High Performance Ethernet Solutions for HPC/AI Clusters

Mohan Kalkunte
Vice President, Architecture, Core Switching Group

Aug 2025

Broadcom AI/ML and HPC Solutions - Overview

HPC/AI Fabric Solutions

Scale Up

Tomahawk Ultra

Tomahawk 6-200G

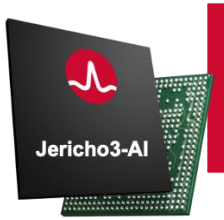
OPEN Compute Project

Scale Up Ethernet (SUE)

Scale Out

Datacenter Scale Out

Region Scale Out



Switch Scheduled

Endpoint Scheduled

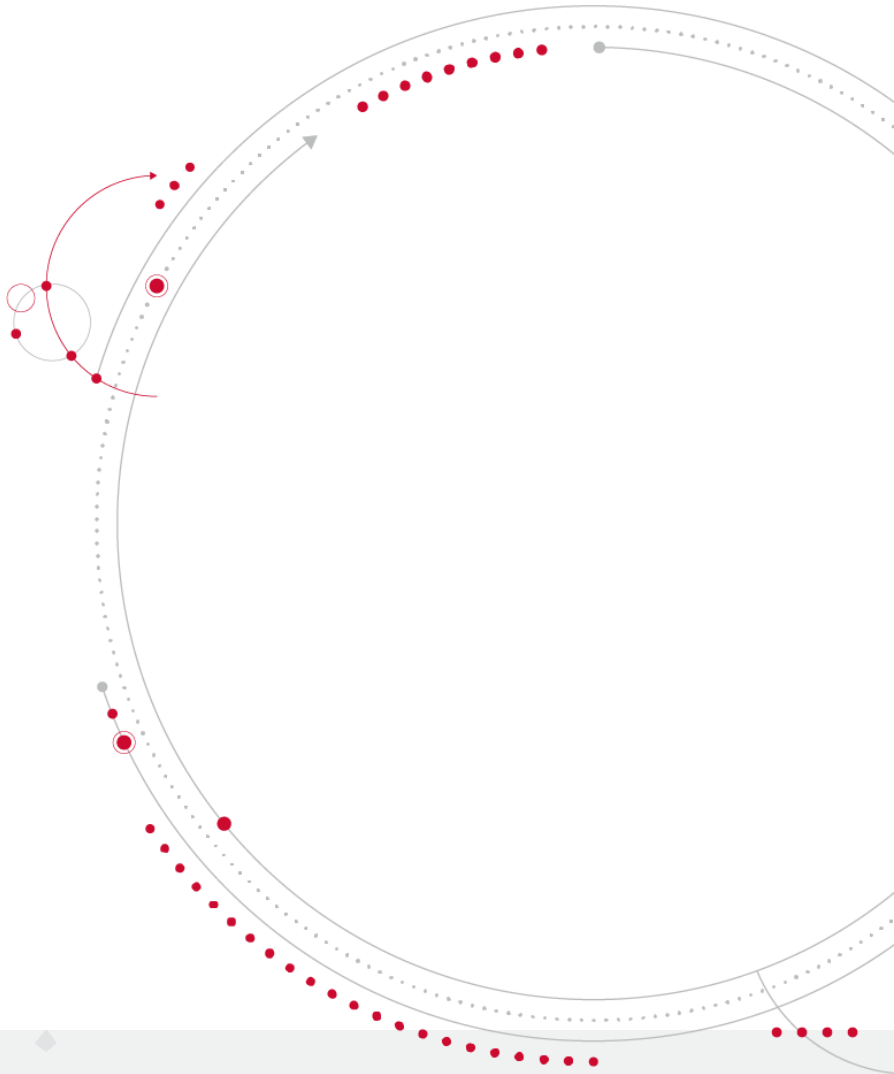
Tomahawk 5

Tomahawk 6-100G

Tomahawk 6-200G

BROADCOM

Scale Up

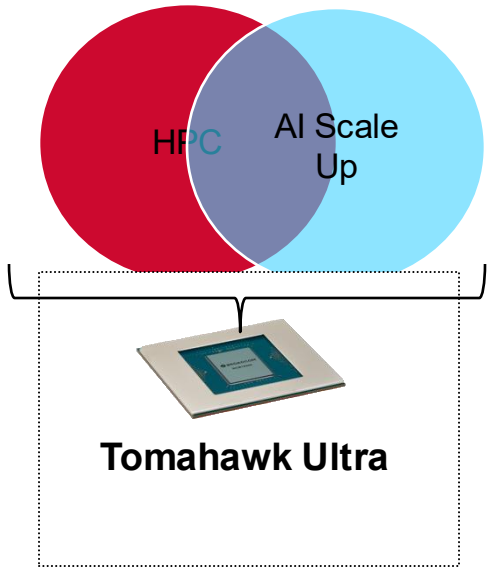


Scale-Up Fabric Characteristics

- Single Tier
  - Simpler and lower cost v/s Multi-tier. Avoids retimers and keeps pod in 1-2 racks.
  - Lower latency.
  - Easier congestion control.
  - Easier to support lossless.
  - Large cluster scale supported today (256+ XPUs) compared to proprietary approaches.

- Multi-Plane
  - Maximize use of switch radix.
  - Add planes per XPU bandwidth requirements.
  - Each plane is independent of others.
  - XPU load balances traffic to different planes.

Tomahawk Ultra Highlights – Covering HPC and AI Scale Up



Performance

250ns latency

51.2T@64B

77 BPPS

Efficiency

In-Network Collectives

Optimized Ethernet Headers

Reliability/Lossless

LLR

CBFC

Programmable Visibility/Telemetry

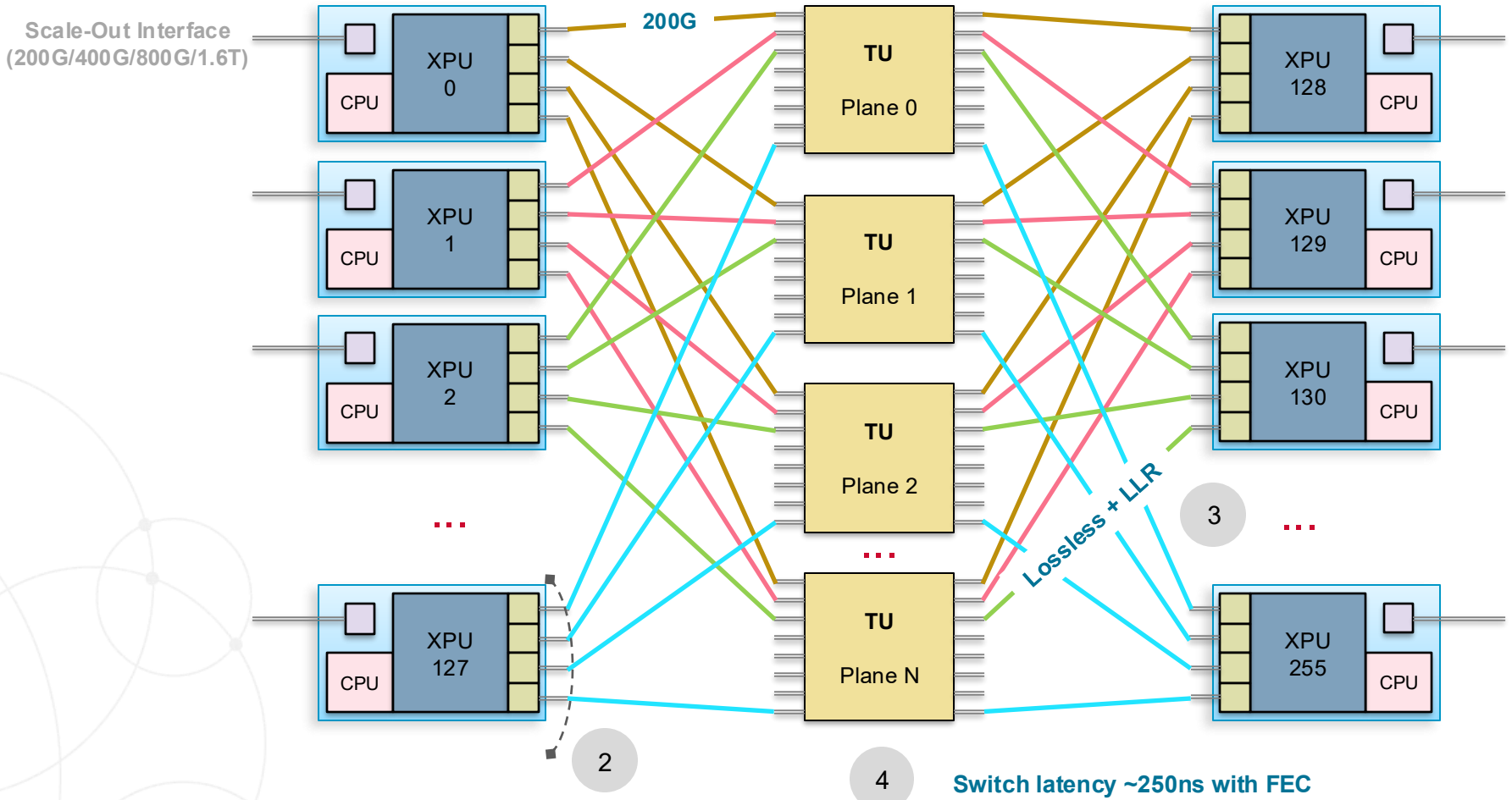
Topology Aware Routing

Congestion Control

BROADCOM

Example 256 XPU Pod with Tomahawk Ultra

1 **Multi-Plane fabric maximizes switch radix. 256 x 200G = 51.2T**

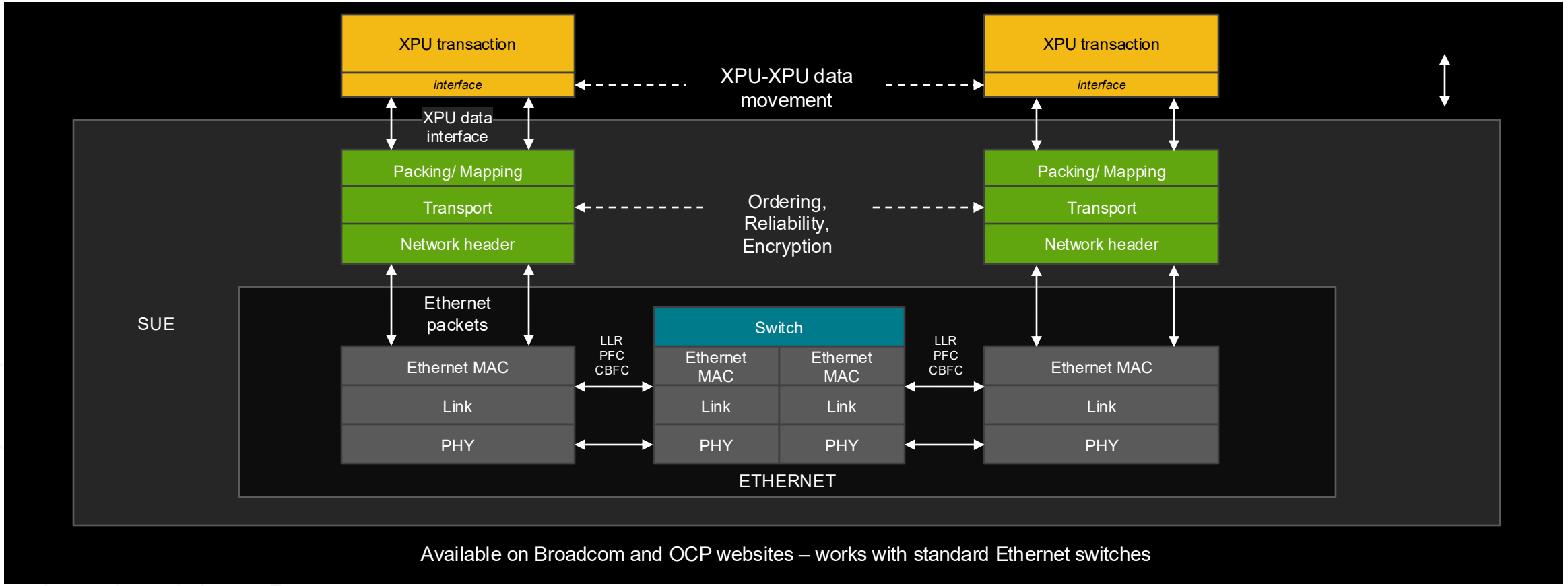


Memory transactions load balanced over N planes.
Transport adapted for small size, low latency transfers.

Switch latency ~250ns with FEC
E2E latency < 400ns

BROADCOM

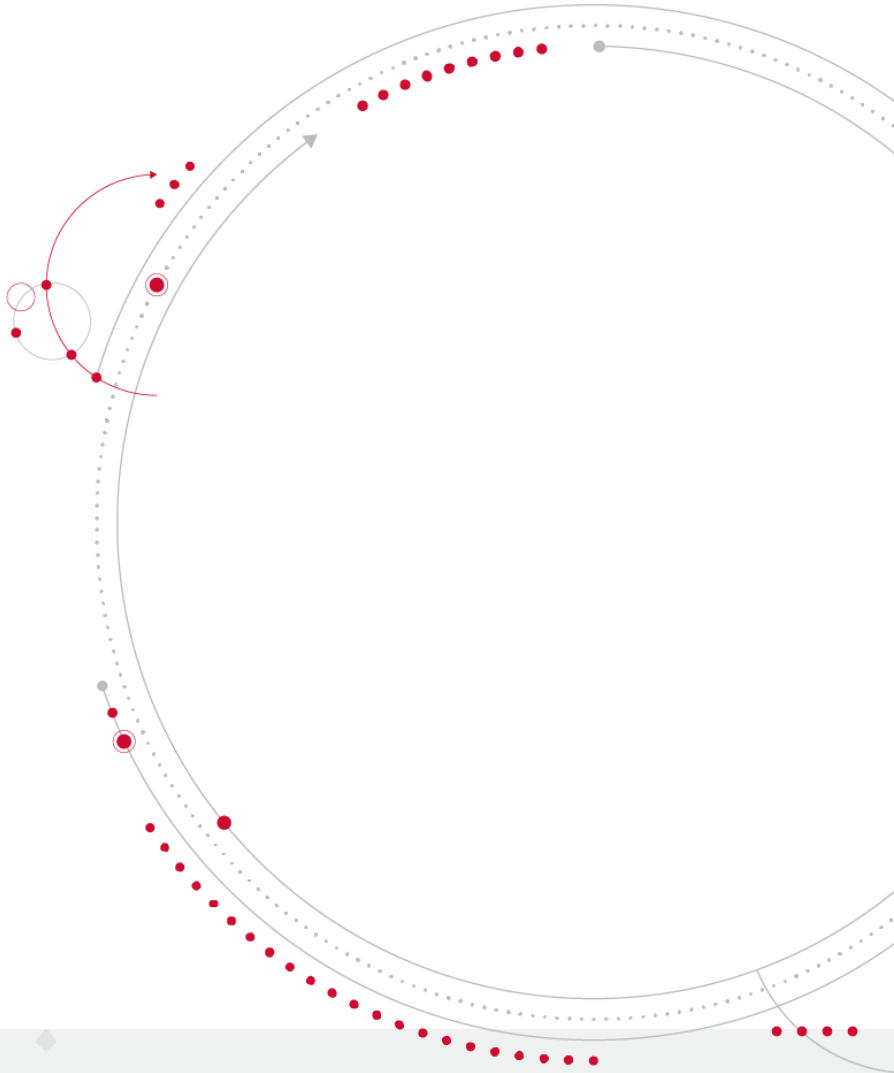
Scale-Up Ethernet (SUE)



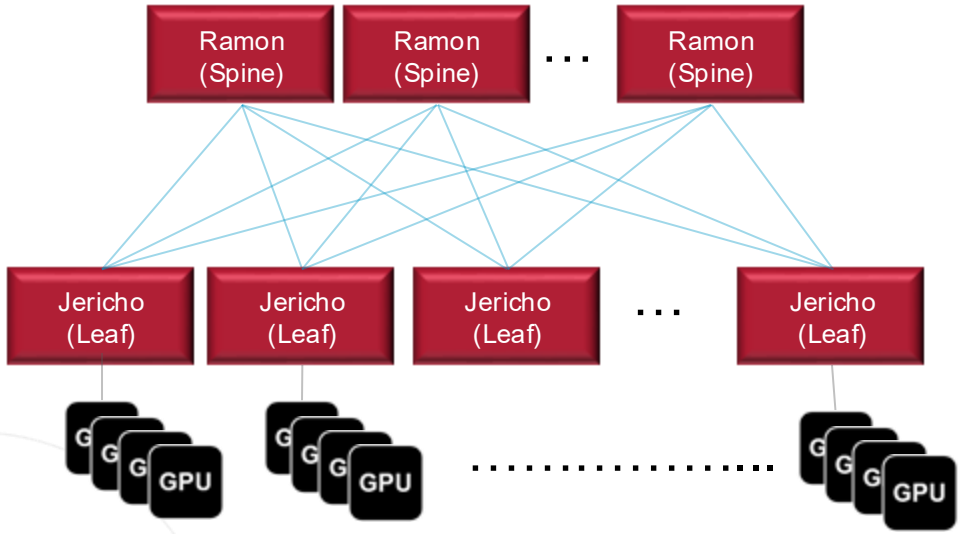
Lightweight Ethernet Interfaces for XPU Scale-Up

BROADCOM

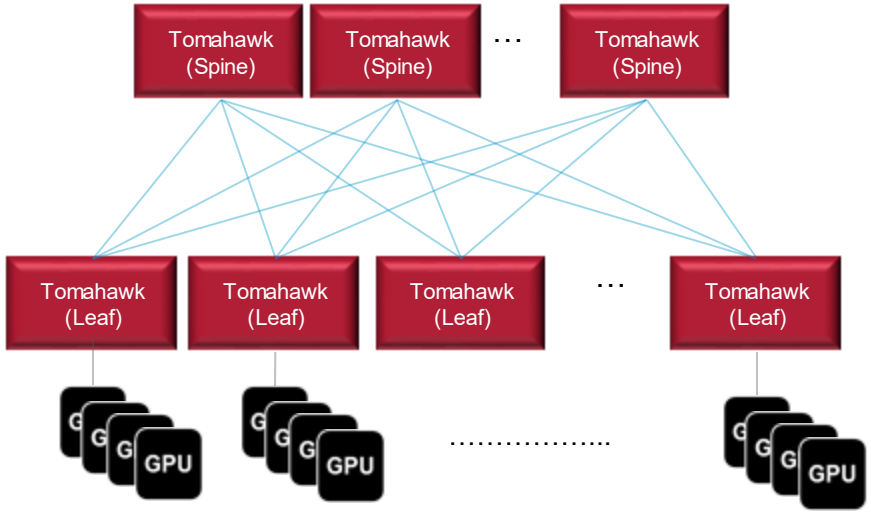
XGS and NIC devices



Broadcom's Scheduled Fabric Solutions



Switch Scheduled



Endpoint Scheduled

Endpoint can be:

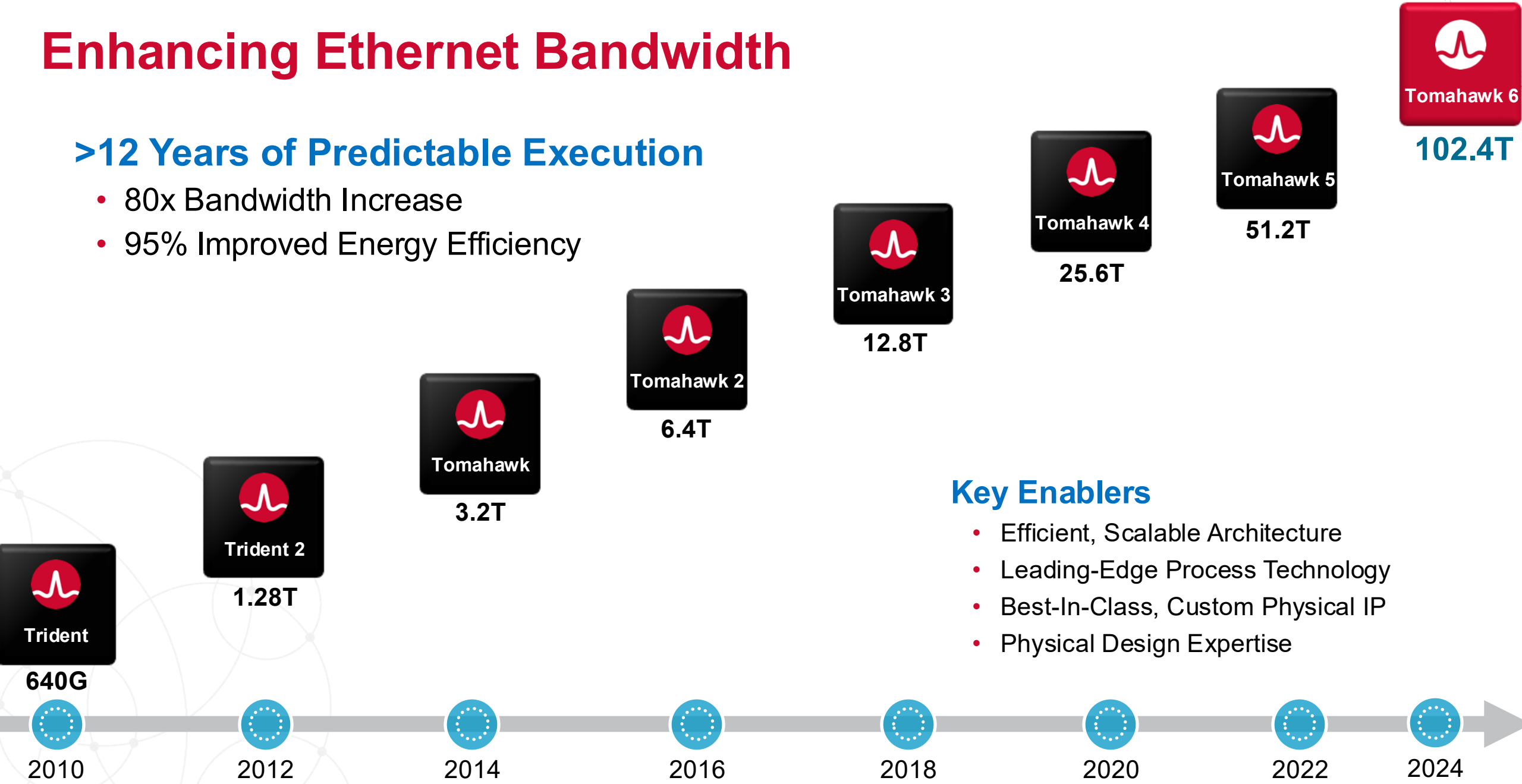
- Broadcom NIC
- Merchant silicon NIC
- Customer NIC
- GPU native Ethernet interface

BROADCOM

Enhancing Ethernet Bandwidth

>12 Years of Predictable Execution

- 80x Bandwidth Increase
- 95% Improved Energy Efficiency



Key Enablers

- Efficient, Scalable Architecture
- Leading-Edge Process Technology
- Best-In-Class, Custom Physical IP
- Physical Design Expertise

BROADCOM

Tomahawk 5: 800GE and AI Workload Acceleration

51.2 Tbps Ethernet Switching Bandwidth

Double the throughput of any other deployable silicon

<1W per 100Gbps

Monolithic 5nm implementation

Low Power & Cost Physical Connectivity

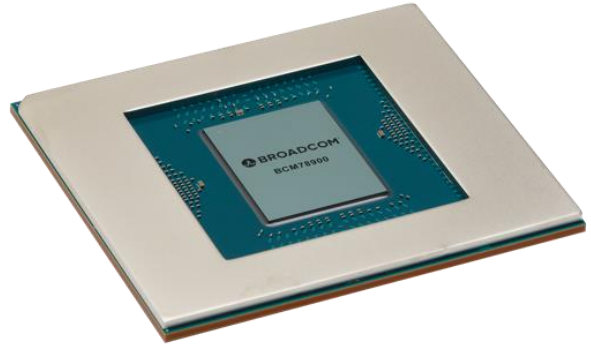
Most flexible, longest reach 100G PAM4 SerDes

Accelerates AI Workloads

Cognitive Routing, advanced telemetry

Resource Virtualization

Improved security, efficient use of massively shared infrastructure



64 x 800GbE
128 x 400GbE
256 x 200GbE
512 x 100GbE

64x800G in 2RU

BROADCOM

Tomahawk 5 – Broadcom Cognitive Routing

- **Suite of capabilities that adds global intelligence to routing decisions**
- **Improves routing for all traffic types**
- **Largest impact for AI flows**
- **Support for all common topologies**
  Clos, Torus, Dragonfly, Dragonfly+, etc.

Improved Network Utilization ⇒ Lowest Tail Latency

Cognitive Routing

BROADCOM

Tomahawk 5 Cognitive Routing

Global Load Balancing

Egress link selection based on global path congestion

Reactive Path Rebalancing

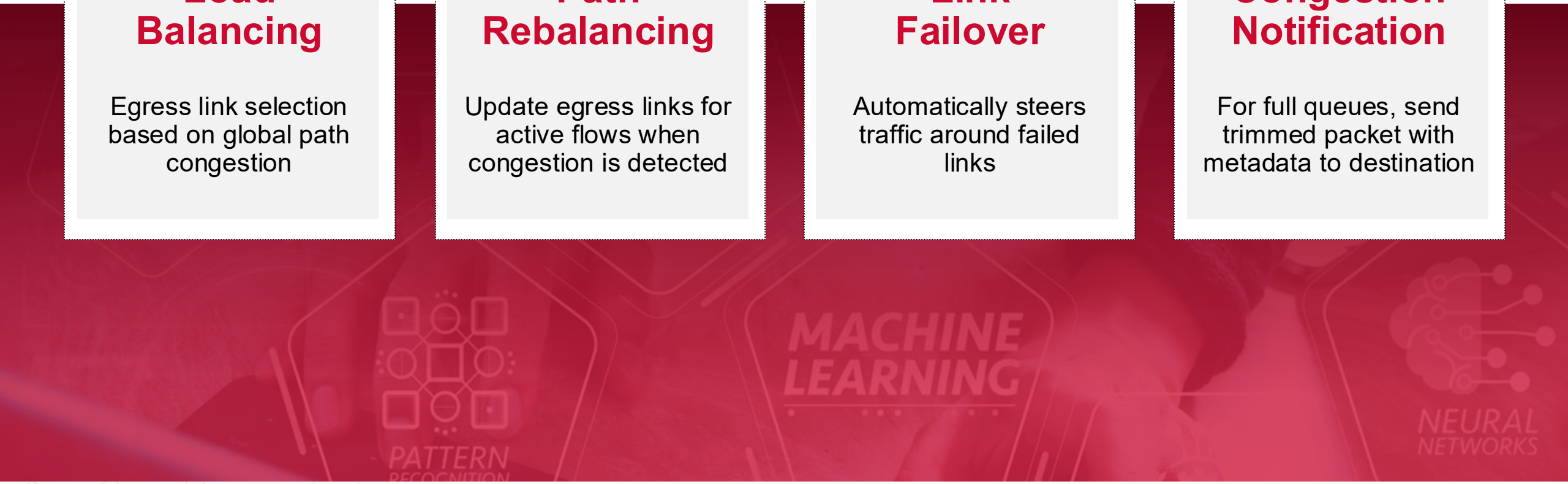
Update egress links for active flows when congestion is detected

Fast Link Failover

Automatically steers traffic around failed links

Drop Congestion Notification

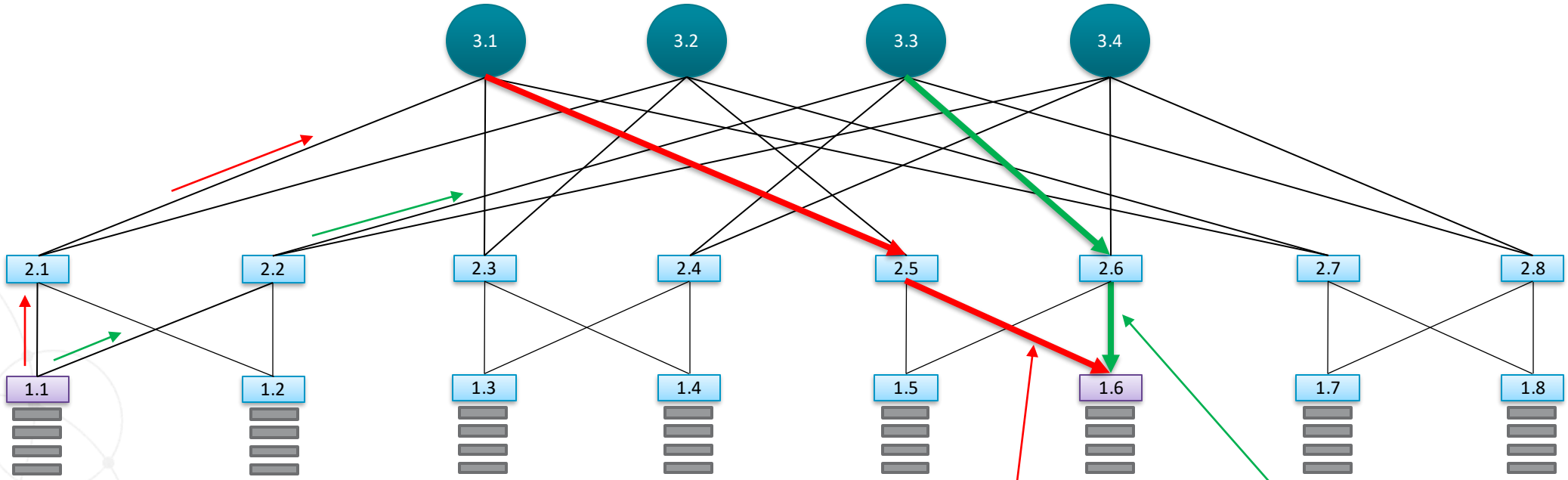
For full queues, send trimmed packet with metadata to destination



BROADCOM

Tomahawk 5 Global Load Balancing (GLB)

- Load balancing based on local link quality is not always optimal
- Global view of congestion is needed for optimal network load balancing



DLB = local load
Does not take into account downstream congested links

GLB = global load
Avoids downstream congested links

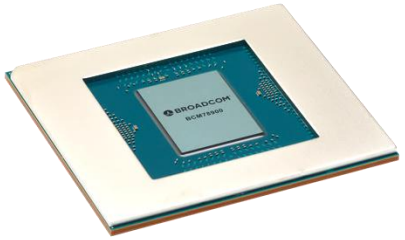
Congested links

Uncongested links

BROADCOM

From Tomahawk 5 → Tomahawk 6

Tomahawk 5



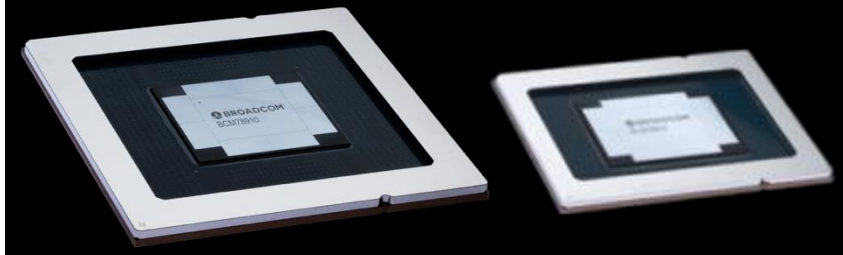
Widely deployed in hyperscale AI scale-out networks

Tomahawk 5 has proven itself in the largest GPU clusters

The only monolithic 51.2 Tbps switch on the market

Predictable performance, fixed latency, lowest power

Tomahawk 6



Tomahawk 6 steps it up in every dimension

- Bandwidth
- SerDes speed and density
- Load balancing
- Telemetry

BROADCOM

Tomahawk 6: Built for AI Scale

World's First 102.4 Tbps Switch Chip

- Double the bandwidth of any other Ethernet switch
- Built to power clusters with 1M+ XPUs

Performance & Power Efficiency

- Cognitive Routing 2.0, Deep Insight
- Advanced 3nm technology

Versatility

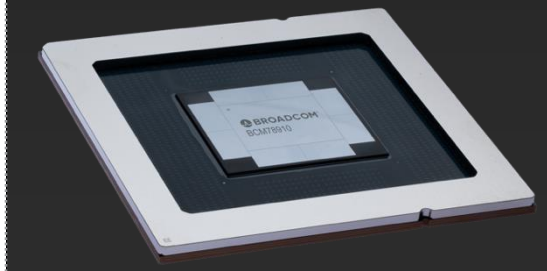
- Scale-up & scale-out, training & inference
- Works with any endpoint, including XPU scale-up interfaces

Industry-Leading SerDes and CPO

- Options for 512x200G PAM4, 1024x100G PAM4, CPO
- Ground-breaking SerDes and optics density

NOW SHIPPING

Tomahawk6-200G

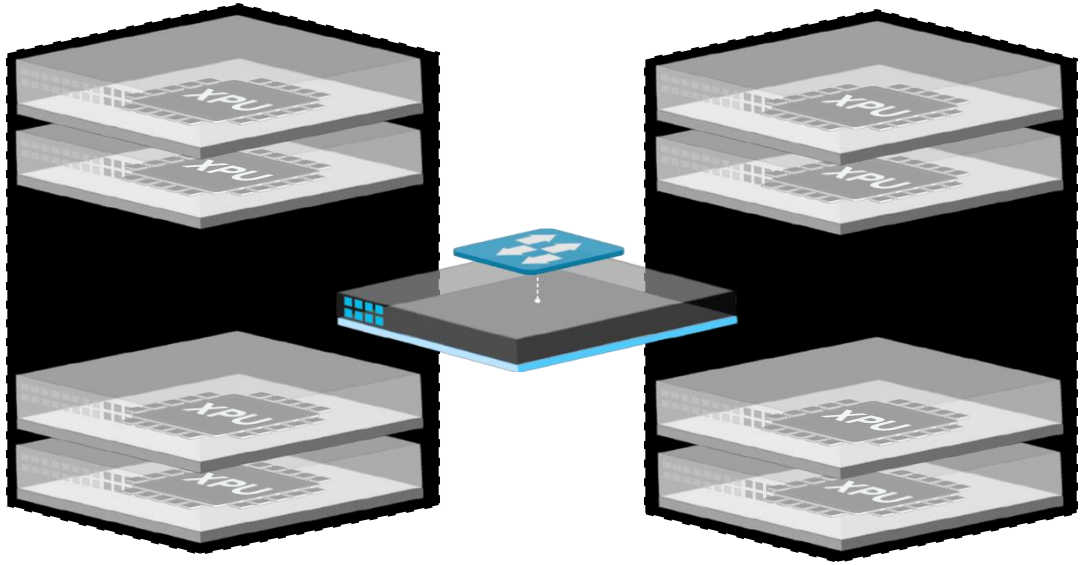


Tomahawk6-100G

Support for 1.6TbE Ports
Ultra Ethernet Compliant

BROADCOM

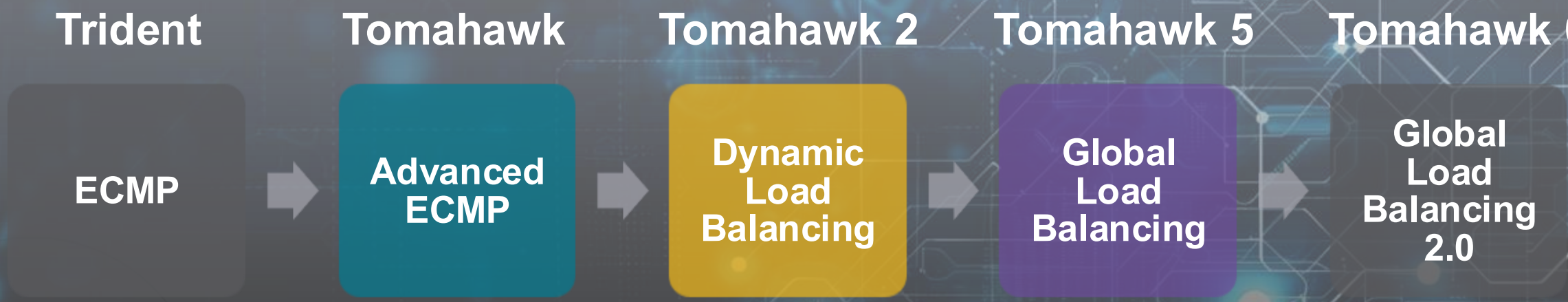
Tomahawk 6 ➔ 512 XPUs in a Scale-Up Cluster



512 XPUs connected in a single hop with 200G PAM4

BROADCOM

Broadcom Innovations in Load Balancing



GLB is one of the core features in Broadcom's Cognitive Routing

BROADCOM

Broadcom Deep Insight ➔ Network & Operational Efficiency

Rich Telemetry

- Fast Congestion Notification Packets
- Back-to-Sender
- CSIG (Congestion Signaling)
- PFC-aware congestion signaling

Real-Time Diagnostics

- Physical link quality monitoring
- Detailed chip pipeline visibility

BROADCOM

3-Pronged Approach to Reducing AI Interconnect Cost and Power

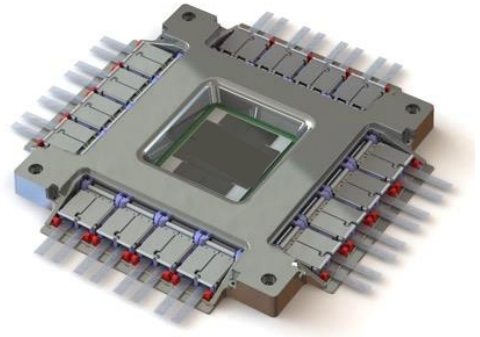
Extended Reach DAC and Cabled Backplanes

- 45+dB @ 200G
- Passive Scale-Up Connectivity

Pluggable Optics

- Support for DSP-Based and LPO/LRO Modules

Co-Packaged Optics



- Lowest Power/Cost Optics
- Fewer Link Flaps

Unmatched Choice and Performance → Lowest TCO & Fewest Link Flaps

BROADCOM

Thor 2 – Feature Highlights

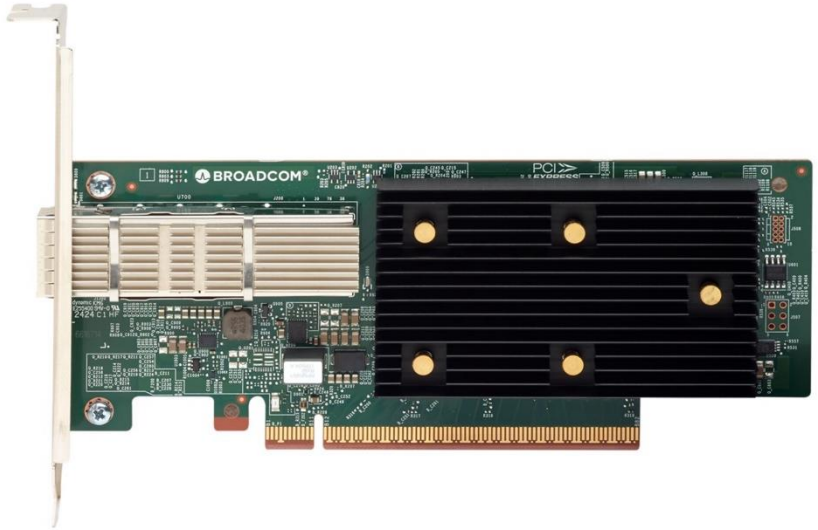
Focus Performance

400G, 230 Mpps

8M+ Truflow offloads at line rate

PCIe Gen5x16

100G PAM4 SerDes



Lowest Power 400G NIC

14-18W Typical Adapter Power, 250 LFM

5nm process technology

Leading SerDes

5-meter DAC

3-Watt 400G Linear Pluggable Optics

Improved AI/ML Scale

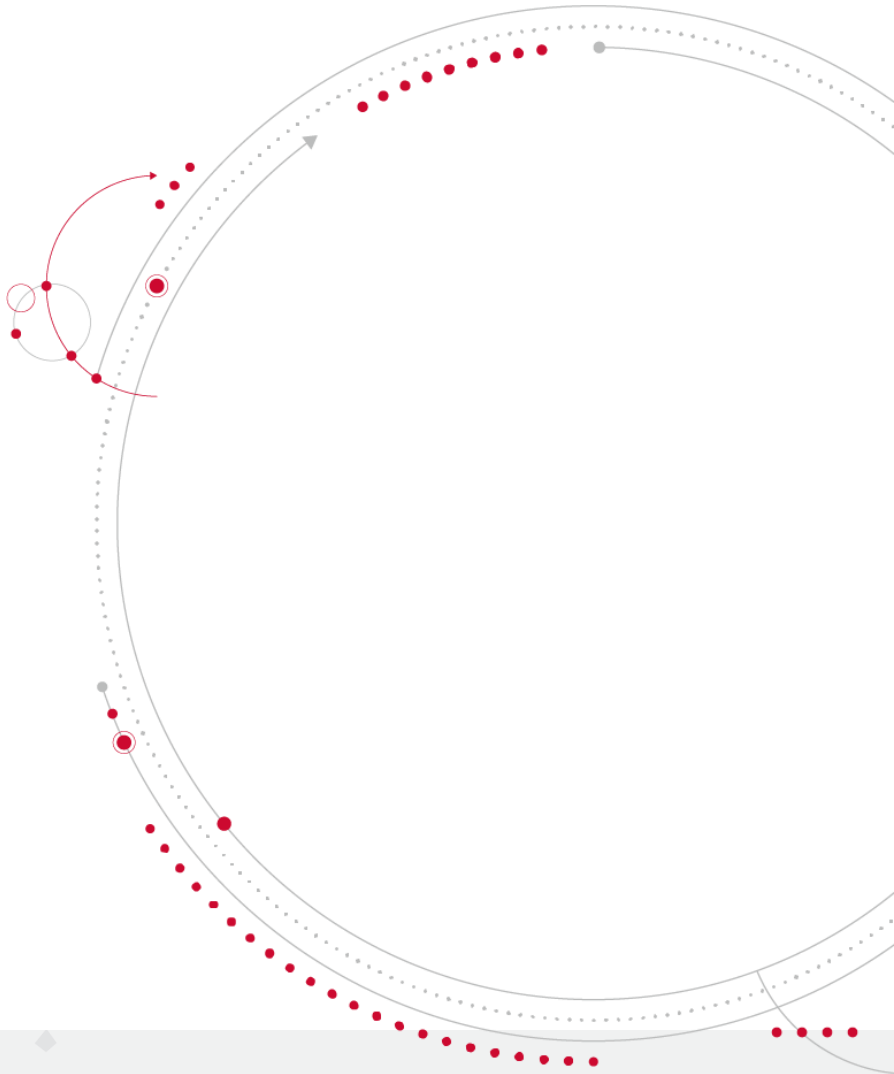
Enhanced DCQCN congestion control

Granular rate control, enhanced ECMP

Hardened RoCE

BROADCOM

DNX devices

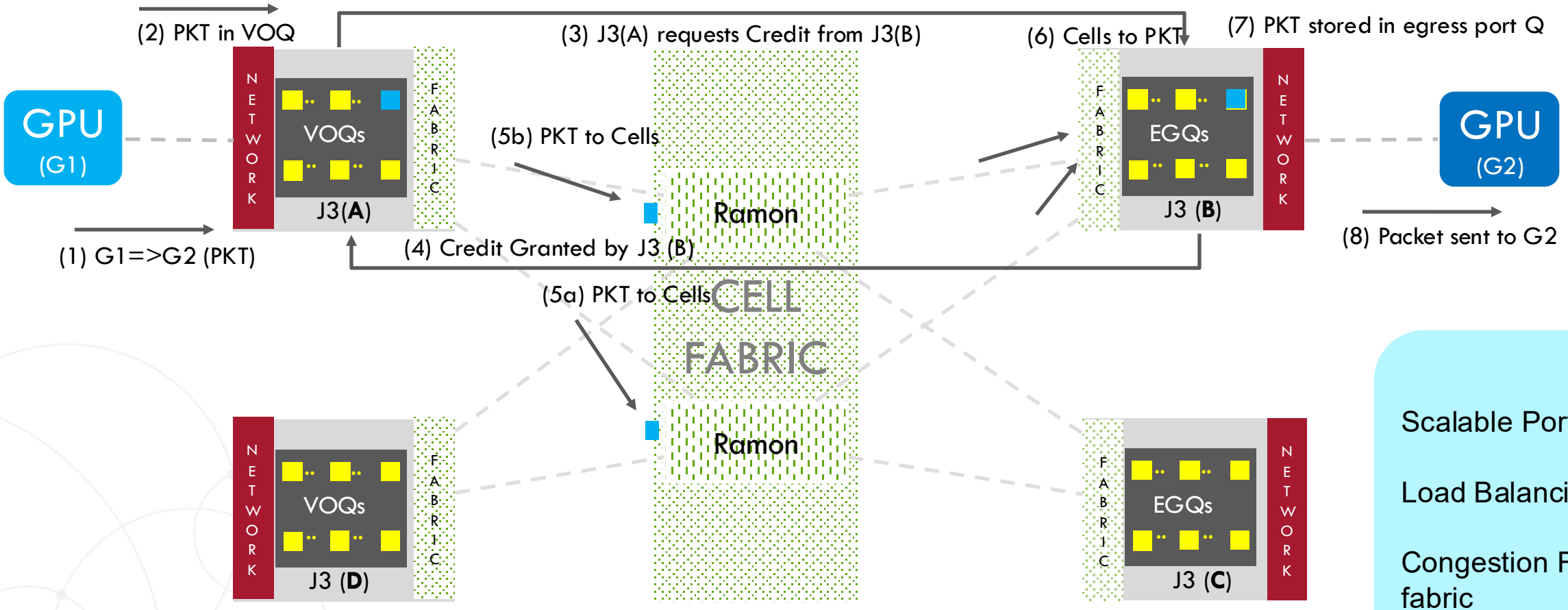


Distributed Switch Fabric

- Leaf Switch – Jericho Packet processor
  - Jericho3 (51.2T) → Jericho4 (102.4T)
  - Standard Ethernet I/O
  - Leaf: switching, forwarding, queuing, scheduling
- Receiver-based scheduling
- Fabric Switch
  - Ramon3 (51.2T) → Ramon4 (102.4T)
  - Cell based Switch
  - Low power

BROADCOM

Distributed Switch Data plane

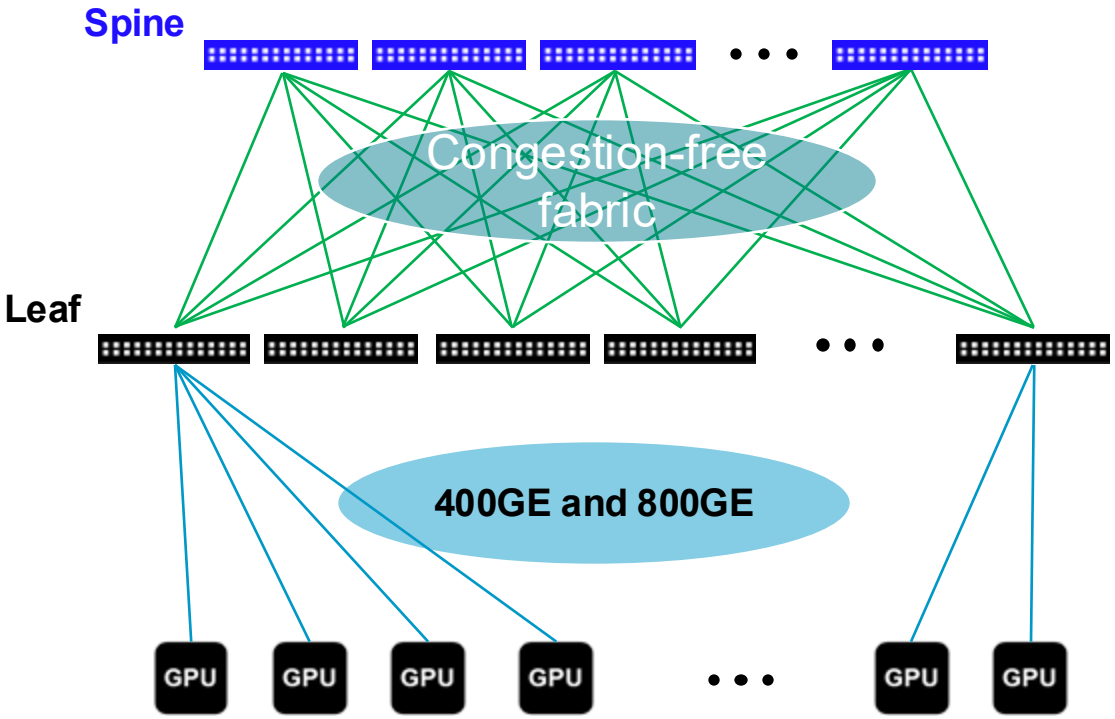


- Scalable Port Rate
- Load Balancing
- Congestion Free fabric
- Self-healing

Jericho3-AI Ethernet Fabric

- Switch scheduled fabric
  - Standard Ethernet I/O
  - Leaf: switching, forwarding, queuing, scheduling
  - Spine: forwarding at low power
  - Receiver based scheduling

- Leaf deployment options
  - ToR/MoR - in the GPU racks
  - In the network rack, with spines

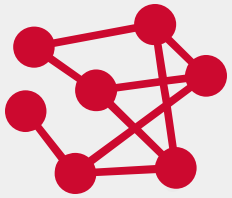


BROADCOM

Jericho3-AI Fabric Innovations for AI Workloads

Perfect Load Balancing

- Equal spraying over all links of the fabric
- Consistent high performance at all network loads

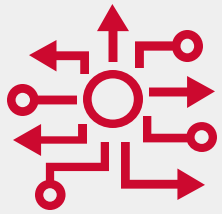


Congestion-Free Operation

- End-to-end scheduled fabric
- No collisions, no jitter

Zero Impact Failover (ZIF)

- Sub-10ns auto-path convergence
- No impact to job completion



Ultra-High Radix

- Massive, flat networks
- 32,000 ports at 800GE single hop/domain
- Scale out to any size with one more layer

Highest Performance Under Load and Scale for Multi-tenant Cloud Networks

BROADCOM

Jericho4: Massive Scale Across Data Centers

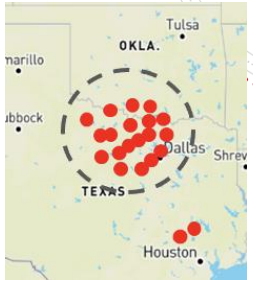
- 36,000 Ports of 3.2T In A System
- 3.2T HyperPort … Fastest Single Port Bandwidth
- 100km+ Lossless Data Center Interconnect
- Integrated Line Rate Encryption

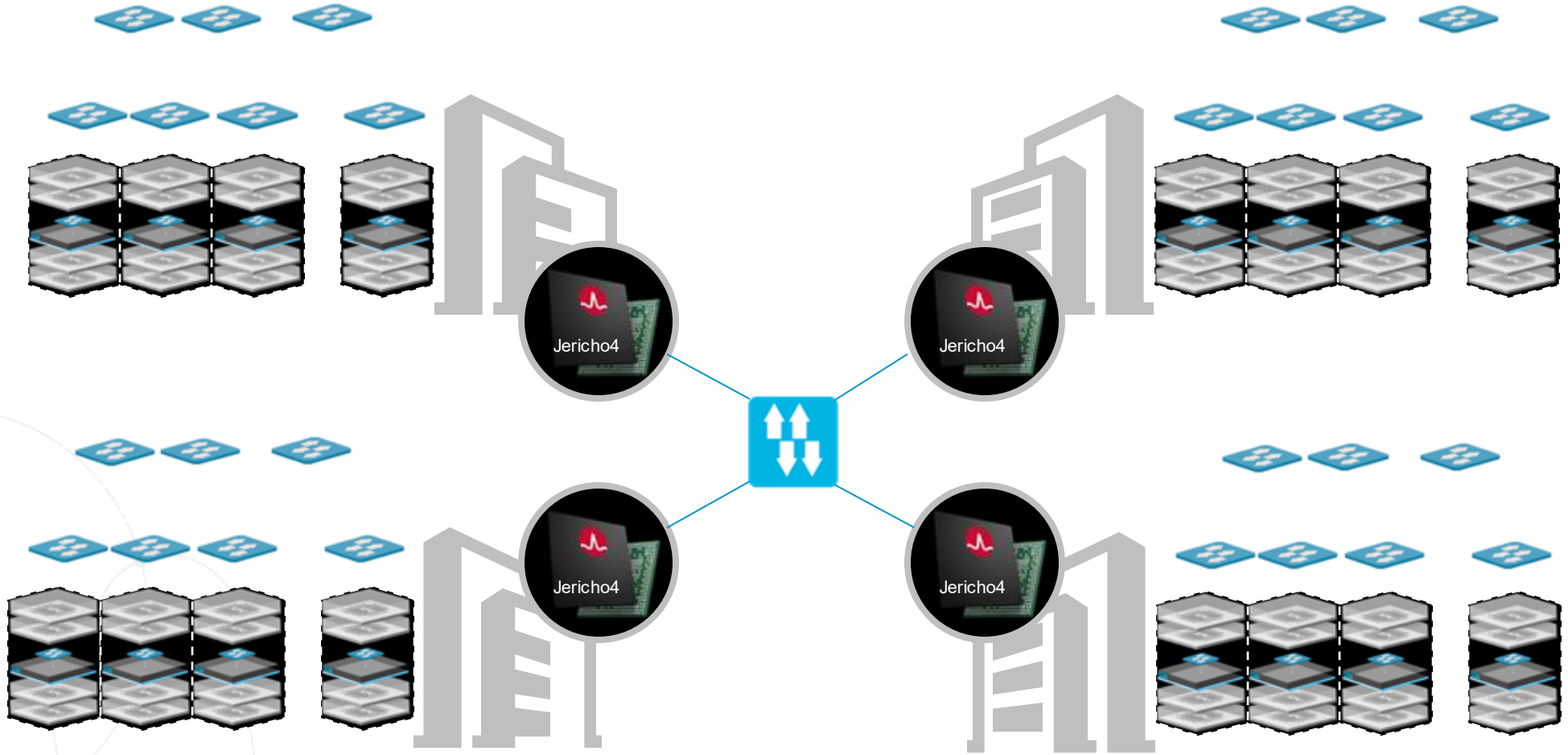
NOW SHIPPING

Jericho4

BROADCOM

Jericho4: 1M+ Accelerators Across Data Centers

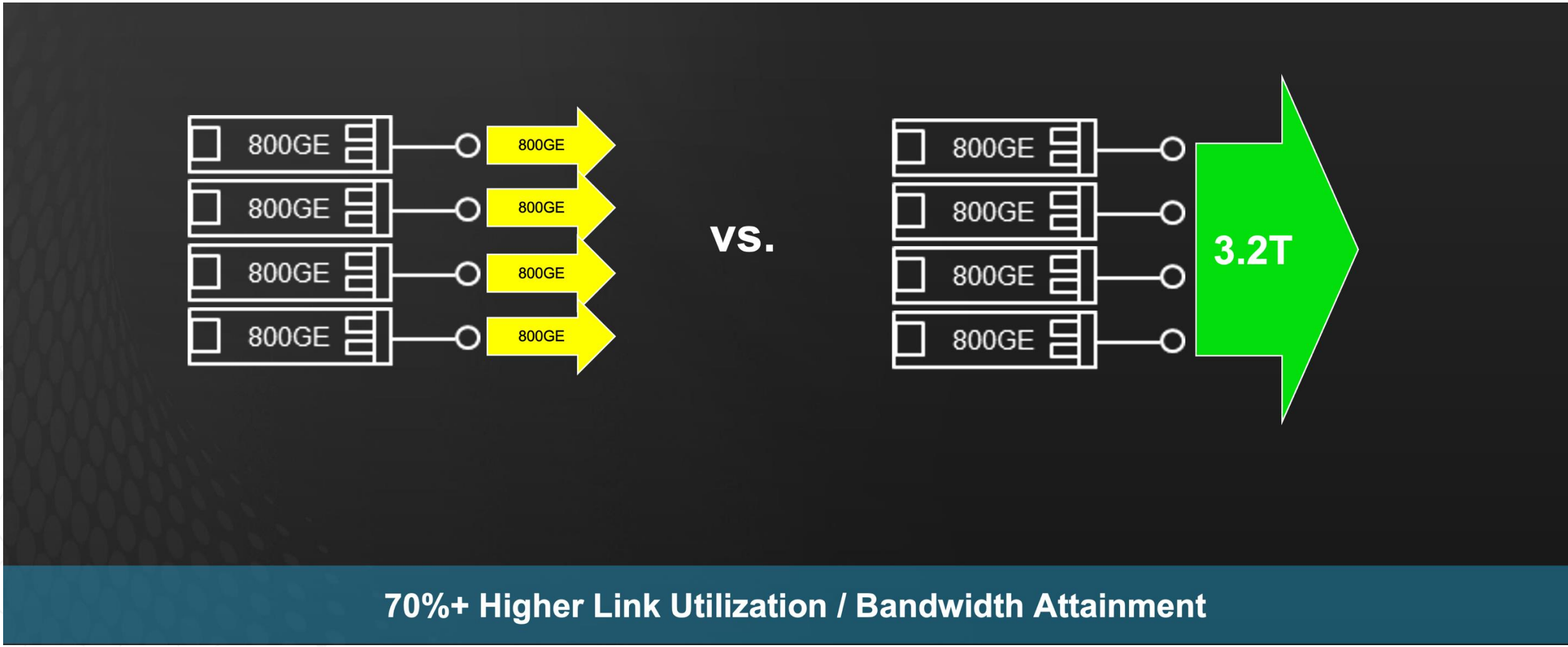




Jericho4 Delivers Scale-Out Network Interconnect

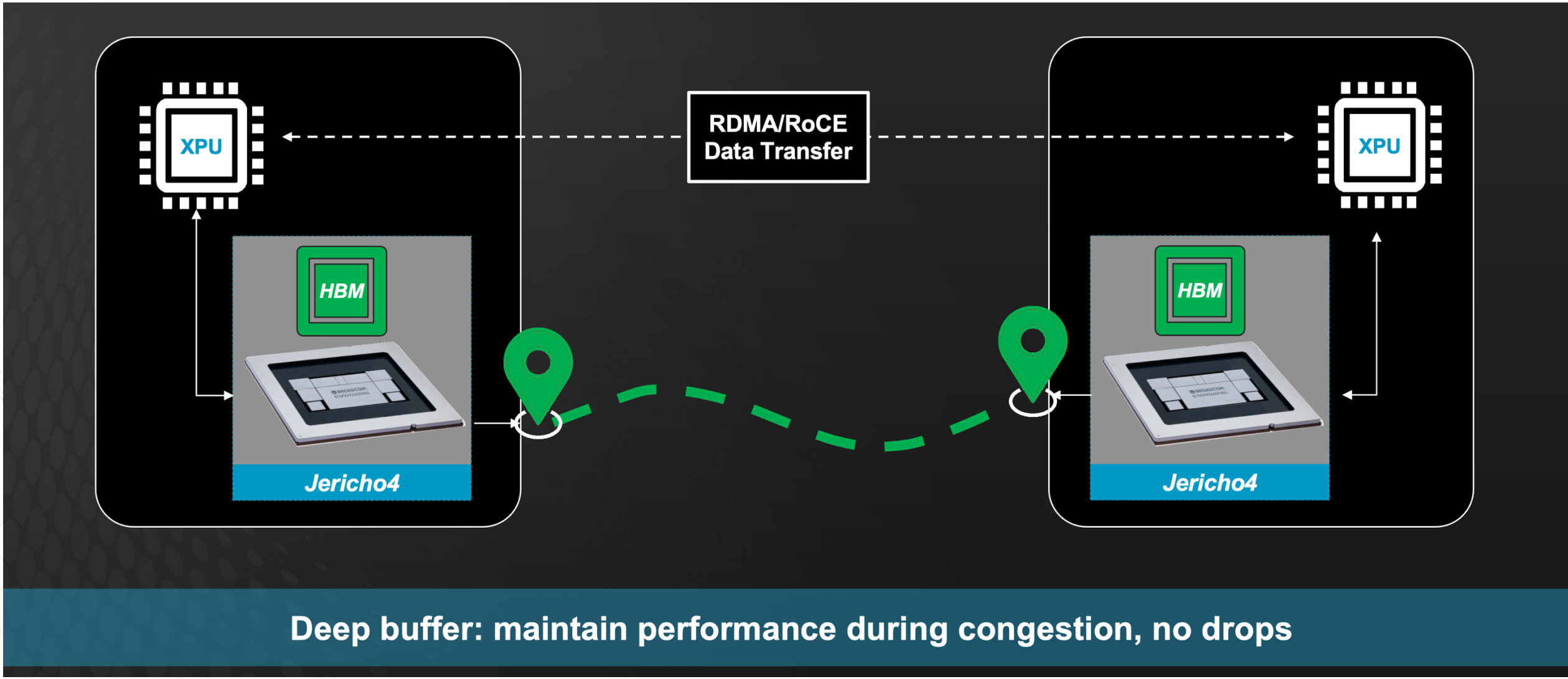
BROADCOM

3.2T HyperPorts Reduce Data Transfer Time



BROADCOM

Largest RoCE Deployment: 100Km+



Deep buffer: maintain performance during congestion, no drops

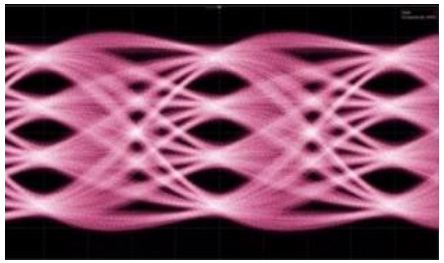
BROADCOM

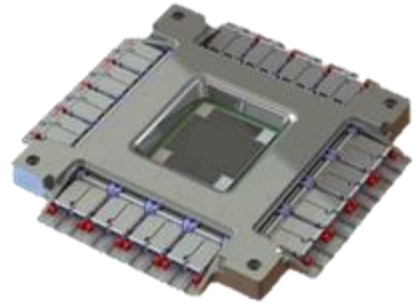
Broadcom Offers Complete Coverage of HPC and AI

HPC	AI Scale-Up	AI Scale-Out	Region Scale-Out
	 **Tomahawk Ultra** 51.2 Tbps	 **Tomahawk 6** 102.4 Tbps	 **Jericho4** 51.2 Tbps

BROADCOM

Broadcom Full-Stack Ethernet AI Innovation





Switches	Endpoints	Physical Layer	Optics
• Tomahawk • Jericho	• Thor NICs • NIC Chiplets • Scale-Up Ethernet Specifications	• Agera Retimers • Sian Optical DSPs	• CPO • VCSELs • Single-Mode Optics

BROADCOM

BROADCOM

BROADCOM®

Thank you