Scaling Out and Up with the Ultra Ethernet Transport

Evolution of UET from Scale-Out Networking to Converged Scale-Up and Scale-Out Fabrics

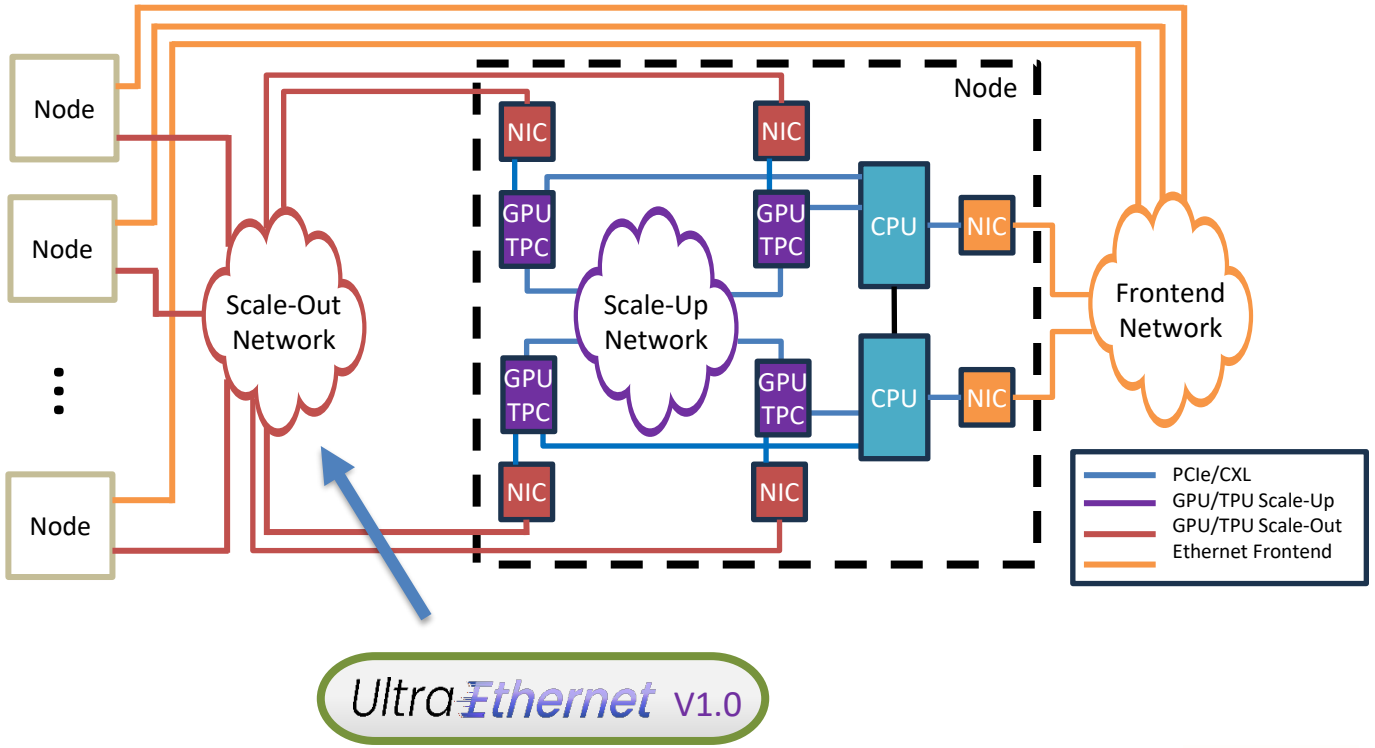
Paul Congdon

IEEE 802 ITU

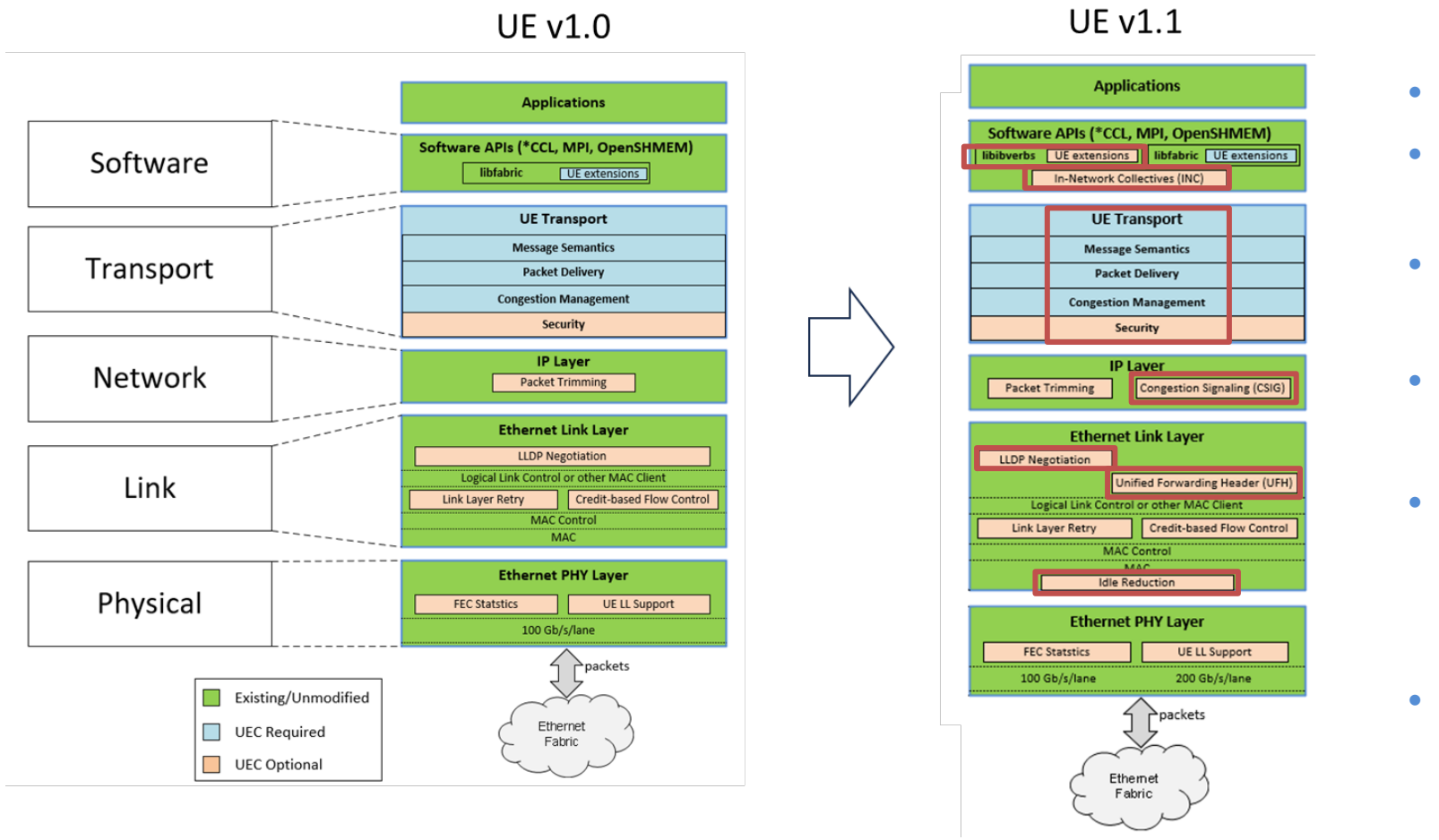
Scale-Up and Scale-Out in Ultra Ethernet

Three complementary network domains

1. Frontend
   a. Connectivity to external services and networks
   b. Traditional TCP/UDP/IP
2. Scale-Out (Backend)
   a. Connects nodes across the cluster
   b. Optimized for large flows, message exchange and collective communications
   c. RDMA based transports
3. Scale-Up (Local)
   a. Connects accelerators within a node or at the rack-level
   b. Optimized for tightly coupled communication and shared memory semantics
   c. Small messages, high packet rates and fine-grained synchronization



UE Evolution

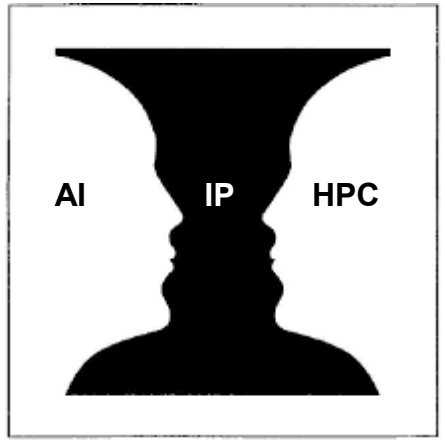


- Verbs mapping
- In-Network Collectives (INC)
- Congestion Signaling (CSIG)
- Inter-Packet Gap Idle Reduction
- Compressed Unified Forwarding Header (UFH) and Scale-up Transport Enhancements
- LLDP Negotiation Enhancements

The Ultra Ethernet Transport (UET)

An RMA protocol for the future

- Multipathing RMA
- Relaxed Delivery Ordering
- Rapid loss recovery
- Modern congestion control for the DC
  - Rapid startup and slowdown
  - multi-path aware



- Runs on IPv4/v6 and Ethernet
- Lossy and Lossless operation
- Ordered and Unordered Delivery
- Design for high scale at low cost
- Day-1 Security

Preserve the applications above, use Ethernet and IP below a new transport in the middle

IEEE 802 ITU

UET Scale-Out Architecture and Concepts

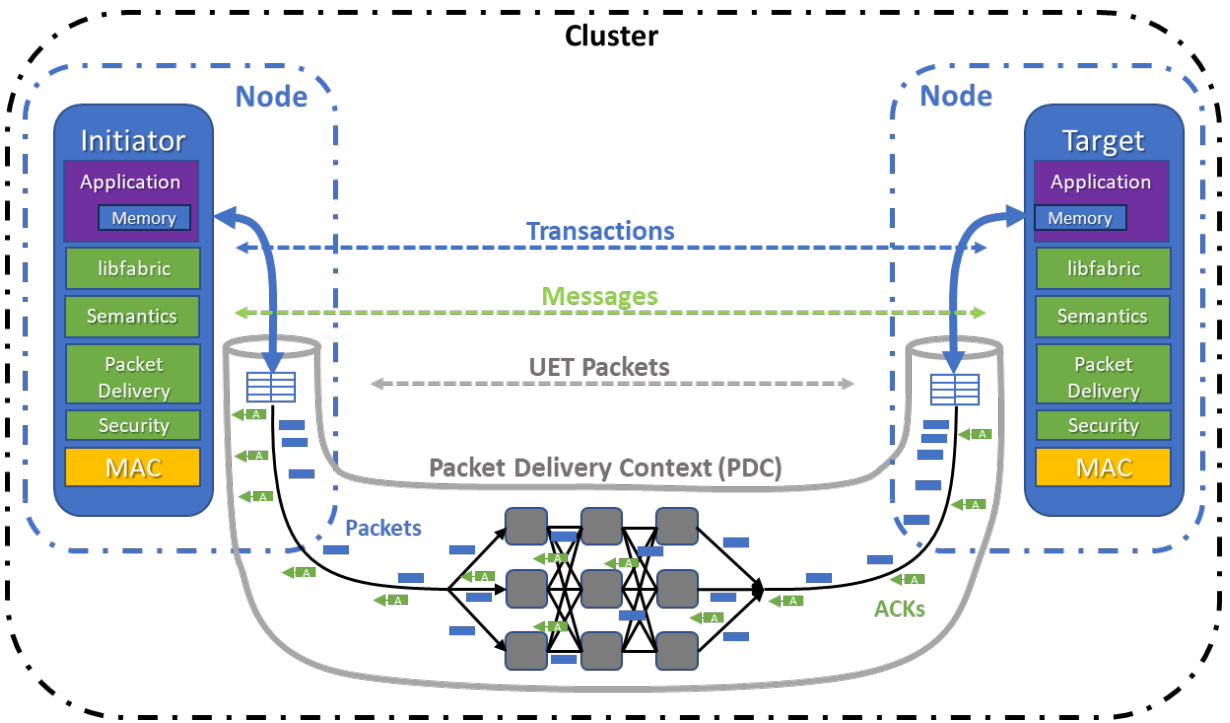
Libfabrics API
- existing, open standard

Semantics
- RDMA message service focused on MPI and *CCL

Packet Delivery
- Multiple delivery/reliability modes
  - ROD, RUD, RUDI, UUD
- On-demand dynamic ephemeral connections

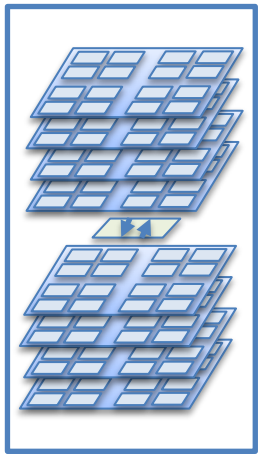
Congestion Management
- Sender-based and receiver-based congestion control
- Packet spraying and trimming support

Security
- Optional end-to-end AES encryption
- Scalable key management

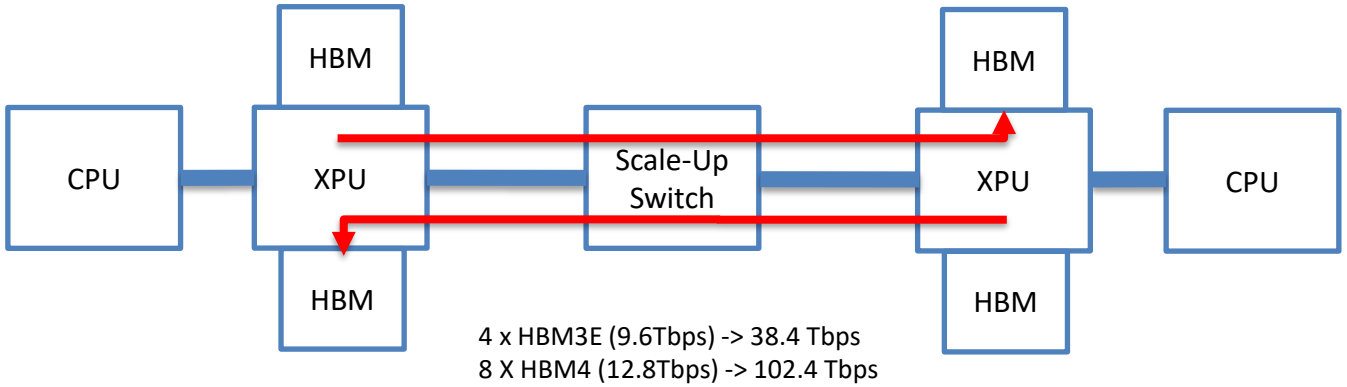


Why Scale-Up Networking?

- The belief is that traditional Ethernet needs optimizations to:
  - "Turn a rack of XPUs into a single coherent compute fabric."
  - "Deliver memory semantics over Ethernet—load/store and atomics at rack-scale."
  - "Deliver 1 µs round-trip latency target for intra-rack XPU pairs."
- Scale-Up = shared memory domain across XPUs with atomic operations, sub-250ns switch latency and deterministic behavior.
- Driven by High Bandwidth Memory (HBM) requirements.



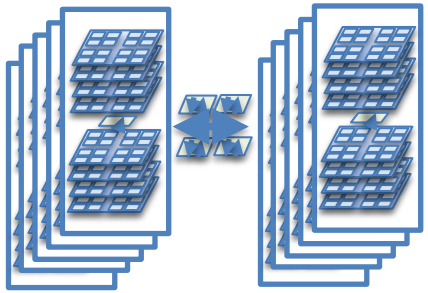
Scale-Up Fabric



- Demands: memory semantics, ordering, fairness, lossless, congestion awareness.

Different Transport Optimizations

Scale-Out (UET v1.0)



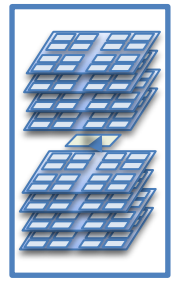
Target Environment

- Large AI/HPC fabrics and clusters
- Multi-hop routed Ethernet fabrics
- Large collective and message transfers
- High network utilization and bounded flow completion time

Transport Optimization

- Packet spraying
- Multipathing
- Congestion management
- Scalable reliability

+Scale-Up (UET v1.1)



Target Environment

- Tightly coupled accelerator fabrics
- Single-tier, low-diameter networks
- Many small memory transactions
- Deterministic low latency and high packet rate

Transport Optimization

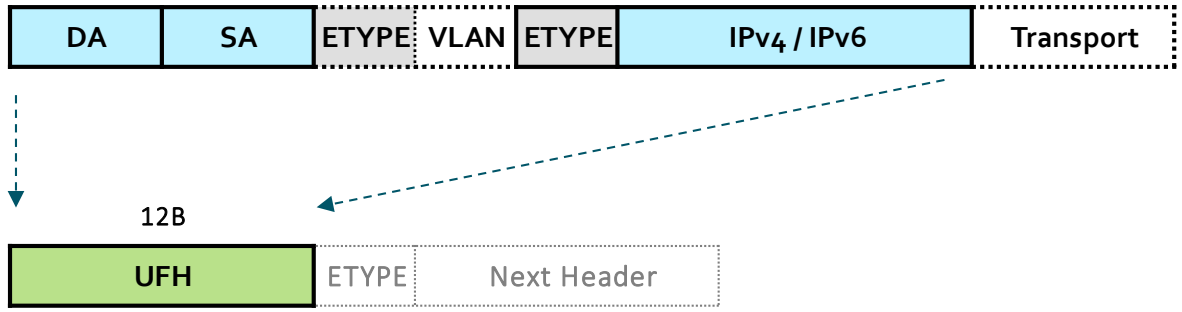
- Wire efficiency
- Small transaction optimization
- Minimal latency
- Minimal transport overhead

UET Enhancements for Scale-up

- Unified forwarding header (UFH)
- New reliable ordered delivery mode for local networks (RODL)
- New direct memory addressing mode
- New optimized semantic headers
- Multiple semantic messages per-packet

Unified Forwarding Header (UFH)

An Ethernet-compatible compressed header that combines L2 and L3 header fields into an Administratively Assigned local MAC Address



Benefits:

- Reduced overhead
- Higher packet rate
- Ethernet coexistence

Key Points:

UFH improves efficiency without abandoning Ethernet

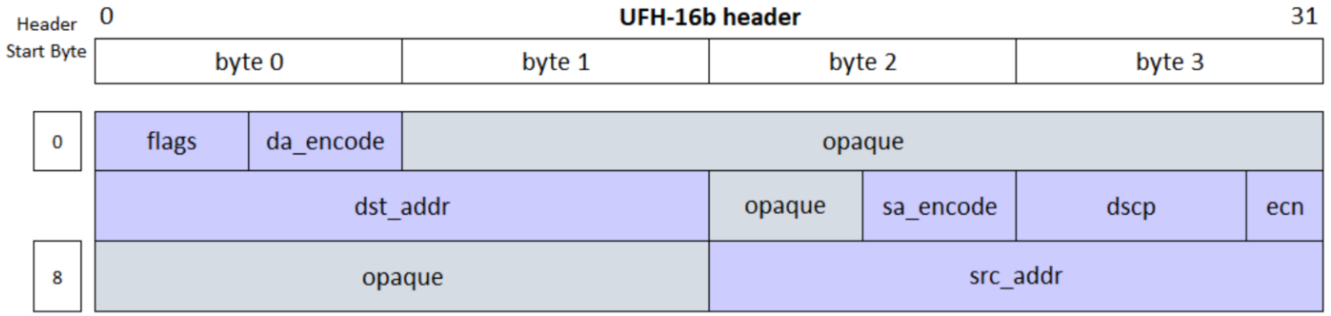
UFH is designed to be used with other transports, not just UET

UET for Scale-up uses UFH with a new PDC mode - RODL

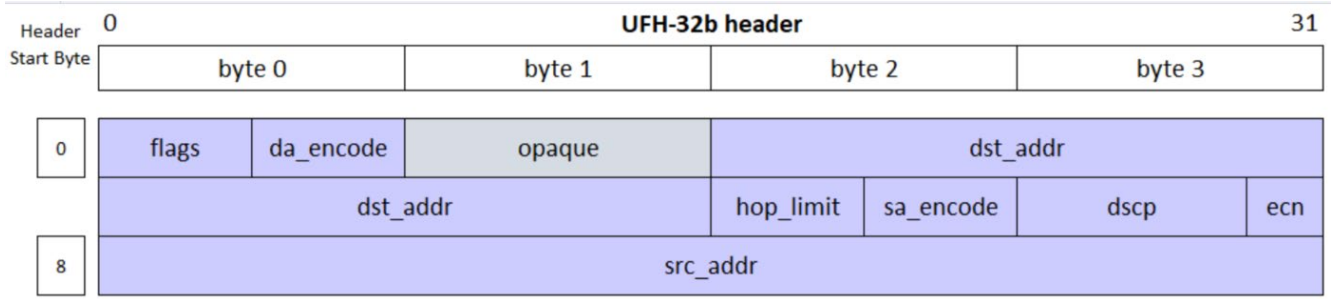
Two UFH Variants

The variants differ by address-container size and the inclusion of additional IP fields. Higher-layer protocols, identified by EtherType, use the opaque space as needed.

- UFH_16b – Optimized for smaller scale single hop networks



- UFH_32b – Suitable for multi-hop topologies



Reliable Ordered Delivery for Local Networks (RODL)

An optimized delivery mode that populates the opaque spaces in the UFH header with compressed PDS header fields.

Operates the same as ROD, but with the following exceptions:

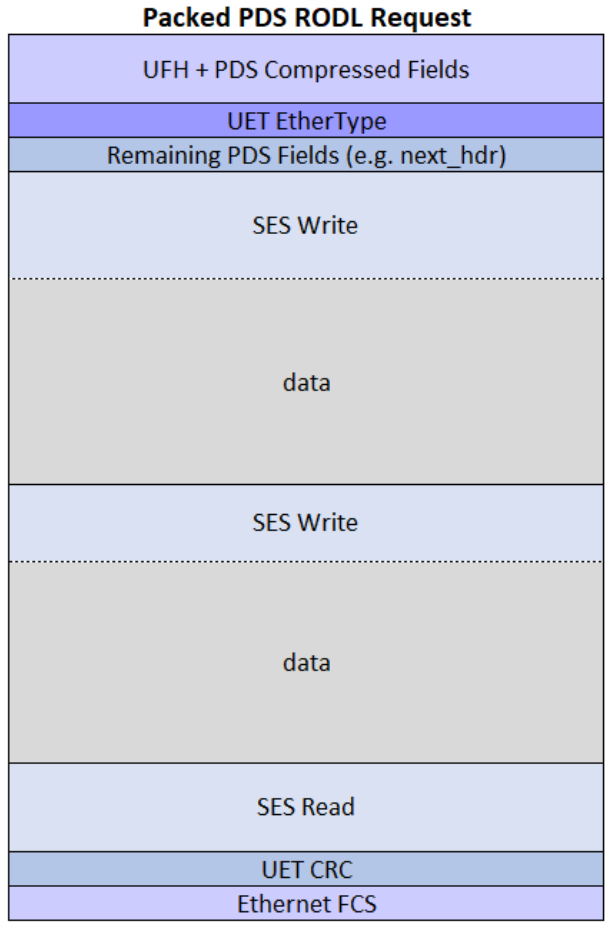
- Bi-directional PDCs allowing both sides to initiate transactions
- Reduced state: smaller PDCID, smaller PSN clear state and compressed flags
- ACKs are carried in requests in the reverse direction
- Retransmissions can be excluded
- Expected PSN is implicitly included in the CRC calculation

Key Point:

RODL is optimized for tightly-coupled accelerator fabrics.

Efficient Small Transaction Processing

- RODL is optimized for small memory read, write and atomic transactions in range of 64B to 256B.
- New compressed semantic headers are defined that use reduced request length, JobID, and direct memory addressing.
- Multiple semantic operations are ‘packed’ into a single PDS Request transmission.
- Multiple semantic responses are ‘packet’ into the PDS Response transmission.



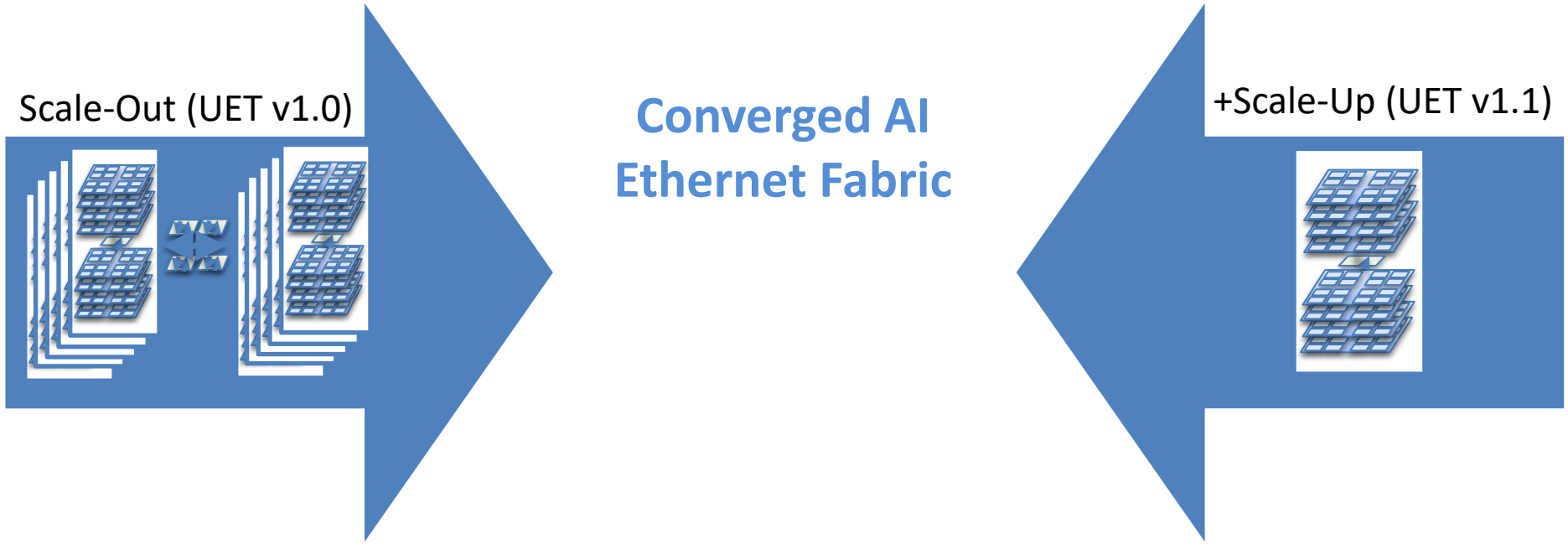
UET Scale-Up Evolution Summary

UET v1.0	UET v1.1 Adds
Scale-out focus	Scale-up optimization
Traditional Ethernet/IP/UDP headers	UFH + Traditional protocols on the same wire
ROD/RUD/RUDI/UUD	RODL
Large fabric support	Optimization for local fabrics
256k Ethernet ports	32 to 4K Ethernet ports
<10 μs one-way latency	<1 μs one-way latency
MPI and AI message transaction efficiency	Load/Store/Atomic transaction efficiency
AI Base, AI Full, and HPC deployment profiles	AI Local deployment profile

Toward Converged AI Ethernet Fabrics

UET evolution is enabling a single Ethernet operational model spanning local accelerator fabrics and cluster-scale fabrics

- Fewer specialized networks and gateways
- Reusable Ethernet tooling: telemetry, security, automation, operations
- Open multi-vendor path from pod-scale to regional AI clusters




IEEE
802


ITU