Ultra Ethernet

ECE/CS598HPN

Radhika Mittal

Acknowledgements: slides made with help from AI (Claude)

Other industry papers on RDMA for AI

- Meta's RoCE deployment for AI (SIGCOMM'24)
  - Centralized traffic engineering, PFC without any congestion control (difficult to tune DCQCN)
    - receiver-driven traffic admission at the library level.
- Alibaba HPN (SIGCOMM'24)
  - Dual plane, two tier topology for AI workloads; path selection at library-level.
- MRC from OpenAI (arxiv paper)
  - Multiplane topology, packet spraying, adaptive load-balancing based on ECN marks, loss detection via trimmed packets, selective retransmits, and no PFC (hard to combine with packet spraying!)
- Vela by IBM (ASPLOS'25), Ailbaba Stellar (SIGCOMM'25)
  - Virtualizing LLM training, use GPU-Direct RoCE

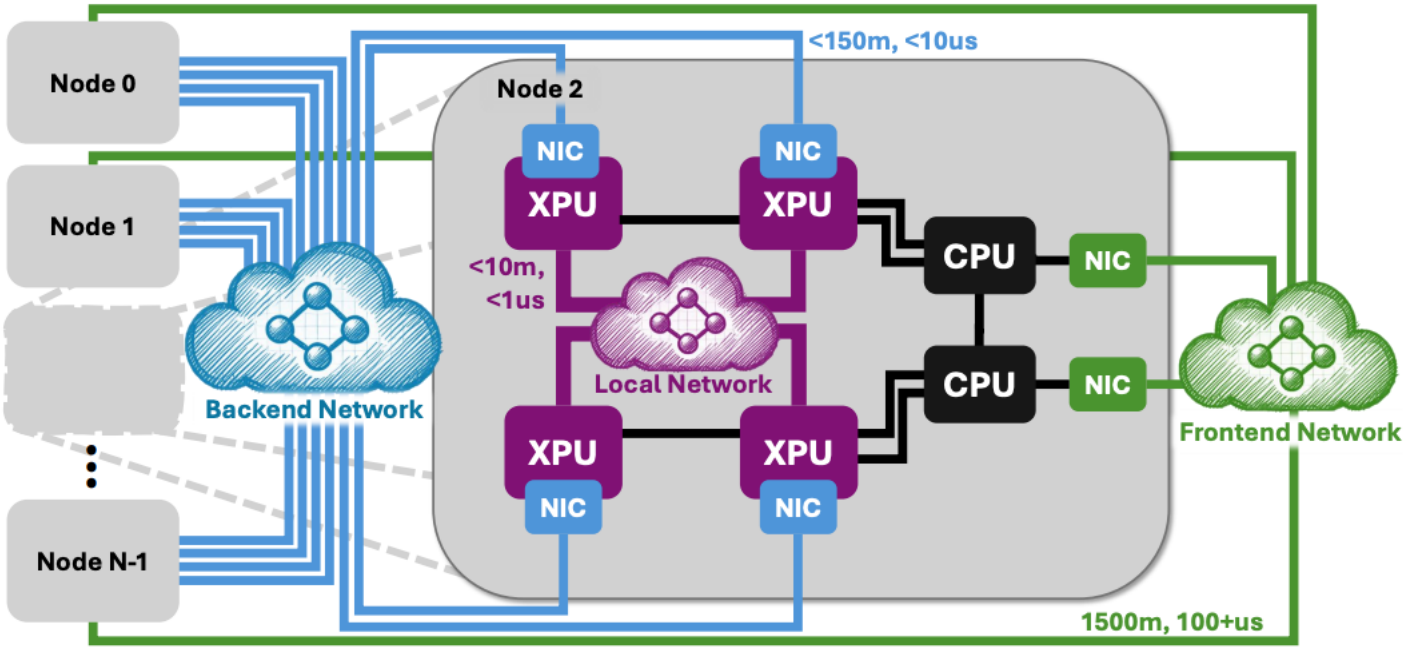
Today's paper: Ultra Ethernet

- RDMA plucked out of Infiniband-based HPC settings and thrown into Ethernet networks.
- Several scaling challenges.
- Clean-slate design for Ethernet?
  - UltraEthernet: collaborative effort (Microsoft, OpenAI, Broadcom, Intel, Cisco, AMD, HPE)
  - Standardize the requirements

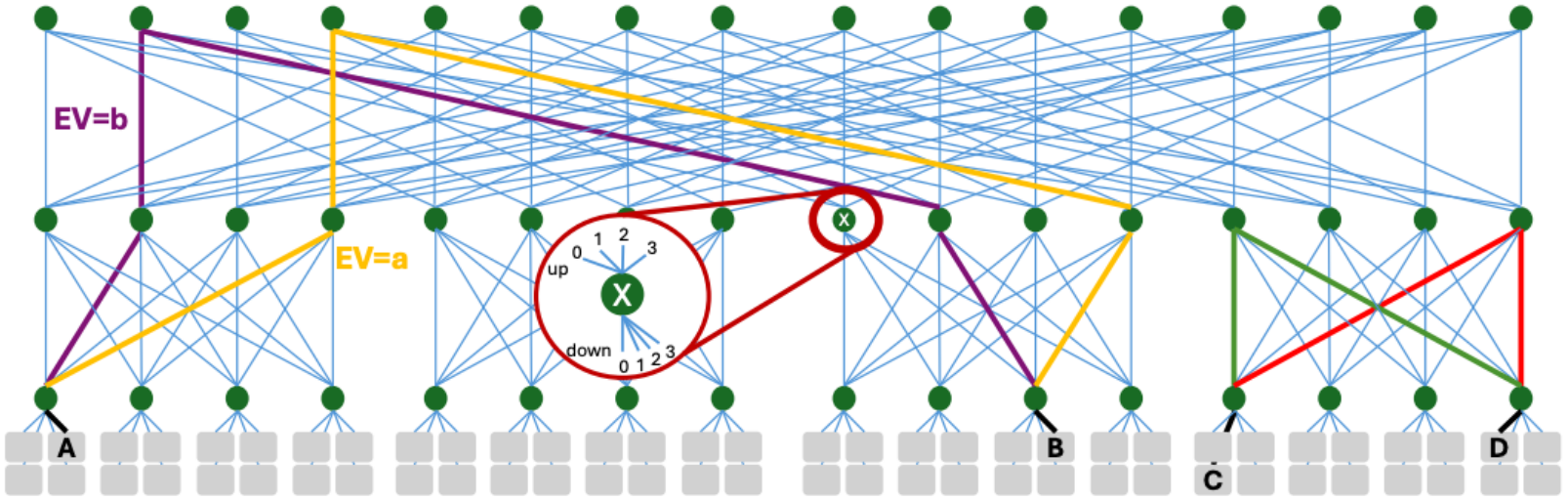
Primary Usecase and Requirement?

- AI workloads
  - direct communication between GPUs (XPUs); CPU-bypass
  - scale to > 100K GPUs
  - effective use all available network capacity to reduce communication time.

Context

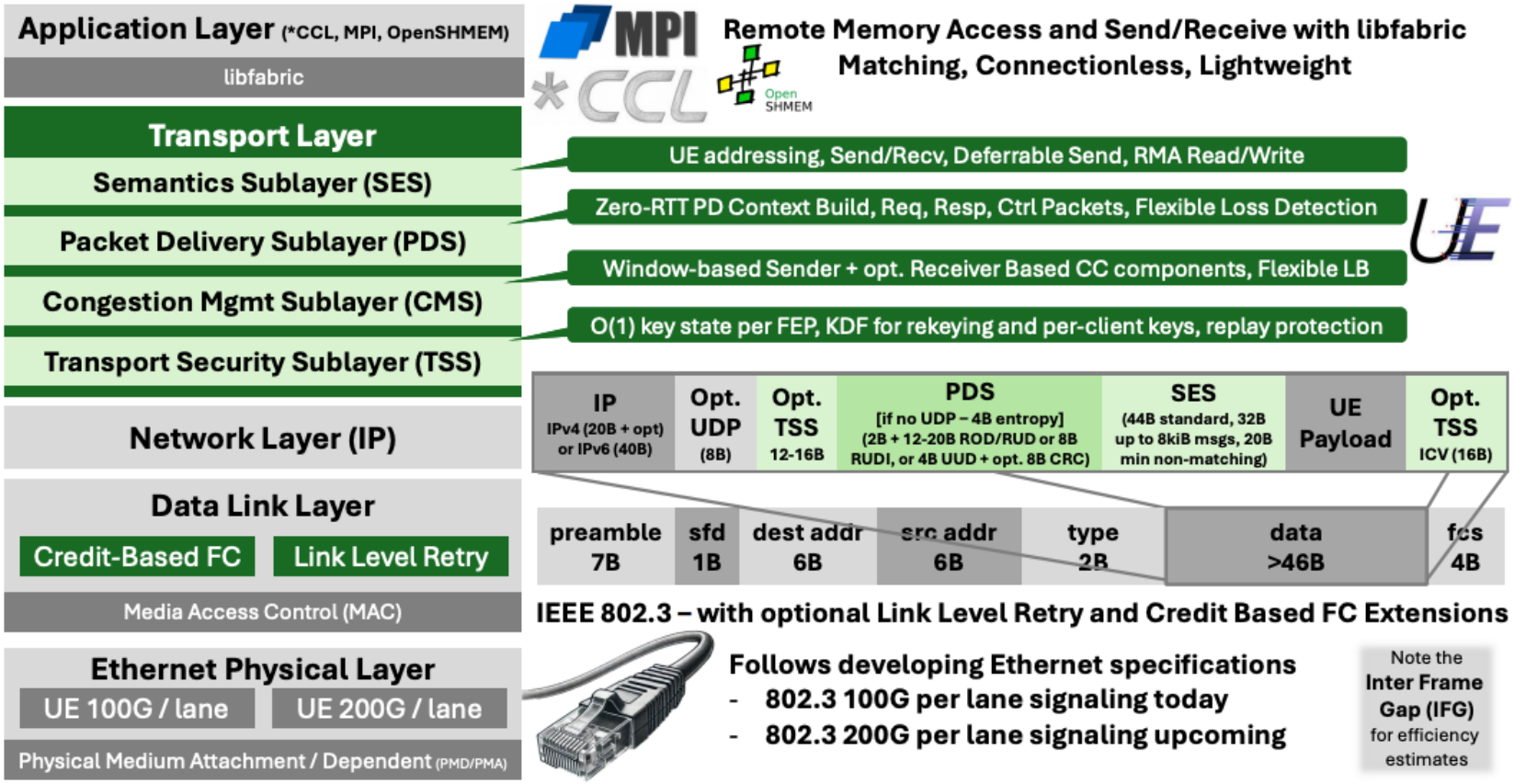

Node 0
Node 1
Node N-1
Backend Network
<150m, <10us
Node 2
NIC
XPU
XPU
<10m,
<1us
Local Network
XPU
XPU
NIC
NIC
CPU
CPU
NIC
NIC
Frontend Network
1500m, 100+us

Context: ECMP and Entropy Values (EVs)



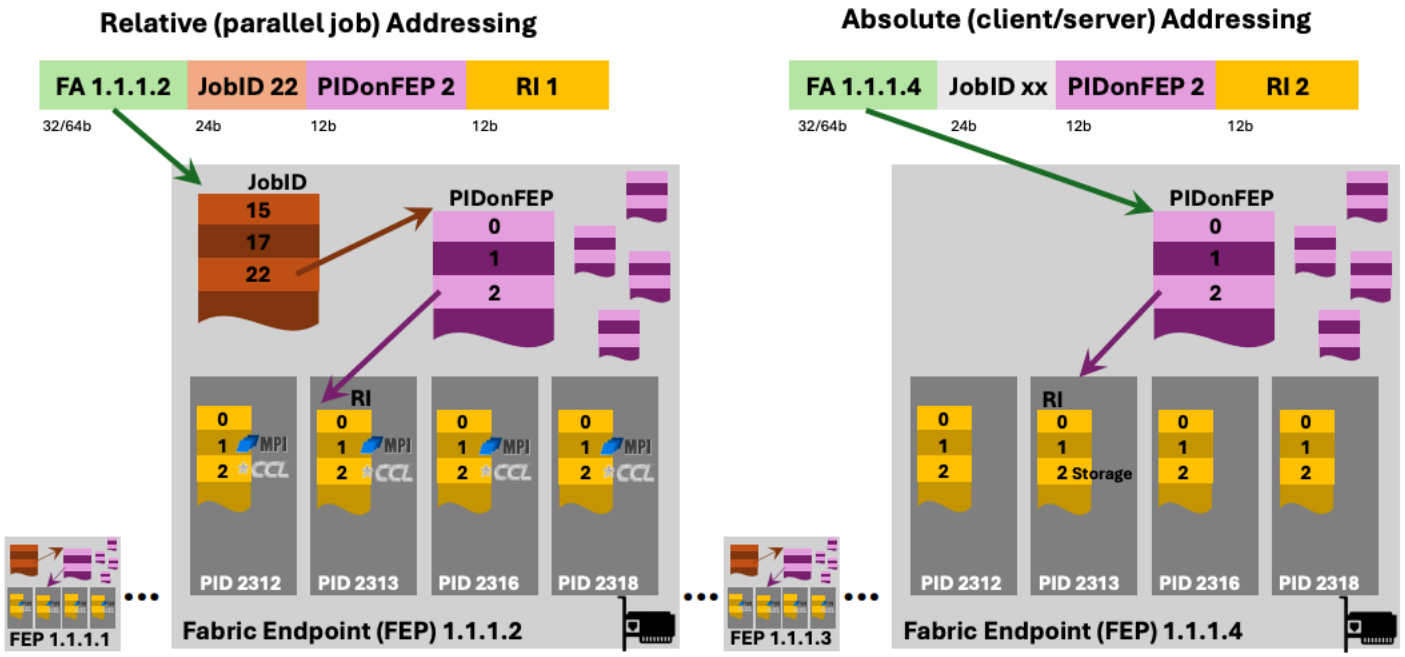
EV: UDP source port
(UDP destination port is fixed)
Also support direct EV value when UDP not used.

Overall architecture



SES: Addressing

- No QPs



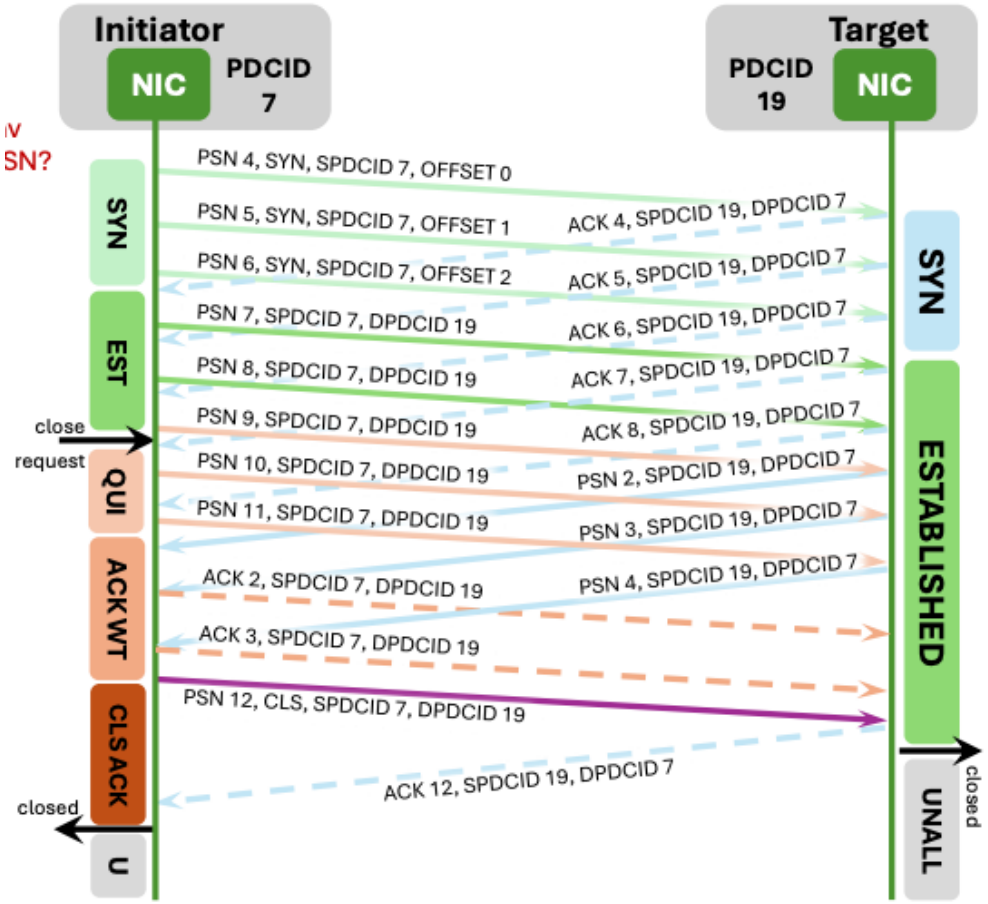
SES: RMA write/read

- Write: Each packet carries target memory address and keys
- Read: single-packet reads only
- Control traffic (e.g. acks is prioritized

Packet Transport Modes

- Reliable Unordered Delivery
- Reliable Ordered Delivery
- Unreliable Unordered Delivery
- Reliable Unordered Delivery for Idempotent Operations

Dynamic Packet Delivery Context Creation



Detecting Packet Losses

- Packet trimming
- Out-of-order count
- Tracking across EVs

Congestion Control

- Network Signal Based:

(1) ECN mark + low RTT indicates that congestion is building up at least at one switch buffer
(2) ECN mark + high RTT indicates that the network is congested, may be overloaded
(3) No ECN mark + low RTT indicates that the network is not congested, may be underloaded
(4) No ECN mark + high RTT indicates that the network congestion is going down

- Receiver Credit Based: only suitable for incast.

Load Balancing

- Change EV every packet (packet spraying)
- Detect low throughput EVs and avoid using them.

Link Layer

- Link-layer retransmits for corruption errors.
- Credit-based flow control (for ROD traffic)

Transport Security Subsystem

- Header authentication.

Other thoughts?

Products

- AMD Pensando Pollara 400G NIC (Dec 2025)
  - UET compatible
  - Also supports P4 programmability!

- Broadcom Ultra Thor 800G NIC (UET compatible)

- Switch LB support just catching up.

New in UET version 1.1

- In-network collectives supports
  - Switches aggregate data.

- Switch telemetry for congestion control
  - CSIG (developed at Google)

Project Report Feedback

- Aim to send by next week.

- Please reach out to me if you feel stuck.