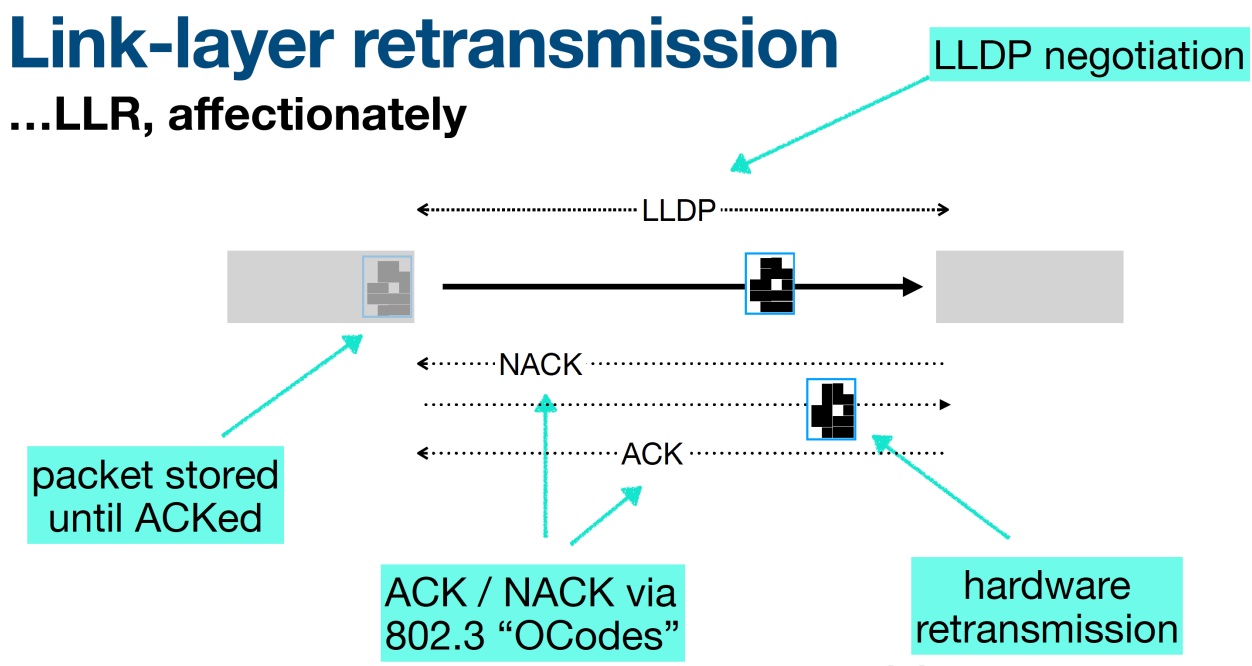

Figure 6. Link-Layer Retransmission (LLR): Figure adapted from NANOG92 [2]

In Network Collectives (INC)

In Network Collectives (INC) is a technology that reduces latency, limits network bandwidth consumption, and speeds up processing by performing collective communication operations at nodes on the network (switch fabric).

INC is generally used as an abbreviation for In-Network Computing (Computation), a technology that aims to improve efficiency and speed by performing calculations within a network, not limited to collective communication operations.UEC's In Network Collectives is also classified as In-Network Computing.

NOTE

About the notation of INC

Currently, in discussions within the UEC SW Working Group, `-` the notation "In Network Collectives (INC)" is used without a hyphen () between "In" and "Network", so this document does not include one either. However, there is a possibility that a hyphen will be included during the process of reviewing the UEC specifications. After the UEC INC specification, scheduled for release in Q1 2025, is released, please follow the terminology defined in that document.

Conversely, the author feels that outside of UEC, hyphens are often used, such as in In-Network Computing, so the hyphen is used. For example, NVIDIA's official documentation uses the term In-Network Computing.

An earlier technology is the Scalable Hierarchical Aggregation and Reduction Protocol (SHARP TM) developed by Mellanox (formerly NVIDIA) , but SHARP only works on InfiniBand and requires an NVIDIA NIC (ConnectX, BlueField).

In contrast, UEC's INC has an open specification and is planned to be able to operate on multi-vendor Ethernet switches (ASICs) and NICs.

Details of UEC INC have not been disclosed, but the differences between it and SHARP are as follows, but conceptually they are very similar:

- Works over Ethernet
- Use UET for data transfer

Therefore, by researching SHARP until the specifications are released, it will be possible to get a rough idea of how UEC INC works as well.

- SHARP Portal Page: "NVIDIA Docs Hub > NVIDIA Networking > Accelerator Software"[17]
- SHARP original paper: IEEE Xplore "Scalable Hierarchical Aggregation Protocol (SHArP): A Hardware Architecture for Efficient Data Reduction"[18][19]
  - It can be downloaded from the NVIDIA website[20]

In addition to the official documentation mentioned above, there are also tutorials and conference talks on YouTube that present research, so please search for them.

NOTE

These are the author's personal impressions of In Network Collectives. I don't fully understand it and I haven't organized my thoughts, so I'm looking for people to research and think about it with me.

- Because UEC is multi-vendor, it is likely that there will be a variety of switch ASICs, server-side NICs, middleware, etc. that support INC.
- Therefore, when considering implementation or evaluating it, it is necessary to organize the points to compare between INC implementations and with SHARP.
- Examples of evaluation perspectives:
  - The evaluation must be performed after confirming or assuming the characteristics of the workload.
  - Compared to HPC, AI (LLM) has a larger message size.
  - Since the memory of the Switch ASIC is limited, when the message size is large, such as in an AI workload, it is necessary to divide it, transfer it, calculate it, combine it, etc.? (hypothesis)
  - There are various possible algorithms for dividing and calculating, so even with the same INC, the algorithm may differ depending on the combination of INC Switch and the library, which could have a significant impact on performance.
  - In order to select the optimal implementation by determining the workload, data size, and the resulting performance, it seems necessary to have a deep understanding of both the implementation of INC Switch and the algorithm that executes MPI CC operations with INC. It is not yet clear whether this part will be standardized by UEC.
- Is this helpful? Session 13 "Designing In-Network Computing Aware Reduction Collectives in MPI"
  - https://www.openfabrics.org/2024-ofa-virtual-workshop-agenda/

Points to consider when using UEC

As a major change from conventional AI/HPC network technology, UEC replaces the transport from RoCEv2 to UET. Therefore, when considering the use of UEC technology, it is necessary to sort out "what requires change (development) and what can be used as is" in terms of both hardware and software, and to proceed while understanding the impact

on systems, services, personnel for construction and operation, etc., as well as the advantages and disadvantages of the change.
In this chapter, we summarize the key points to consider when considering the use of UEC technology from the following perspectives:

- Changes and points to note when using UEC
- Points to consider when using UEC
- UEC specifications, product roadmap, and UEC-compatible products (which vendors will provide compatible products and when?)

summary

- UEC replaces RoCEv2 with UET, "improvements across the stack"
  - It is important to understand the scope of impact, including AI applications and middleware
  - Technical verification is required in collaboration with various vendors such as GPU, NIC, Switch ASIC, Switch OS, etc.
- The timing of the release of each elemental technology of UEC varies
  - Step-by-step technical verification is necessary
- Many of the link and network layer features are optional
  - You can start with verification using existing Ethernet switches

Changes and points to note when using UEC

Changing technology stack

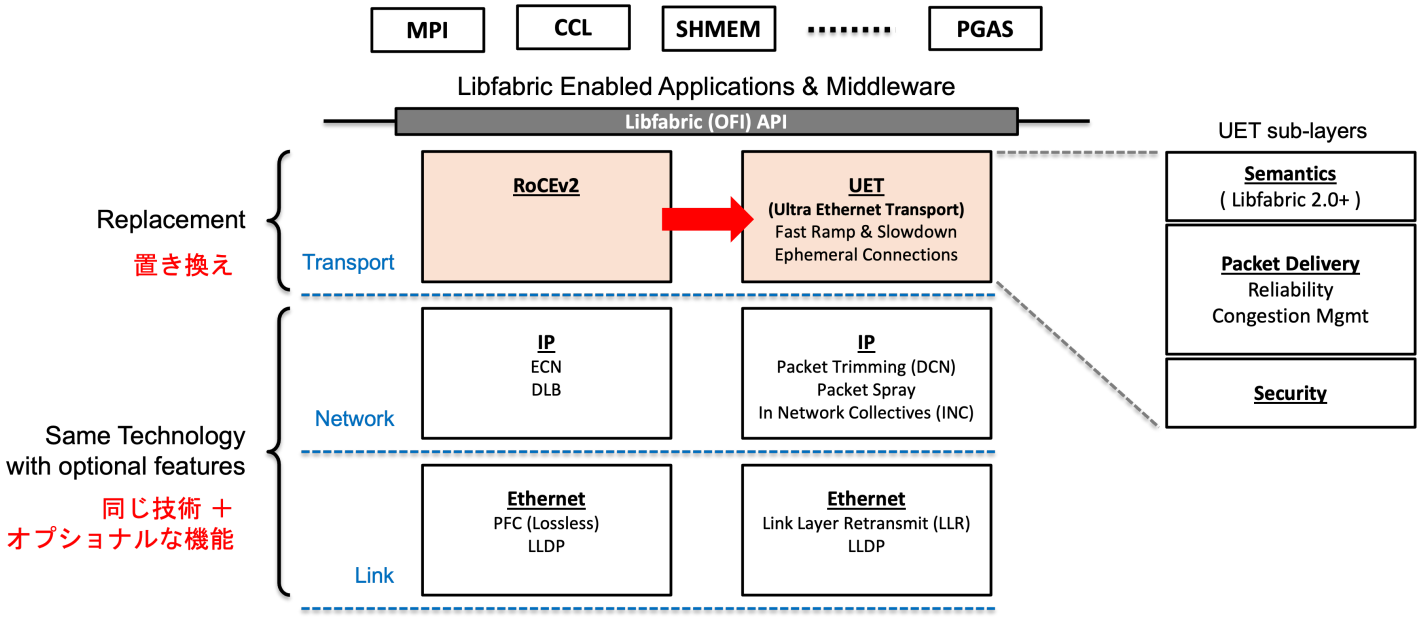

Figure 7. UEC technology stack and comparison with RoCEv2

The biggest difference between UEC's AI/HPC network technology using RoCEv2 is that the transport layer has been replaced with the newly defined UET (Ultra Ethernet Transport) from RoCEv2, as shown in Figure: Comparison of UEC's technology stack with RoCEv2 . UET adopts and extends the libfabric API, a modern RMA API for HPC that is widely used in HPC applications.

In other words, applications and middleware that use libfabric may be able to use UEC technology by supporting API extensions, but applications and middleware that do not use Libfabric may need to make major changes (program modifications) to support UET.

> NOTE
>
> Support for APIs other than libfabric is sometimes discussed, but there is currently no clear roadmap. If there are many users who require support for APIs other than libfabric, the discussion may move forward, so it would be a good idea to provide input to UEC member companies, such as use cases and libraries that would be useful to support them.

On the other hand, since extensions to IP and Ethernet are only optional functions, UET can be used with existing switches. However, to get the most out of UET, it is necessary to at least support IP layer functions such as Packet Trimming.

In addition, since Lossless Ethernet is no longer necessary, tuning of PFC buffer settings is no longer necessary, so network layer settings are expected to be simpler than before. Conversely, the stability and performance of Libfabric Provider, which implements UET, may differ for each NIC, including the driver. Therefore, the need for server-side tuning is unknown, and at least in the initial stages, verification is expected to be necessary for each NIC.

Comparison of available technologies by device configuration

The available technologies vary depending on the device configuration, not only when using UEC, but even in the case of conventional RoCEv2.

Figure: AI/HPC Hardware Matrix summarizes the technologies available for each GPU/NIC/Switch combination. Please refer to this when considering the technologies and equipment you want to use.

For example, to use Adaptive Routing, you need to purchase equipment from NVIDIA. Also, as of January 2025, InfiniBand is required to use INC, but in the future, INC will be available over Ethernet by selecting UEC-compatible NICs and switches.

	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 Routring?	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 technology and related components (Summary)

We have summarized the components involved when using each UEC technology.

Basically, when using UEC, you need to consider all of the following: NIC, Switch ASIC, Switch OS (NOS), and AI applications. Please refer to this when you want to narrow down the products and vendors to check whether each technology is supported.

Regarding "AI applications," since the scope of measures required varies depending on the app itself, middleware, etc., it may be necessary to categorize them in more detail and confirm the impact.

Technology	Description	Switch ASIC	Switch OS	NIC (driver)	AI Apps
UEC Transport (UET)	A transport protocol that replaces RoCEv2. Supports out-of-order packet reception, best effort (non-lossless), etc.	optional	optional	YES	YES
Packet Trimming	Also known as Drop Congestion Notification (DCN). When congestion occurs, packets are trimmed to a certain size and forwarded with high priority to inform the receiving node of the congestion. This eliminates the delay caused by the drop detection timer.	YES	YES	YES	NO
Link Level Retry (LLR)	Manages transmission and retransmission to prevent packet loss at the Ethernet link layer. Credit Based Flow Control (CBFC)	YES	YES	YES	NO
LLDP Negotiation	Negotiates capabilities between nodes, such as LLR support. UEC extensions are made to the conventional LLDP.	YES	YES	YES	NO
Packet Spray, Ordered(ROD) and un-ordered(RUD)	By spraying packets to all available paths, packet-level load balancing is performed, improving the utilization of the Ethernet fabric. This is a UET function that embeds an ID in the packet (message) that identifies where in memory it should be stored, eliminating the need for packet reordering (buffering) at the receiving node.	YES?	YES?	YES	NO
Ephemeral connections	Putting session information in the first packet eliminates the need for handshakes. (PDC, PDS) Eliminates handshake overhead in workloads that involve repeated burst data transfers, improving throughput by shortening transfer times for each burst.	NO	NO	YES	NO
In Network Collectives (INC)	Offloads collective communication operations to switches in the Switch Fabric. The expected benefits are improved performance (reduced congestion and data transfer volume, reduced latency), improved resource utilization (processors, accelerators, memory), and reduced power consumption.	YES	YES	YES	??

Points to consider when using UEC

I have listed some points that came to mind to consider when using UEC.

The points that should be emphasized will likely differ depending on the organization (user, manufacturer (vendor), system integrator), role (network engineer, server engineer, application developer), use case, budget, etc.

This article lists only what the author could think of as of January 2025, so please provide feedback if you have any suggestions, such as "There is another perspective to this," or "When you look at it from this perspective, the advantages and disadvantages are different."

Why Ethernet over Infiniband?

Before considering UEC, you need to consider whether to choose InfiniBand or Ethernet. For example, the following points should be considered:

- Stability: Are latency and throughput stable? (The final indicators are shortening job completion time and stability.)
- Flexibility: Is it easy to change configurations?
- price
- Ease of securing human resources and training costs
- The Need for Multitenancy
- Will it be used for traffic other than AI/HPC?

The perspectives that are important vary depending on one's position and situation, but from the perspective of a service provider, the LINE Yahoo! blog "GPU Cluster Network and Its Design Philosophy (Rethinking AI Infrastructure Part 2)"[21] published in January 2025 clearly outlined the concept.

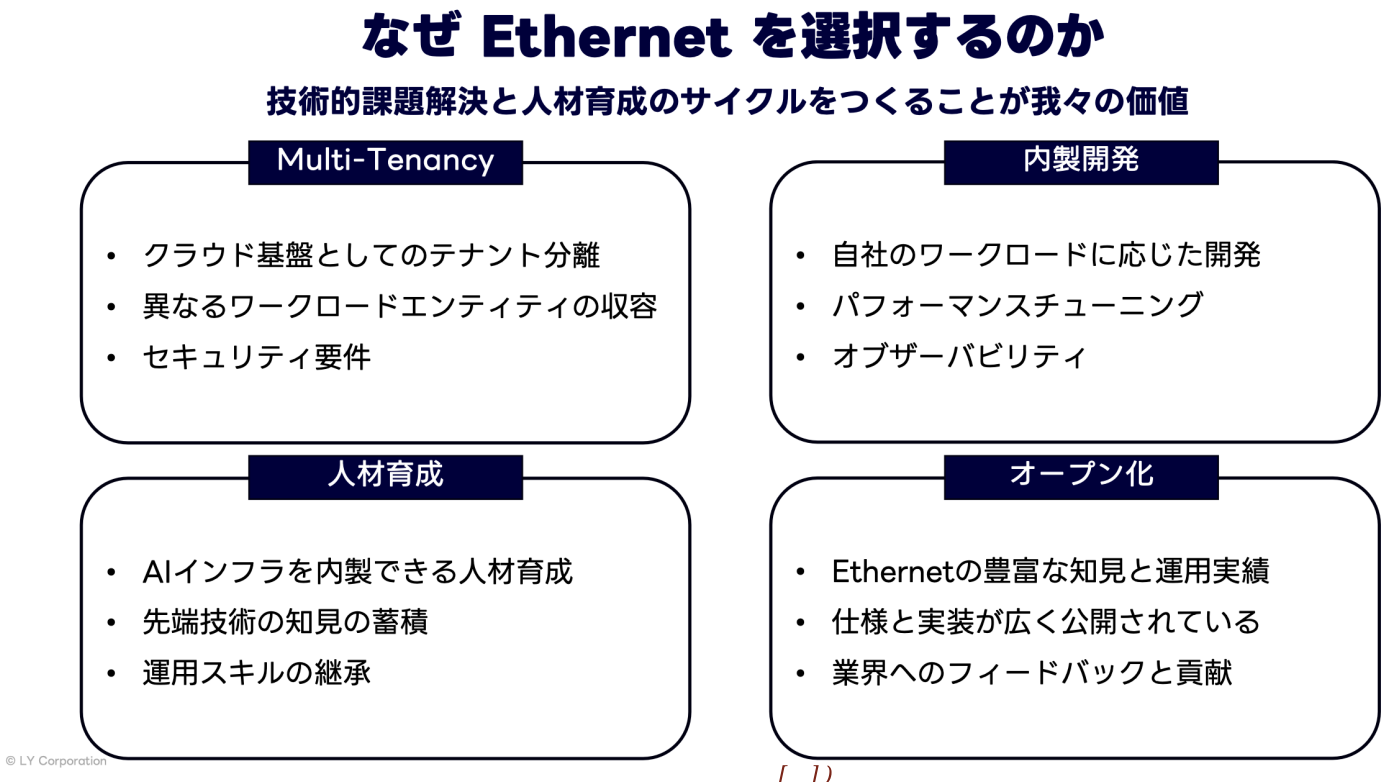

Figure 9. Why choose Ethernet? (Quoted from LINE Yahoo! BLOG[3])

On the other hand, if you are using it for large-scale training, you can learn different perspectives by referring to blogs and papers published by hyperscale companies such as META.

Why not use the NVIDIA Suites?

Given the maturity of the surrounding libraries and software ecosystem, in the AI field the first choice will often be to use NVIDIA equipment for all GPUs, NICs, and switches.

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

Since NVIDIA is also a member company of UEC, it is possible that support for NVIDIA libraries such as NCCL will progress in the future, but it is not expected to be supported in the initial release. Therefore, it is important to clarify "Why not choose the NVIDIA suite?" when selecting equipment or technologies other than NVIDIA, not just UEC, and the following points can be considered, for example.

- price
- deadline
- Use of AI accelerators other than NVIDIA GPUs, such as AI-specific chips
- Avoid vendor lock-in

Comparison of UEC and RoCE

I've listed the main points.

The content written here is not necessarily correct, and there are probably many other perspectives to consider, so we would like to continue to organize and add to the information through discussions at JANOG55 and other forums.

- merit
  - Open technology and implementation (access to library source code?)
  - No need for Lossless Ethernet ⇒ Simple switch fabric that does not require tuning
  - Expanding accelerator options
- Disadvantages
  - The hassle of migrating from the NVIDIA ecosystem and the risk of it not working
  - We can't rely on NVIDIA (maybe a vendor will provide UEC system packs? DELL, HPE?)
  - Risk of verification and compatibility issues associated with system integration

The need for tuning in existing environments

In order to achieve the performance limits of lossless Ethernet, which is the premise of RoCEv2, it is said that tuning according to the workload is necessary, such as buffer size and thresholds. However, the balance and priority between the performance you want to achieve and the cost of tuning (time and equipment) varies depending on each service (system) and organization.

For this reason, we believe that it is possible that, for example, considering the following points, it may be decided that "it is better not to perform tuning." Therefore, whether "no tuning is necessary" is a reason to introduce UEC will be determined by taking into account various benchmarks, indicators other than performance, and convenience, depending on the accelerator used.

- Performance with default settings
- The scale of the service (small improvements can have a big impact on large services)
- Tuning talent costs
- Time required for tuning (opportunity loss during that time)
- Cost balance of networks, servers, etc.
- Maturity of the technology stack of the accelerator (GPU) used

UEC specifications and product roadmap

The timing of the release of UEC specifications and products is an important factor in considering use.

Figure: "UEC Target Timelines, 2024-10-15" is a roadmap announced at the OCP Global Summit held in October 2024. According to this, UEC specifications and products are expected to be released at the following times.

- UEC specification (v1.0): Around March 2025 (first quarter)
- UEC-compliant products: 2025 (probably mid-2025 or later?)

However, as of 2023, it was announced that the specifications would be released in 2024, so the roadmap may change in the future. Let's proceed with our consideration while checking the latest trends through UEC press releases, blogs, and announcements from vendors.

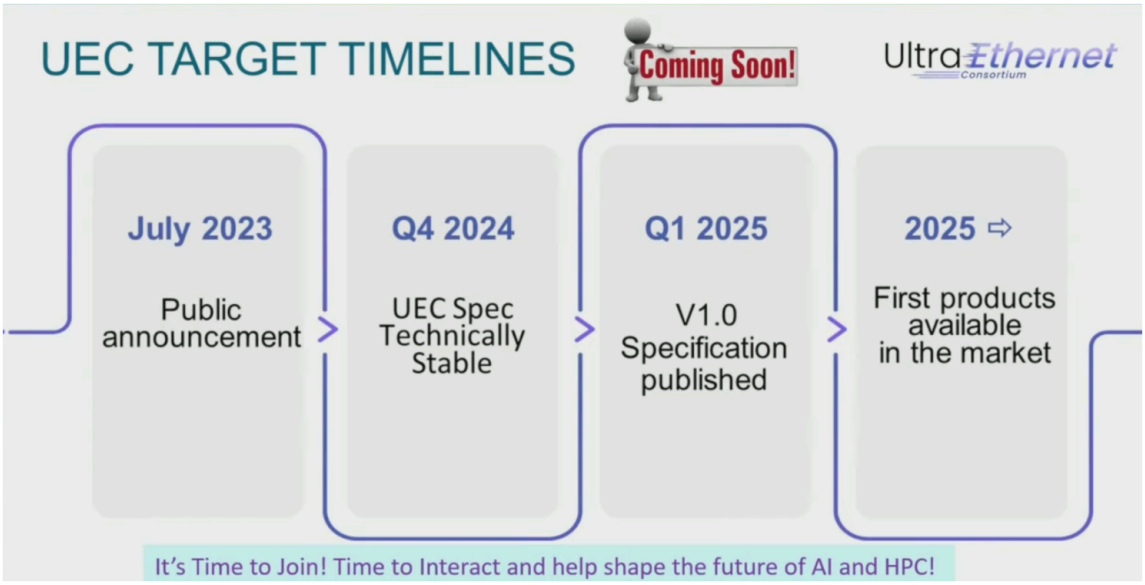

Figure 10. UEC Target Timelines, 2024-10-15, Quoted from "Accelerating AI/HPC: OCP and UEC's Collaborative Vision for High-Performance Networking, Uri Elzur, OCP2024" [4]

UEC compliant products (including pre-shipment products)

In the second half of 2024, various companies announced UEC-compatible products one after another. Here we list UEC-compatible products. For products that have been announced, we provide links to them, and for vendors that have not yet announced but are likely to release UEC-compatible products, we provide only the company or product name as a placeholder.

In addition, data center products, not just UEC products, tend to be allocated to companies that purchase in large quantities, such as hyperscalers, at first, and then become available for purchase by large to medium-sized businesses. Therefore, it is important to note that there is often a lag between the product release and when it actually becomes available.

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) Support
- Jericho 3-AI: ???

Marvell Technology

Marvel NIC

- Any chance of releasing a SmartNIC?

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 Platform

- Arista Etherlink AI Platform
  - 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
  - All Etherlink switches support the newly established Ultra Ethernet Consortium (UEC) standard, which is expected to provide even greater performance benefits when UEC NICs become available in the near future.

Asterfusion

After the UEC specifications were released, it was announced that UEC support would be implemented in the future.

- 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 center 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 finalized, 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
  - Since it says "Line-rate programmability to support evolving UEC (Ultra Ethernet Consortium) standards", it can be interpreted as meaning that it will be compatible with UEC in the future (but does not yet support it).
- 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. Click "Download White Paper" on the UEC homepage.
7. **https://ultraethernet.org/blog/**
8. **https://ofiwg.github.io/libfabric/**
9. Click "Download White Paper" on the UEC homepage.
10. Click "Download White Paper" on the UEC homepage.
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/SAI/blob/master/doc/SAI-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. RL 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**

Last updated 2025-01-28 16:23:26 +0900