Optical Scale-Up – System Reliability and Resiliency to Flaps

Future Technologies - Short Reach Optical Interconnect



Moderated by Edi Roytman, Sr. Distinguished Engineer,

Marvell Technology (ex Microsoft)

OCP FUTURE TECHNOLOGIES SYMPOSIUM

October 14, 2026
San Jose, California

MARVELL

Scale-up Optical Reliability and Fabric Resiliency to Optical Link Flaps

Lenin Patra
Senior Vice President – Data Center Architecture

October 2026

Forward-looking statements

Except for statements of historical fact, this presentation contains forward-looking statements (within the meaning of the federal securities laws) including statements related to future revenue, future earnings, and the success of our product releases that involve risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “projects,” “believes,” “seeks,” “estimates,” “can,” “may,” “will,” “would” and similar expressions identify such forward-looking statements. These statements are not guarantees of results and should not be considered as an indication of future activity or future performance. Actual events or results may differ materially from those described in this presentation due to a number of risks and uncertainties.

Forward-looking statements are only predictions and are subject to risks, uncertainties and assumptions that are difficult to predict, including those described in the “Risk Factors” section of our Annual Reports on Form 10-K, Quarterly Reports on Form 10-Q and other documents filed by us from time to time with the SEC. Forward-looking statements speak only as of the date they are made. You are cautioned not to put undue reliance on forward-looking statements, and no person assumes any obligation to update or revise any such forward-looking statements, whether as a result of new information, future events or otherwise.

© 2026 Marvell. All rights reserved.

Scale-X: Scale-in vs. scale-up vs. scale-out vs. scale-across networking

Tray scale

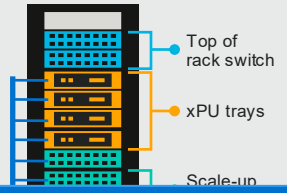


Scale-in

Tray level scaling

Extreme performance
Lowest latency
Highest bandwidth
Dragonfly networking

Rack scale
In rack



Scale-up

Vertical scaling

Extreme performance
Low latency
System operates as one unified fabric

Pod scale
Across racks



Scale-out

Horizontal scaling

Mili-second level of latency
Greater flexibility
Distributed workload

Data center scale
Across data centers



Scale-across

Regional scaling

Extends AI fabrics across regional distances to connect geo-distributed data center facilities

Pod scale
Across racks



Scale-out

Horizontal scaling

Mili-second level of latency
Greater flexibility
Distributed workload

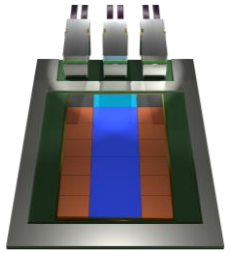
Connectivity: The nervous system of AI infrastructure

Binding compute, memory, networking into one AI fabric

© 2026 Marvell. All rights reserved.

Optics tailored for **scale-x architecture** use case

Coh and Coh-lite | QAM16
800G | 1.6T



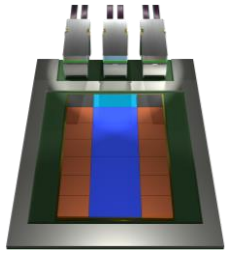
- Speed: 800G/lambda → 1.6T/Lambda
- Arch: Coh &Coh-lite SerDes → EIC → PIC

"Ultra-fast"
Fastest speed per channel
Single Lambda | Dual Lambda

"Scale-across and DCI "

SiPho| TFLN
Optics

Fast | PAM
200G, 400G



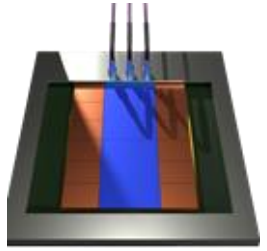
- Speed: 200G/lane → 400G/Lane
- Arch: SerDes → EIC → PIC

"Fast"
Faster speed per channel
Fewer wavelengths (DR, FR)

"Scale-up and scale-out"

SiPho| TFLN | EML
Optics

Flat | NRZ
50G, 100G



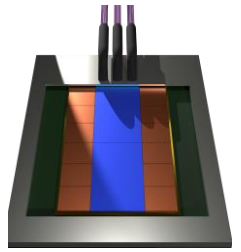
- Speed: 50G/lane → 100G/Lane
- Arch: D2D →SerDes-Lite → EIC → PIC

"Flat"
Lower power
More wavelengths

"More for scale-up "

Multi-wavelength DWDM
optics (OCI class)

Slow and Wide | NRZ
2G, 50G



- Speed: 2G/4G → 50G
- Arch: D2D → Gearbox → EIC → PIC

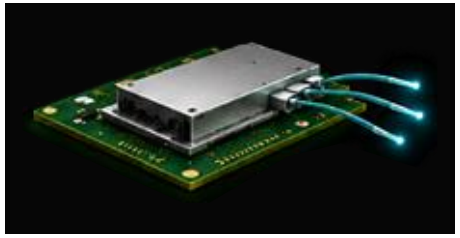
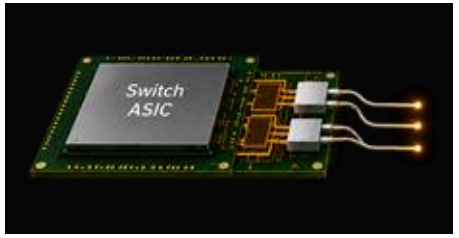
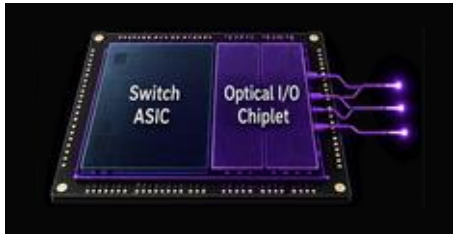
"S&W"
Lowest total pJ/bit
Massively parallel

"Primarily for scale-in; overlaps with scale-up"

µEmitter optics
µLED | µVCSEL

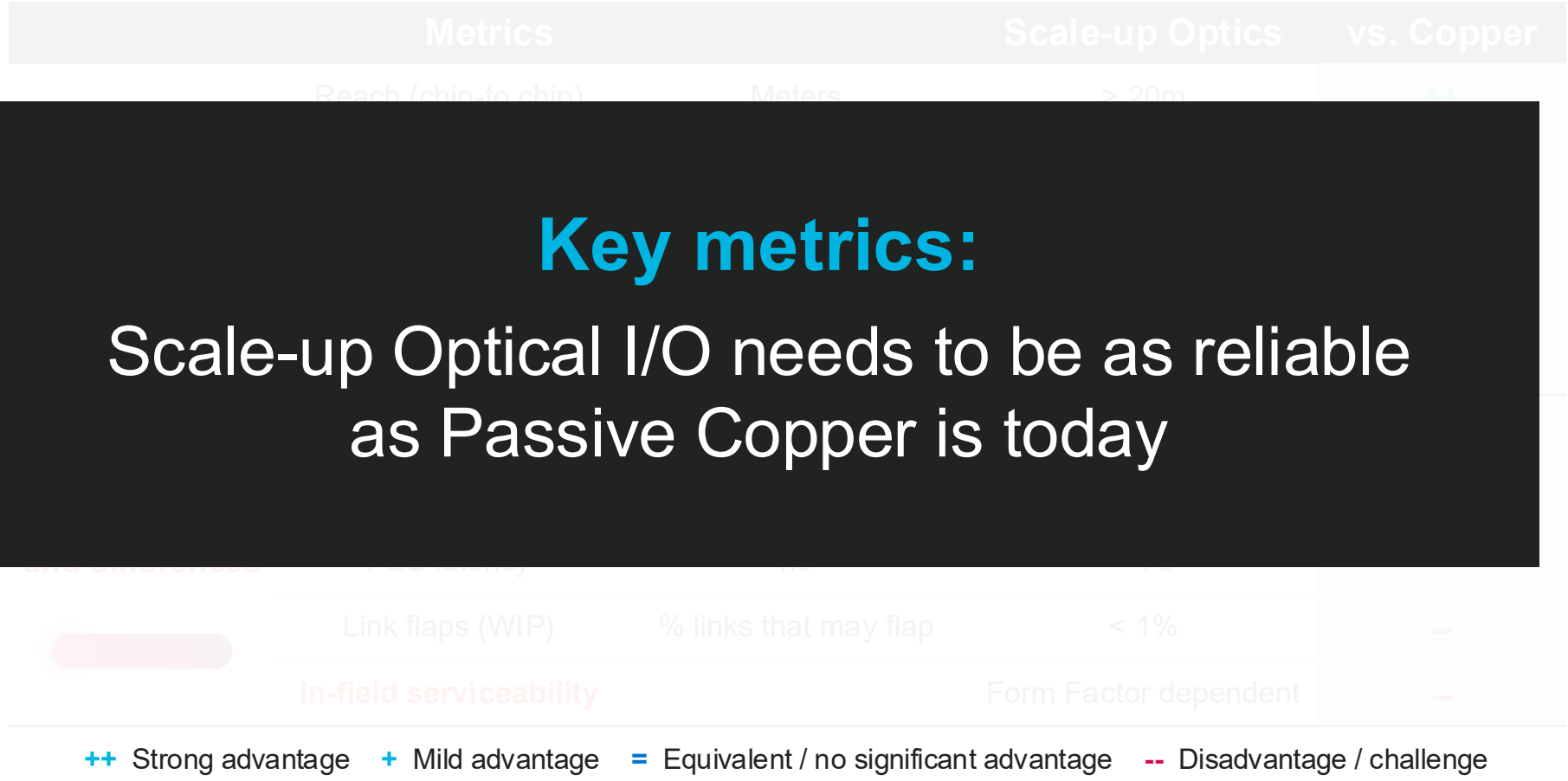
© 2026 Marvell. All rights reserved.

Optical **form factor**: Serviceability and co-integration

	XPO | OSFP pluggable optics Pluggable transceiver module	**Near-Packaged Optics (NPO)** Optics near the ASIC	**Co-Package Optics (CPO)** Optics co-packaged with the ASIC	**On-die optical I/O chiplet** Optical I/O integrated on die (chiplet)
				
Integration level	**Lowest**	**Low**	**High**	**Highest**
Electrical interconnect length	Long (10s of cm)	Short (mm-cm)	Very short (mm)	On die (μm)
Power efficiency	Lowest →	Better →	High →	Highest
Bandwidth density	Lower	Better	High	Highest
Thermal consideration	Lower	Moderate	High (co-design)	Highest (on-die)
Serviceability	Hot swapable ←	Replaceable module ←	Not field replaceable ←	Not field replaceable
Deployment Domain	Scale Out Scale up	Scale up Scale out	Scale up Scale out	Scale In Scale up

© 2026 Marvell. All rights reserved.

Scale-up optics | Can it meet the fidelity of copper?



Key metrics:

Scale-up Optical I/O needs to be as reliable as Passive Copper is today

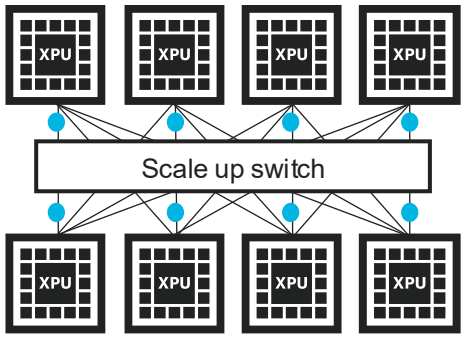
Source: Based on industry specs

© 2026 Marvell. All rights reserved.

Scale-up network topologies

Deployed in data center

Switch

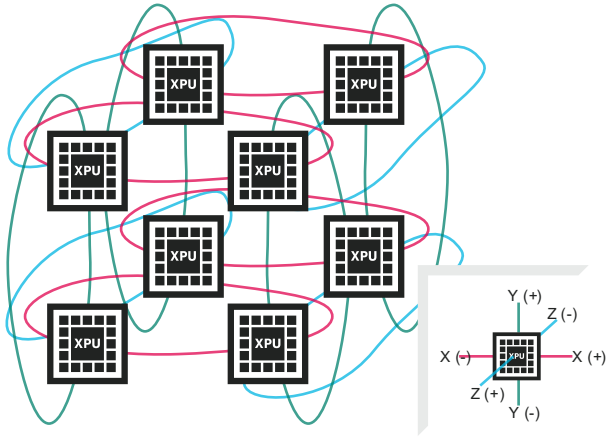


Latency, bandwidth

Scalability, complexity

72 XPU in one rack – copper

2D / 3D Torus

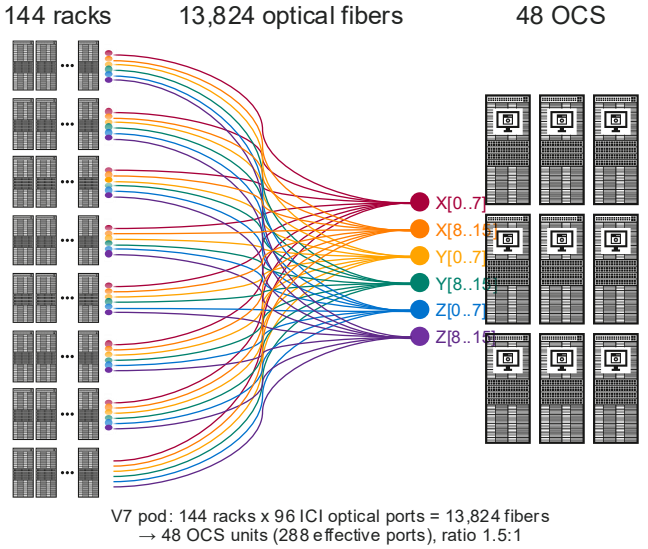


Latency, bandwidth

Scalability, complexity

64 XPU in one rack – copper

OCS in PoD

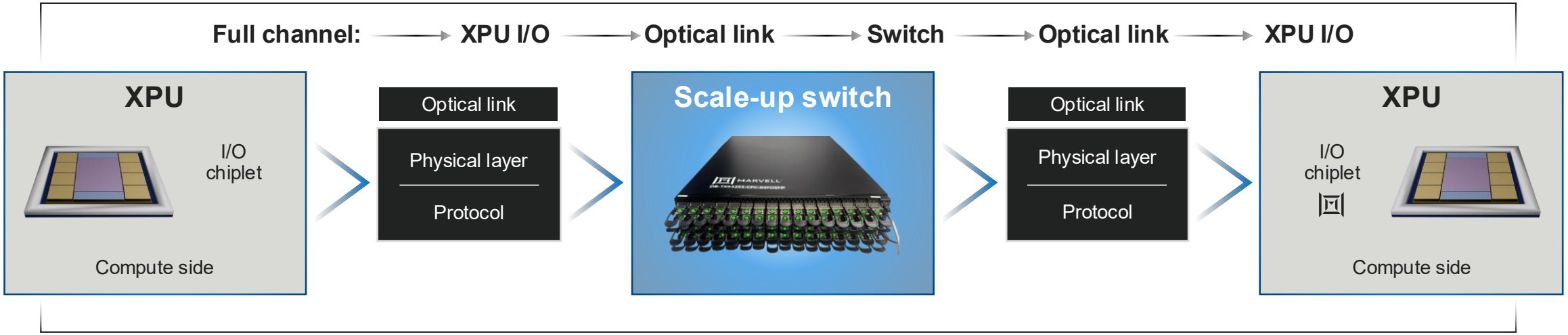


V7 pod: 144 racks x 96 ICI optical ports = 13,824 fibers
→ 48 OCS units (288 effective ports), ratio 1.5:1

9K XPU in one super POD- optics

© 2026 Marvell. All rights reserved.

Scale-up network solutions
Co-engineered from end-to-end



Design for link performance

Optimize channel margin, signal integrity, power and latency with redundancy

Architect for rapid bring-up

One owner for qualification, accelerating time to in-service

Operate with Link resiliency

End-to-end telemetry, link health visibility and network security

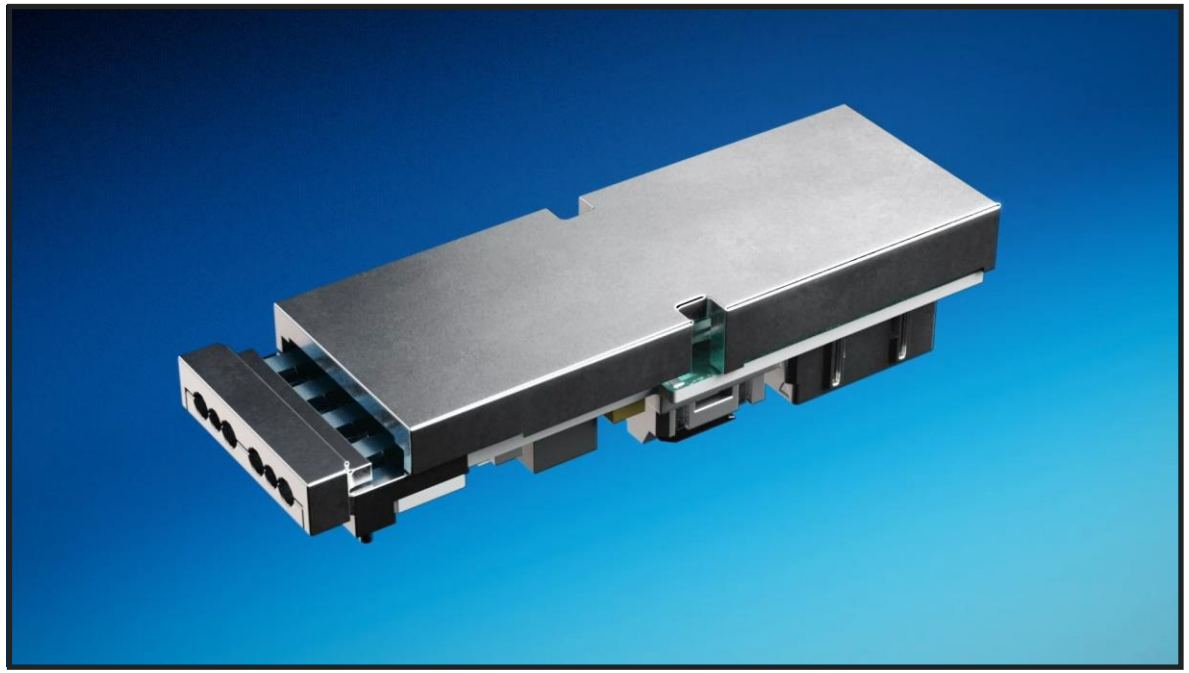
Built up with multi-layer link resiliency with scale-up networking

© 2026 Marvell. All rights reserved.

Deployment of optics in the scale-up network with multi-layer resiliency

NPO now, CPO next
Near-Package Optics

Physical layer: Copper → Optics



Lower failure blast radius

Field-replaceable optics

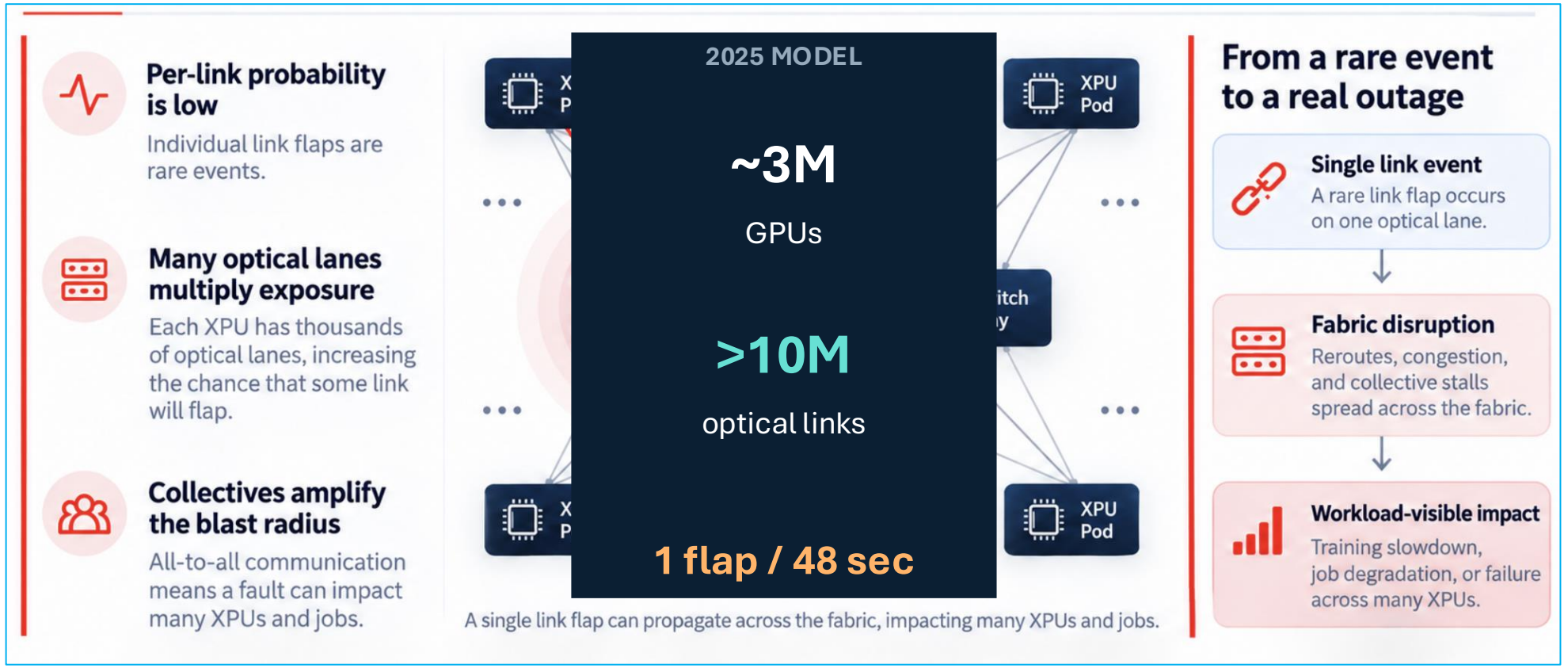
Flexible and configurable

Different priorities
Different architectures

One goal: optimize the system for TCO and serviceability

© 2026 Marvell. All rights reserved.

Ghosts in the AI factory: When rare link flaps become fabric-wide events



Link resiliency is a must for scale-up networking

© 2026 Marvell. All rights reserved.

A **practical 3R blueprint** for flap-resilient optics

Links may flap but AI workloads must never stop

01 Redundancy

Survive

Independent data paths / planes

Isolate failure domains and preserve alternate routes when a link or plane fails. Architect the Optics with redundant Optical Links

02 Resiliency

Recover

Local hardware detection + failover

Detect link faults and shift traffic to healthy capacity without waiting for software.

03 Reliability

Prevent

Telemetry + fleet remediation

Correlate FEC/BER, optics health and flap history to eliminate recurring failures.

Success criteria

No job abort

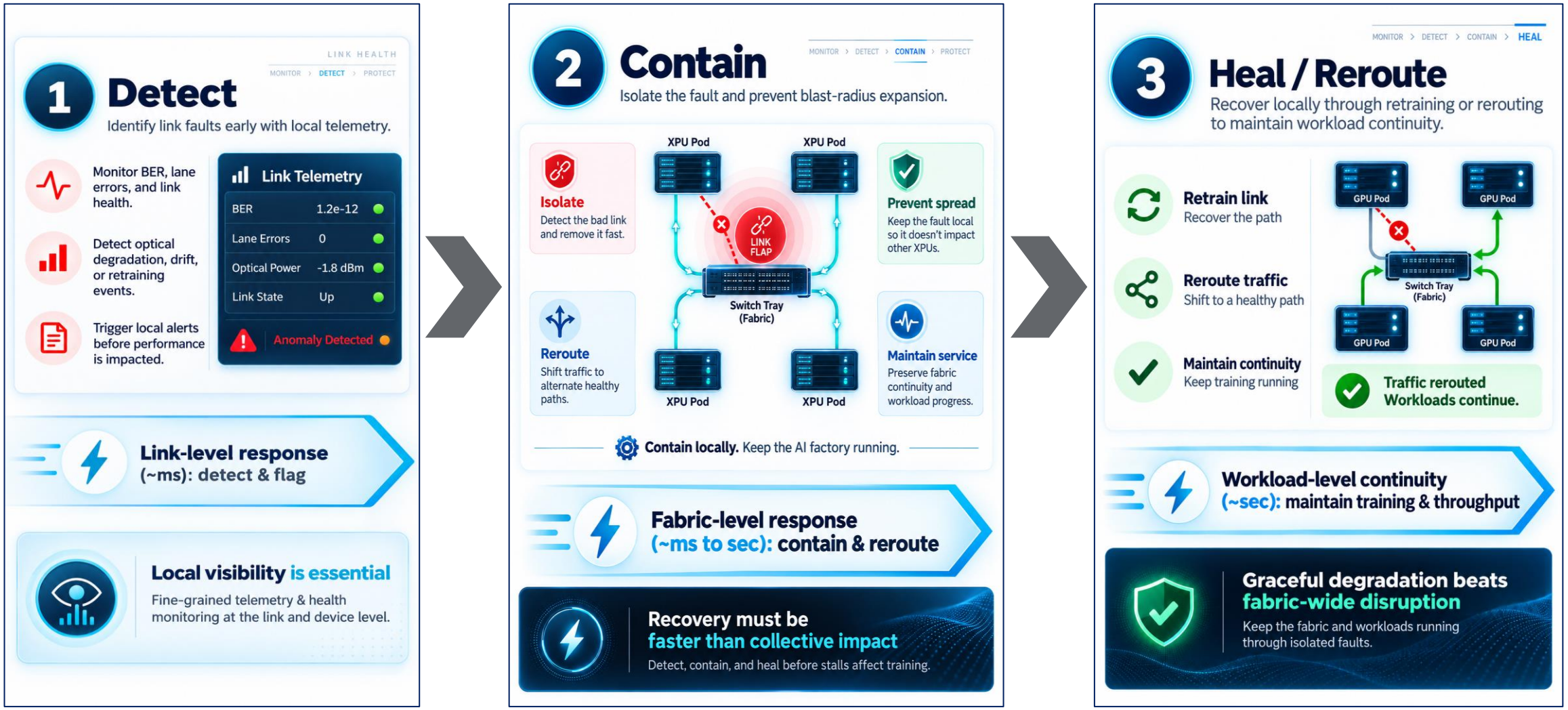
Fast, capacity-aware failover

Fewer repeat flaps

Graceful degradation | Hardware-local recovery | Closed-loop prevention

© 2026 Marvell. All rights reserved.

Link flap resiliency is **must** for scale-up networking



© 2026 Marvell. All rights reserved.

Introducing Marvell® RELIANT™ telemetry: Reliable Link Analytics and iNTelligence

From silicon to rack-scale system – real-time insights for a resilient AI fabric

Detect | Contain | Heal | Keep going

End-to-end visibility across the stack

System layer
Fleet health and orchestration

Fleet-wide telemetry
Health, performance, anomalies

Application layer
Workload resilience
Fast recovery using shadow state

AI driven insights
Predict, optimize, automate

Link layer
Traffic management

Link resiliency
Load balance, reroute, maintain flow
CBFC: Prevent Congestion and maintain Link health

Link level retry
Fast recovery

Rapid link recovery
Retry failed transactions
Rapid Link Init and Link training

Physical layer (PCS + Optical PMD)
Signal integrity, Optical impairments

Link health monitoring
BER, lane errors, signal quality, FEC bins, Equalizer adaptation statistics

Faster recovery at every scale

>1 second
- Fleet / workload scale
- Maintain cluster health

~500 milliseconds
- Application level
- Resume workloads

< 1 millisecond
- Link level
- Resume workloads

Near instantaneous
- Link retry
- Recover failed transactions

Real-time
- Physical layer
- Detect and correct errors

© 2026 Marvell. All rights reserved.

Key takeaways

Predictive optical intelligence is becoming mission-critical infrastructure for million-XPU AI clusters

Multi-layer link resiliency is a MUST to enable the fleet-level operation for scale-up networking

Marvell RELIANT Telemetry enables it @ every layer of the optical network

RELIANT platform turns optical connectivity from a reactive debug problem into a predictive operations layer for hyperscale AI fabrics.

© 2026 Marvell. All rights reserved.



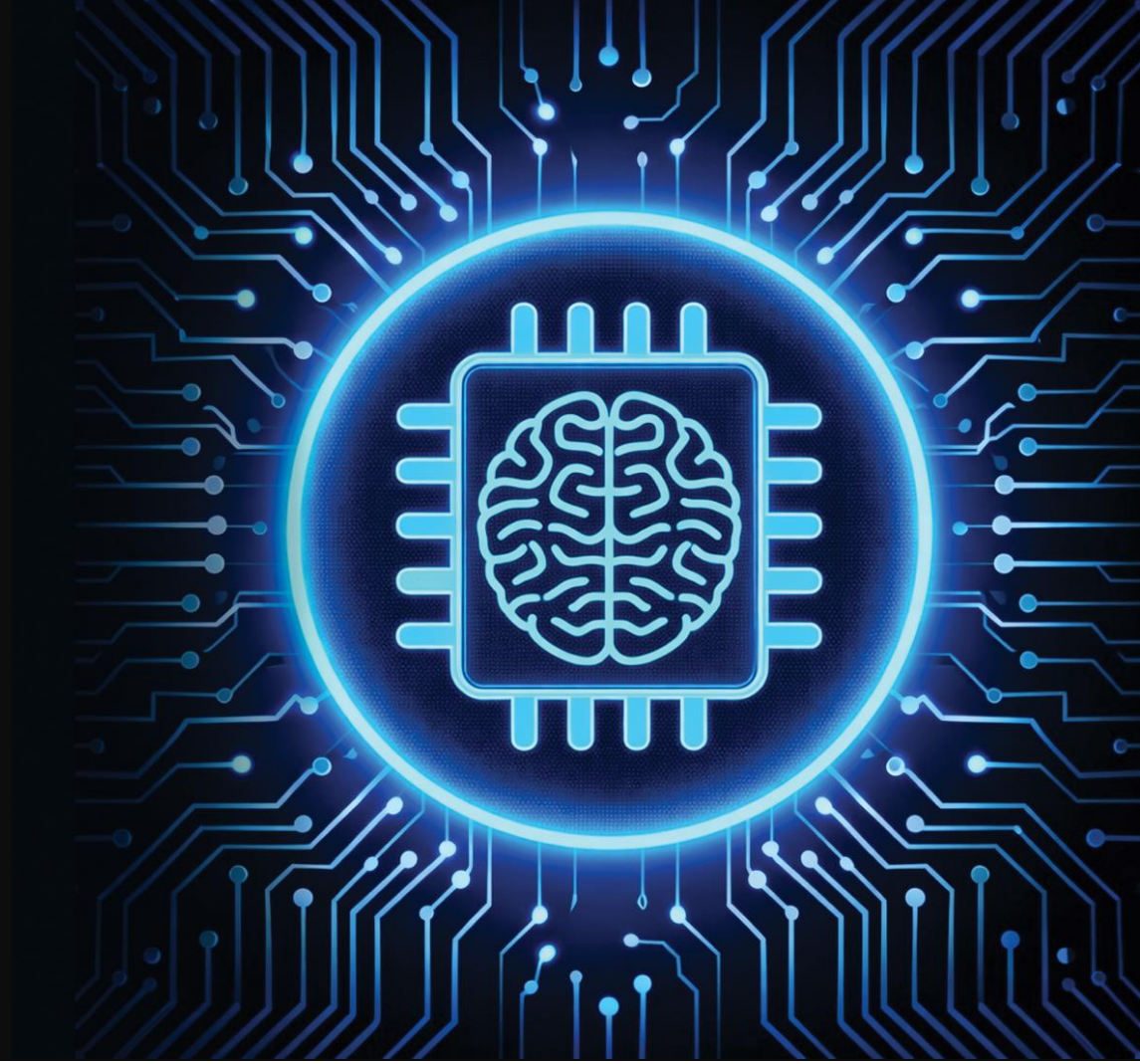
Thank You



Essential technology, done right™

RANOVUS®

26-42.3c Scale-up optical reliability and GPU-system resiliency to optical link flaps



Jeff Hutchins, Director, CTO office

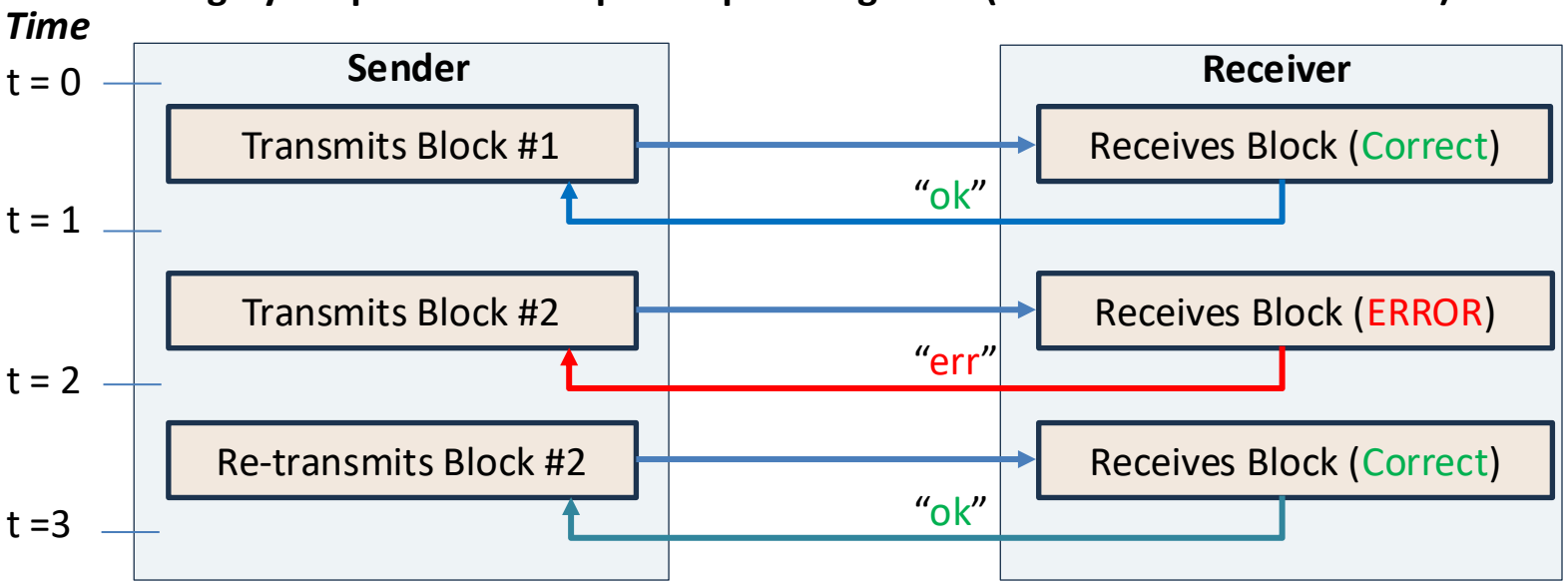
Ranovus

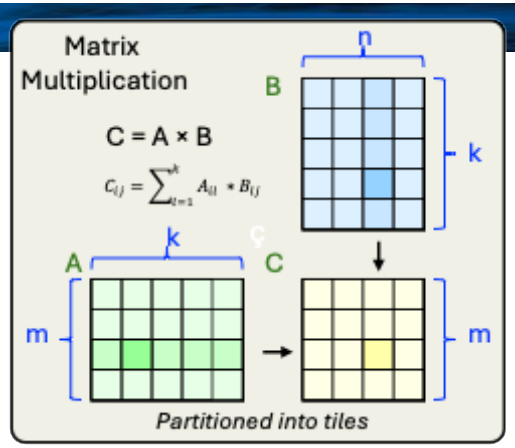
14 Oct 2026

RANOVUS®
Lighting Up the Future of AI®

Why reducing link errors for scale-up important

Highly simplified scale-up example using CBFC (credit-based flow control)





Job partitioned into tiles assigned to accelerators (exchanging large blocks of data)

- The following block is only transmitted once the current block is received correctly
- Periodic retransmissions drive tail latency & reduces effective interconnect BW
- Link flaps occur when the number of retries exceeds the retry limit → focus on minimizing burst errors

Minimizing burst errors to prevent link flaps is a key objective to reduce tail latency for scale-up

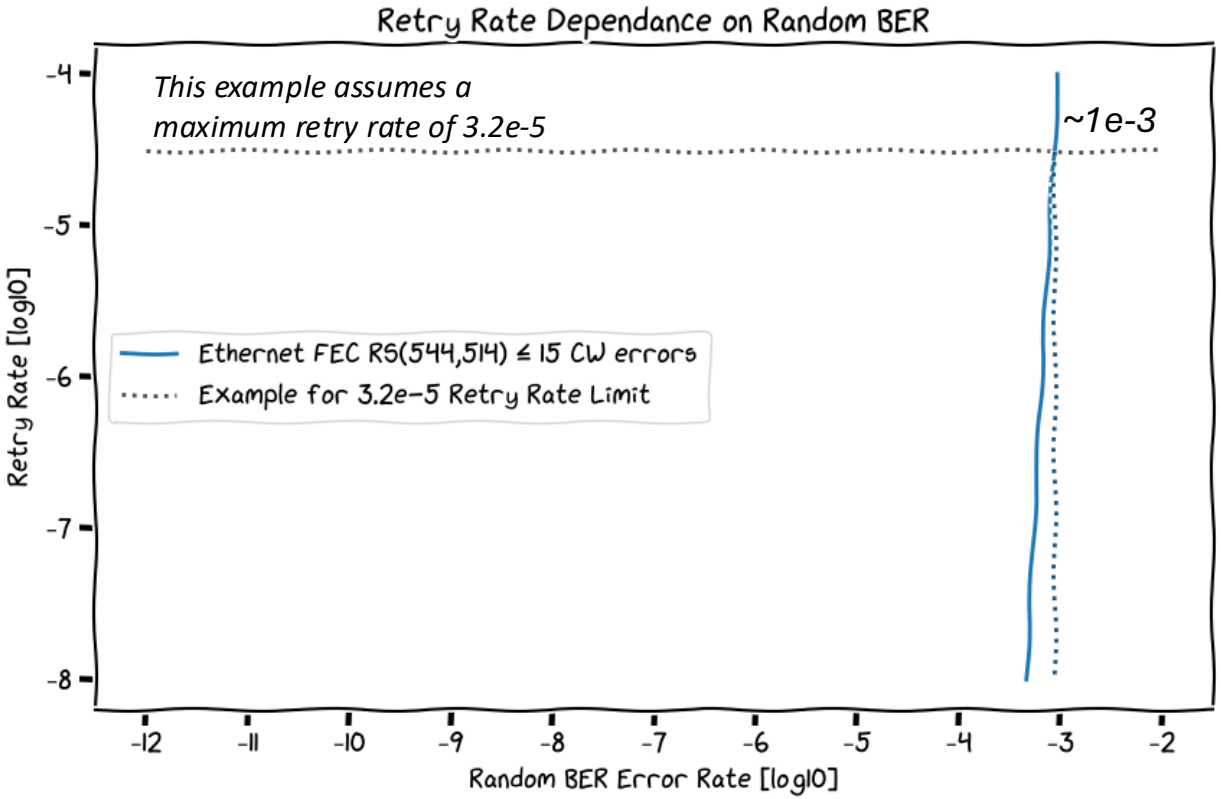
Forward error correction and random bit error rate

Retry rates due to FEC choice and assuming only random BER

An approximate analysis for some different FECs implementations

Ethernet FEC can correct up to 15 corrupted code words (CW)

- **A retry is need when more the 15 CWs are corrupted**



Note that the publication reports that the tested CPO based Ethernet implementation had no retries!

Meta [1] *reported no UCWs (uncorrectable code words) for > 1e6 400G-port-device-hours for a CPO switch!*

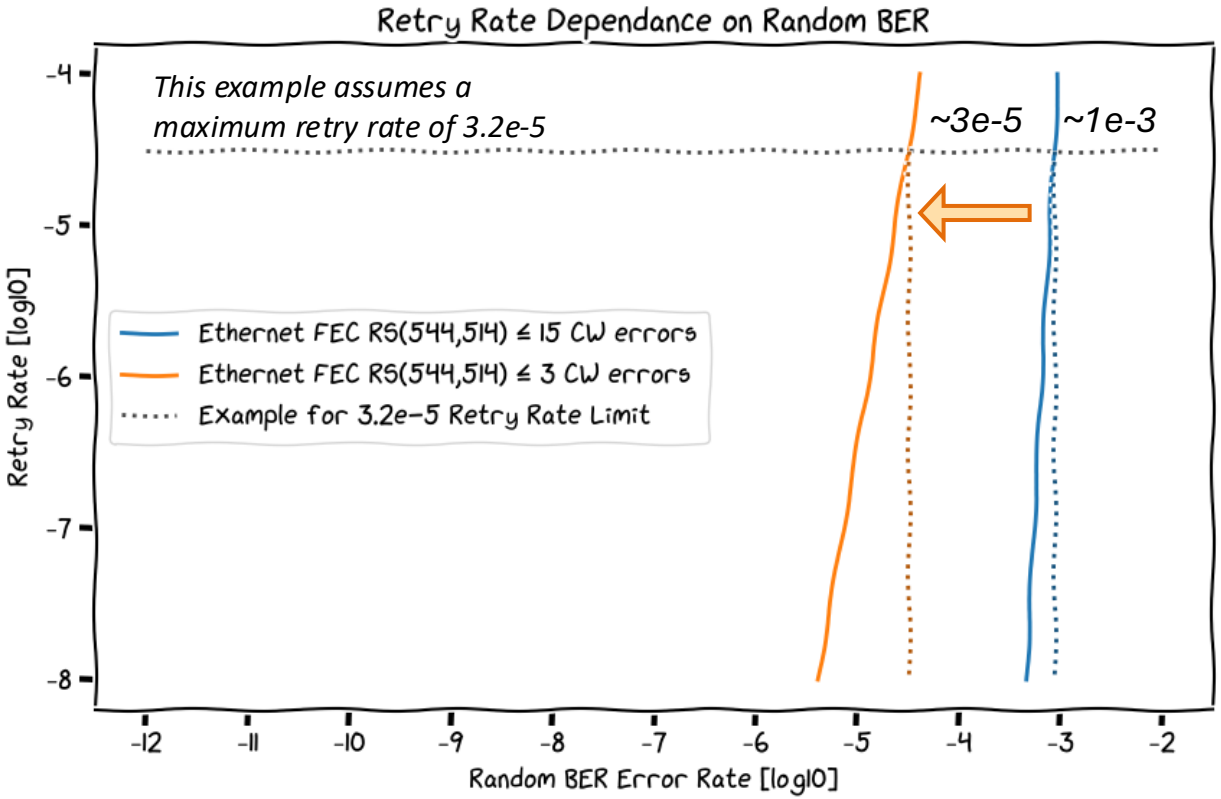
[1] ECOC25, Tu.01.03.3: Co-packaged Optics Technology Evaluation for Hyperscale Data Center Fabric Switches, Siamak Amiralizadeh et. al;, Oct 2025, Meta

Forward error correction and random bit error rate

Retry rates due to FEC choice and assuming only random BER

An approximate analysis for some different FECs implementations

- **However, optical systems exhibit burst errors**
- **Adding margin can reduce the likelihood of burst errors from causing a retry**



Meta [1] *reported no UCWs (uncorrectable code words) for > 1e6 400G-port-device-hours for a CPO switch!*

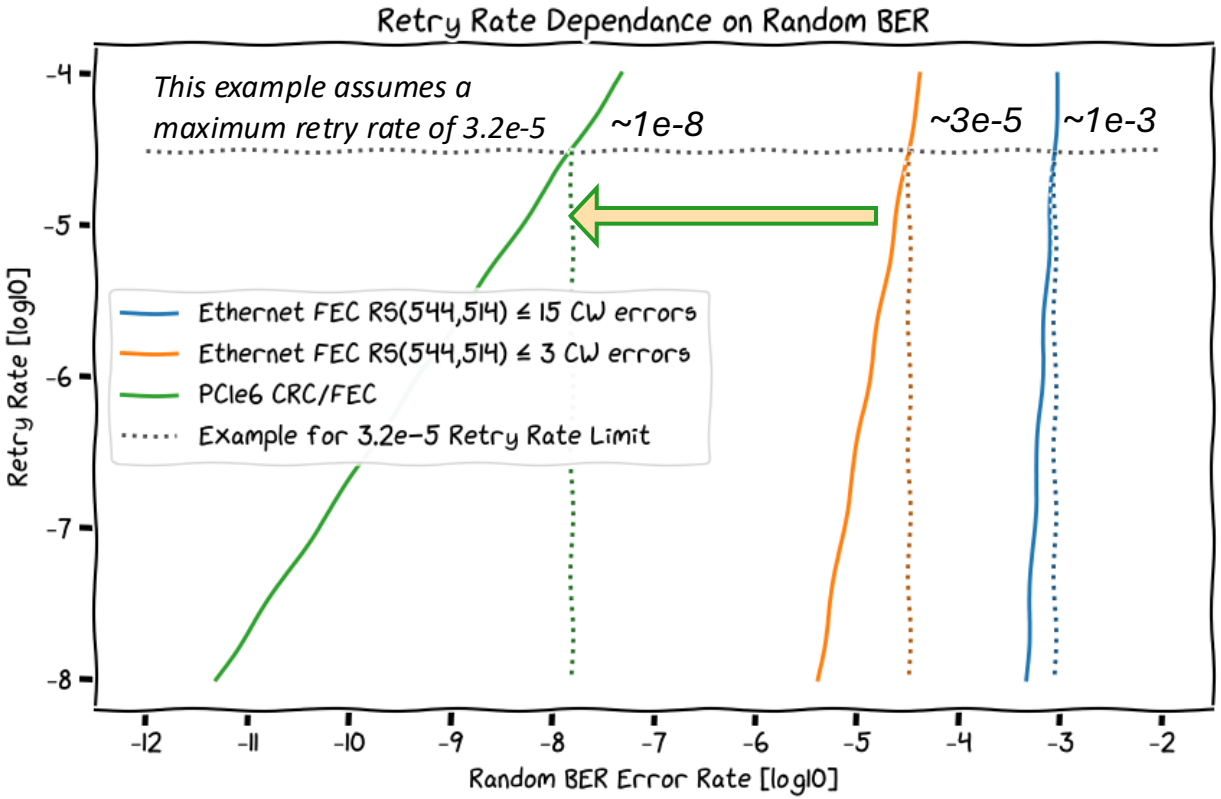
[1] ECOC25, Tu.01.03.3: Co-packaged Optics Technology Evaluation for Hyperscale Data Center Fabric Switches, Siamak Amiralizadeh et. al;, Oct 2025, Meta

Forward error correction and random bit error rate

Retry rates due to FEC choice and assuming only random BER

An approximate analysis for some different FECs implementations

- **Scale-up is looking to lower latency links**
- **Reduced FEC coding gain leads to lower latency**
- **Lower gain requires a lower random BER**
- **One way to lower BER is a a lower data rate**
  - **A recent example is the OCI-MSA**



- **Scale-up requirements are looking toward lower latency links**
- **Lower latency drives reduced FEC coding gain and therefore lower random BER**
- **Lower random BER drives lower data rate**

Potential causes of burst errors

	Possible Burst Error Contributors	Susceptible Configurations
Environmental (OE)	Mechanical shock, vibration, & warpage	Potentially all OEs
	Humidity, thermal variation, & EMI	
	Thermal & power supply transients	
	Transient high optical power effects on Rx	
Pattern Sensitivity (Baseline Wander)	Laser biasing operating point dynamics	For direct modulation implementations
	Modulator operating point dynamics	For external modulation implementations
Optical Interface & Fiber	Loss variation (optical connectors)	Due to mechanical or dust
	Optical reflections (Tx & Rx I/O)	Connectors and I/O ports
	Short fiber MPI	Multi-path interference for short fibers
	Fiber movement (mechanical vibration & shock)	
Electrical Interface (High-speed)	Reflection interference for short electrical traces	High-speed channels, (not D2D)
	Humidity, thermal variation, & EMI	
	Stability of electrical contacts / reliability changes	For socketed implementations
Laser Input Stability	Impact of reflection variability into laser cavity	Pre-isolator, or non-isolator implementations
	Impact of stress, thermal variation for laser die	
	Fiber (mechanical vibration, shock, & MPI)	For external laser implementations
Firmware/Management	Varied causes	
Other (Higher level)	Timing issues / synchronization failures / auto negotiation	Especially systems with different retimer/SerDes implementations
	Interoperability, incompatibility, & configuration issues	
	Stability, degradation of PCBA components	For example, power supply components

- Brainstormed list of potential causes of burst errors which can lead to link flaps
- Short periods of burst errors can yet lead to retries yet have a minor impact on the random BER (depending on the measurement period)
- Some of these causes are
  - Detectable well before failure
  - Other causes may appear suddenly

Each implementation should use design techniques to minimize the potential for burst errors

What can the industry do from a module perspective?

RANOVUS®

Define the application requirements:

- Provide a target acceptable retry rate for each application

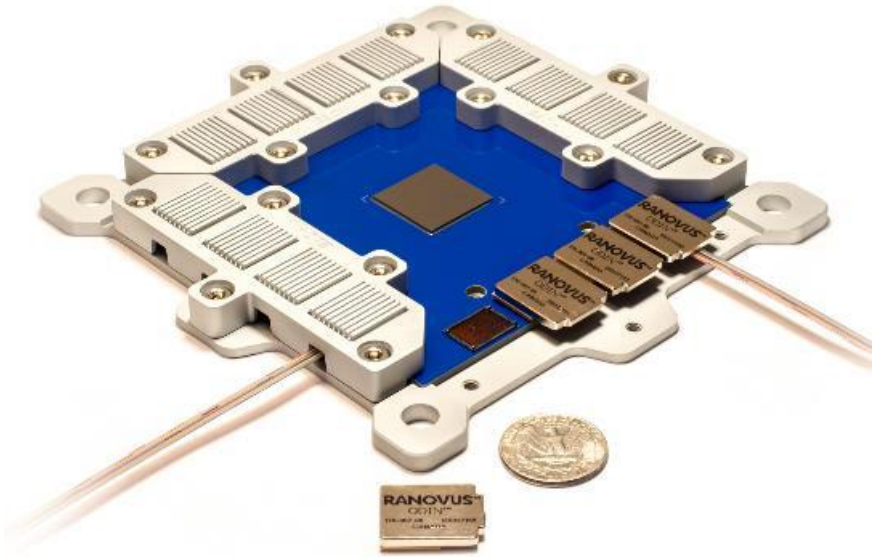
Education:

- Highlight the importance of minimizing burst errors and link flaps at the transceiver level
- Suggest telemetry approaches (gradual appearance) and types of test methodologies which could be employed (sudden and gradual)

Standards:

- Standardize telemetry approaches
- Standardize on test methodologies for burst errors leading to optical link flaps (gradual and sudden transition to burst errors)
- Develop specifications to accompany the test methodologies

Thank You!



RANOVUS®

Lighting Up the Future of AI®

Optical Scale-Up – System Reliability and Resiliency to Flaps

Eduard (Edi) Roytman & Dr. Greg Link

Microsoft



OCP FUTURE TECHNOLOGIES SYMPOSIUM

October 14, 2026
San Jose, California

Motivation

- Optical Scale-Up allows 1-tier Pod size expanded up to the radix of the switch. 1024+ GPUs/XPUs per Pod are possible.
- Optics has its benefits and challenges. One challenge is Flaps.
- Flaps are intermittent link-down events, uS – minutes in duration.
- Some lanes might flap all the time and be persistently “down”.

OK, why do we care?

- Network Switches know how to deal with flaps for 30+ years. GPUs/XPUs don’t…
- If every flap causes a job crash, **we could see tens to hundreds of job crashes a day… not acceptable!**

Some example stats…MTBF and Time (hrs) to 1st Flap

- Sweep over Flap-FIT rate (10-1M) and Pod size (16-1024 GPUs/XPUs)

MTBF (hrs)	16 GPUs	32 GPUs	64 GPUs	128 GPUs	256 GPUs	512 GPUs	1024 GPUs
10 Flap FIT	48828.13	24414.06	12207.03	6103.52	3051.76	1525.88	762.94
100 Flap FIT	4882.81	2441.41	1220.70	610.35	305.18	152.59	76.29
1000 Flap FIT	488.28	244.14	122.07	61.04	30.52	15.26	7.63
10000 Flap FIT	48.83	24.41	12.21	6.10	3.05	1.53	0.76
100000 Flap FIT	4.88	2.44	1.22	0.61	0.31	0.15	0.08
1000000 Flap FIT	0.49	0.24	0.12	0.06	0.03	0.02	0.01

Hours to 1st failure with P-fail = 0.5 (50%)	16 GPUs	32 GPUs	64 GPUs	128 GPUs	256 GPUs	512 GPUs	1024 GPUs
10 Flap FIT	33845.1	16922.5	8461.3	4230.6	2115.3	1057.7	528.8
100 Flap FIT	3384.5	1692.3	846.1	423.1	211.5	105.8	52.9
1000 Flap FIT	338.5	169.2	84.6	42.3	21.2	10.6	5.3
10000 Flap FIT	33.8	16.9	8.5	4.2	2.1	1.1	0.5
100000 Flap FIT	3.4	1.7	0.8	0.4	0.2	0.1	0.1
1000000 Flap FIT	0.3	0.2	0.1	0.0	0.0	0.0	0.0

Example highlighted (**NOT** indicative of field data):

- With a Flap-FIT rate of 10K, 1K GPU pod might hit 1st flap within first 30min!

Lane Failures Lead to Job Complexity

- Workloads are generally dependency graphs of operations and data transfers
- If data transfers are blocked,
  the dependency graph cannot be completed – job hangs, likely aborted

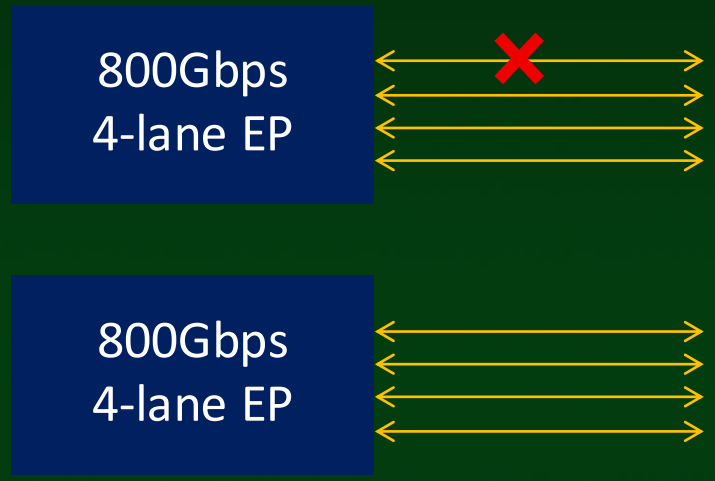
- Various layers could mitigate these dynamic lane failures:
  - TCP will want to re-transmit … when it comes back up?
  - MRC might re-transmit on a different path
  - Switching fabrics can mitigate with ECMP
- Industry seems to be trying different solutions to find the right blend

- Today is an opportunity to better align objectives and potential solutions

GPU/XPU resiliency to flaps

- Key Q – when lane(s) flap, how we avoid a job abort?

Example:

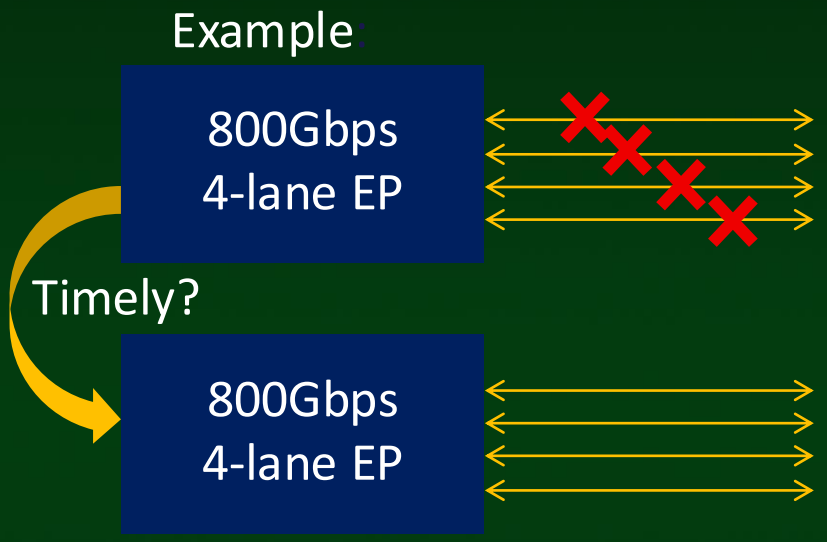


What happens if 1 lane flaps?

- 1x800 mode ➔ Job abort
- 4x200 mode ➔ degrade to 3x200 or 2x200

GPU/XPU resiliency to flaps

- Key Q – when lane(s) flap, how we avoid a job abort?



What happens if 1 lane flaps?

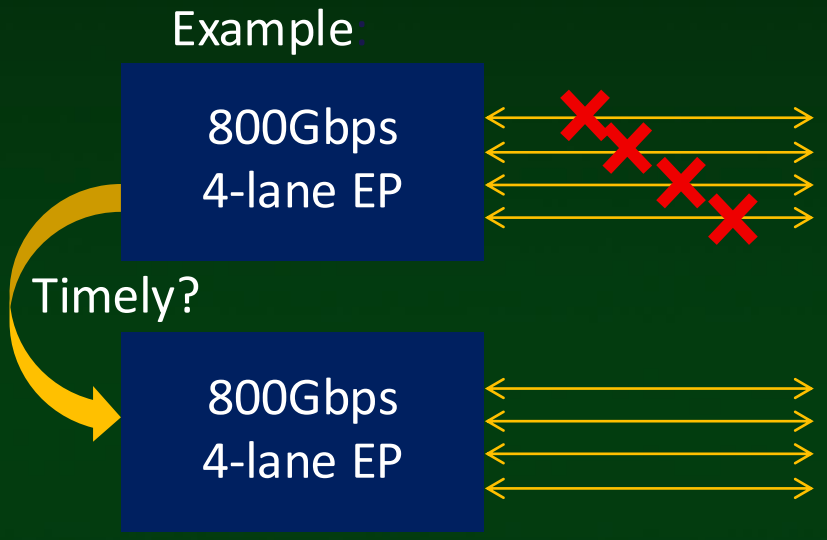
- 1x800 mode ➔ Job abort
- 4x200 mode ➔ degrade to 3x200 or 2x200

What happens if all 4 lanes flap together?

- EP becomes blocked – no valid path
- Is MRC going to help?
- If traffic can be moved in ***timely*** fashion to another EP, job survives, otherwise job crashes.

GPU/XPU resiliency to flaps

- Key Q – when lane(s) flap, how we avoid a job abort?



What happens if 1 lane flaps?

- 1x800 mode ➔ Job abort
- 4x200 mode ➔ degrade to 3x200 or 2x200

What happens if all 4 lanes flap together?

- NIC becomes blocked – no valid path
- Is MRC going to help?
- If traffic can be moved in ***timely*** fashion to another NIC, job survives, otherwise job crashes.

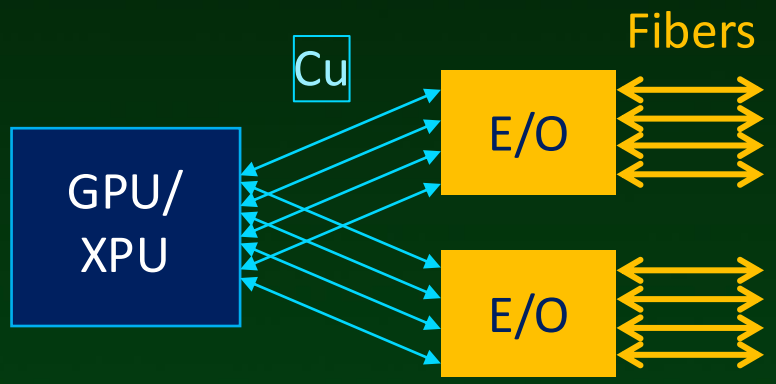
What is a likelihood of all 4 lanes flapping together?

- It depends on whether <u>all lanes of a given EP</u> belong to the same flap fault domain
- Multi-lane flaps do exist and can be quite common ***within the same flap fault domain.***
- If all lanes per port are in ***different flap fault domains***, with P as prob of 1-lane flapping, prob of all 4-lanes flapping per port is ~P^4 - a much smaller value.

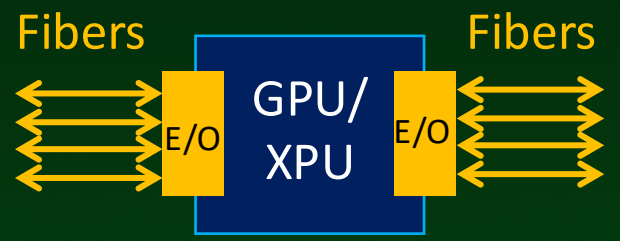
Optical XCVRs

- Integration continues: QSFP➔OSFP ➔32-64 lane OBO, XPO, CPO….
  - More lanes per optical reliability and flap fault domain
  - More lanes per laser
  - More lanes per connector
  - Closer to ASIC: from Pluggable to on-board to co-packaged
- Data and analysis we want from optical vendors:
  - Laser and whole module reliability – especially over time
  - Connector tolerance to dust and assembly variation
  - Advanced in-band and out-of-band link-state telemetry & control
  - Flap stats from Pod-scale full-fledged PoCs with realistic optics, switches and fibers

Co-boarded optics…. Going into Co-packaged optics



Pluggable or Co-boarded Optics



Co-packaged Optics

Pluggable or Co-boarded Optics	Co-packaged Optics
E/Os are independently serviceable	E/O are not independently serviceable
Flap fault domains: On-board Cu swizzle	Flap fault domains: On-die swizzle required
Higher end-to-end power	Lower end-to-end power
Datarate is Cu limited	Datarate is not Cu limited
BW/shoreline is Cu SERDES limited	Much higher BW/shoreline is possible

Call to Action: End-to-End Flap Strategy is required

- **Proactive Measures (before flap happens)**
  - High-Quality Optical Technology
  - Flap-aware link deployment, cleaning, servicing procedures
  - Flap-resilient system design practices to reduce blast radius and maximize nodes in service
  - Smart telemetry to observe, predict, log and manage flapping links per policy
- **Resiliency Measures (while flaps are active)**
  - Platform/Package level measures
  - GPU/XPU SoC HW/FW level measures
  - Multi-ported NICs that ensure multiple valid paths between endpoints
  - Software stack that allows submitting traffic to alternative NICs to ensure dependency completion
  - Collectives that can accommodate dynamic lane events
- **Reactive Measures (after flaps are gone)**
  - Smart telemetry to take-out highly offending links and log service requests
  - Dampen/Quarantine occasionally flapping links per policy
  - Routing algorithms that can handle suddenly unbalanced traffic patterns
  - Job migration if #active links below customer service agreement

How to get involved

- **OCP Networking - Optical Reliability workstream**
  - ocp-optics-reliability@OCP-All.groups.io | Home
- **OCP Short Reach Optical Interconnect**
  - By Bobby Lu (Lightintelligence) & Ron Swartzentruber (AMD)
- **More to come…. Stay tuned!**

OCP FUTURE TECHNOLOGIES SYMPOSIUM

October 14, 2026
San Jose, California



Closing Thoughts

- **AI scale-up networks have become mission-critical infrastructure.**
- **At pod scale, reliability matters as much as bandwidth.**
  - **Every link event can translate directly into lost GPU productivity.**
- **The solution is end-to-end: optics, silicon, systems, networking, software, and operations must work together.**
- **We must evolve from fault detection to fault prediction through richer telemetry and fleet-scale learning.**
- **The true metric of success is not link uptime, but uninterrupted AI job execution.**
- **Call to Action:** *Let's align as an industry on common reliability metrics, telemetry standards, and resiliency architectures to make optical scale-up fabrics trustworthy at AI-factory scale.*

OCP FUTURE TECHNOLOGIES SYMPOSIUM

October 14, 2026
San Jose, California

