

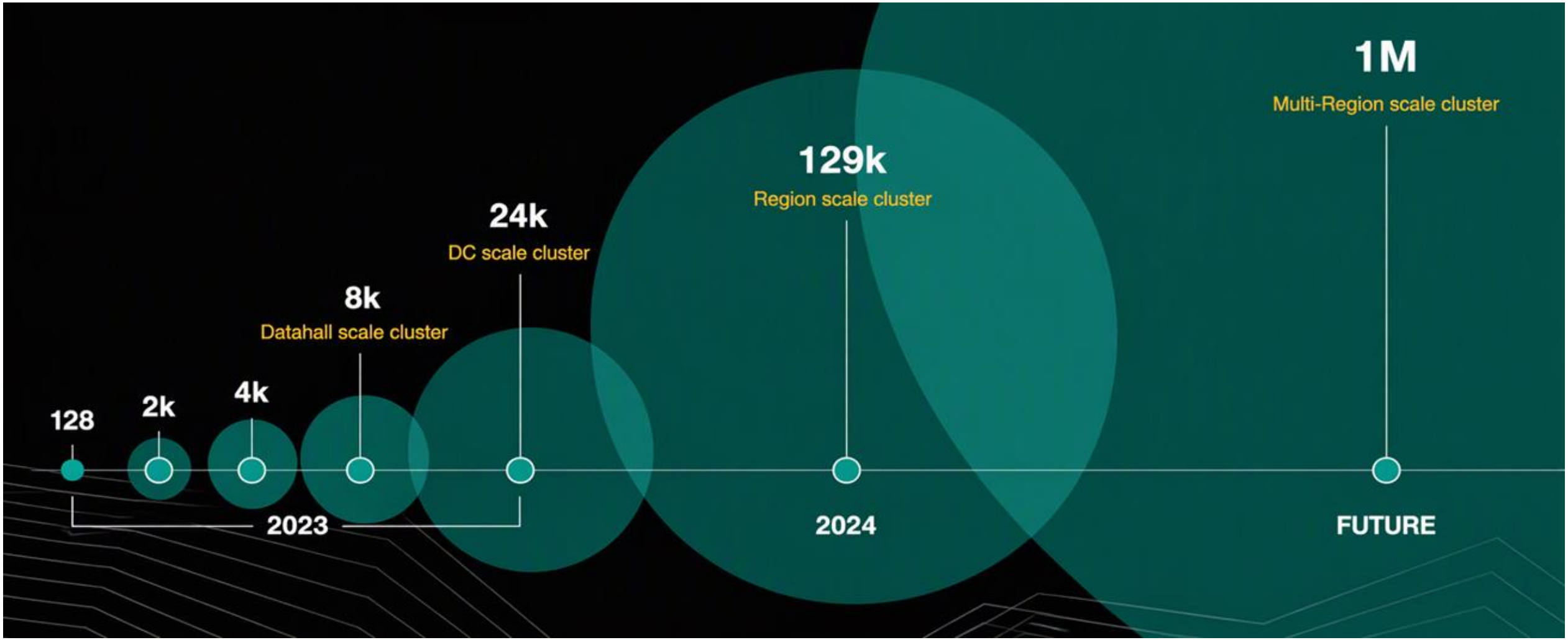
ESUN: Open Ethernet for the AI Scale-Up Era

Manoj Wadekar *AI Network Technologist, Meta*

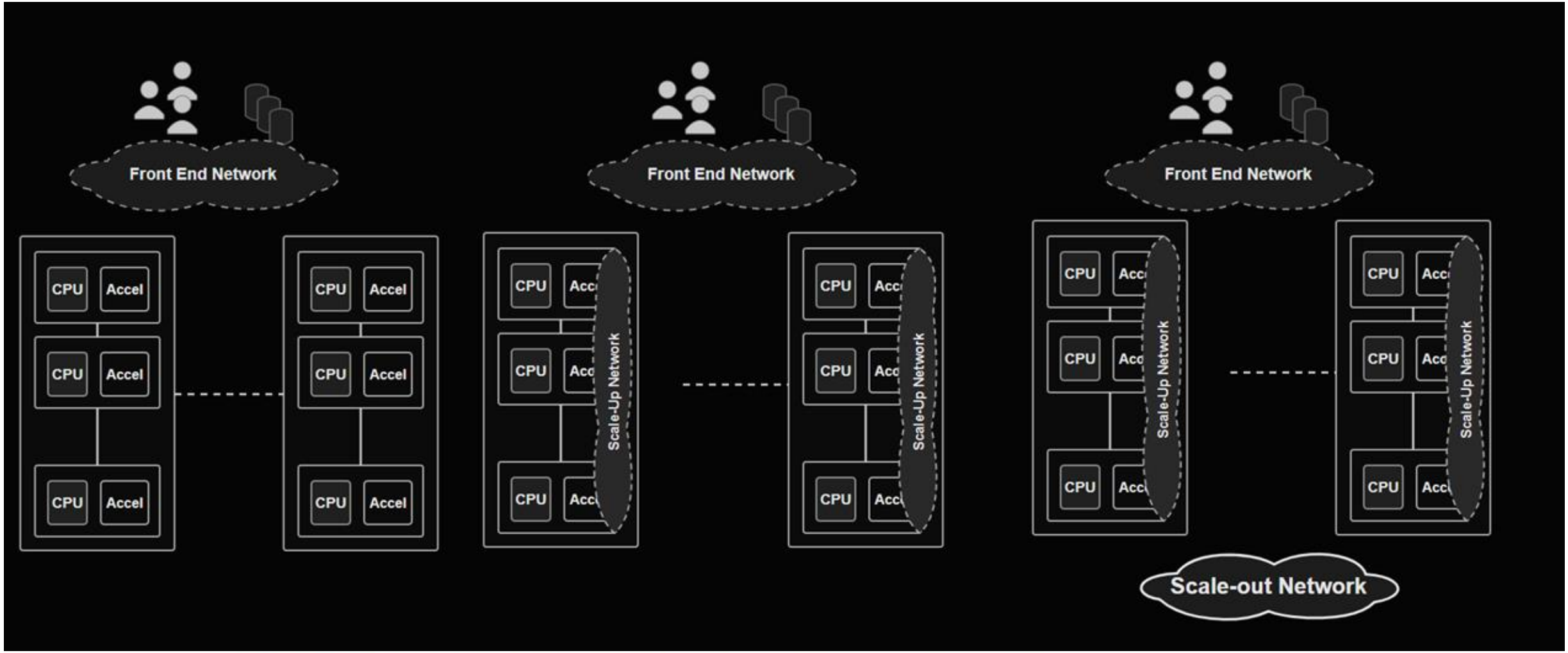
Pratik Marolia *AI System Architect, Microsoft*



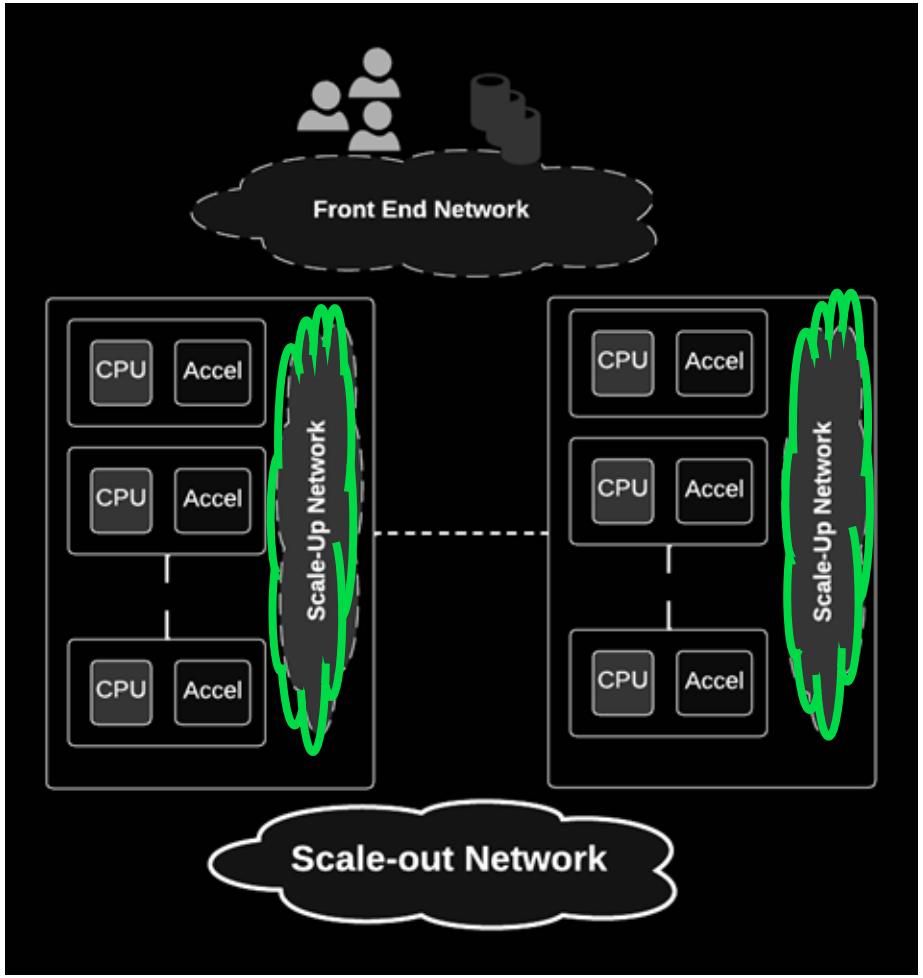
AI Clusters Growing

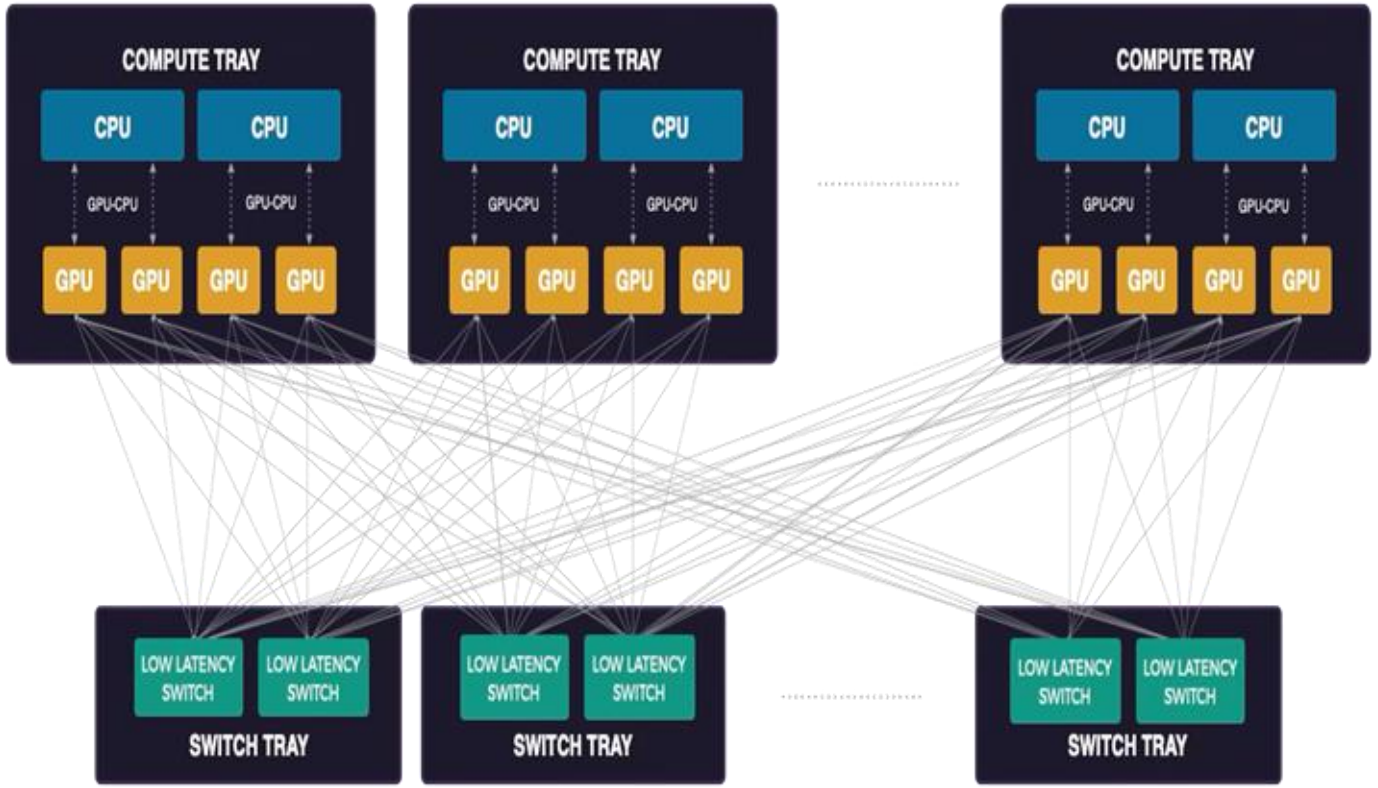


AI Networks: Trillions of wires, Multiple networks



Scale-Up Networks : Focus of ESUN





- Number of GPUs: 8 → 72 → 576 → 1K
- Bandwidth per GPU (GBps): 450 → 1200 → 1800 → ..
- **ESUN was formed to create an open solution**

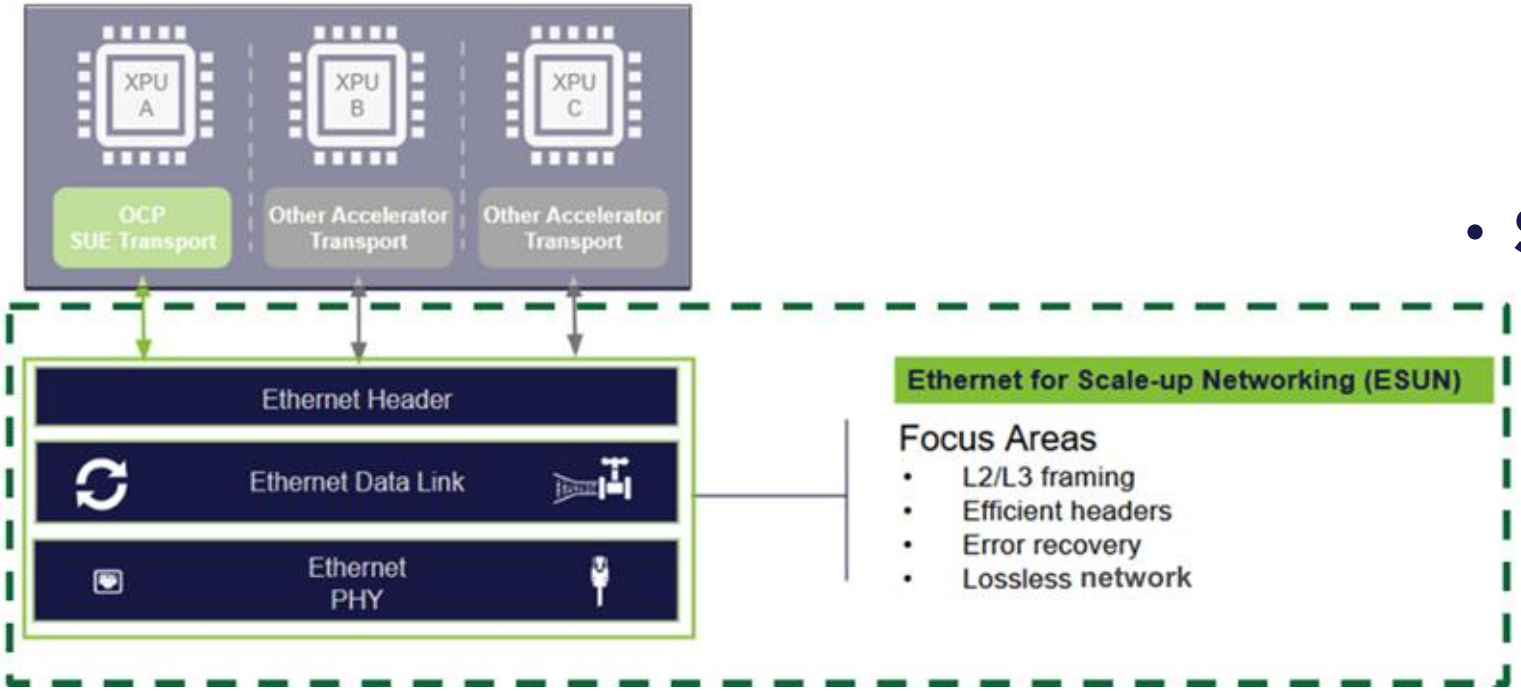
ESUN Progress

- **Mission:** Transitioning scale-up AI interconnects from proprietary protocols to open Ethernet standards

  - Champions Ethernet as an **Open interconnect for Scale-Up** by specifically identifying critical features.
  - **Leverage strong Ethernet eco-system** of NOS, operational and performance tools
  - Establishes clear **Operator/End User requirements**



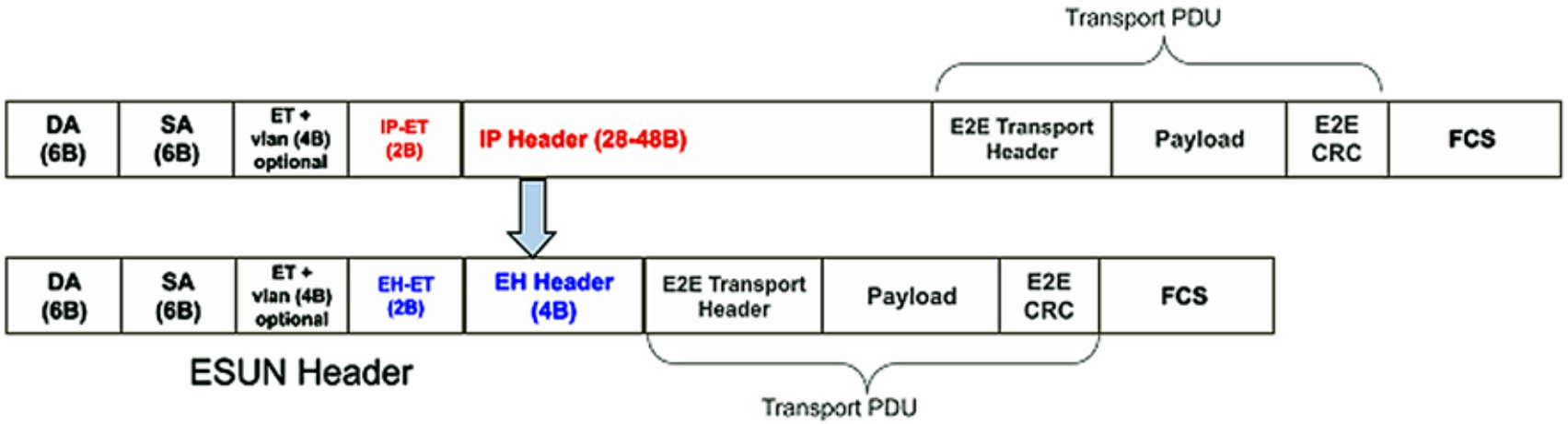
ESUN Technical Focus



- **Focus On Fabric**
  - For wider adoption while allowing implementation flexibility for GPUs
- **Standards Coordination:**
  - Leverages Standards work in IEEE, UEC

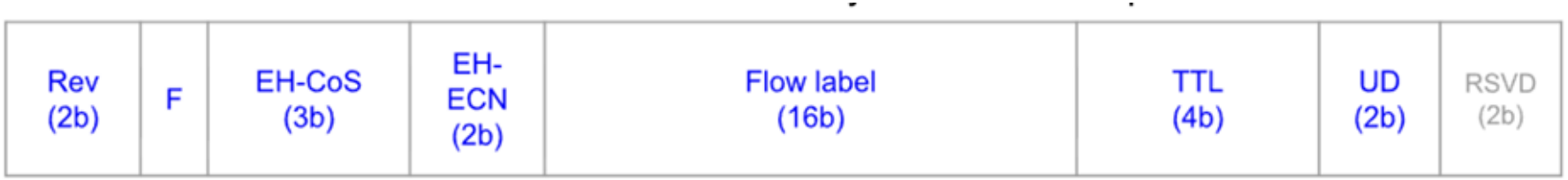
ESUN 1.0 Spec Released March 26!

ESUN 1.0: Compact Framing



- Traditional IPv4 was designed for dynamic, routable, and large-scale networks
- Scale-up network is fixed topology. ESUN header replaces IPv4/UDP to improve header efficiency - gives 24 ~ 44 Bytes savings
- Pros: higher bandwidth efficiency for smaller messages

ESUN 1.0: Header



ESUN header

- **ESUN** header drops IP header, but carries forward useful features to support to build a multi-hop fabric
  - EH-ECN: provides congestion feedback
  - EH-COS: identifies differentiated traffic classes
  - Flow label: fabrics can use this to achieve load balancing
  - TTL: detects unintended loops
  - UD: endpoint defined bits, opaque to the fabric

ESUN 1.0: Reliable Network Fabric

- Link level retry (LLR) provides localized, faster recovery & lowers endpoint buffering
- CBFC enables tighter buffer sizing, not packet drops due to buffer overflow
- Pros:
  - Enables simpler transports, simpler endpoints
  - Shallow buffer switches

Note: LLR and CBFC are defined by UEC

ESUN 1.1 - New Work Items in progress

ESUN 1.1 Specification

"Switch Profiles"

Meg Lin (Marvell) / Ramesh Sivakolundu (Cisco)

Items Being Worked for ESUN 1.1

- **Optimized Headers (UFH):**
  UFH 0.9 to be contributed to UEC 1.1 — **WIP**
- **IPG Optimization:**
  IPG Reduction 0.9 to be contributed to UEC 1.1 — **WIP**

ESUN 1.0 Errata

Future Release Candidates

Low Latency FEC Modes

Depends on UEC WG timeline

Reduced Lane Mode

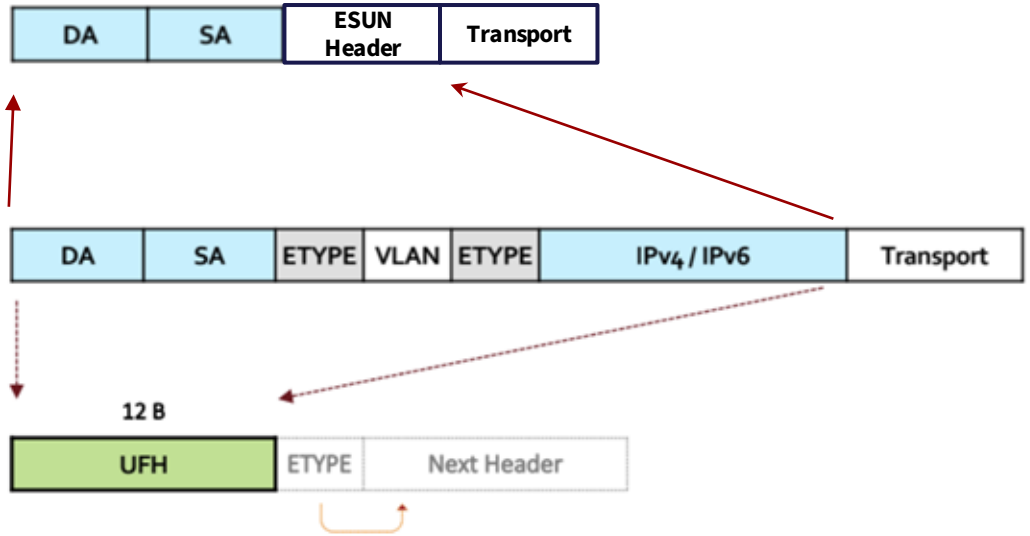
Depends on UEC WG timeline

In-Network Collectives (INC)

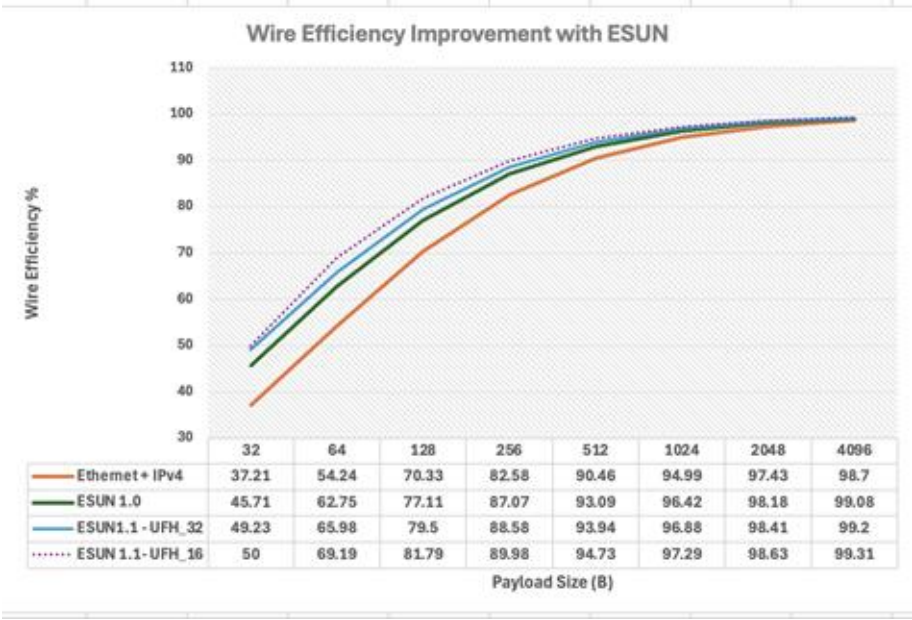
TBD. Started educational presentations from UEC

ESUN 1.1: UFH 0.9

ESUN Header



UFH Header



	32	64	128	256	512	1024	2048	4096
Ethernet + IPv4	37.21	54.24	70.33	82.58	90.46	94.99	97.43	98.7
ESUN 1.0	45.71	62.75	77.11	87.07	93.09	96.42	98.18	99.08
ESUN1.1 - UFH_32	49.23	65.98	79.5	88.58	93.94	96.88	98.41	99.2
ESUN 1.1- UFH_16	50	69.19	81.79	89.98	94.73	97.29	98.63	99.31

'Unified Forwarding Header' is superimposed on the Ethernet MAC header

- Remove redundant L2/L3 addressing and IP header fields not required in a homogenous environment.
- Header contains source and destination address, DSCP, ECN, Hop Limit.
- Header is self-identified by encodings in the DA. The 'Ethertype' field continues to identify the next header.

ESUN 1.1: IPG Idle Reduction

- IEEE Standard requires average of 12 bytes inter-packet-gap
  - Also, must pad out after terminate to 8B boundary with idles
    - Minimum IPG will be 1-8 bytes depending on frame size
  - Deficit idle count may force additional 8B blocks of idle
- Proposed option to only require one IPG byte (terminate symbol)
  - Retain frame always starting on 8B boundary
  - Resulting in 1-8 bytes IPG to reach frame boundary
- Defined in upcoming UltraEthernet specification

Roadmap & Working Group Priorities

ESUN Announcement
October 2025

ESUN Group Formation
December 2025

ESUN 1.0 Spec Published
March 2026

ESUN 1.1 Spec Target
Oct - Nov 2026*
(WIP, UEC Updated Spec dependent)

ESUN 2.0 Spec Target
TBD

Call to Action

- Problem to solve
  - Jointly advance Ethernet fabrics optimized for scale-up networking
- How to get involved in the project/sub-project community
  - Project Wiki: https://www.opencompute.org/w/index.php?title=Networking/ESUN
  - Sign up for mailing list: OCP ESUN Mailing List
  - Join bi-weekly call: OCP Networking Project Calendar

Thank You!

