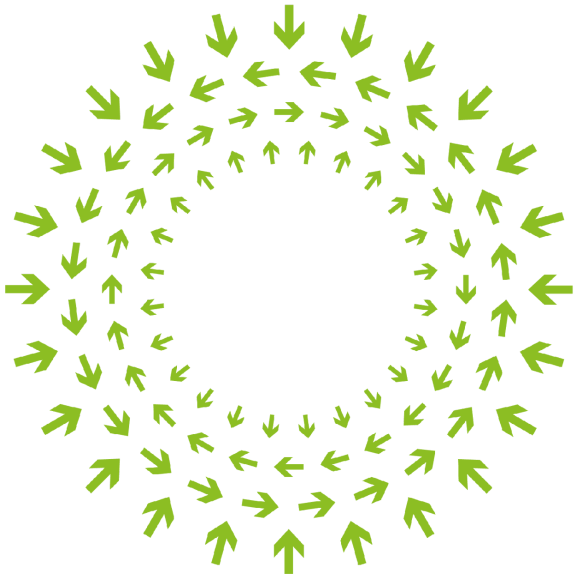


OPEN

Compute Project

Networking Project: ESUN Workstream

OCP ESUN - Network Operator Requirements - Base Specification

1.0

Effective February 9, 2026

Author: Manoj Wadekar, Meta
Author: Rajesh Sankaran, Microsoft

Table of Contents

1. License **3**
1.1. Open Web Foundation (OWF) CLA 3
1.2. Acknowledgements 4
2. Compliance with OCP Tenets **5**
2.1. Openness 5
2.2. Efficiency 6
2.3. Impact 7
2.4. Scale 7
2.5. Sustainability 7
3. Scope **9**
4. Trends Driving Scale-Up Networks **9**
4.1. Low Latency: 9
4.2. High Performance: 9
4.3. Stability 10
5. Current Ethernet vs. ESUN **10**
5.1. Header Efficiency 10
5.1.1. Current Ethernet Header 10
5.1.2. ESUN Header (EH) 10
5.2. New ESUN Header Definition 11
5.3. Lossless, differentiated and resilient Hierarchical Topologies 12
5.3.1. Current Ethernet capabilities: 12
5.3.2. ESUN Enhanced capabilities 13
5.4. Security 13
6. Product Requirements **14**
6.1. End Point changes for ESUN 14
6.2. Switch Side changes for ESUN 14
7. ESUN Requirements **14**
8. Summary of Proposed Enhancements in ESUN **15**
9. References **16**

1. License

1.1. Open Web Foundation (OWF) CLA

Contributions to this Specification are made under the terms and conditions set forth in **Modified Open Web Foundation Agreement 0.9 (OWFa 0.9)**. (As of October 16, 2024) ("Contribution License") by:

Meta and Microsoft

Usage of this Specification is governed by the terms and conditions set forth in **Modified OWFa 0.9 Final Specification Agreement (FSA)** (As of October 16, 2024) **("Specification License").**

You can review the applicable Specification License(s) referenced above by the contributors to this Specification on the OCP website at https://www.opencompute.org/contributions/templates-agreements.

For actual executed copies of either agreement, please contact OCP directly.

Notes:

> The above license does not apply to the Appendix or Appendices. The information in the Appendix or Appendices is for reference only and non-normative in nature.

NOTWITHSTANDING THE FOREGOING LICENSES, THIS SPECIFICATION IS PROVIDED BY OCP "AS IS" AND OCP EXPRESSLY DISCLAIMS ANY WARRANTIES (EXPRESS, IMPLIED, OR OTHERWISE), INCLUDING IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, OR TITLE, RELATED TO THE SPECIFICATION. NOTICE IS HEREBY GIVEN, THAT OTHER RIGHTS NOT GRANTED AS SET FORTH ABOVE, INCLUDING WITHOUT LIMITATION, RIGHTS OF THIRD PARTIES WHO DID NOT EXECUTE THE ABOVE LICENSES, MAY BE IMPLICATED BY THE IMPLEMENTATION OF OR COMPLIANCE WITH THIS SPECIFICATION. OCP IS NOT RESPONSIBLE FOR IDENTIFYING RIGHTS FOR WHICH A LICENSE MAY BE REQUIRED IN ORDER TO IMPLEMENT THIS SPECIFICATION. THE ENTIRE RISK AS TO IMPLEMENTING OR OTHERWISE USING THE SPECIFICATION IS ASSUMED BY YOU. IN NO EVENT WILL OCP BE LIABLE TO YOU FOR ANY MONETARY DAMAGES WITH RESPECT TO ANY CLAIMS RELATED TO, OR ARISING OUT OF YOUR USE OF THIS SPECIFICATION, INCLUDING BUT NOT LIMITED TO ANY LIABILITY FOR LOST PROFITS OR ANY CONSEQUENTIAL, INCIDENTAL, INDIRECT, SPECIAL OR PUNITIVE DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING

NEGLIGENCE), OR OTHERWISE, AND EVEN IF OCP HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

1.2. Acknowledgements

The Contributors of this Specification would like to acknowledge the following companies and individuals for their feedback:

Company	Representative
Altera / Intel	Navneet Rao
AMD	Jaime Edwards
Ampere	Sumiran Shubhi
Arista	Hugh Holbrook
Arm	Luke Culpepper
Astera Labs	Philip T. Mazzara
Broadcom	Ricky Chun, Robin Grindle, Mohan Kalkunte, Niranjan Vaidya
ByteDance	Sean Kao
Cadence Design Systems	Shane Calnan
Celestica	Mark Sanders
Centec	Junjie Wang
Cisco	Jeffrey Lo, Ramesh Sivakolundu
Concordleading Holding	Daniel Lin
Congdon Consulting	Paul Congdon
Dell	Joe White, Claudio DeSanti, Amnon Izhar, Venkatesan Mahalinga, Srini Irigi
Delos Data	Edward Doe
Eridu AI	Dan Biederman, Eric Yeh
Exaware Routing LTD.	David Zelig
Grovf	Levon Mirzoyan
HPE	Mark Pearson
Individual	Durgesh Srivastava, Hesham Elbakoury, Jesin Kim, Julissa Benavente, Li Jieyu, Nickie Lee, Raymond Law, Runlong Hu, Sean Chen, Sree Chadalavada, Sudhakar M
Keysight Technologies	Abhijit Lahiri

Lenovo	Jeff Guo
Lightmatter	Bijan Nowroozi
Liquid	Mitch Wright
Marvell	Kelvin Vivian, Meg Lin
Mediatek	Nagabhushan Chitlur
Meta	Arvind Srinivasan, Ashutosh Grewal, Dharmesh Jani, Jasmeet Bagga, Mikel Jimenez Fernandez, Ola Tørudbakken
Microsoft	Pratik Marolia,Prashant Ranjan, Nilesh Kamdar
Montage Technology	Rober Jin
Nexthop Systems	Juuso Lehtinen
Nokia	Christian Kuhtz
Nvidia	Gilad Shainer
Ohio State University	Pankaj Mehra
Pegatron	Max Cheng
Sambanova Systems	lakshmi narasimhan seshan
Samsung	Vitorio Cargnini, Mandar Sawant
Seagate	Hongjian Fan
Sk Hynix	Jungmin Choi
Sudo Information Technology	Lihua Fang
Synopsys	Priyank Shukla
UpScaleAI	Sandeep Banerjee
Viavi Solutions	Asim Rasheed
Xcena	Brian Hirano
XDPnet	Tom Herbert

2. Compliance with OCP Tenets

2.1. Openness

The specification for OCP ESUN (Ethernet for Scale Up Networks) defines the requirements for network controllers and switches that support scale-up GPU connectivity, clearly outlining the features necessary for at-scale deployment in AI infrastructure. By publishing these

details, we expect silicon and system vendors to develop ESUN-compliant devices that can be easily integrated by multiple customers, eliminating the need for custom hardware or firmware development.

This specification represents an initial version that establishes a functional baseline architecture for ESUN. As operational experience grows and other OCP members contribute feedback, the specification will evolve, with a commitment to maintaining backward compatibility with version 1.0 to protect prior investments.

Openness is measured by the ability of third parties to build, modify, or personalize the device, platform, or network from this contribution. The ESUN specification provides a baseline architecture that enables a diversity of implementations, supporting differentiated traffic, hierarchical topologies, and robust congestion management. Openness is achieved by removing barriers that prevent an open networking platform. In this context, the ESUN specification aims to remove such barriers by providing a consistent definition for deployment of lossless, differentiated, and resilient scale-up networks.

2.2. Efficiency

The OCP specification for ESUN (Ethernet for Scale Up Networks) advances efficiency by standardizing scale-up GPU connectivity and hierarchical network topologies. By defining ESUN header requirements, congestion management, and differentiated traffic handling, it eliminates costly custom solutions and streamlines deployment. Optimized packet headers reduce protocol overhead, improving goodput and lowering latency, which maximizes compute utilization and reduces energy consumption. Enhanced congestion management and load balancing further boost operational efficiency, minimizing idle time for expensive GPU resources.

The specification also preserves existing hardware and software investments by maintaining backward compatibility and providing a consistent framework for resilient, differentiated networking. This approach reduces adaptation costs and avoids unnecessary duplication of effort, directly supporting the Open Compute Project’s tenet of efficiency by enabling sustainable, cost-effective growth in AI infrastructure.

2.3. Impact

The ESUN specification aligns with the Impact tenet of the Open Compute Project by introducing a new networking technology poised to transform the industry. For modern AI platforms, there is a pressing need for more efficient, resilient, and scalable interconnects to support large GPU domains. ESUN addresses these barriers by standardizing enhancements to Ethernet, including lossless hierarchical topologies, differentiated traffic handling, and robust congestion management.

Our approach leverages widely adopted Ethernet technologies, making scale-up networking practical and accessible without requiring proprietary solutions. This standardization achieves the transformative impact sought by the Open Compute Project, empowering diverse vendors and customers to deploy high-performance AI infrastructure at scale.

2.4. Scale

This specification adheres to the Scale tenet published by the Open Compute Project, which requires that contributions be capable of serving Fortune 100 and large hyperscale customers. With the increasing demands for high-performance, low-latency connectivity in AI and compute platforms, ESUN meets this scale tenet by defining a standardized approach to scale-up networking using widely available Ethernet technologies.

This approach unlocks compute scaling for hyperscale environments, ensuring that network infrastructure can grow seamlessly alongside compute resources to meet the needs of the largest customers in the industry.

2.5. Sustainability

This specification for ESUN (Ethernet for Scale Up Networks) complies with the Open Compute Project’s Sustainability tenet in several ways.

First, by providing a standardized and efficient method for scale-up GPU connectivity, ESUN enables data centers to optimize network resources and reduce the need for excess

hardware. Enhanced congestion management, differentiated traffic handling, and lossless hierarchical topologies minimize compute idle time and maximize utilization, thereby limiting the environmental impact associated with overprovisioning and underutilized infrastructure.

Secondly, ESUN's robust and flexible specification leverages widely adopted Ethernet technologies, allowing for incremental scaling and interoperability across diverse compute platforms. This flexibility ensures that operators can better utilize existing hardware, extend the lifecycle of network equipment, and avoid unnecessary replacement or customization, supporting responsible resource management.

Overall, the ESUN specification offers a clear path toward sustainable and efficient data center operations. By reducing operational overhead, enabling scalable growth, and promoting the use of open, interoperable technologies, ESUN drives innovation while supporting environmental stewardship in the broader technology ecosystem.

3. Scope

This document provides Base Specification for ESUN.

OCP ESUN (Ethernet for Scale Up Networks) is about evolving Ethernet to meet the unique demands of Scale-Up AI infrastructure. ESUN (used interchangeably as “OCP ESUN”) builds on Ethernet by introducing evolutionary enhancements that improve goodput, latency, and reliability to meet the demands of Scale-Up GPU connectivity.

4. Trends Driving Scale-Up Networks

Scale-up networks for GPUs are high-performance interconnects designed to facilitate efficient, low-latency communication among multiple GPUs within a tightly coupled cluster. The primary traffic on Scale-up networks is Model Parallel Traffic which demands High Bandwidth and Low Latency.

4.1. Low Latency:

The scale-up network enables the model developer to treat all the GPUs in the domain as a single large GPU, resulting in improved developer efficiency. Consequently, this makes it very important to have a stable network to ensure GPUs in the domain are not stalled. Increasing the size of the scale-up networks gives a larger memory domain for the model developers to work with, which in turn lowers communication latency and provides end-to-end training and distributed inference speedups. Essentially, the application now is trending towards smaller message sizes and requires lower latency across a larger domain. This is critical for the success of future Models such as LLAMA, GPT-OSS, DeepSeek, etc.

4.2. High Performance:

Our goal is to reduce compute Idle time for GPUs (an expensive resource). A highly performant scale-up network will allow us to achieve this goal. Optimal compute/network interleaving can achieve near-peak performance. This drives the exploration of fused kernels, integrating data movement control and compute-network interleaving for lower latencies, increasing the need for more performant scale-up networks. A lower network overhead can lead to increased goodput.

4.3. Stability

Scaling up these networks is challenging. Until Llama4, scale-up was 8 GPU's contained within a node. Next-gen platforms feature dedicated scale-up networks, expected to grow to >=1K GPUs, expanding beyond a single "box" across racks.

Stability and scalability issues with increased GPU cluster size can impact training uptime. A 20% bandwidth drop in a small 128-accelerator domain means 5% e2e performance loss and higher operational cost/year.

5. Current Ethernet vs. ESUN

5.1. Header Efficiency

5.1.1. Current Ethernet Header

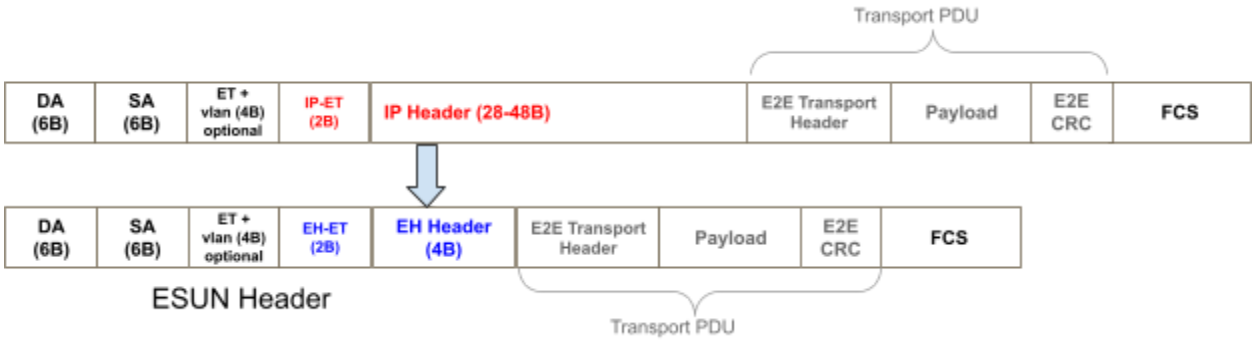
The following header demonstrates use of standard IPv4 or IPv6 based protocol on standard Ethernet.



Standard Ethernet IP Header

5.1.2. ESUN Header (EH)

ESUN optimizes the packet headers for smaller topologies by removing IP headers and using L2 addresses for forwarding. Functionality that is lost due to this removal of IP headers (E.g. ECN, DSCP and higher entropy due to IP headers and sockets) is reintroduced through a new smaller header called "ESUN Header" or EH.



This new proposed header allows use of L2 encapsulation to carry all the SU traffic with additional 4B and removal of IP headers (saving 20-40B). EtherType EH-ET (value pending assignment) will be used to identify ESUN Headers. **ESUN provides higher Goodput efficiency for all the message sizes with this change**.

5.2. New ESUN Header Definition

ESUN defines an ESUN Header to reintroduce many of these lost capabilities (when the IP header is removed).



ESUN Header Definition

- **Fabric Feedback - EH-ECN (2b)**: This is a 2-bit field that is defined in the exact same way it is currently defined as ECN in the IP TOS field (It encodes ECN capability of the protocol and actual state feedback from the fabric when it experiences congestion).
- **Differentiation - EH-CoS (3b)**: This is a reduced version of the DSCP field in the IP-TOS header. This mimics 802.1p header that has a 3-bit CoS definition that identifies Traffic Class for the traffic. The reason for introducing this in the ESUN Header is that in the Scale-Up domain that doesn't use VLANs, one can achieve higher goodput efficiency by trading off 4B 802.1 Header with these 3 bits in ESUN Header. NOTE: UEC™ Specification for CBFC also defines a Traffic Class field. In the case of ESUN header presence, EH-COS shall take precedence.
- **F (Flow Label Valid)**: This bit specifies the presence of valid Flow Label present in the ESUN Header for device to use for load balancing. In case of this bit being TRUE, packets shall be forwarded in deterministic fashion using the Flow Label.

- **Load Balancing - Flow Label (16b)**: This source specific field is recommended to be used by the fabrics to achieve load balancing in the network. Flow label provides additional entropy in hashing (e.g. {DA, SA, Flow label})
- **TTL- Time To Live (4b)**: This is another field moving from IP header to EH providing control over any unintended loops in the topology. This field will be decremented by each switch during forwarding, and the packet will be discarded when TTL reaches 0.
- **UD - User Defined (2b)**: This is end-to-end information to be populated and consumed by the transport implementation at endpoints.
- **Rev – Revision Number (2b)**: This provides future revisions to be made to the headers while allowing backward compatibility. OCP ESUN group will continue to monitor development in standards bodies (such as UEC™, IEEE™). As and when these standards are firm, the ESUN group will review to evolve the future specifications. (Value used for this specification will be '00')

  NOTE: Modification of EH during forwarding (e.g. ECN and/or TTL) requires recomputation of FCS.

5.3. Lossless, differentiated and resilient Hierarchical Topologies

Radix limitation of switches as well as number of lanes per GPU will lead to deployment of hierarchical topologies for SU networks. Hierarchical topologies demand support for transient as well as persistent congestion control to avoid packet loss and high latencies. Resilience is required in these networks to deal with link level and path level failures. And congestion avoidance is achieved through load balancing techniques (e.g. ECMP, Dynamic Load Balancing).

5.3.1. Current Ethernet capabilities:

	Link level CM (Transient Congestion	Fabric Feedback (Persistent Congestion)	Traffic Differentiation	Resilience and Load Balancing	Multi-Hop Topology
Standard Ethernet	PFC	IP based ECN	DSCP, 802.1p	No LLR, ECMP, IP Headers	TTL, IP based ECN

Existing congestion control mechanisms (like ECN and PFC) have helped Standard Ethernet achieve the desired goal of congestion management. However, in the previous section we mentioned that Goodput improvement required removal of IP header which took away ECN, DSCP information from the packet (that was carried in the TOS field) and also removed the fields that were

used for achieving higher entropy for path selection to gain desired load balancing.

Ethernet currently doesn't support any link level error recovery beyond Forward Error Correction but allows upper layer protocols to provide that functionality.

5.3.2. ESUN Enhanced capabilities

ESUN proposes use of a feature, Link Level Retry, defined by Ultra Ethernet Consortium (UEC™) to achieve link resiliency against link level errors. (This feature can be helpful to make high speed links resilient against packet drops due to high BER). The goal for LLR is to improve the quality of the channel by reducing exposure to BER. Packet drops outside of the BER-induced errors will still need to be handled by end-to-end recovery mechanisms.

ESUN also recommends an additional link level congestion control mechanism for Credit Based Flow Control (CBFC), defined by UEC™, that can provide more memory efficient designs for the ESUN switch devices.

With these enhancements ESUN provides best of the breed Fabric capabilities:

	Link level CM (Transient Congestion	**Fabric Feedback (Persistent Congestion)**	**Traffic Differentiation**	**Resilience and Load Balancing**	**Multi-Hop Topology**
Standard Ethernet	PFC	IP based ECN	DSCP, 802.1p	No LLR, ECMP, IP Headers	TTL, IP based ECN
ESUN	PFC, CBFC	EH-ECN	EH-COS	LLR, EH-Flow Label	EH-TTL, EH-ECN

5.4. Security

Security is an important consideration. As ESUN header replaces IP header, it is no longer compatible with IPSec.

MACSec can be used alongside ESUN to add confidentiality, integrity and authentication at Ethernet Layer 2 level. This provides point-to-point link level encryption.

End-to-end encryption operates at higher levels in the stack and is outside the scope of ESUN.

6. Product Requirements

6.1. End Point changes for ESUN

1. Support for PFC. Optional support for LLR/CBFC
    a. Enable configurability of LLR and CBFC per port if supported.
2. ESUN Header support
    a. Specifying CoS, Flow label, TTL
    b. Responding to ECN and announcing capability

6.2. Switch Side changes for ESUN

1. Shall support PFC, CBFC and LLR
    a. Shall enable configurability of LLR and CBFC per port.
2. ESUN Header support
    a. Ethernet DA based (48b) forwarding
        i. This is assuming that there is no IP header, and switches shall forward packets based on Ethernet Destination Addresses in the ESUN header that are configured statically in the switches (there is no learning or aging of MAC addresses).
        ii. Switches should not broadcast ESUN packets with unknown addresses.
    b. CoS mapping to TC:
        i. In the presence of CoS in 802.1Q tag (if enabled by some implementation) or CoS in CBFC and EH: EH-COS should take precedence.
    c. Flow label should be used for hashing in load balancing decisions if F bit is set.
    d. ESUN switches should be standard ECN capable and should indicate congestion via ECN bits when appropriate.
    e. TTL field shall be decremented during forwarding and shall be used for removing loops.

7. ESUN Requirements

Feature	Requirement
PFC	Required
CBFC	Required (Switch), Optional (Host)

LLR	Required (Switch), Optional (Host)
EH-ECN	Required
EH-CoS	Required
F, EH-Flow Label	Required
EH-TTL	Required

8. Summary of Proposed Enhancements in ESUN

- Goodput improvement through:
  - Replacing 20-40B IP Header with smaller (4B) ESUN Header
  - New ESUN Header definition to carry desired ECN, CoS, TTL and Flow label information
- Feature support for lossless, reliable delivery(defined in UEC):
  - Improved link level congestion management with CBFC
  - Link resilience with LLR

These changes enable ESUN to be the best in class Scale Up network option:

Lossless, differentiated and resilient Hierarchical Topologies	Standard Ethernet	ESUN
Link Level CM (Transient Congestion)	PFC	PFC, CBFC
Fabric Feedback (Persistent Congestion)	IP based ECN	EH-ECN
Traffic Differentiation	DSCP (802.1p*)	EH-CoS (802.1p*)
Resilience (Link)	No LLR	LLR
Load Balancing	ECMP	EH-Flow Label
Multiple Switch Hops	Supported, IP-TTL	Supported, EH-TTL

9. References

[1] "Introducing ESUN: Advancing Ethernet for Scale-Up AI Infrastructure at OCP", 2025, OCP Blog
[2] IEEE 802.1Q-2018, IEEE™ Standards Association
[3] IEEE 802.3-2022, 21 Sept 2022, IEEE™ Standards Association
[3] Ultra Ethernet™ Specification v1.0.2