

MetaRoCE RDMA Transport

Arvind Srinivasan, *Meta Platforms*
Balakrishnan Raman, *AMD*
Sandeep Nagaraja, *Meta Platforms*



AI Network Transformation

The network is ***in the critical path*** for AI Clusters — no longer auxiliary but an integral part of infrastructure

SCALING CHALLENGES ARE REAL

Massive Cluster Size

Hundreds of thousands of accelerators today, across multiple data centers and regions — designed for a million.

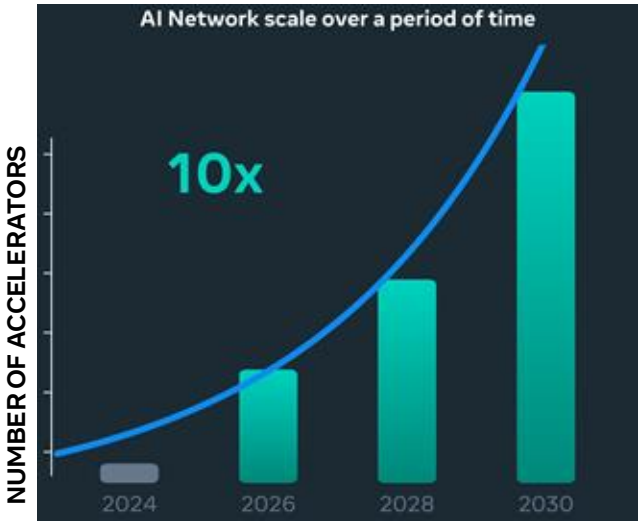
Training Bottlenecks

All-reduce and all-to-all synchronization — performance gated by the slowest node.

Inference Latency

Latency between distributed model shards directly impacts response times for hundreds of millions of users.

AI NETWORK SCALE OVER TIME



Meta-RoCE Transport

Optimized RDMA Transport for Meta's Network Endpoints & AI Accelerators

PERFORMANCE @SCALE

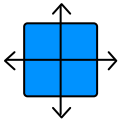


>90% network efficiency in non-blocking manner

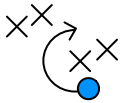


Per packet load balancing & spray

RESILIENCE @SCALE

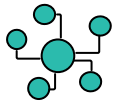


Reliably scaling across multiple network planes with transparent failover

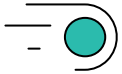


Per-path reliability coupled with congestion control

OPERATIONS @SCALE



Operates on commodity Ethernet Infrastructure



Frictionless adoption with full verb compatibility

Meta-RoCE Transport

Multipath RDMA transport for AI Clusters

- Per-path reliability and congestion control
- Receiver rate hints (RCN) for fast incast convergence
- Tolerates packet loss and reordering
- Single QP carries multiple streams over multiple network paths
- Runs on commodity Ethernet: ECN + ECMP, no switch modifications

Fundamentals

Key Architectural Decisions

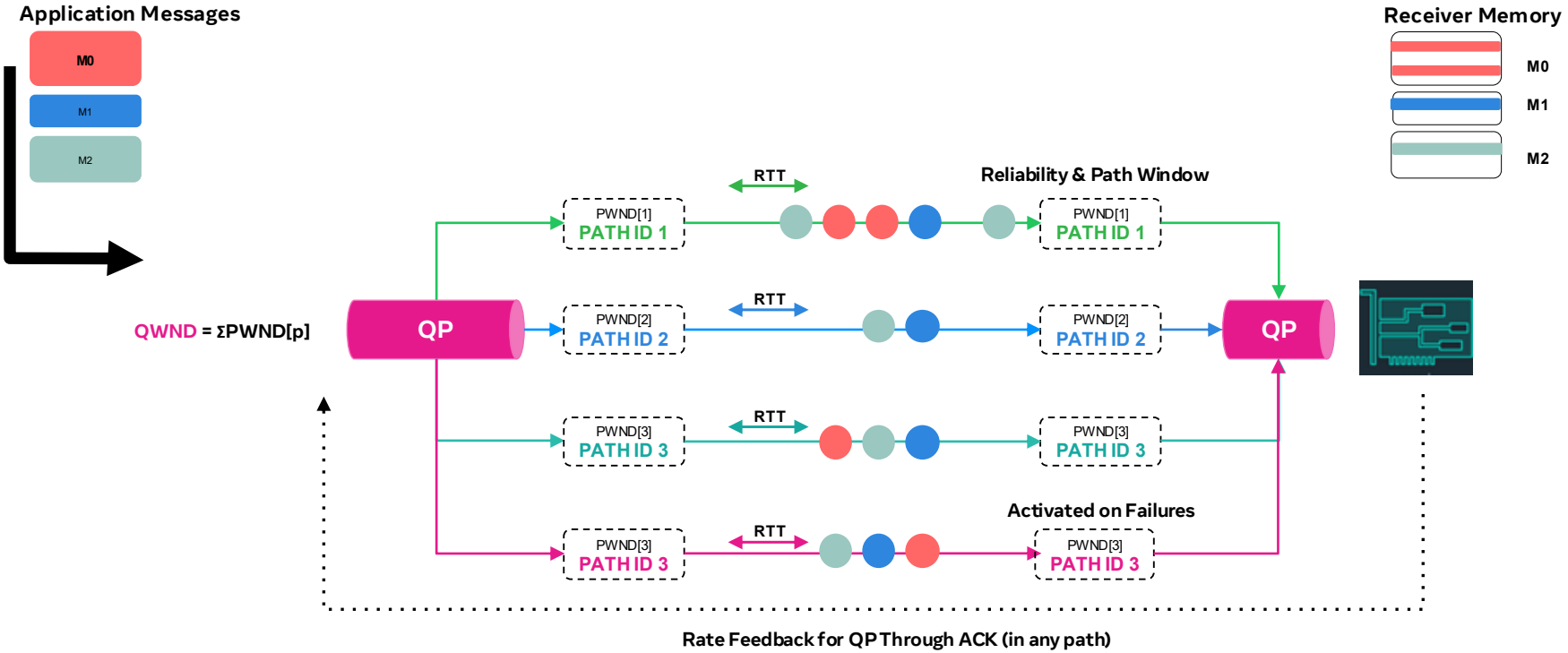
Use Commodity Ethernet Switch

- ECN and ECMP only — no trimming, adaptive routing, or in-network telemetry
- Trimming adds switch and NIC complexity; SACK detects loss without either

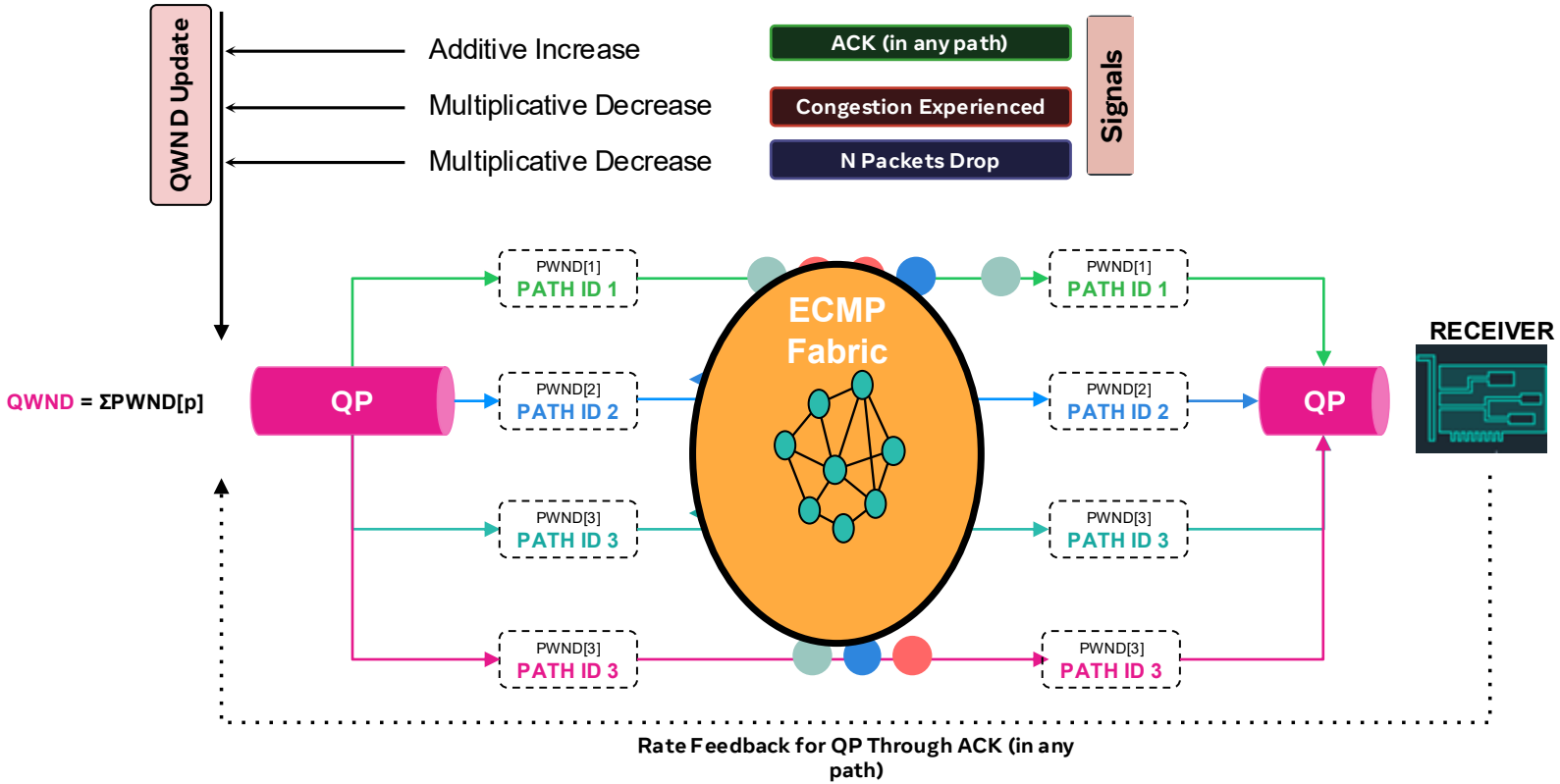
PFC, go-back-N, and reorder buffers do not scale

- PFC headroom grows with distance; deadlock risk in large fabrics
- Go-back-N wastes bandwidth; reorder buffers cost SRAM that grows with BDP
- Instead: Order completions, Not delivery — OOO placement, no reorder buffer

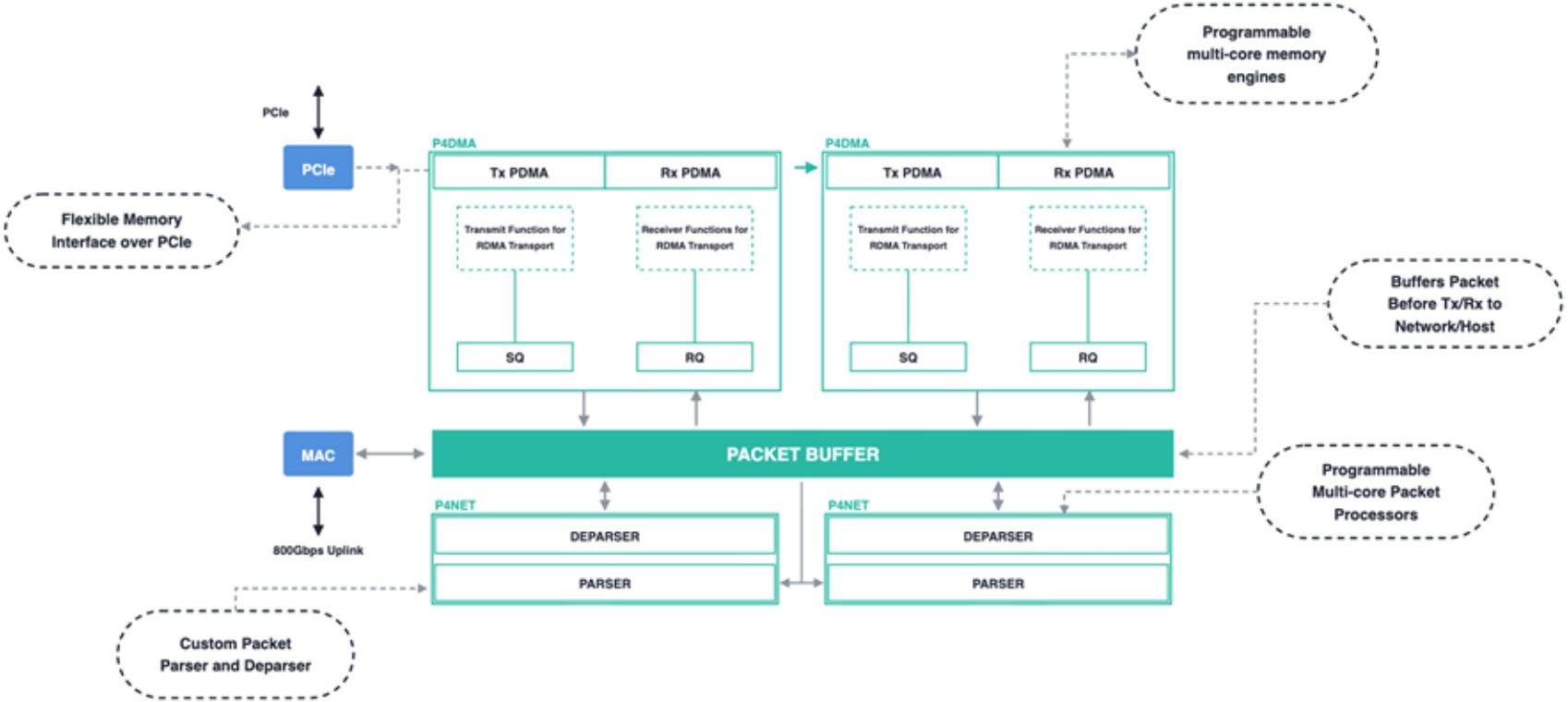
MetaRoCE Paths: Building Block for Multipath



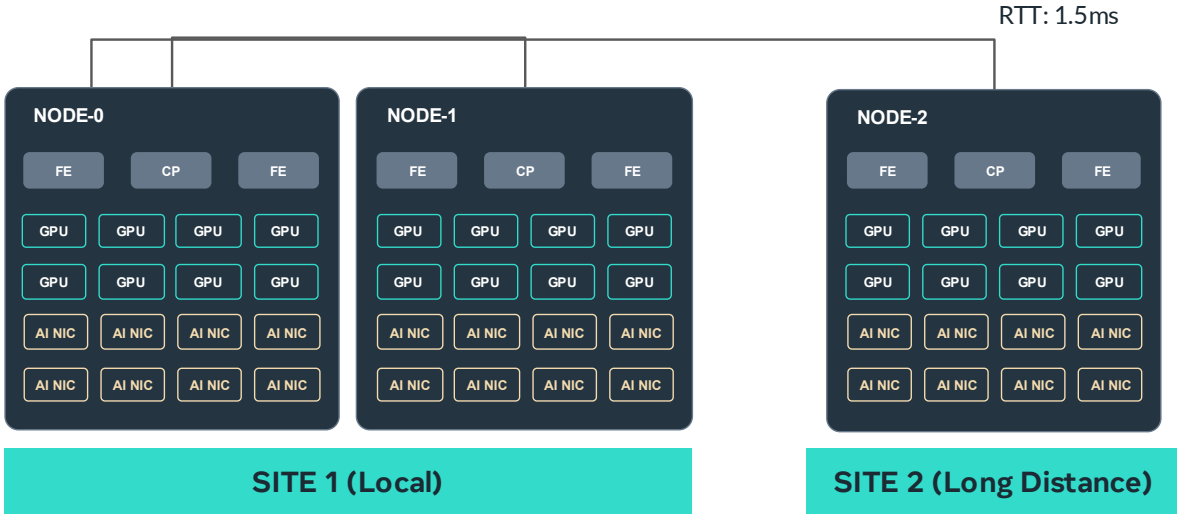
Congestion Control & Reliability



AMD Programmable AI NIC

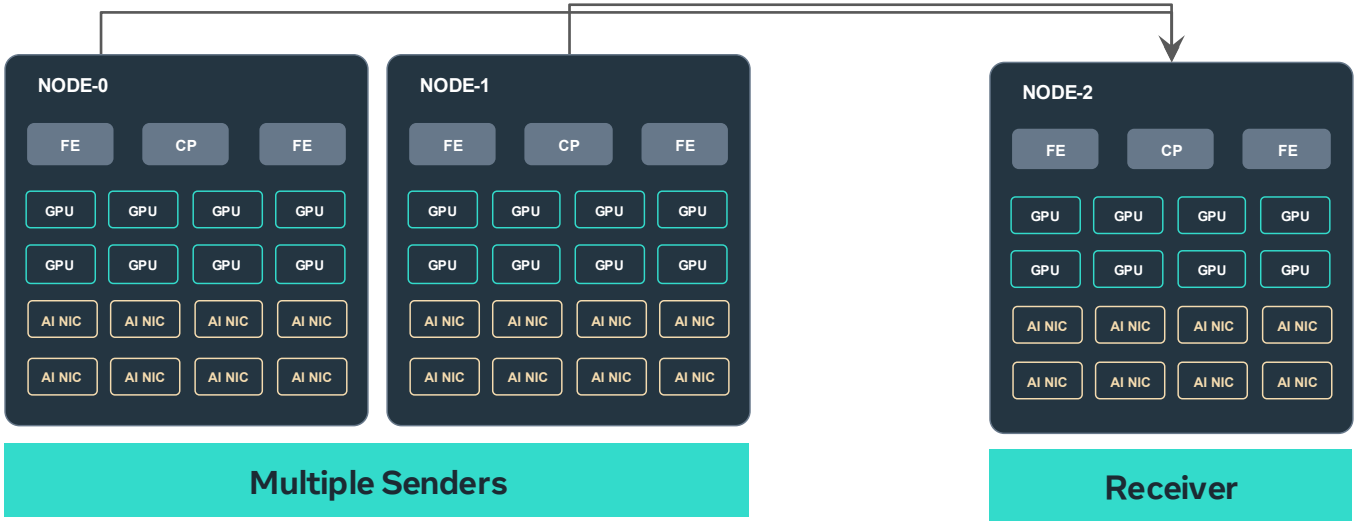


Simultaneous Scale-out and Scale-across



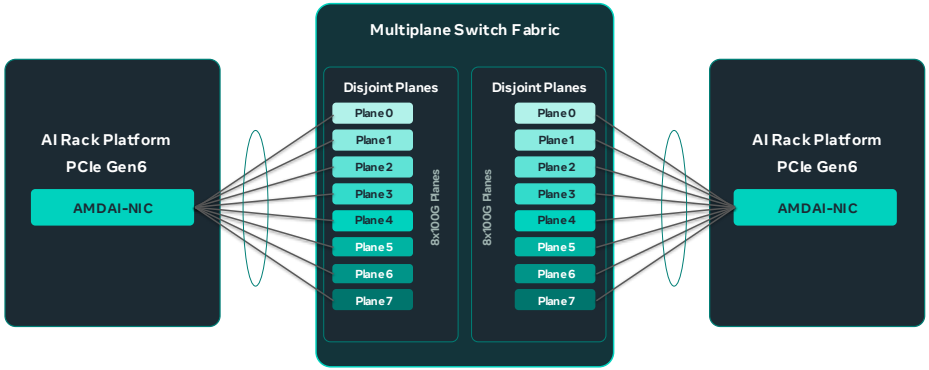
Scale-Out	Scale Across
RTT: 5-15 micro-seconds	RTT: Often reaches milli-seconds i.e. 50-100x
Max Cwnd: ~75; Max Path: ~12 (4k MTU packets); Line rate performance	Max Cwnd: 3500 to 7000 Max Path: ~30 - 50 (4k MTU packets) Line rate performance

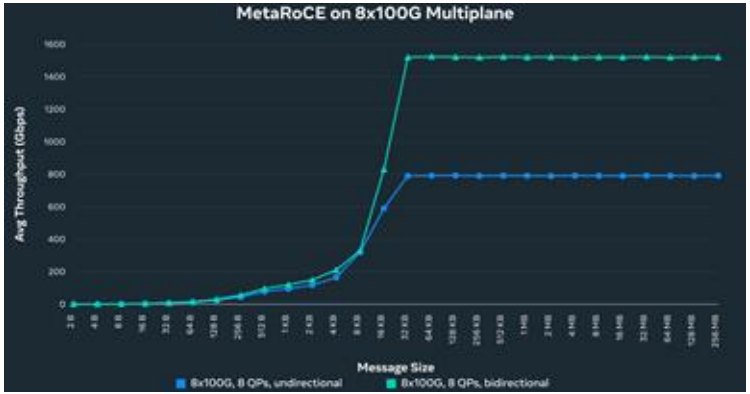
MetaRoCE Congestion Control: Incast Handling

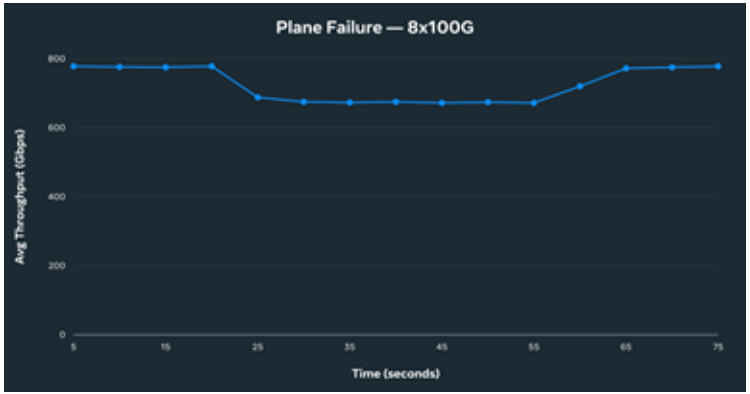


- Congestion window rate-hints(in Gbps) : 200 Gbps per sender

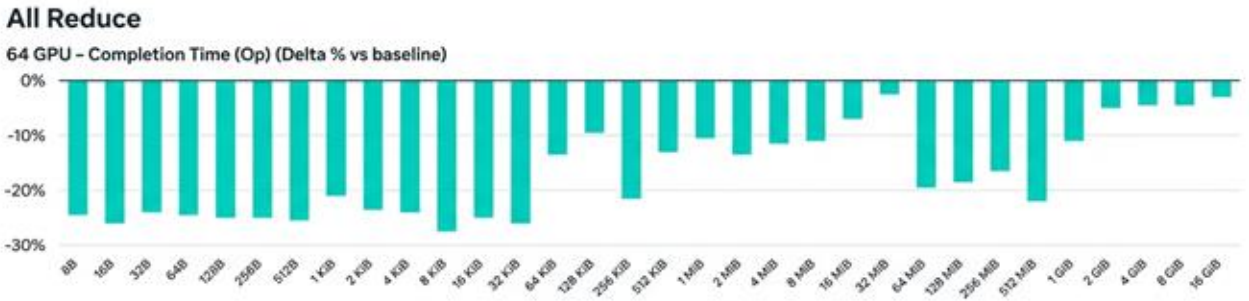
MetaRoCE on a Multiplane FPF Network

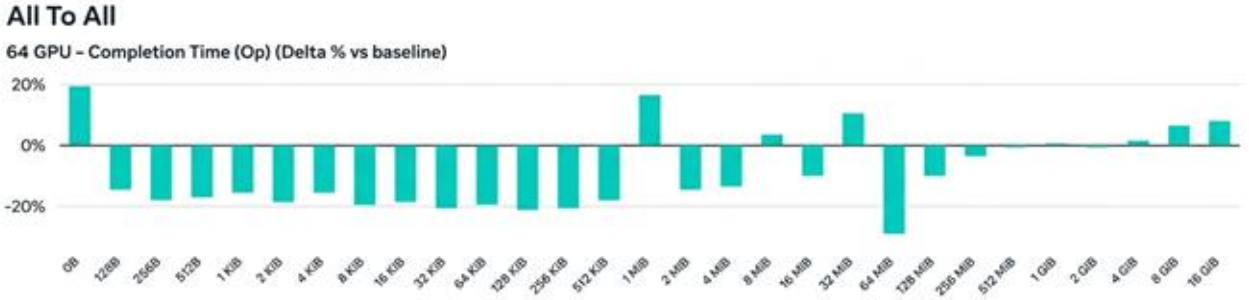




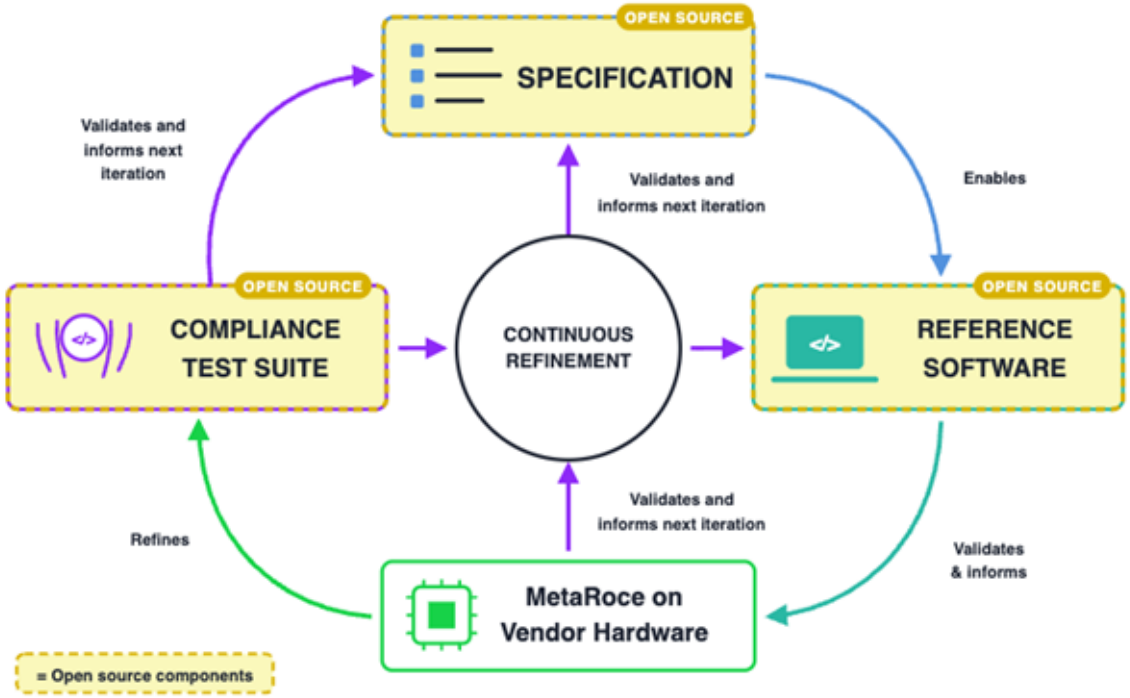


MetaRoCE vs RoCEv2 Collective Completion Times





Enabling OCP community on MetaRoCE development



Call to Action

Make large-scale AI faster, cheaper, and truly open for everyone.

1. Adopt & Contribute

Join us in making MetaRoCE an open, multi-vendor AI fabric.

- Adopt the open spec using Ref Implementation.
- Implement & test against SmartNICs, DPUs, switches, and ASIC designs.
- Publish your results.

2. Validate & Standardize

Help achieve broad hands-on validation, interoperable hardware, and shared operational experience.

Together, we can turn a purpose-built transport into an industry standard.

MetaRoCE GitHub

https://github.com/facebookincubator/metaroce

Scan to open

Thank You!

