Tahan: An Open- Ethernet-Based Scale-Up Switch for Meta's MTIA Accelerators

Vimal Vasudevan *Technical Program Manager, Meta*

Ashutosh Grewal *Software Engineer, Meta*

Agenda

01 Industry Problem

02 Inflection Point

03 Solution - Introducing Tahan

04 Bandwidth and Latency

05 Scale Up Features

06 Next Gen Scale Up Fabric

07 Open Standards for Scale Up

08 Call to action

Industry Problem

- Proprietary and vendor-locked
- Closed, Bundled ecosystems from a single vendor
- Limited flexibility
- Constrained roadmaps
- Little leverage over cost or supply.

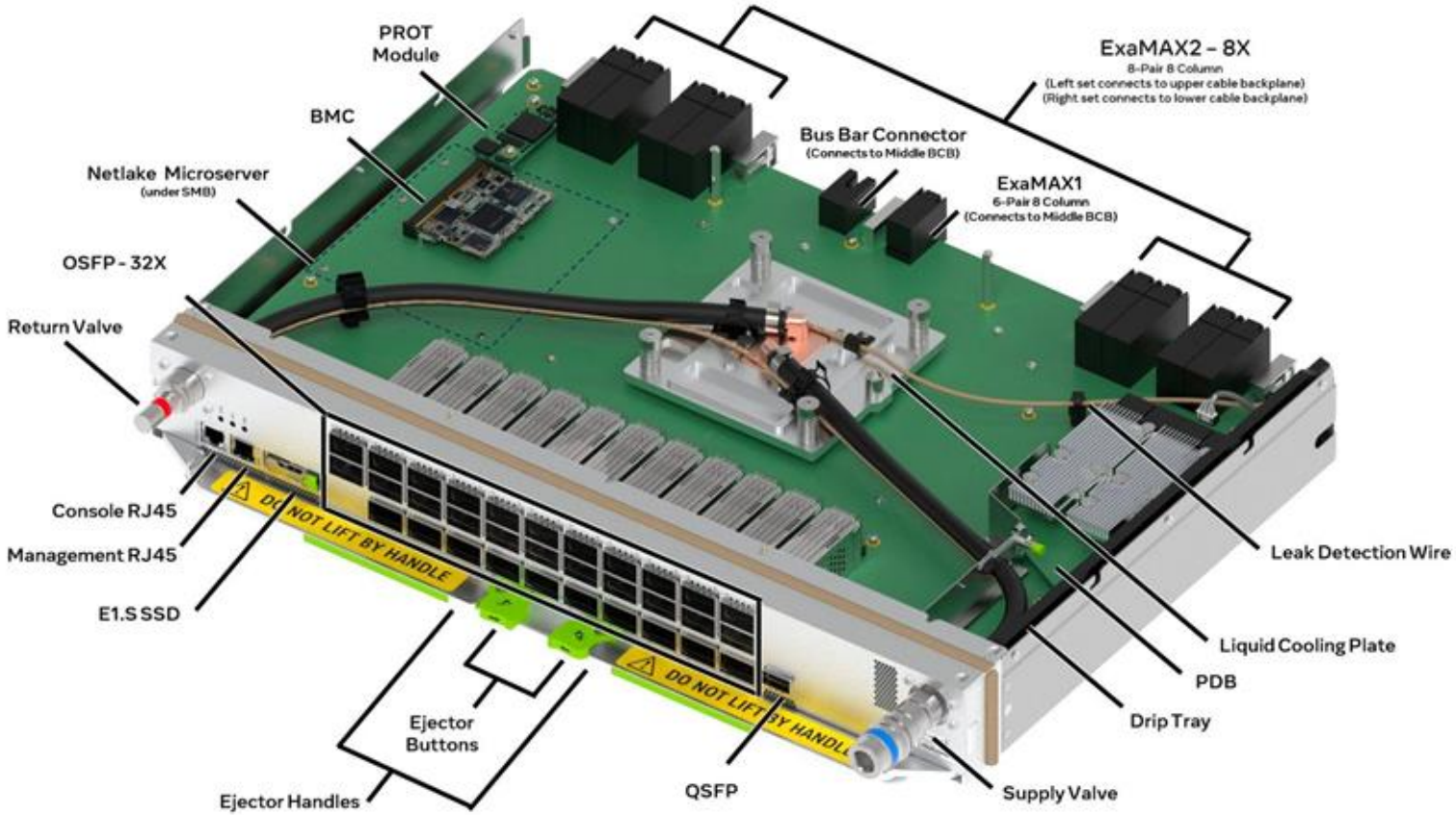
Inflection Point - Meta Custom Silicon

- MTIA - Proprietary scale-up model no longer fit
- Giving up Control and Efficiency
- **Vendor-locked scale-up solutions were not a viable path for an open, purpose-built platform.**

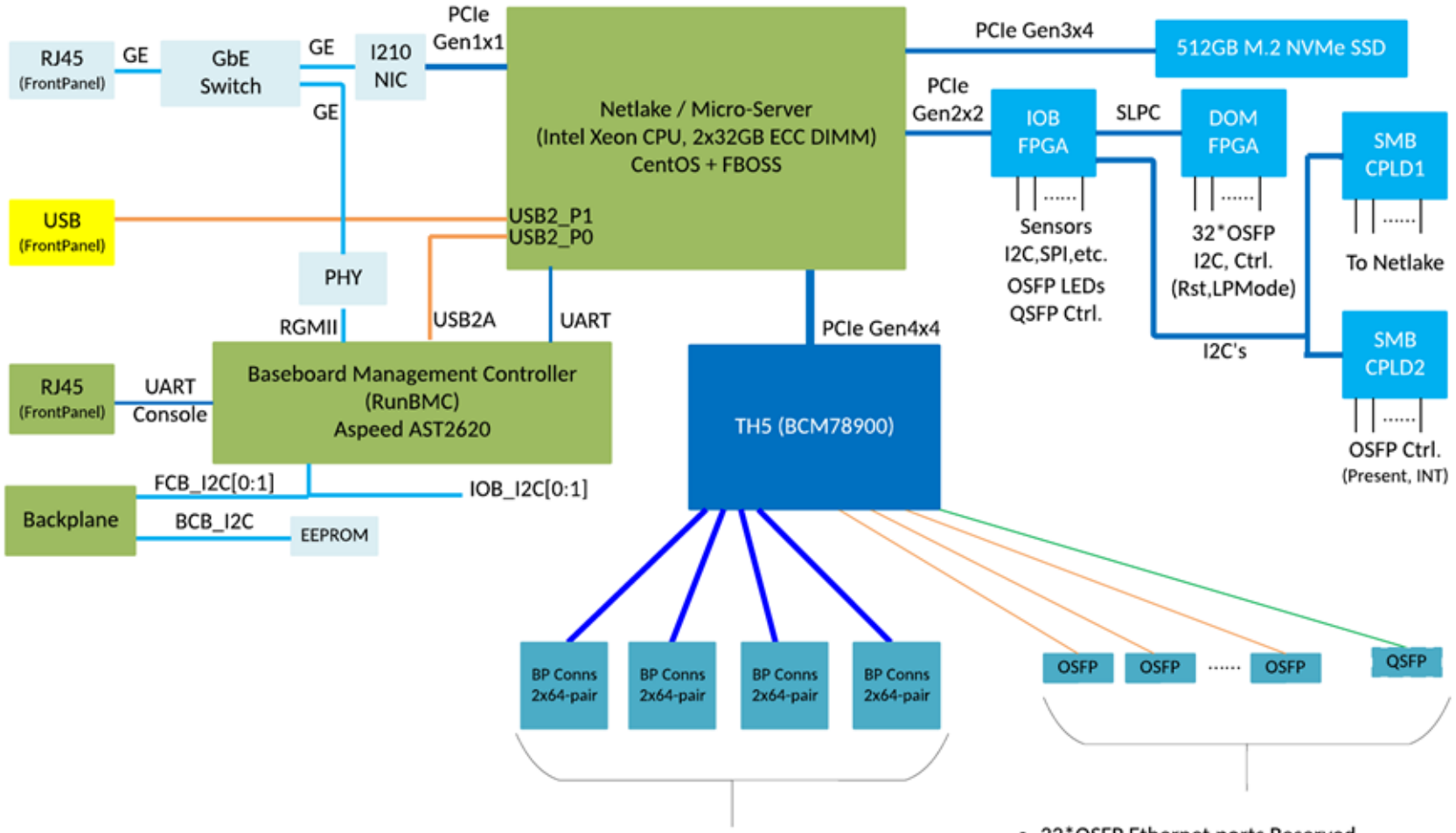
Solution - Introducing Tahan

- Open, Ethernet-based scale-up switch, Tahan designed specifically for MTIA300 Training Accelerators.

- Vendor independence
- A mature, open ecosystem
- Full-stack control
- Scale and Cost Efficiency



Tahan - System Overview

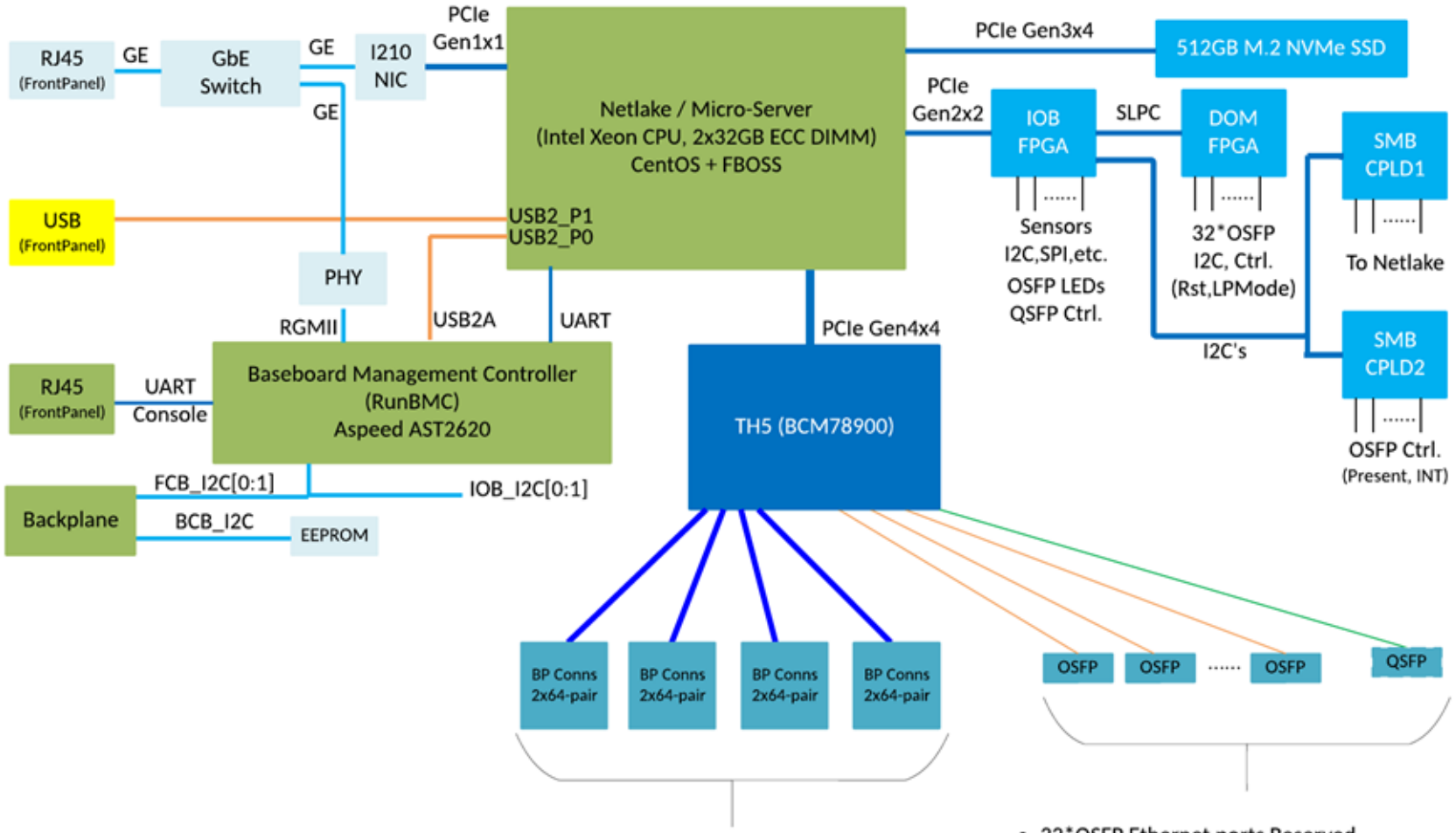


- 32*800GE Ethernet ports towards cable backplanes
- 32 ports go to 16 Accelerators, 2 per Accelerator;

- 32*OSFP Ethernet ports Reserved
- 1*QSFP Ethernet port towards ToR switch

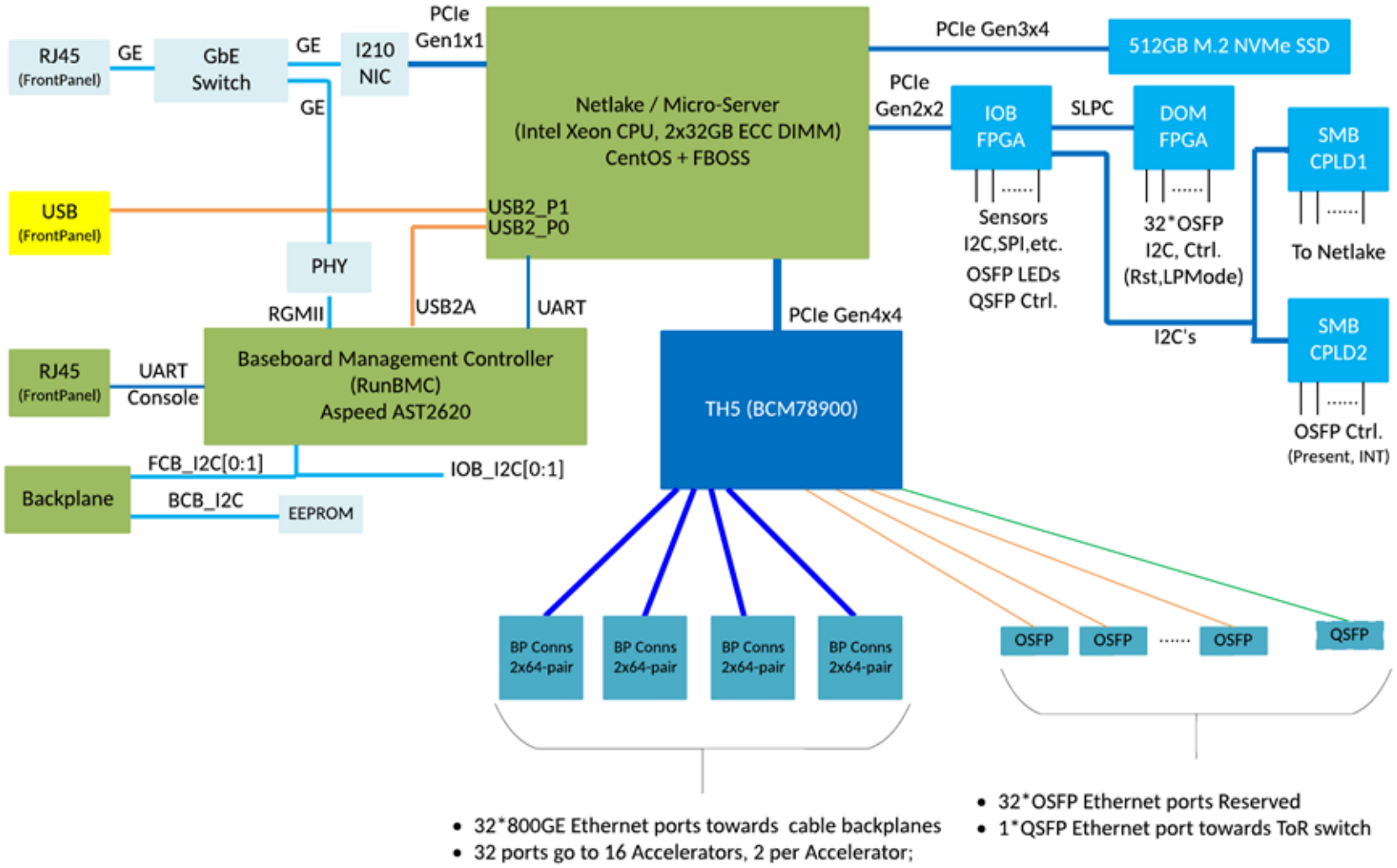
Tahan - System Overview

- 1OU, Hybrid liquid/air cooled
- Based on 51.2 Tbps Broadcom Tomahawk5 ASIC



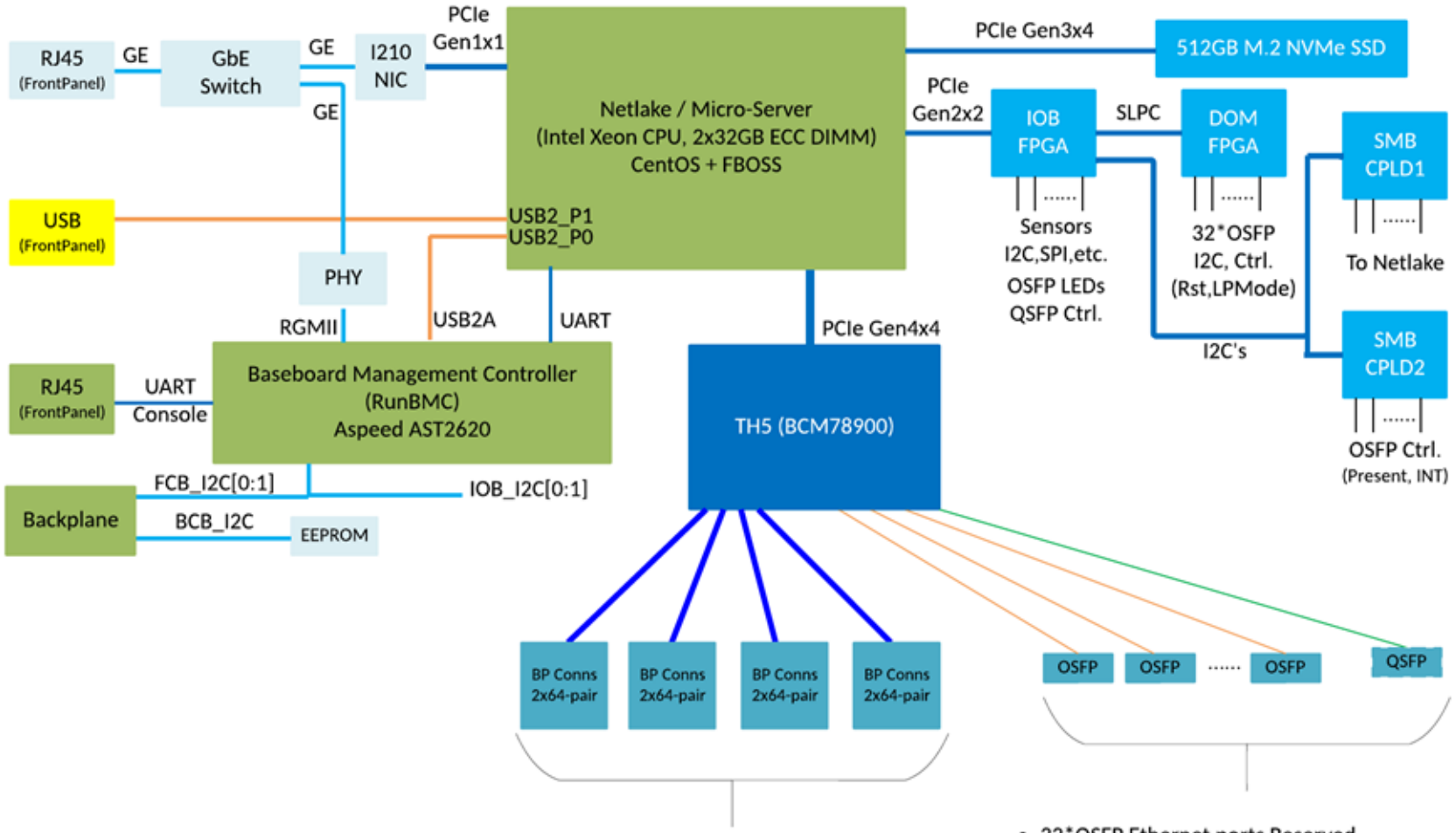
Tahan - System Overview

- PHY-less design with flyover cables
- Features 32x 800GE towards backplanes, 32x OSFP ETH(800GE) on front panel



Tahan - System Overview

- Deployed in ORv3-HPRv1 racks



Bandwidth per Accelerator

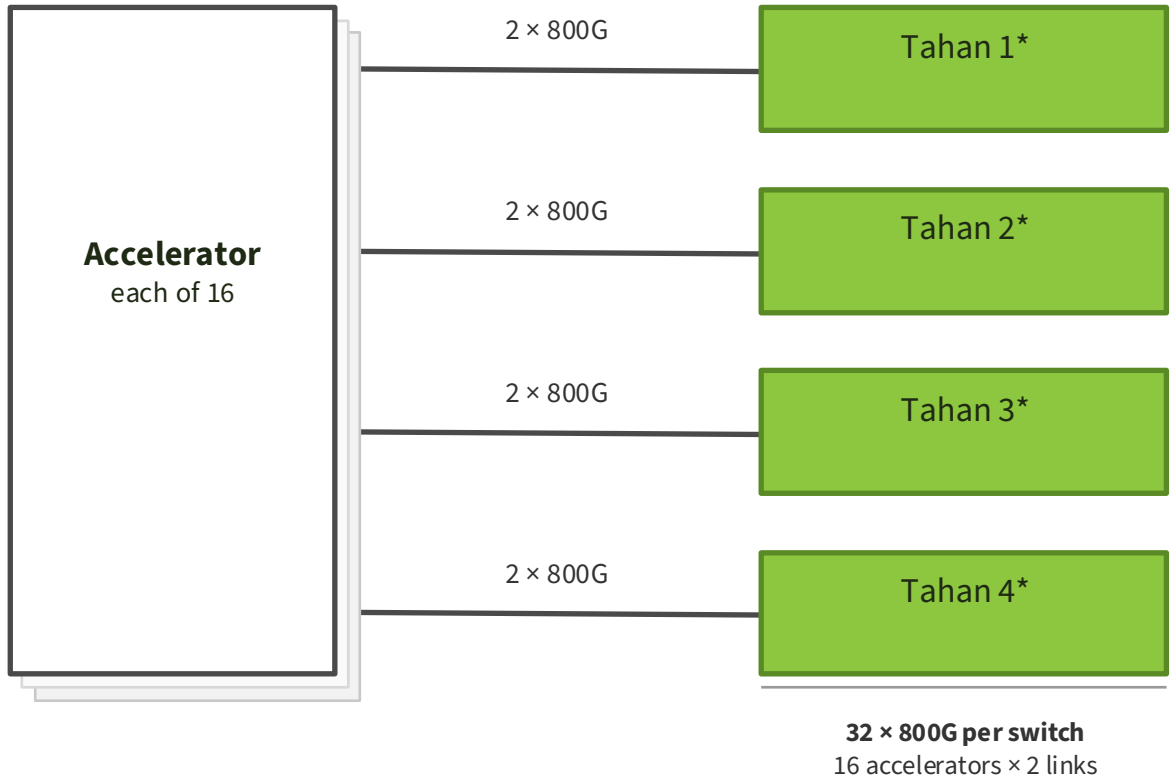
6.4 Tbps

per accelerator, per direction

16 accelerators, 8 × 800GbE each · 4 switches

Non-blocking

All of it on merchant Ethernet silicon



* Broadcom Tomahawk5, 51.2 Tbps merchant Ethernet switch

Latency: One Hop, Always

1 hop

always
every accelerator pair - no second tier

< 1 µs

per switch hop today
store-and-forward

–25%

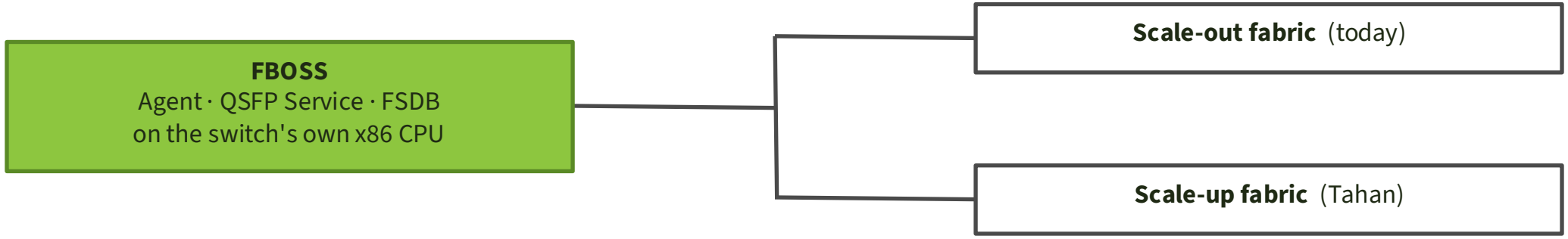
latency, cut-through
egress on the header, not the full frame - under exploration



One switch traversal, always. Across the fabric path the cables are ~1.5% - **the switch is the budget.**

Fabric path only - excludes NIC, PCIe and host software

One NOS, Two Fabrics



- **FBOSS already underpins Meta's scale-out fabrics - the same NOS evolves to run the scale-up fabric**
- Same repair and upgrade philosophy
- Same monitoring and telemetry

A Subset of the Scale-Out Feature Set

Reused

Forwarding
IPv6 dataplane, interface routes

Lossless
PFC lossless classes, AI-tuned QoS

Turned off

Routing Protocols
BGP routes

Multipathing
ECMP, Dynamic load balancing

Next-Gen Scale-Up Fabric

72 accelerators, 24 × 400G each · 8 switches

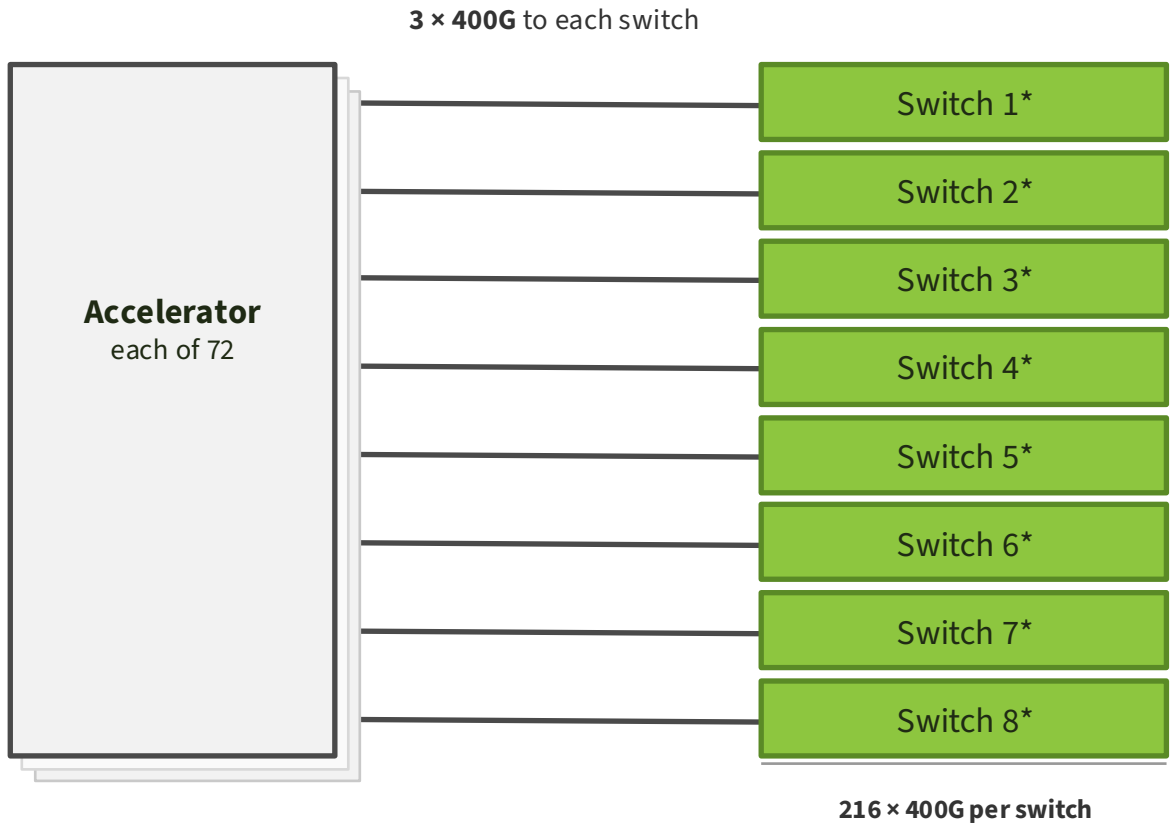
9.6 Tbps

per accelerator, per direction

Non-blocking · **Single hop**

1.5× Tahan
on merchant Ethernet silicon

* Broadcom Tomahawk6, 102.4 Tbps merchant Ethernet switch



Open Standards for Scale-Up

ESUN

Ethernet, minus the header tax

48 B of IPv6 + UDP → ~4 B; more goodput per frame

LLR

Link Layer Retry

Replay at the link, one link RTT; transport never sees the loss

CBFC

Credit-Based Flow Control

Send only against receiver credits; lossless, less buffer than PFC

ESUN: OCP base spec Rev 1.0 approved · LLR and CBFC: UEC specs, required by ESUN

Call to Action

Take it

The complete Tahan hardware spec

Will be contributed to OCP

Join it

ESUN at OCP; LLR and CBFC at UEC

Rev 1.0 approved, driven with Microsoft

Build it

Scale-up on merchant Ethernet

One NOS, one operational model

Thank You!

