BROADCOM

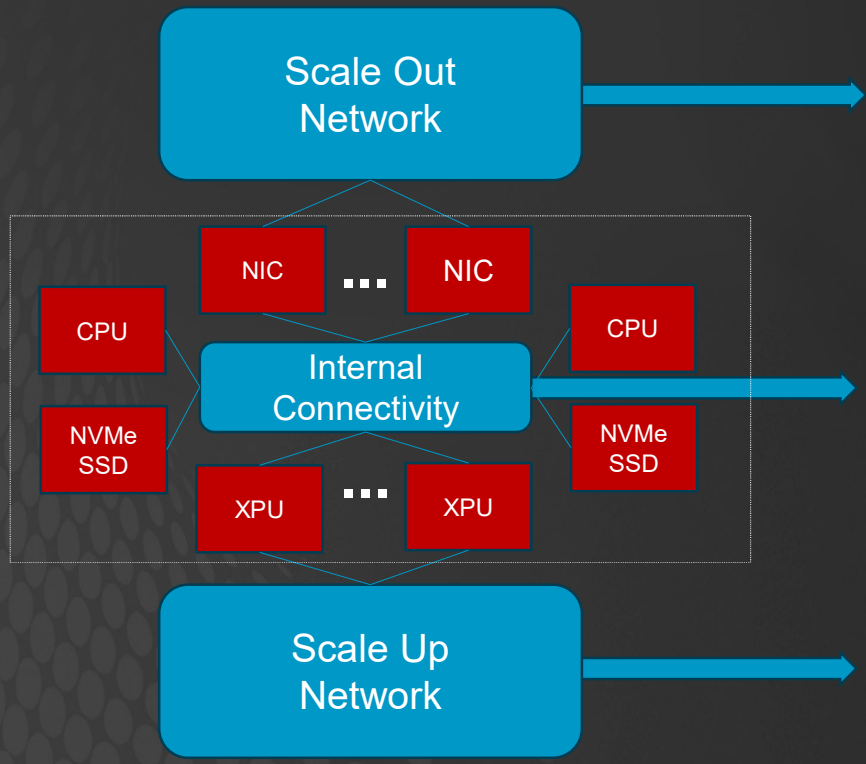
ENABLING AI Infrastructure

RoCE Enhancements for Large Scale Multi-Path Ethernet Networks

Hemal Shah, Distinguished Engineer and Architect
Core Switching Group, Broadcom Inc.
MVAPICH2 User Group Conference, August 19, 2025

OPEN // SCALABLE // POWER EFFICIENT

AI Infrastructure Connectivity



Ethernet

- Cost-effective, flexible, and scalable ecosystem
- High-Bandwidth → 800G and above
- Large scale → 100K-1M XPUs
- RoCE transport → Low latency, high bandwidth & low overhead

PCIe

- Ultra Low Latency
- Standard interface on AI server devices
- Enables peer-2-peer data transfers

Options

- 1. Ethernet → Performance & Scale
- 2. NVLink / UALink
- 3. PCIe → Simplicity & Cost

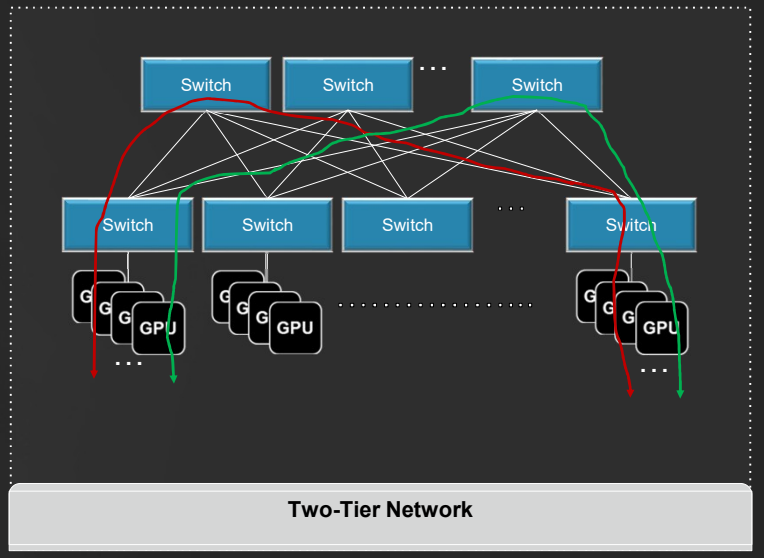
BROADCOM

RoCE for AI/HPC Networks

- Default RDMA transport for AI training and inference clusters
- Enables direct data transfer without involving compute engines and/or OS → reduces communication times
- Provides efficient high-bandwidth transfer → crucial for distributing massive datasets & parallel processing
- Optimizes XPU-XPU communication → RDMA between XPU memories & high XPU utilization
- Scales to 100K-1M+ XPUs → supports large-scale AI/HPC cluster deployments

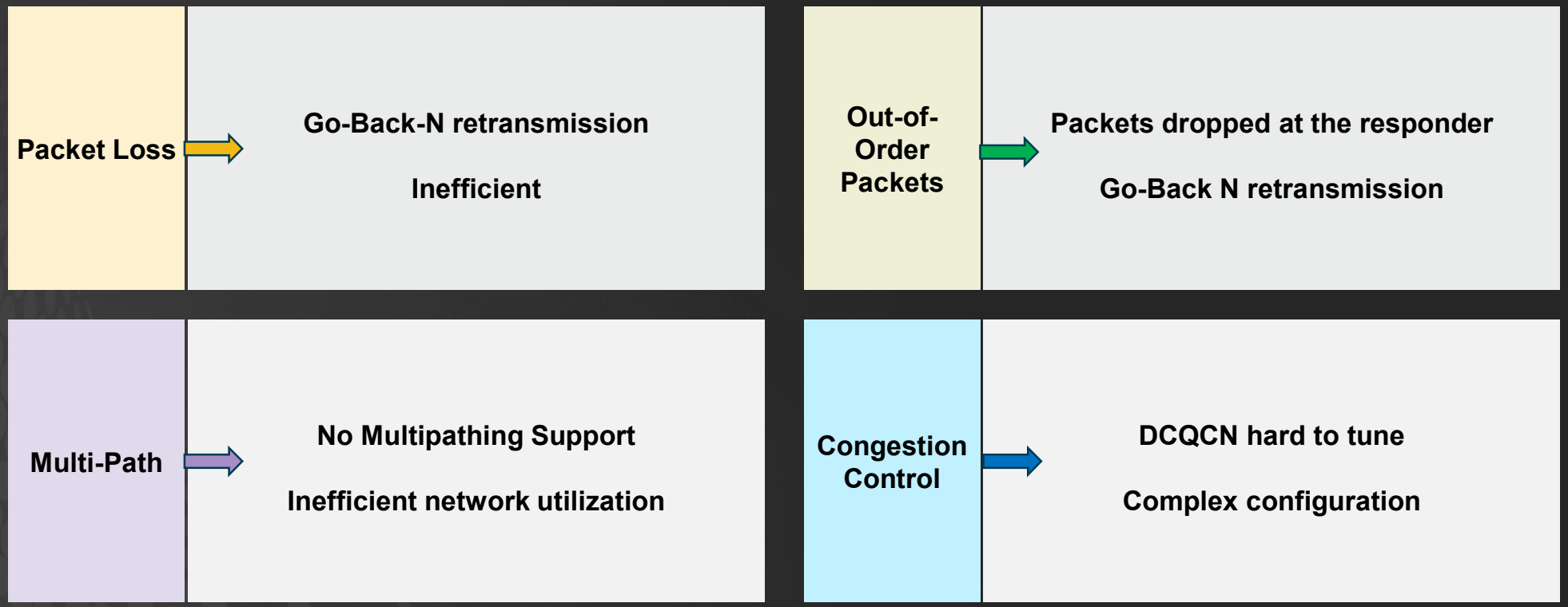
Multi-Pathing in AI/HPC Ethernet Networks

- Large scale AI/HPC networks are N-tier
- Multiple paths available in the network
- Benefits of multi-pathing
  - Network load balancing
  - Fault-tolerance
  - Congestion reduction
- Out-of-order packet arrival due to packet spraying and drops

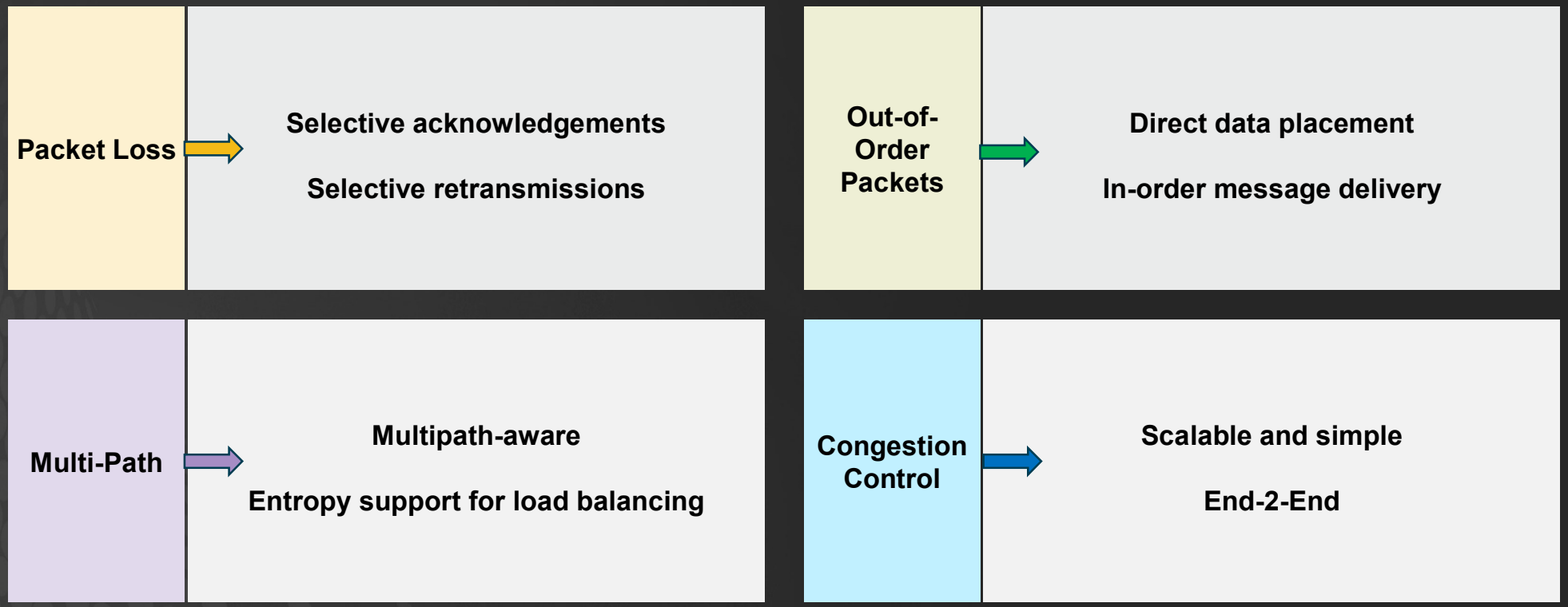


BROADCOM

RoCE Challenges in Large Scale AI/HPC Ethernet Networks



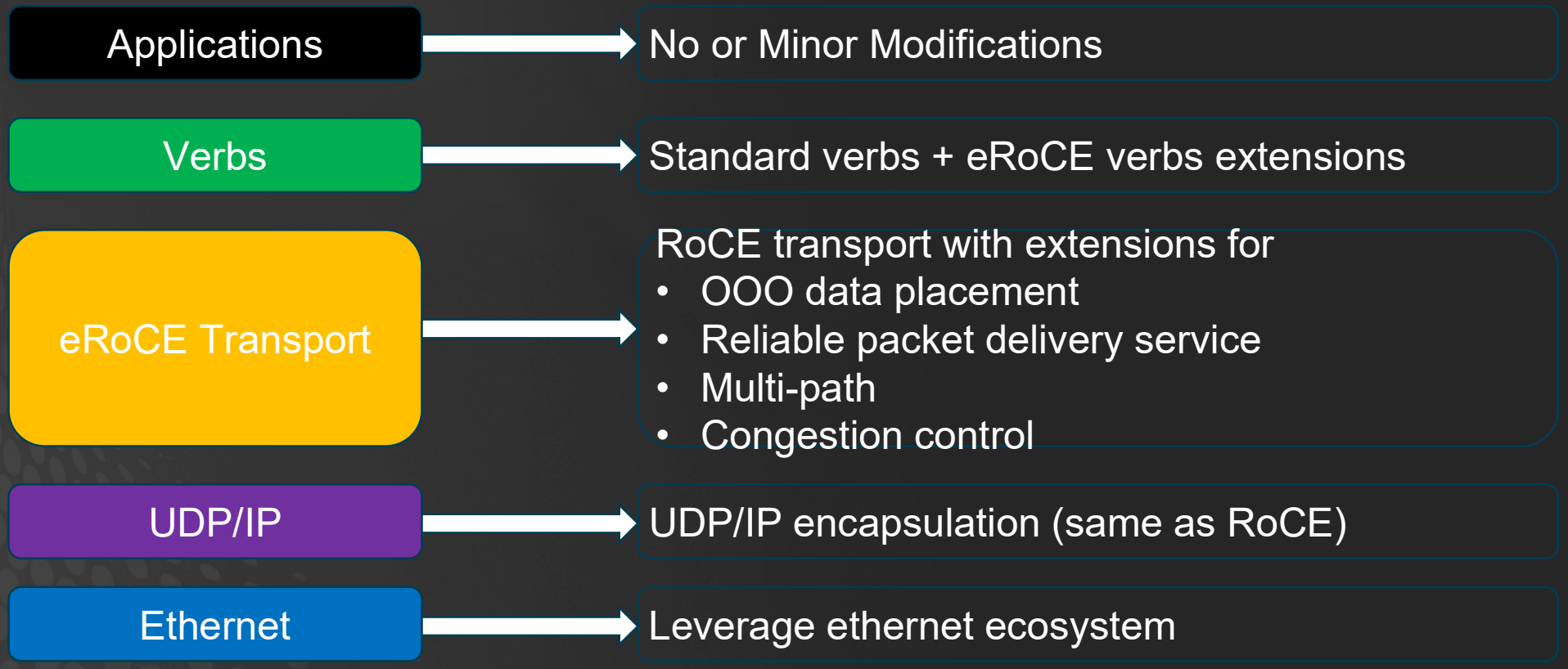
Enhanced RoCE (eRoCE) for Large Scale AI/HPC Ethernet Networks



eRoCE Features for AI/HPC

Feature	Description
Multipath	• Header entropy variation (per QP or per packet) • Switch load balancing support: ECMP, E-ECMP, DLB flowlet, DLB spray • NIC packet spraying across multiple planes • Re-ordering at the target NIC
Out-Of-Order (OOO) Placement	• Multi-pathing and packet drops → OOO packets • OOO placement of RDMA Write & Read responses (Write w/ Imm delivered in-order) • In-order placement/delivery of Sends, RDMA Read Requests, and Atomics • Better utilization of packet buffers & PCIe
Reliable Delivery	• SACK and NACK • Hardware based selective retransmissions • Ordered and unordered delivery
Congestion Control	• End-2-End credit-based congestion control • Programmable • Path probing
Telemetry	• ECN for marking • Packet trimming for drop notifications • CSIG and DCN

eRoCE Stack



Summary

- Ethernet with RoCE has been the widely deployed in AI/HPC networks
- RoCE faces challenges (go-back-n, no multi-pathing, congestion control..) for large scale Ethernet networks
- eRoCE addresses these challenges with reliability, multi-path & congestion control enhancements to RoCE
- eRoCE extends RoCE while preserving programming model, RDMA semantics, and baseline RoCE transport
- eRoCE supports large-scale AI/HPC cluster deployments