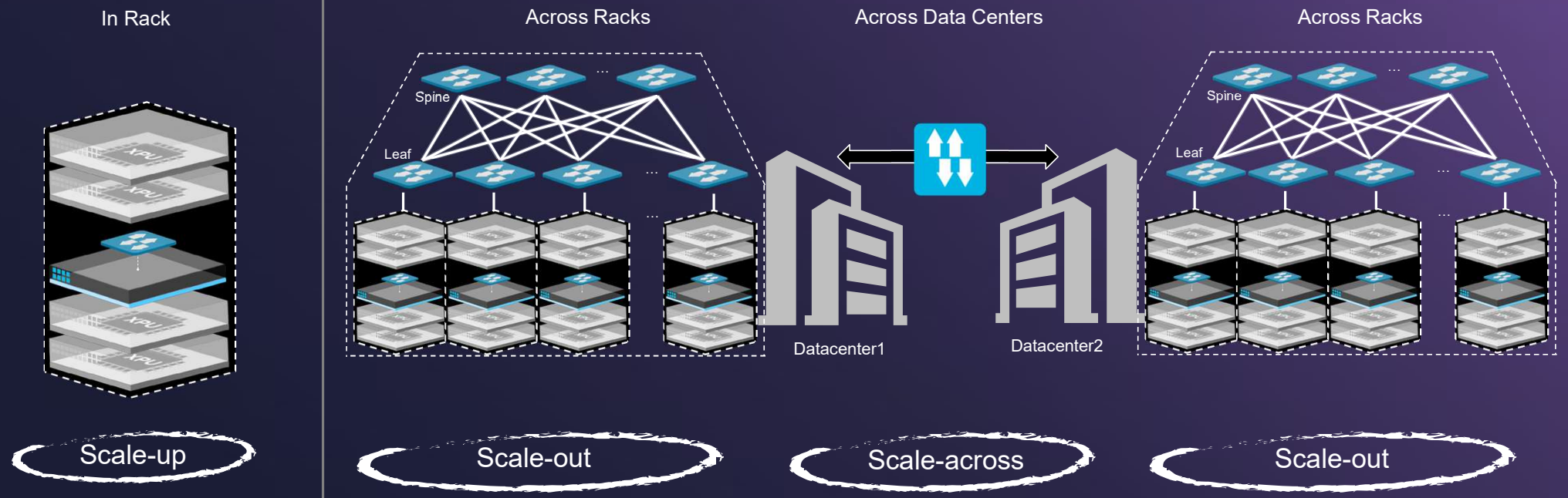
Thor Ultra NIC for AI and HPC Network Infrastructures: An Architectural Overview

Hemal Shah, Distinguished Engineer and Architect, Broadcom Inc.

Ethernet Networking for AI and HPC



Ethernet Everywhere

Ethernet NIC Requirements for AI and HPC

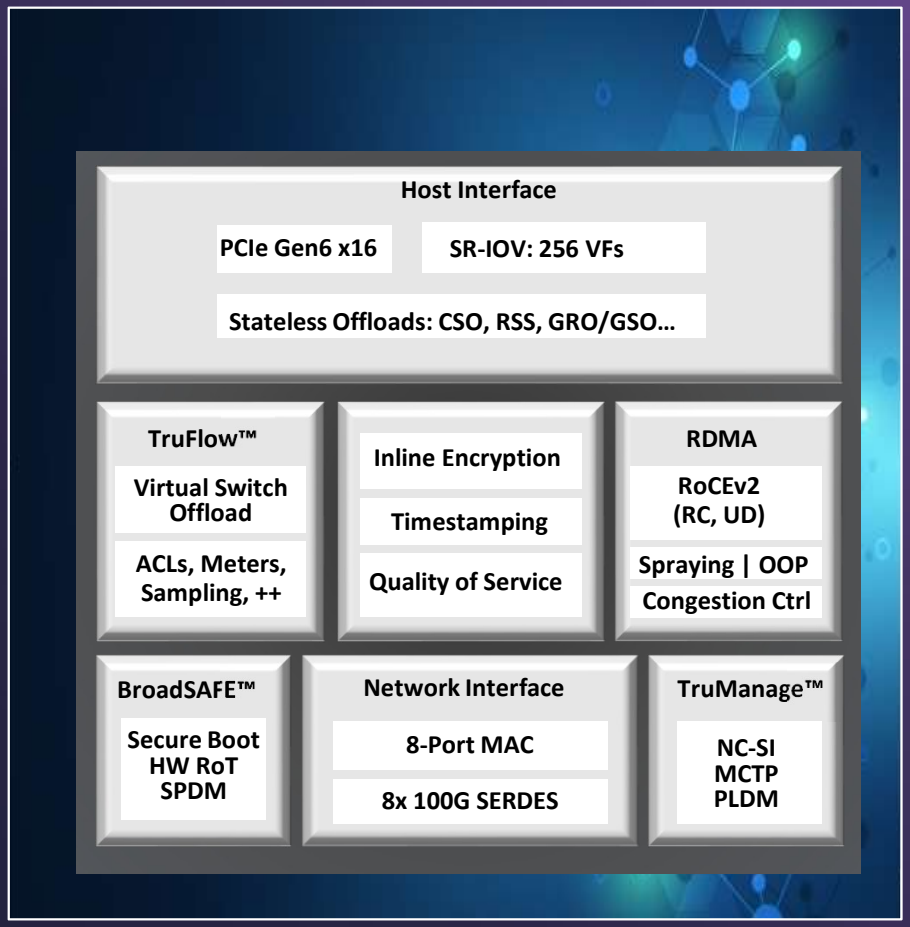
- High bandwidth – 800 Gbps now → 1.6-3.2 Tbps in future
- Scalable – maintain high throughput as NIC resources scale
- Low latency – especially for a category of HPC applications

- Transport - RDMA with multi-path, reliability & congestion control enhancements
- QoS – multiple traffic classes and traffic shaping
- Virtualization – to support compute as a service

- Timing support – to enable time accurate services
- Security – to enable trusted compute infrastructure
- Standards based management – operational management of NIC

Introducing Thor Ultra Ethernet NIC

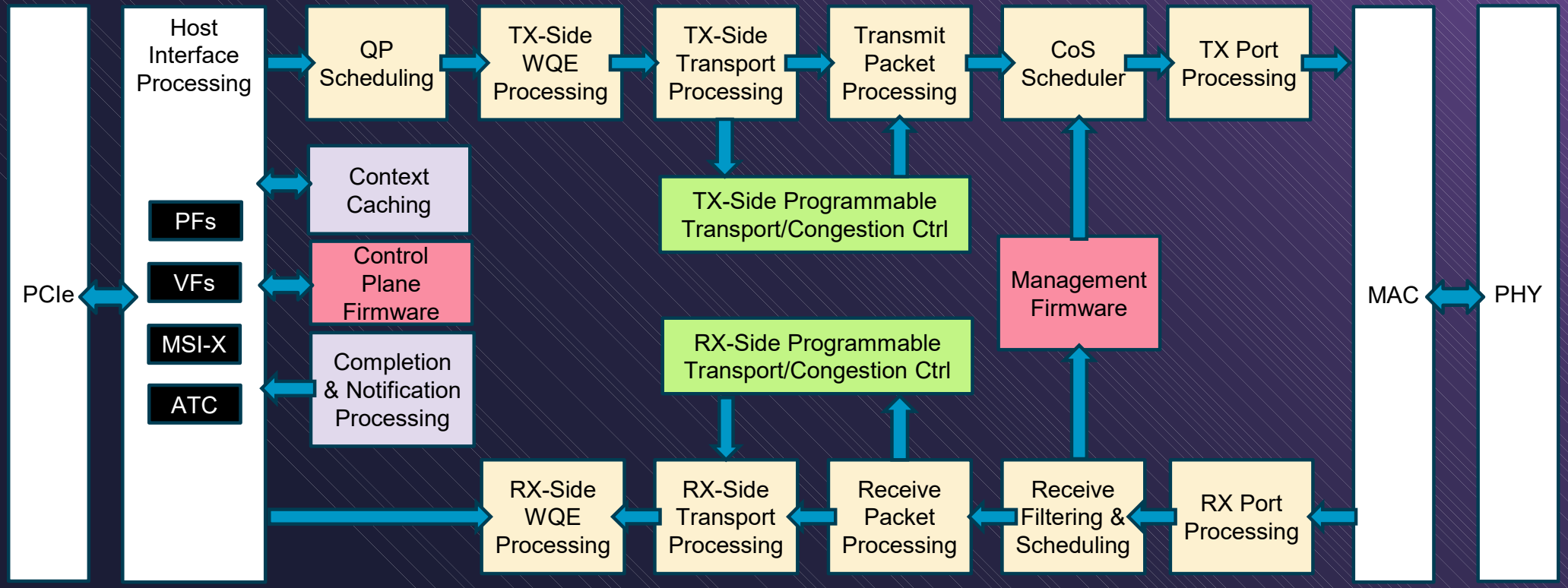
- PCIe Gen6 x16 Host Interface
- 8x 100G PAM4/NRZ Serdes
- Enhanced RoCE with 64K+ QPs
- Hierarchical QoS & SR-IOV
- Timing, Security & Manageability Support



Thor Ultra NIC Boards

Description	PCIe	OCP NIC 3.0
Form Factor	CEM HHHL	TSFF
Connector Configurations	1x OSFP112 and 2x QSFP112	
Ethernet Port Configurations for OSFP112	1x 800G, 2x 400G, 4x 200G, 8x 100G	
Ethernet Port Configurations for 2x QSFP112	2x 400G, 4x (2*2x) 200G, 8x (2*4x) 100G	
Cable/Module Support	DAC, LPO, AEC, and Optics	
Cooling Options	Air flow and liquid cooling	
Max Board Power (w/o optics)	50-55W	

Thor Ultra NIC Pipeline



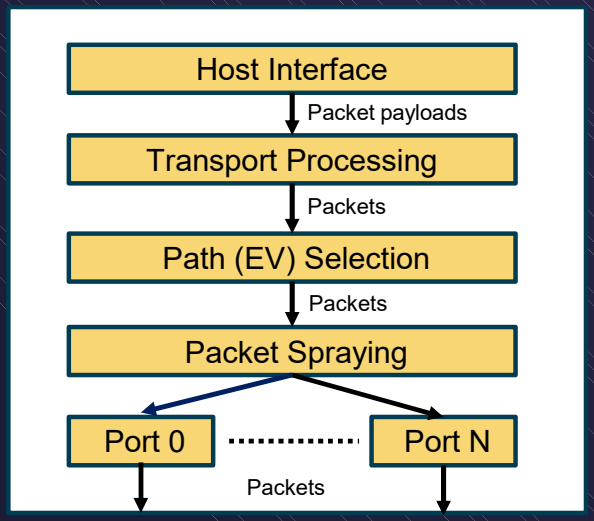
 |

Thor Ultra enhanced RoCE (eRoCE) Features

Feature	Description
Multipath	• Header entropy variation per QP or per packet • Switch load balancing modes supported: ECMP, E-ECMP, DLB flowlet, DLB spray • NIC packet spraying across multiple ports (planes) – up to 8 planes
Out-Of-Order (OOO) Placement	• OOO placement of RDMA Write & Read responses (Write w/ Imm delivered in-order) • In-order placement/delivery of Sends, Read Requests, and Atomics
Reliable Delivery	• Selective Acknowledgement (SACK) & Negative Acknowledgement (NACK) • Selective retransmissions • Probes to evaluate the quality of a path
Congestion Control Enhancements	• Receiver Credit Based Congestion Control (RCCC) - baseline • Programmable Congestion Control
Telemetry	• ECN for deterministic or probabilistic marking • Packet trimming for drop notifications • CSIG

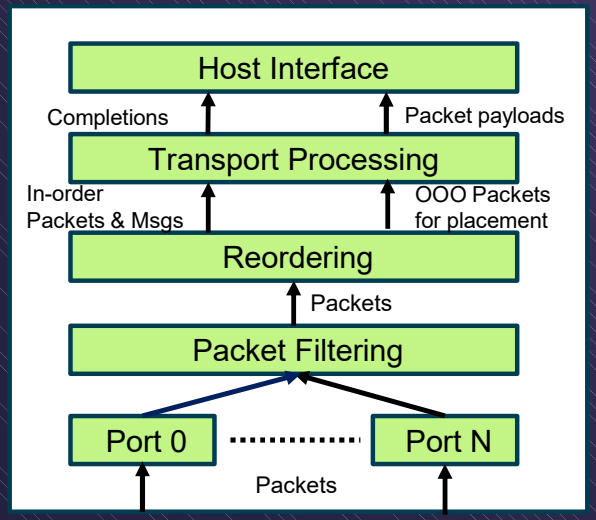
eRoCE = Multipath Reliable Connection (MRC) ++

Thor Ultra eRoCE Implementation



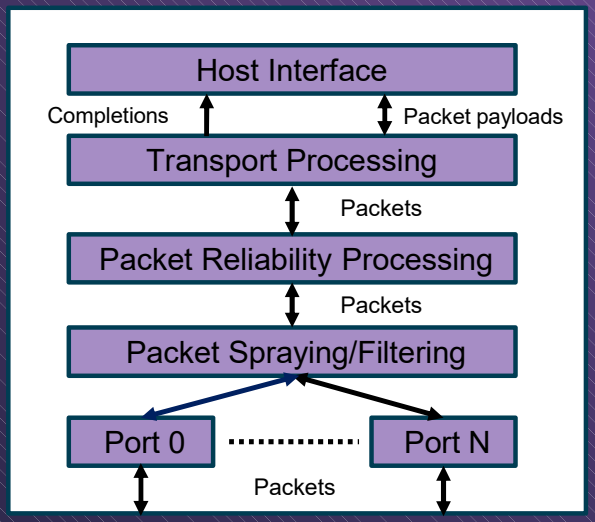
- Path selection per packet
- Paths tracked per QP
- Packets sprayed across ports
- Multiple load balancing modes

Multipath



- Ext. headers for placement info
- Re-ordering logic at the receiver
- Bitmaps to track received pkts
- In-order message delivery

OOO Placement



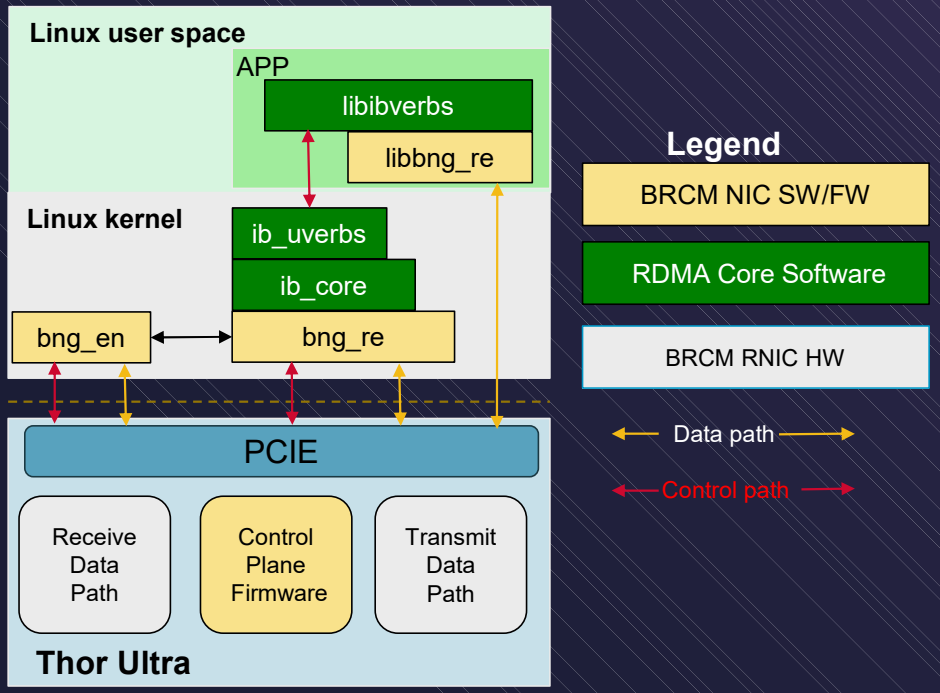
- SACK bitmaps per QP
- Configurable SACK policies
- Packet timers with exp backoff
- NACK for congestion signaling

Reliable Delivery

Thor Ultra Congestion Control Implementation

- Receiver Credit-based Congestion Control (RCCC) – baseline for eRoCE
  - Receiver managed credit allocation to active senders
  - Speculative credits for line rate startup
  - Credit scheduling is dynamically adjusted at the receiver
  - Telemetry (ECN, trimming, and CSIG) aware
- Programmability
  - Match-action (P4-like) programmable engines
  - Enables user defined and/or customizable algorithms

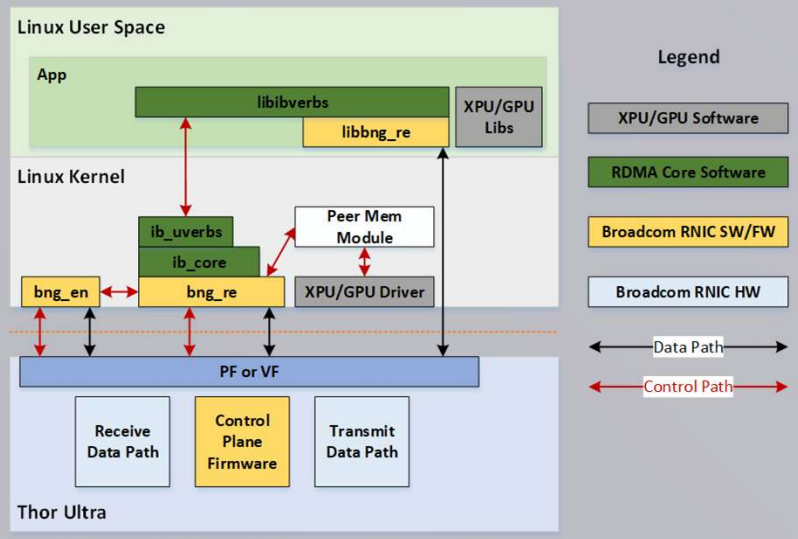
RoCE Software Components for Linux OS



- Broadcom Linux RoCE Components
  - RoCE User Library (libbng_re)
  - RoCE driver (bng_re)
  - NIC driver (bng_en)
- RoCE Driver (bng_re)
  - Netdev Event based interface with L2
  - Kernel mode verbs support for RC and UD
  - Kernel mode direct verbs for eRoCE
- RoCE User library (libbng_re)
  - User mode verbs support for RC and UD
  - Direct verbs APIs for eRoCE

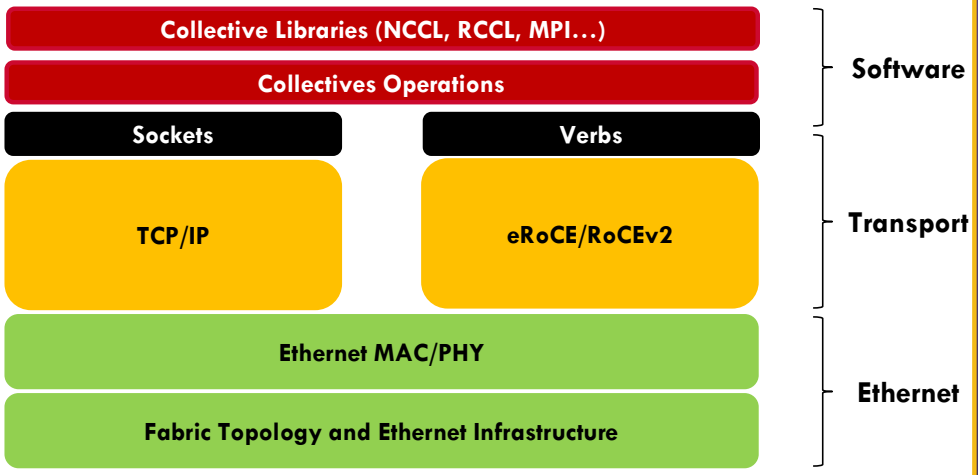
XPU/GPU Infrastructure Support

Peer Memory Direct



- Peer memory model
- Linux dma-buf based (upstream)
- XPU/GPU agnostic

Collectives

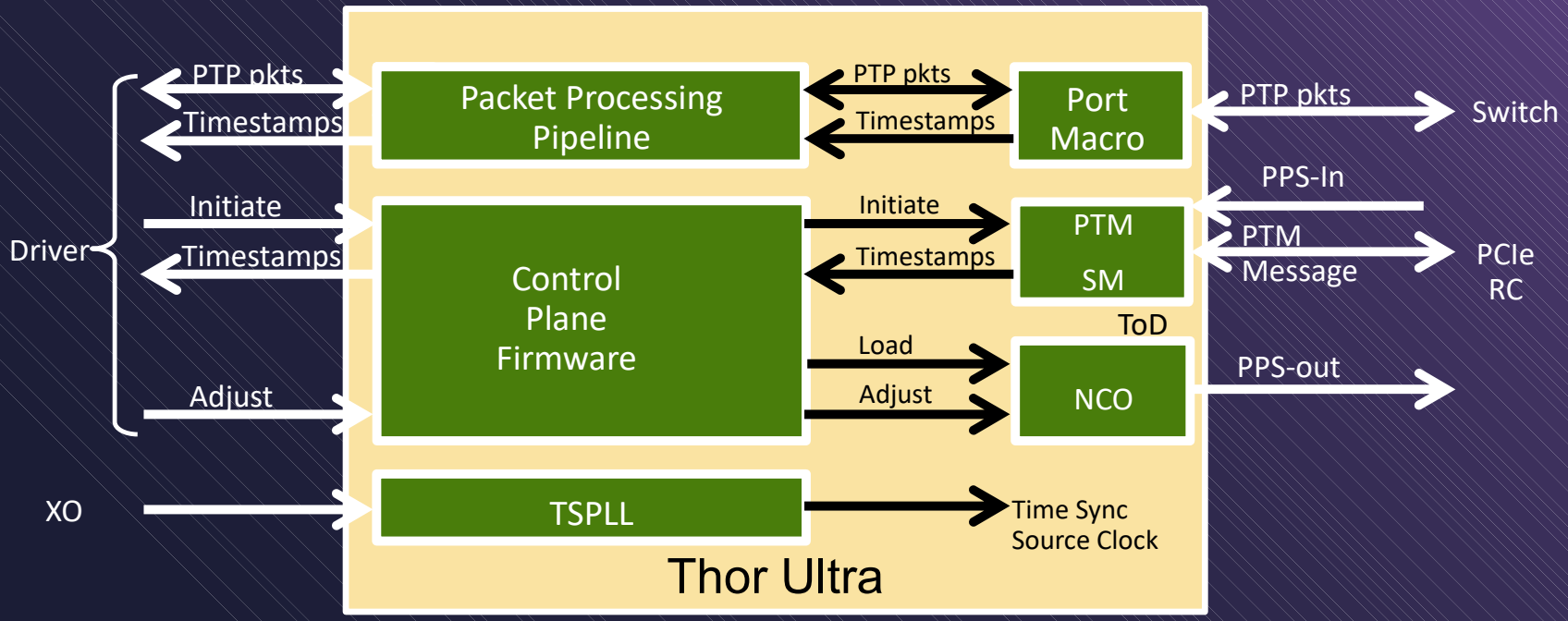


- Enabled by Verbs Provider (upstream) & xCCL Plug-ins
- Unmodified applications
- Multiple Collective Libraries supported
  - xCCL, NCCL, RCCL, MPI….

Thor Ultra Multi-tenancy Support

- QoS
  - Hierarchical scheduling – QP & CoS
  - Traffic shaping in the hardware
  - Rate limits – QP, PCIe function, port
  - Flow level meters with drop
- Virtualization
  - SR-IOV
  - Function level isolation
  - NIC resource virtualization
  - Address Translation Service (ATS)

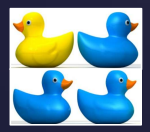
Thor Ultra Timing Support



- Precision Time Protocol (PTP) v2
- Precision Time Measurement (PTM)
- Packet timestamping
- Time & Freq PHC adjustment
- Time accuracy in ns
- External TCXO

Thor Ultra NIC Security

ROM Bootstrap Loader
FUSE-based Silicon Root-of Trust



Protected device-unique Identifier

Hardware Datapath Protection & Isolation

Authenticated firmware updates

SPDM based device attestation

Secure Firmware Loading and Validation

PSP based Encryption/Decryption

Thor Ultra NIC: Enabling Stability, Control and Visibility

Health Monitoring

Continuous health monitoring

Chip/Optics temp monitoring

Statistics and error reporting

Monitor NIC resource utilization

Diagnostics

Logging of adapter states

Crash and core dumps

Eye diagrams

Self diagnostics tests

NIC Control

Gain insights into NIC config

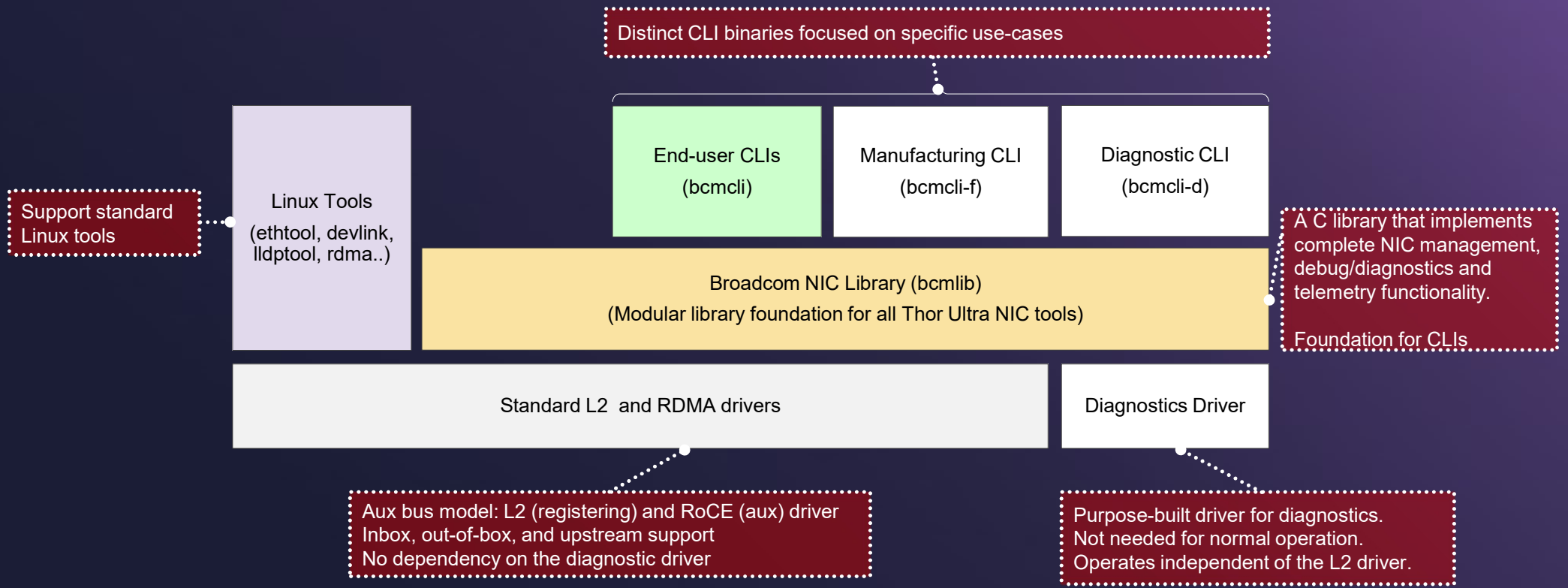
Link, Media, and Pre-boot settings

Authentication & attestation

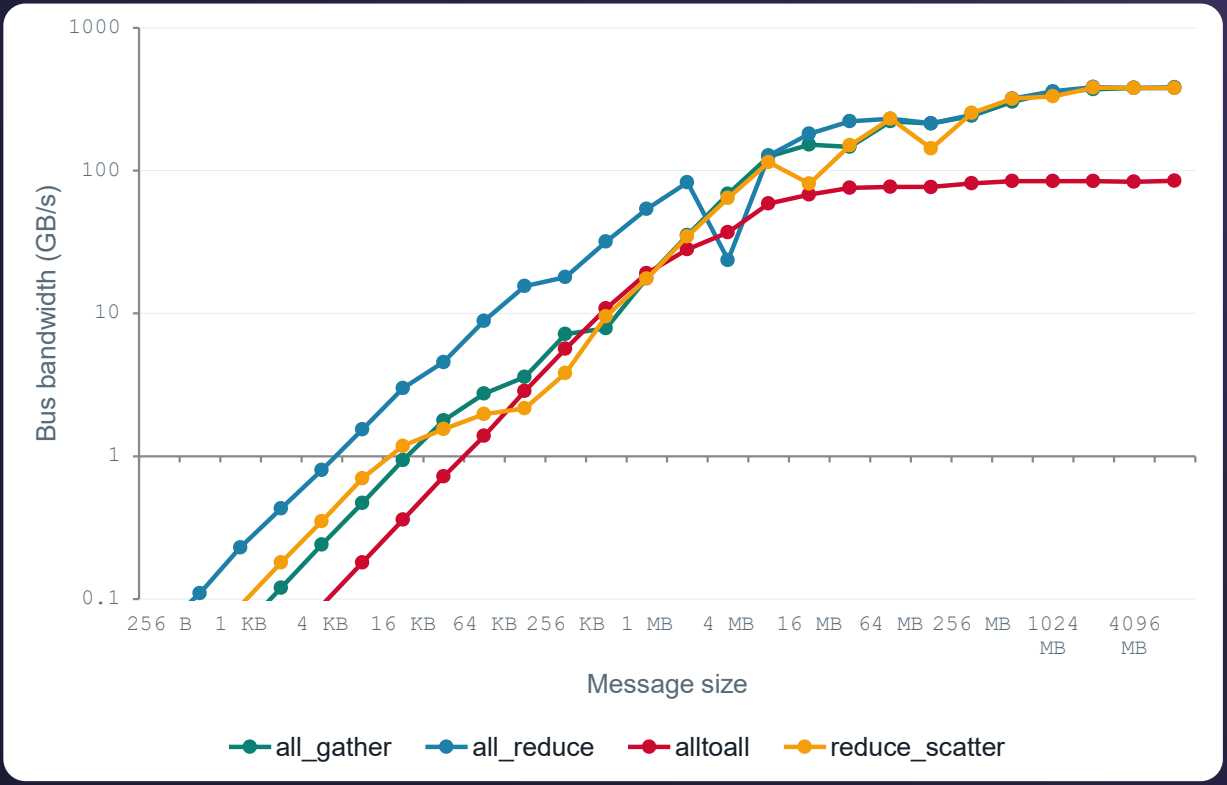
Firmware inventory & update

DMTF and OCP Standards Based Manageability

Thor Ultra Linux Tools



Collective Performance – Initial Data on a Gen5 GPU Platform



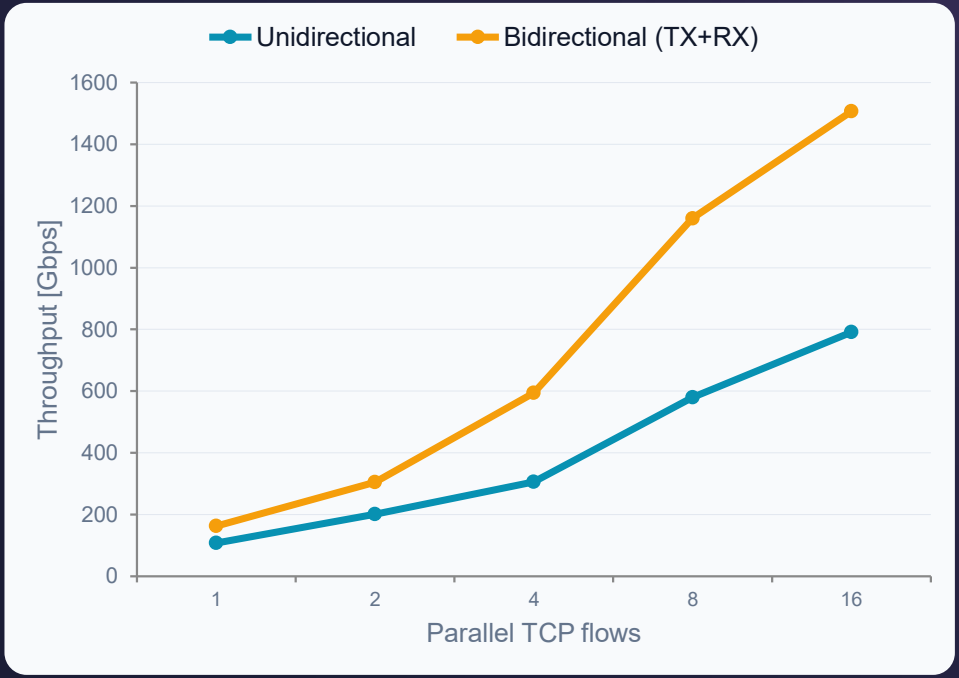
Collective	Peak GB/s	@ size	vs ceiling
all_gather	381.92	8GB	400 GB/s
all_reduce	383.93	2GB	400 GB/s
alltoall	84.62	8GB	93 GB/s
reduce_scatter	380.23	2GB	400 GB/s

- 2 nodes
- 8 Gen5 GPUs per node
- 16 ranks
- 16 x 400G connectivity to Switch

Near Line-rate

Bus bandwidth > 96% — for gather, reduce, reduce_scatter

Thor Ultra TCP Throughput



CPU / memory	**Gen6 CPU · 512 GB DDR5-8000**
OS / kernel	**Ubuntu 24.04.1 · 6.17**
NIC driver	**1.140.1.85**
TCP test	**iperf3 3.21+ (tuned buffers)**

Near line-rate TCP

Unidirectional iperf3 hits 791 Gbps at 16 flows — 98.9% of the 800G link

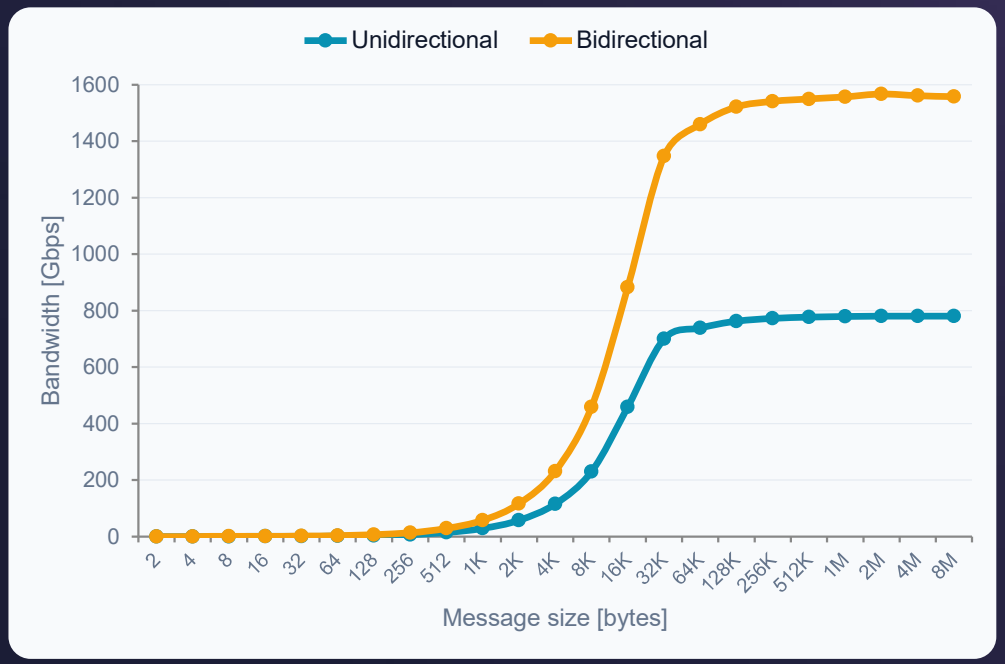
Full-duplex headroom

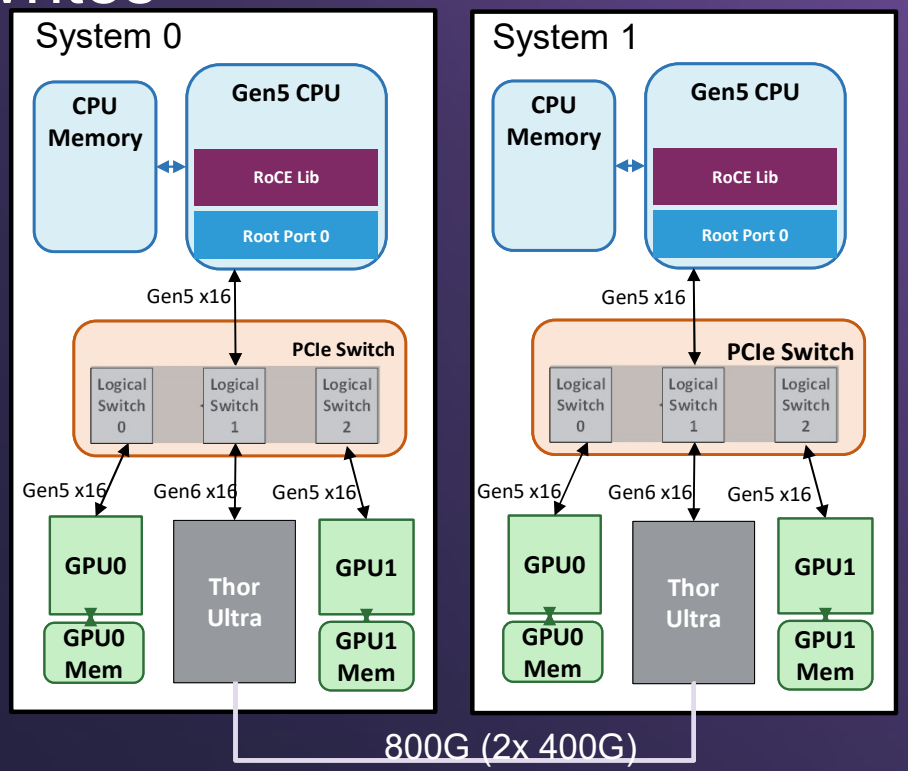
Bidirectional aggregates to ~1.51 Tbps; TX and RX each sustain ~750 Gbps

Scales with flows

Throughput climbs near-linearly to 8 flows, then tapers as the link saturates.

Thor Ultra Throughput for RDMA Writes





Line-rate RDMA

Unidirectional saturates at 781 Gbps — 97.6% of 800G — for large messages.

Full-duplex ≈ 2X

Bidirectional reaches 1558 Gbps (97.9% of 1.6 Tb/s), roughly double unidirectional.

Fast ramp

Unidirectional reaches ~88% of line rate by 32 KB and within ~3% of peak by 128 KB.

Summary

- Thor Ultra → 800G NIC for AI/HPC markets
- Thor Ultra data path → designed to support enhanced RoCE (MRC++)
- QoS and Virtualization features → enable XPU/CPU as a service
- Timing features → enable synchronization of AI/HPC systems
- Industry leading security and manageability → easy to use and trusted


BROADCOM®