



TECHNOLOGY
EXPLORATION
FORUM 2024
ETHERNET IN THE AGE OF AI

TEF 2024: Ethernet in the Age of AI

...

October 22-23, 2024
Santa Clara Convention Center, CA, USA

 ethernet alliance

The State of Today's Industry SDO's and MSAs

October 22-23, 2024

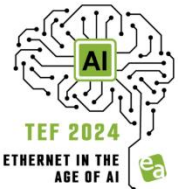




This presentation has been developed within the Ethernet Alliance, and is intended to educate and promote the exchange of information. Opinions expressed during this presentation are the views of the presenters, and should not be considered the views or positions of the Ethernet



Per IEEE-SA Standards Board Bylaws
“At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position of IEEE.”

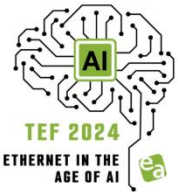


MODERATOR & PANELISTS

- Moderator - John D'Ambrosia, EA TEF Chair (Futurewei)
- Panelists
 - Mark Nowell, IEEE P802.3dj, Vice-Chair (Cisco Fellow)
 - “Progress update on IEEE’s next generation Ethernet project: IEEE P802.3dj - 800 GbE & 1.6 TbE”
 - Nathan Tracy, OIF, President (TE Connectivity, Technologist)
 - “OIF Addresses The Paths To Next Generation AI”
 - Uri Elzur, UEC Technical Advisory Committee Chair
 - “Introduction and proposed Ethernet Enhancements for AI”
 - Tom Palkert, SFF Transceivers Group Chair (Samtec)
 - “Migration of SFF Connectors and Modules”
 - Bijan Nowroozi, Open Compute Project Foundation, CTO
 - “The Open Compute Project (OCP) Open Systems for AI Initiative: Towards Open Systems for AI”

Progress update on IEEE's next generation Ethernet project: IEEE P802.3dj - 800 GbE & 1.6 TbE

Mark Nowell, Cisco Fellow
IEEE P802.3dj Vice-Chair



Acknowledgements

I'd like to acknowledge the contribution and collaboration with the 802.3dj project leadership in developing many of these slides:

- John D'Ambrosia - Chair, P802.3dj
- Matt Brown - Chief Editor, P802.3dj
- Kent Lusted - Secretary and Electrical Track Chair, P802,3dj
- Gary Nicholl, Logic Track Chair

Per IEEE-SA Standards Board Bylaws

“At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position of IEEE.”

Moving Beyond 400 GbE (4x100 Gb/s)

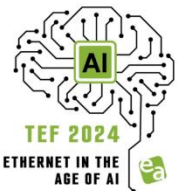
Initially one project it became a two-step approach

- Step 1 - IEEE Std 802.3dfTM-2024 - 400 Gb/s and 800 Gb/s Ethernet
 - Defines the 800 GbE MAC
 - Defines copper and optical PMDs based on mature 100 Gb/s per lane signaling technology
 - Approved 15 Feb 2024
- Step 2 - IEEE P802.3dj 200 Gb/s, 400 Gb/s, 800 Gb/s, & 1.6 Tb/s Ethernet Project
 - Defines the 1.6 TbE MAC
 - Define electrical / optical signaling based on new 200 Gb/s per lane signaling technology
 - Define copper and optical PMDs based on $\geq$ 200 Gb/s per lane signaling

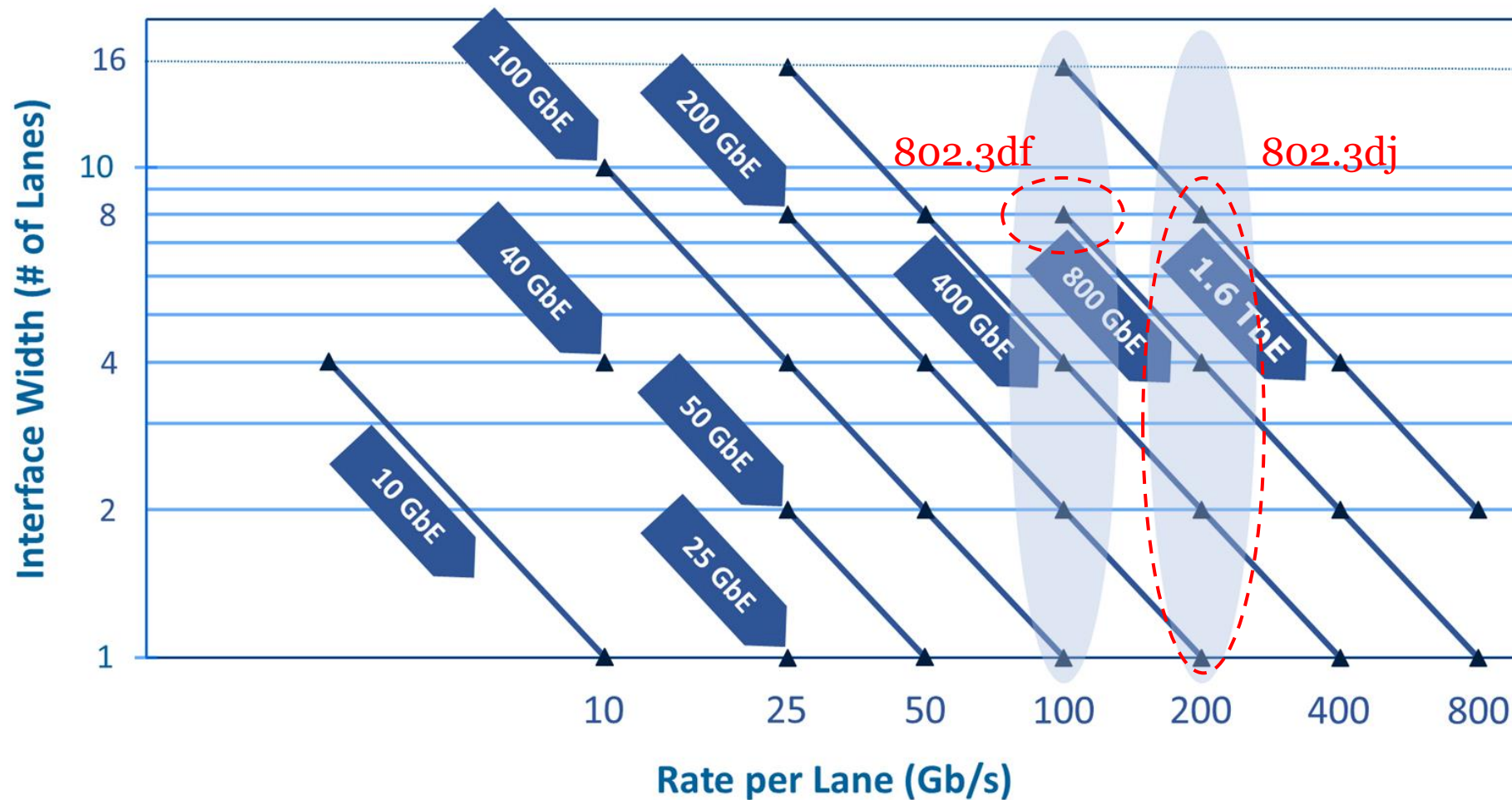
complete

In progress

10/28/2024



Ethernet's High-Speed Solutions - reuse, reuse, reuse



P802.3dj: Scope of the project

Objectives: set by assessing market requirement and technology potential

Key themes to notice:

- 200 GbE, 400GbE, 800GbE and 1.6TbE Ethernet interfaces (N x 200Gb/s)
 - Consideration of Breakout use-case
- Electrical interfaces based on 200 Gb/s
 - Copper Cables
 - Backplanes
 - Chip to module (AUI) interfaces
- Optical Interfaces
 - Only single mode fiber (Duplex and Parallel)
 - 500m to 10km using 200 Gb/s per lane IMDD technology
 - 10km to 40 km using 800 Gb/s per lane (16QAM) coherent technology

IEEE P802.3dj Objectives

Ethernet Rate	Signaling Rate	AUI	Back-plane	Copper Cable	Single-mode Fiber				
					500m	2km	10km	20km	40km
200 Gb/s	200 Gb/s	200GAUI-1 C2C C2M	200GBASE-KR1	200GBASE-CR1	200GBASE-DR1	200GBASE-DR1-2			
400 Gb/s	200 Gb/s	400GAUI-2 C2C C2M	400GBASE-KR2	400GBASE-CR2	400GBASE-DR2	400GBASE-DR2-2			
800 Gb/s	200 Gb/s	800GAUI-4 C2C C2M	800GBASE-KR4	800GBASE-CR4	1.800GBASE-DR4 2.800GBASE-FR4-500	1. 800GBASE-DR4-2 2. 800GBASE-FR4	800GBASE-LR4		
	800 Gb/s	Electrical Interfaces PAM4			Optical IMDD Interfaces PAM4		800GBASE-LR1	800GBASE-ER1-20	800GBASE-ER1
1.6 Tb/s	100 Gb/s	1.6TAUI-16 C2C C2M					Optical Coherent Interfaces 16QAM		
	200 Gb/s	1.6TAUI-8 C2C C2M	1.6TBASE-KR8	1.6TBASE-CR8	1.6TBASE-DR8	1.6TBASE-DR8-2			

What Will P802.3dj provide?

Flexibility of implementations

- Copper cables
- Multimode Fiber – 100m (future project)
- Single Mode Fiber inside DC – 500m & 2km
- Single Mode Fiber Campus – 10 km
- Outside plant, DCI – up to 40 km

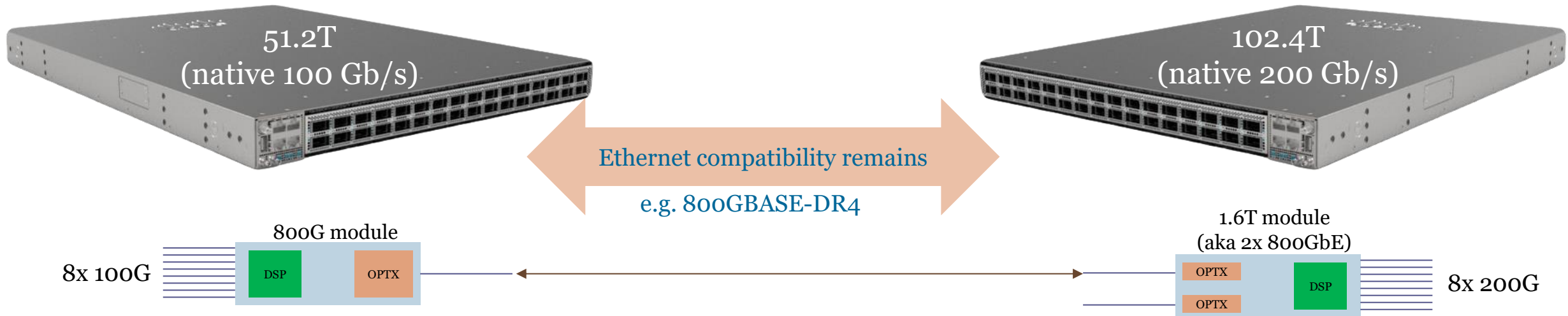
Compatibility with known implementations was a strong factor in P802.3dj decisions

- Compatible with current CMOS
- Compatible with popular system configs or pluggable modules



What Will P802.3dj provide?

Compatible with legacy systems: Maintaining Ethernet's backwards compatibility



P802.3dj's logic and architecture ensures compatibility with "legacy" systems and implementations.

Any 800 GbE capable port will operate with any other 800 GbE capable port:

- As long as the same PMD interface is used
- For example, 800GBASE-FR4 in both ports

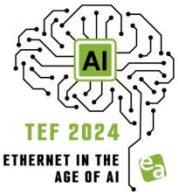
Inside P802.3dj

Architecture

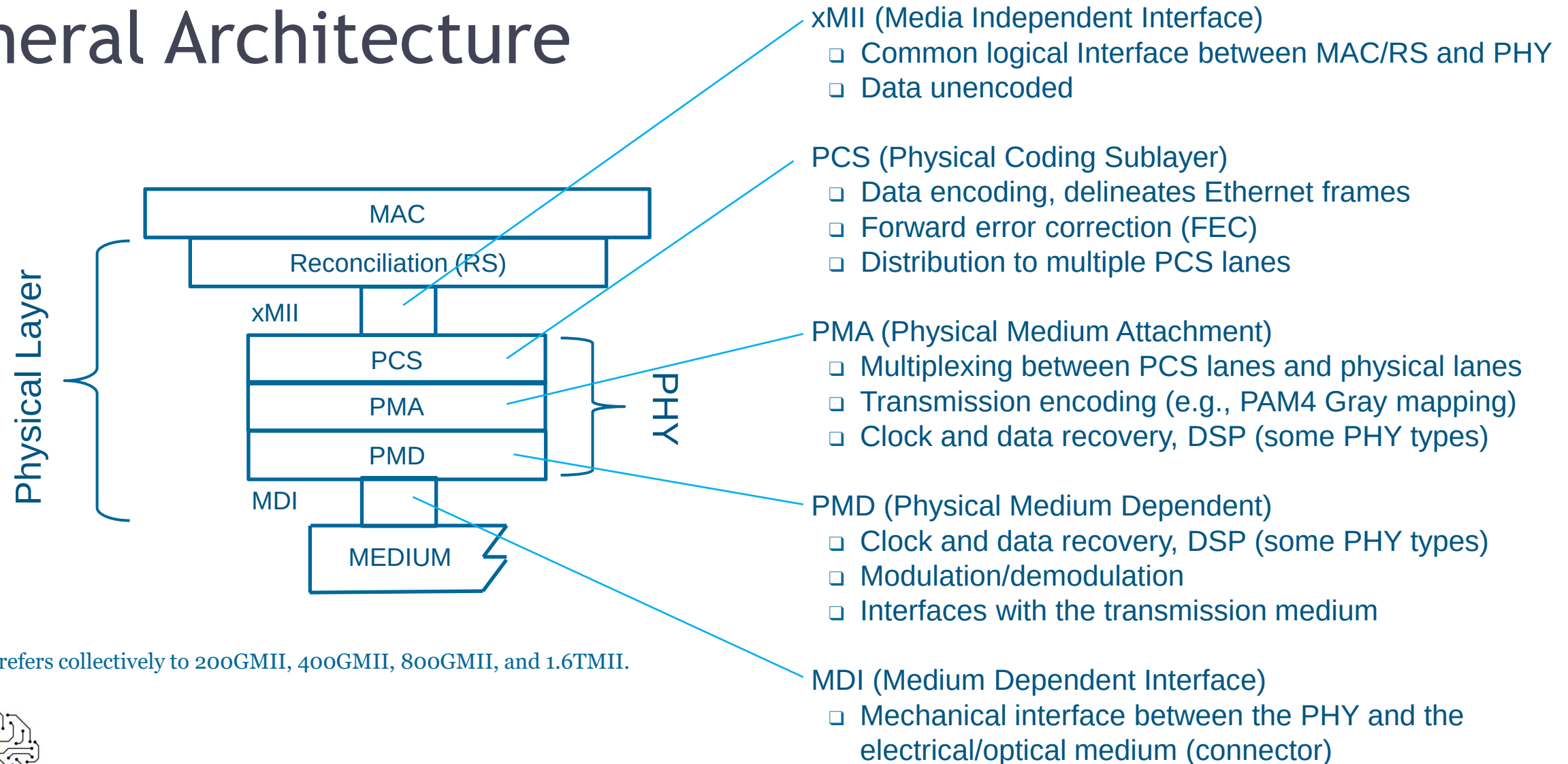
- Maintain Ethernet architecture and compatibility
- Extend architecture to support new capabilities

Technology

- 200 Gb/s is hard!!
- New FECs
- New optimizations and improvements

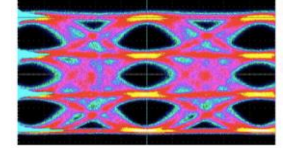


General Architecture



xMII refers collectively to 200GMII, 400GMII, 800GMII, and 1.6TMII.

Key elements of 200 Gb/s technology



Compatible with current CMOS, SiPhotonics, EML, Cable, Fiber technology

**200 Gb/s
Electrical and
Optical
technology
(PAM4)**

**New FECs and
compatible
FEC
architectures**

Multiple FEC schemes for different reaches. Common architecture to enable compatible implementations

Implementations can use same technology for 8x, 4x, 2x or 1x variants

**Re-use
compatible
technology**

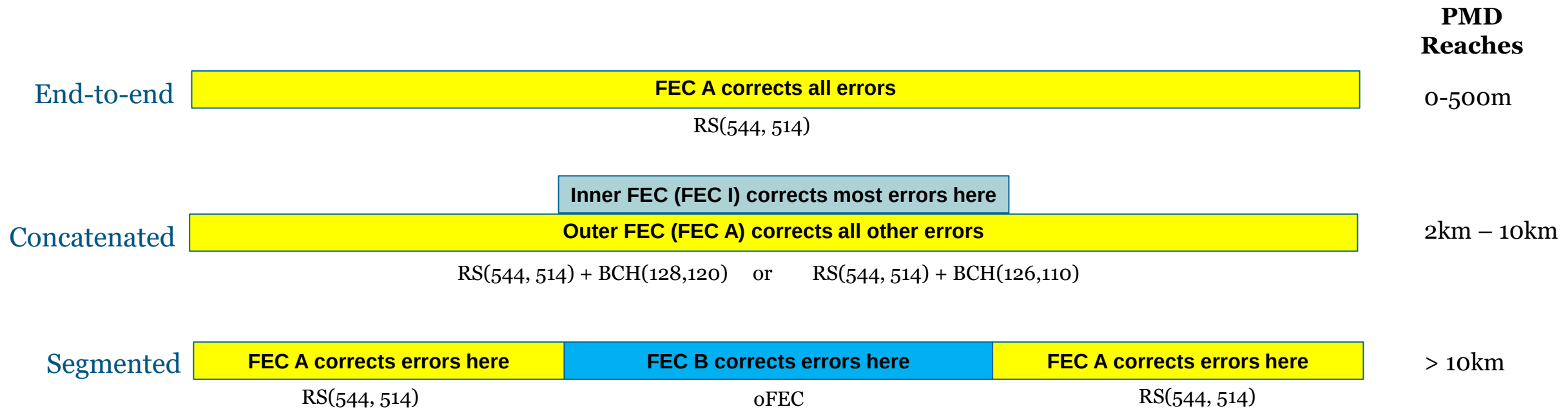
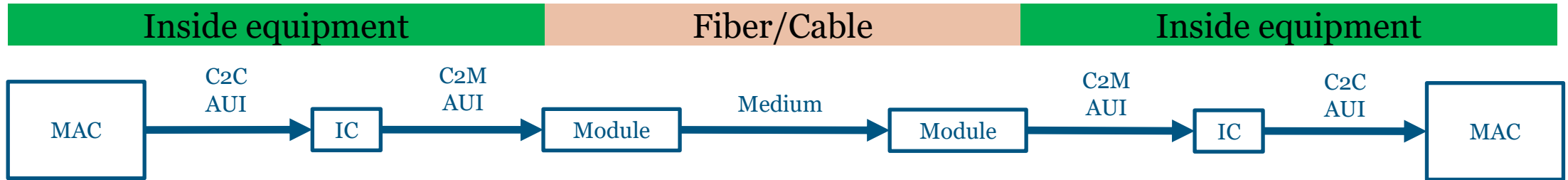
**Advanced link
operation**

New (optional) Optical Link Training Updated and proposed Auto-negotiation

Each speed step is more challenging!

P802.3dj: Multiple FEC Schemes Needed

C2M = chip-to-module
C2C = chip-to-chip



Impact on latencies for P802.3dj PHYs

Data Rate	Cu/IMDD reaches	Latencies (ns)					
		RS-FEC	Additional Codeword interleaving*	BCH FEC	Interleaving	Total	Media (per m)
200 GbE	0-500m	88.2	51.2	-	-	139.4	~5
	2km-10km			23.5	256	418.9	
400 GbE	0-500m	62.6	25.6	-	-	88.2	
	2km-10km			23.5	128	239.7	
800 GbE	0-500m	62.6	-	-	-	62.6	
	2km-10km			23.5	51.2	137.3	
1.6 TbE	0-500m	49.8	-	-	-	49.8	
	2km-10km			23.5	25.6	98.9	

* (burst error mitigation)

Ref: https://www.ieee802.org/3/dj/public/adhoc/optics/0423_OPTX/brown_3dj_optx_01b_230413.pdf
https://www.ieee802.org/3/dj/public/adhoc/optics/0423_OPTX/he_3dj_optx_02_230427.pdf

200+ Gb/s per lane comes with challenges!

Electrical Interfaces

- At 200 Gb/s, the Chip-to-Module interface is as complicated as a 100 Gb/s backplane interface
- Need to find margin in every place possible to meet budget
- Update methodology and advanced equalization
- Ensure equipment can be built

IMDD Optical

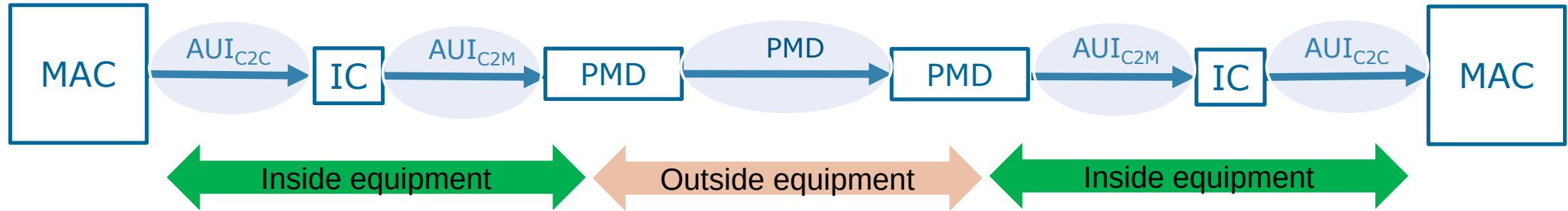
- Optimize per reach
- Optical channel starting to impair (dispersion etc)
- Different solutions for different reaches
 - E2E & Concatenated FEC
 - λ grid
- Update methodologies to address known issues

Coherent Optical

- Leverage technologies from longer-haul applications
- Multiple options for different use-cases
- Concatenated FEC
- Segmented FEC
- Focus on ensuring interoperability

Investigate latest implementations and packaging.

Making 200 Gb/s easy (easier) to use



① **FEC** — Integrated host FEC for all variants — **FEC**

② Extra FEC for different PMDs as needed

FEC — **FEC**

③ — Auto-negotiation between devices (Cu) —

④ — Electrical and optical link training between devices —

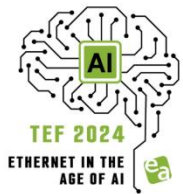
NEW

Summary

- IEEE P802.3dj is in flight, estimated Standard publication: 2H 2026
- New foundational industry development of 200 Gb/s technology
 - PAM4 lives for another day
- Numerous challenges and new technology being developed to overcome and maintain the same expectations that an Ethernet standard provides:
 - Guaranteed interoperability
 - Support for existing (and new) use cases
 - Backwards compatibility
- Broad, cross-industry, convergence of Physical Layer experts working collaboratively to solve these challenges as fast as possible.

OIF Addresses The Paths To Next Generation AI

Nathan Tracy
OIF, President
TE Connectivity, Technologist
October 22, 2024



Outline

- New data rate isn't a surprise
- Industry needs have evolved
- Will this continue at next rate?
- What are the levers we can consider for a next rate
- Make decisions carefully - consequences on end users and suppliers
- An initial thought on feasibility
- OIF 448G Framework Project

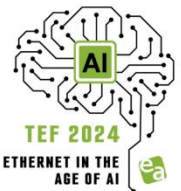
New Data Rates are Not a Surprise

OIF Name	Rate per pair	Year Published	Activities that Adopted, Adapted or were influenced by the OIF CEI
CEI-224G	224Gbps	2024	Several channel reach projects in progress, kicked off in 2022
CEI-112G	112Gbps	2022	Several channel reach projects are complete, one channel project is in progress
CEI-56G	56Gbps	2017	InfiniBand, T11 (Fibre Channel), Interlaken, ITU
CEI-28G	28 Gbps	2012	10GFC, SATA 3.2, SAS-4, 10GBASE-KR4, CR4, ITU
CEI-11G	11 Gbps	2008	10GBASE-KR, 10GFC, 16GFC, SAS-3, RapidIO
CEI-6G	6 Gbps	2004	4GBASE-KR, 4GBASE-KR4, InfiniBand DDR, SATA 3.0, SAS-2, RapidIO v2, HyperTransport 3.1, Interlaken, ITU
SxI5	3.125 Gbps	2002-3	Interlaken, FC 2G, InfiniBand SDR, XAUI, 10GBASE-KX4, 10GBASE-CX4, SATA 2.0, SAS-1, RapidIO v1, ITU
SPI4, SFI4	1.6 Gbps	2001-2	SPI-4.2, HyperTransport 1.03

**448Gbps
Conversations
Have Started**

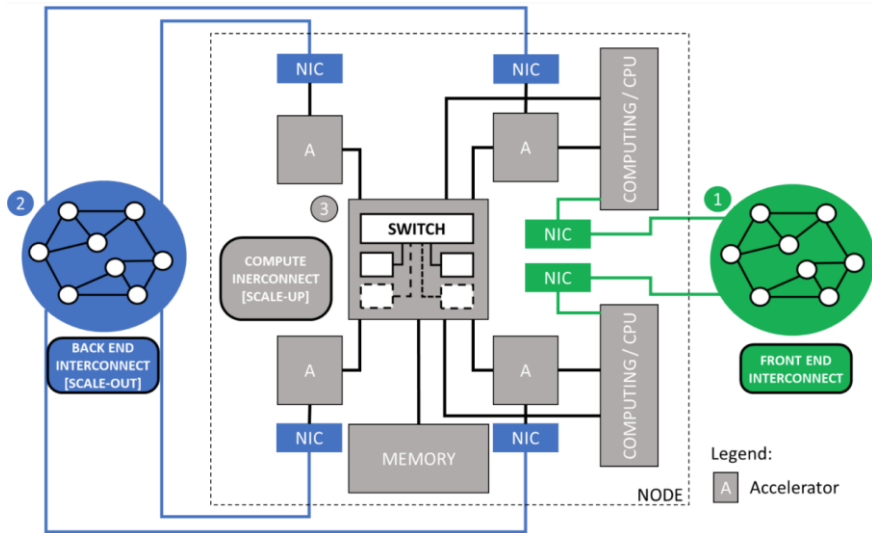
OIF Process/Tools

- Workshops (50Gbps, 100Gbps)
- Framework Projects
- Network Operators/PLUG sessions
- Interops (OFC, ECOC)
- Broad membership including entire ecosystem
- Outreach for requirements



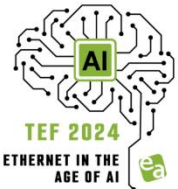
OIF Energy Efficient Interconnect Project, Requirements Document

<https://www.oiforum.com/wp-content/uploads/OIF-EEI-Requirements-RD-01.0.pdf>



“Requirements” Project completed in less than 6 months

Each ‘network’ within the Network of ‘networks’ has unique requirements



Improving energy efficiency is a critical goal of OIF’s EEI activity

Front End Interconnect

Scale Out Interconnect

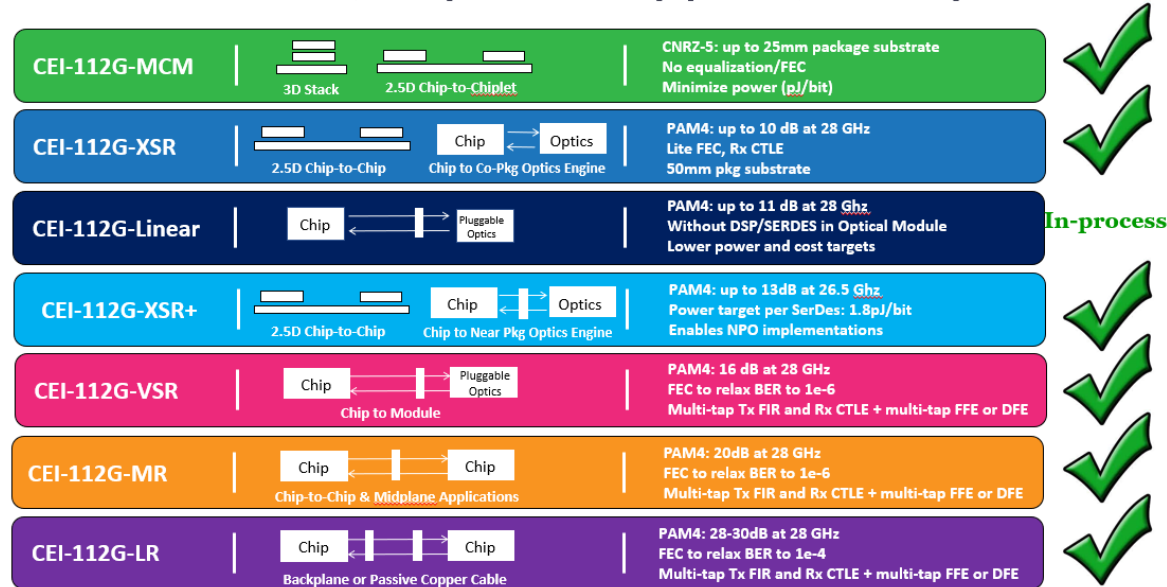
Compute Interconnect

Requirement	Data Center Network / Front end network to AI clusters	High-Performance Computing and AI GPU interconnect	Large Compute Node/ Compute Disaggregation
	① Front End Interconnect	② Back End Interconnect	③ Compute Interconnect
Description of use case	This use case describes the optical interconnections that are used to ease servers communications. (Server to servers, servers to storage, servers to edge, etc.) This is about how servers communicate with the outside world. They may be general purpose datacenters servers or specialized servers for AI, analytics, HPC systems etc.	Specialized interconnect for AI intra-cluster Accelerator communication, HPC intra-cluster communication, and others. This is to scale-out the computing infrastructure. Going to 1,000-10,000 nodes. This use case describes the optical interconnections that are used to ease Accelerator communication to execute the AI models or train the models. (Accelerator refers to GPU, TPU, Analog Computers, etc.)	Pooled memory allows multiple nodes to access the same memory reducing the system memory consumption. (Extending to pooled memory and pooled IO)
End user deployment	Next Generation Ethernet Infrastructure. (Front end network)	AI/ML and HPC (Back end network)	Within nodes. Comp-IO, Comp-Mem, Comp-Comp. Note : (The Disaggregated node may be much larger than nodes today.)
Timeframe	NOW (Already in pervasive use)	Emerging Now (2023)	2025-2027
Form factor(s)	Traditional Ethernet traffic (Server to server, server to storage and server to edge communication)	Pluggable, NPO/CPO	Co-Packaging / AOC
Traffic Type	Traditional Ethernet traffic (Server to server, server to storage)	ND	PCIe Express Gen 5/6 and CXL 3.0 or later
Latency	Consistent with End-to-End KP4 FEC or <20ns + d*5ns + (FEC delay)	<5ns + d*5ns + (FEC delay)	Target < 1 us or less write, target < 2 us or less read
End-to-end FEC	KP4, optionally, KP4-IEEE defined concatenated FEC	ND	PCIe + FEC
Data rate	200G / 100G per lane 50Gper lane for backward compatibility 800GbE (Main use case), 1.6TbE for future 400GbE for backward compatibility	400GbE, 800GbE, and future 1.6TbE 200G / 100G per lane 50Gper lane for backward compatibility 800GbE (Main use case), 1.6TbE for future 400GbE for backward compatibility	PCIe Express Gen 5/6
Expectation of capability for the host	Unclear	200G / Lane linear channel for CPO (TBD)	ND
Expectation for compliance (@TP1x, @TP4)	IEEE Standards: SRx, DRx, FRx, etc.	Compliance to TPx to be defined in electrical link	ND
Expectation for compliance (@TP2, @TP3)	IEEE standard 800GE-FR4, 400G-FR4 and 200G-FR4	Compliance with TPx to be defined in optical link	ND
Target IL for link (for optical links)	Unclear	2.5dB, 500m	ND
Degree of interoperability	Full interoperability with all module types and technologies	Fully interoperable with pluggable modules desirable, but depends upon power/spec trade-off.	ND
Universal host requirement	ND	ND	PCIe, UCIe
Energy efficiency for link	<10pj/bit for OE+Laser	<<4pj/bit for OE+Laser	ND
Energy efficiency for link + host	ND	ND	ND
Power saving target (modules only)	ND	ND	ND
Power saving target (end-to-end)	ND	ND	Target < 75% current implementation
Density (Beachfront, Areal)	>1Tbps/mm (End goal in a complete CPO evolution)	>1Tbps/mm	
Link accountability (out-of-service)	System and module parameter measurements will result in a determination of the location of the fault.	System and module parameter measurements will result in a determination of the location of the fault.	ND
Link accountability (in-service)	Easy to identify the link element with issue	Easy to identify the link element with issue	SW assumes reliable link to memory for CPU/GPU Loads/Stores
Link margin requirements and allocation	Sufficient link margin to match pluggable interoperability desirable	Sufficient link margin to match pluggable interoperability desirable, but	ND
Compliance verification	PAM-4, NRZ	Ability to determine compliance of each link element	ND
Modulation format	ND	PAM-4, NRZ	ND
Designated Management Interface	Definition to be developed by CMIS	Definition to be developed by CMIS	ND

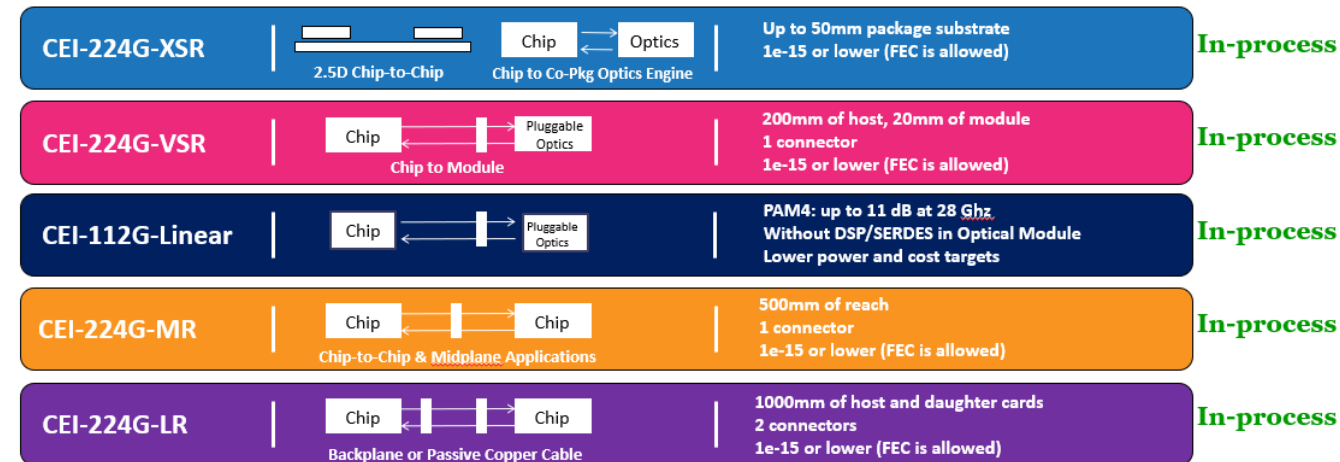
OIF Electrical Channels Supporting Power Optimization

- CEI (Common Electrical I/O) IA (Implementation Agreement) has always tailored electrical SerDes and channel reach specifications to optimize SerDes power

112 Gbps SerDes/Reach Family

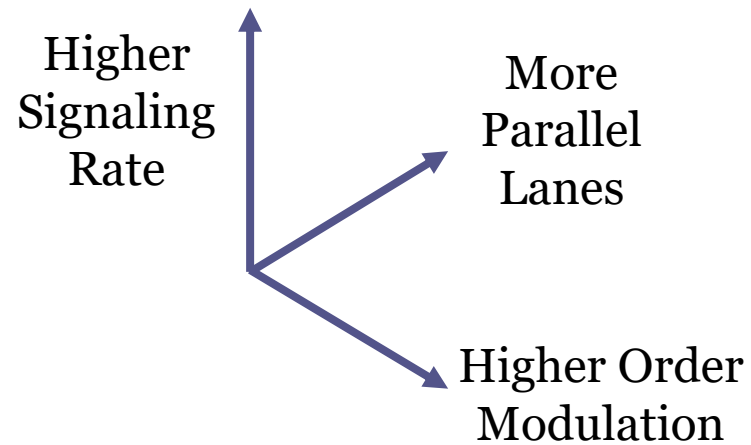


224 Gbps SerDes/Reach Family (so far ...)

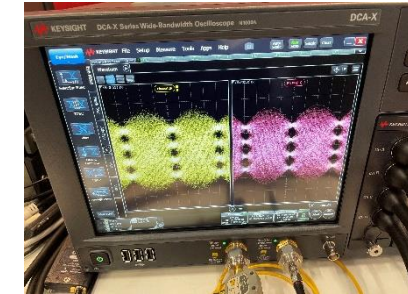
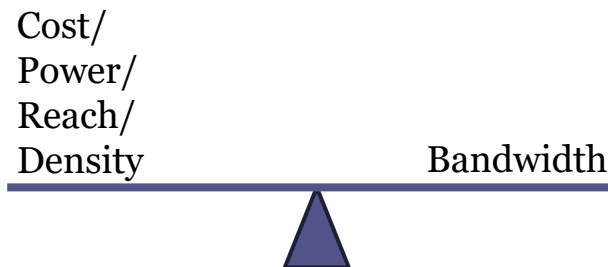


What does 448G mean to you?

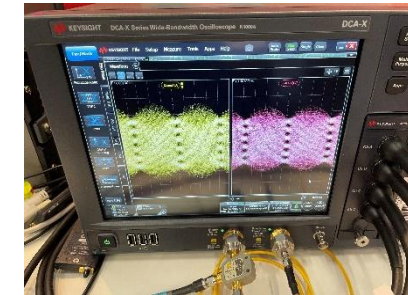
- Some of the 3 or 4 axis to consider



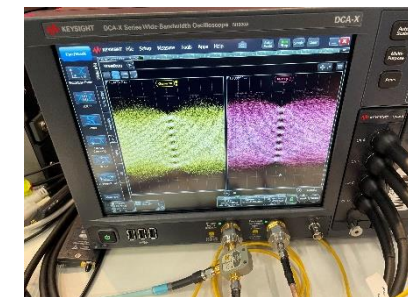
- The tradeoffs - how do we stay in balance



PAM4
112 GHz



PAM6
86 GHz



PAM8
75 GHz

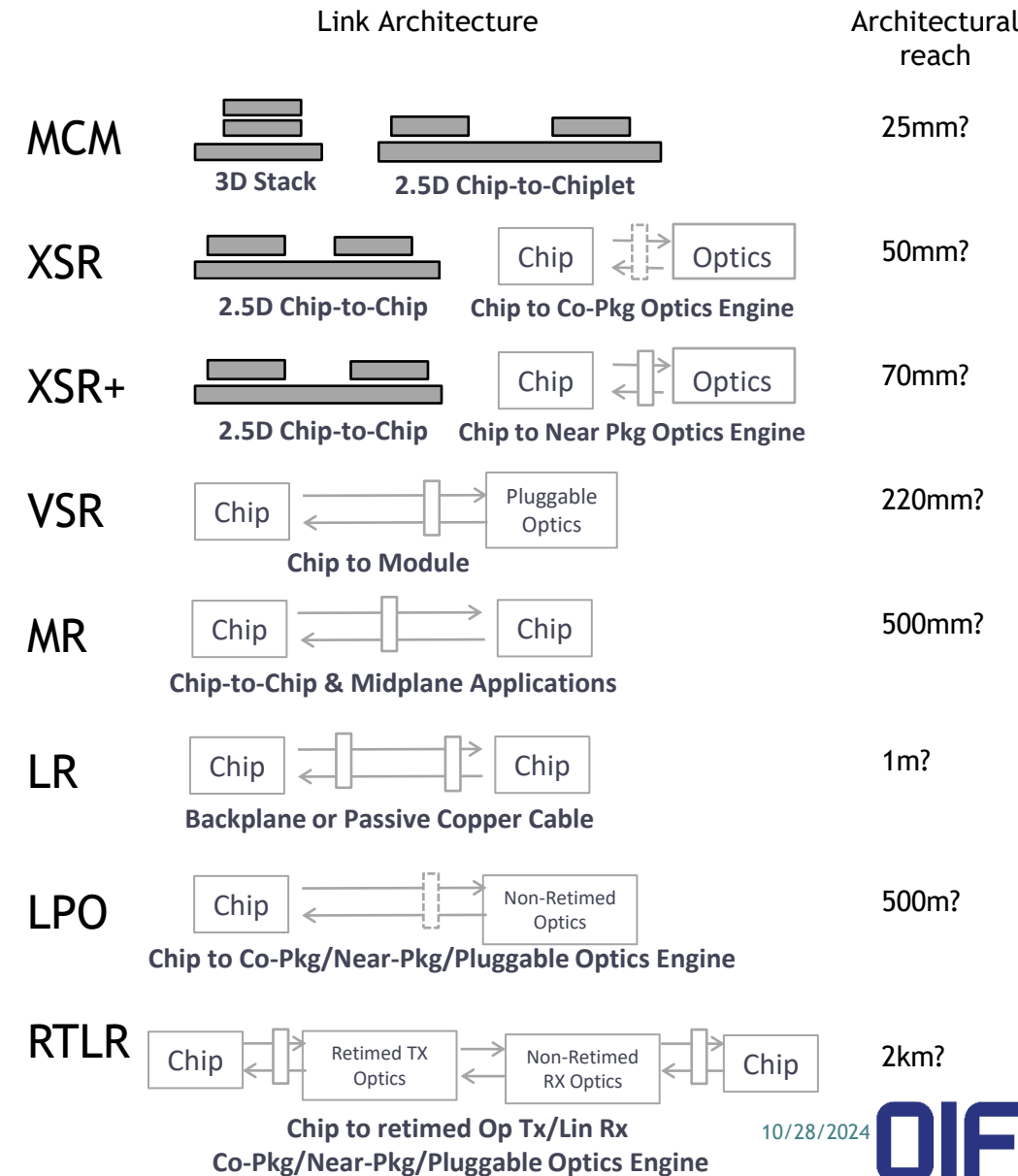
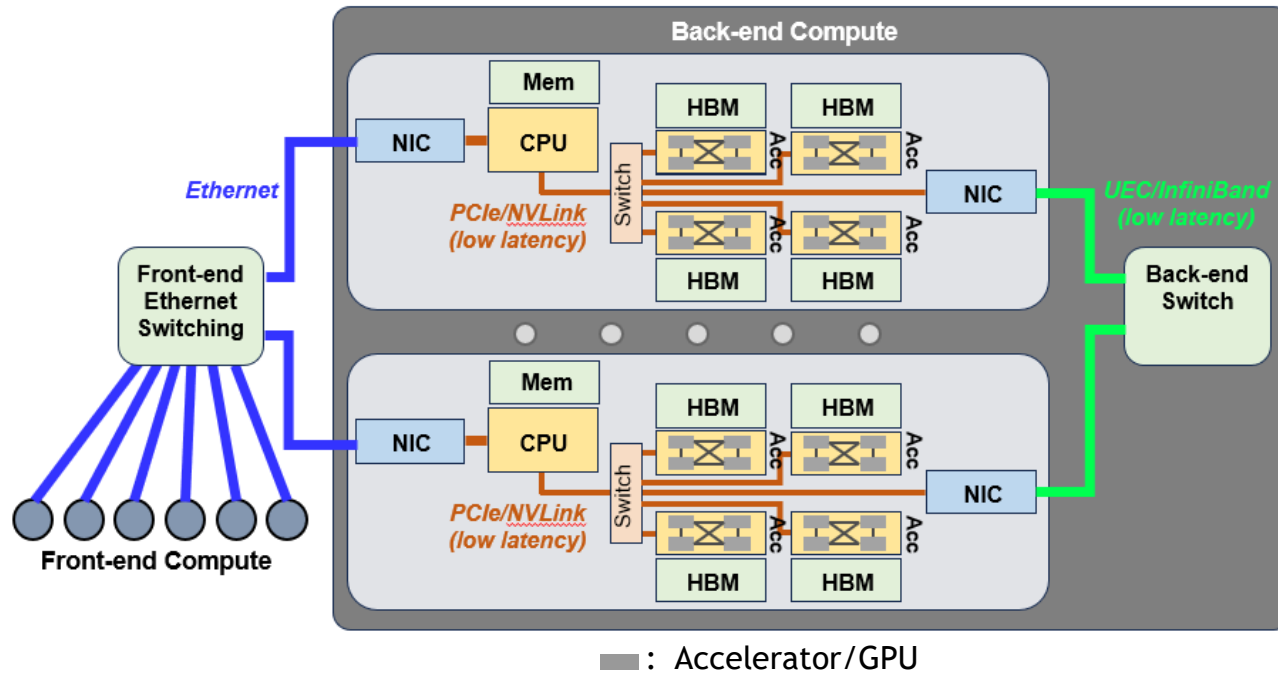


QAM
Bidirectional
DMT

Industry Needs. How Do We Determine Priority?

- Not singular set of requirements, how does we deal with this?
- Evolving needs and methods
- What's the current / 448G requirements?
- Is it possible lower “aggregate cluster power, ie, rack power” comes with higher SerDes power?
 - i.e., choose a modulation that requires more silicon power, but enables longer reaches and more electrical links (less optical links)
- Long electrical links, high density, managed latency, power management. What's the goal(s)?

Architecture Drives Needs or Vice Versa?



Industry Focus Today: AL/ML Requirements

A current voice:

Next Gen Large-Scale AI Cluster

- GB200 Rack Scalable Unit
- 72x B200 GPUs interconnected with both IB and NVLink5
 - Rack acts as one large GPU
- GPU Connectivity, use both:
 - Network: 1x 800G IB/Eth per GPU **over optics**
(same ratio as before, 2x the BW)
 - GPU Fabric: 7.2Tb/s NVLink5 per GPU **over copper**
- If NVL5 interface was retimed optics...
 - = 14.4Tb/s per GPU @ 15W per 800G (retimed)
 - = 20kW per rack
 - = 9 Mega Watts over 32K GPUs



A call to Action:

We must increase reach without a 4x increase in power

Metric	IB/Eth	NVL	HBM3
Bandwidth (TBps)	0.8	3.6	36
B/W Ratio	1	4.5	45
Reach	>100m	1.5m	0.005m
Power Eff. (pJ/b)	20	5	5
Cost Eff. (\$/Gbps)	0.5	0.05	(almost) Free

<15pJ/bit = nice, we'll take it, but doesn't significantly move the needle

<10pJ/bit = drives system design re-think

<5pJ/bit = drives architecture re-think

Slides from OIF at OFC "Energy Efficient Interfaces, Reining in power consumption" presented by Craig Thompson, Nvidia

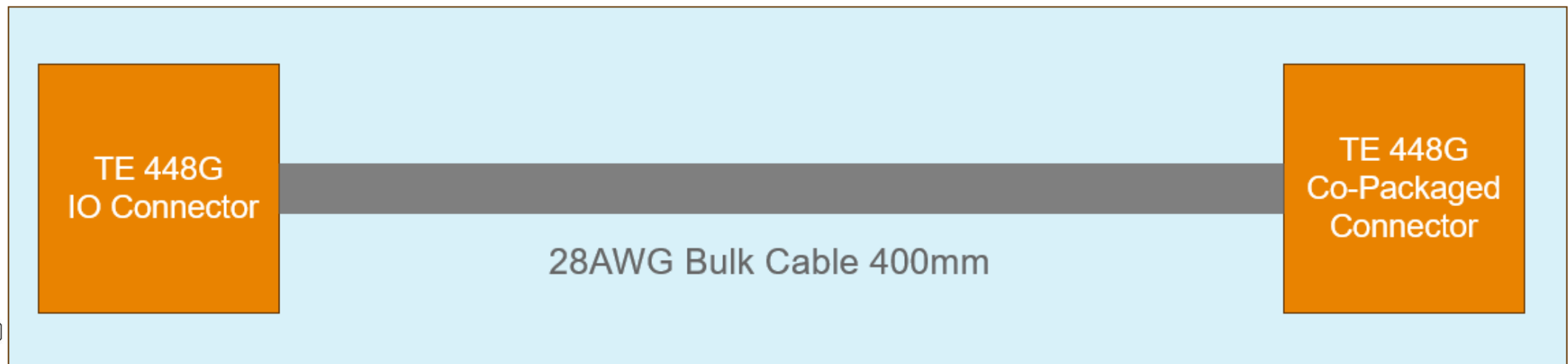
- Focus on electrical channel reach and energy efficient optics

Consequences of Decisions

- Be careful what you ask for, you may be forced to live with it
- Balanced pain (everyone must have shared pain, but get it in balance)
- How do we determine priority of requirements?
- Need to ensure we have operator's voice included

Technical Feasibility for 448G Electrical

- Can we get to the faceplate? First indications are positive
- Preliminary concept work at TE shows a path to 15 inches, reasonable VSR chip-to-module channel loss, based on PAM6 or PAM8 signaling
- Depending on modulation, copper cable IO could be viable

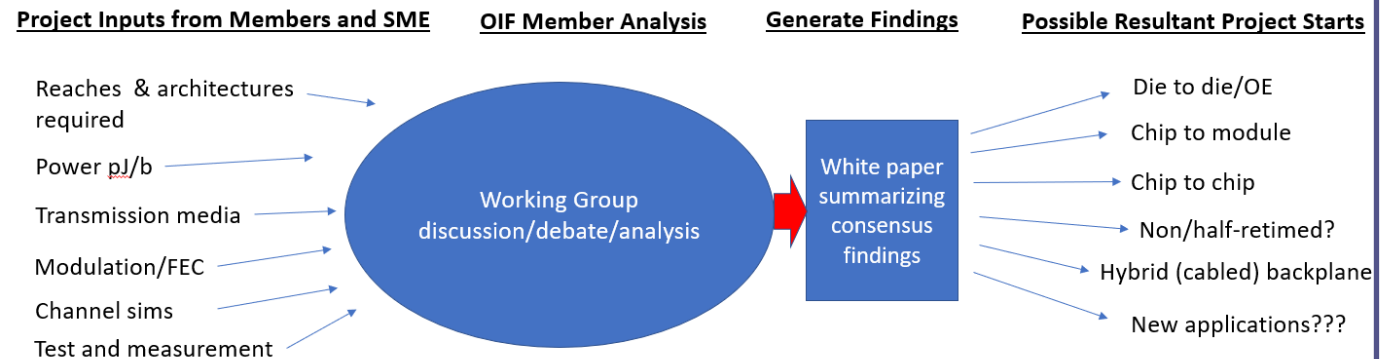


OIF 448 Gbps Framework Project is Kicked Off

- Opportunity to establish a common industry baseline on Needs, Opportunities and Challenges
- Debate the multiple tradeoffs
- Work to establish a prioritization of Requirements
- Kick off 448G projects

Expected Output

The intended output of this project is to build a body of knowledge that will be summarized in a technical white paper which will enable several project starts leading to future CEI clauses addressing specific reaches and architectures



OIF

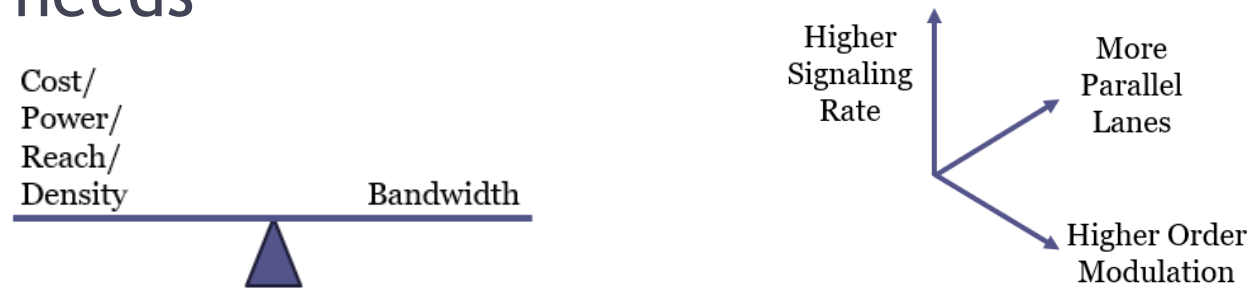
Copyright © 2024 OIF

oif2024.419

5

Summary

- OIF has a methodology
- OIF membership has the necessary broad cross section of industry stakeholders
- 448Gbps will bring bigger challenges than we have seen in the past
- Requires the whole ecosystem's inputs
- More than ever before, requires absolute focus on developing the prioritized needs

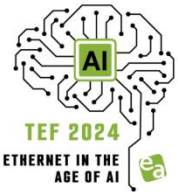


Other OIF Speakers at the TEF: Cathy Liu, Mike Li, Jeff Hutchins

UltraEthernet

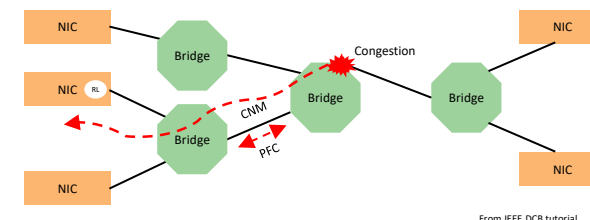
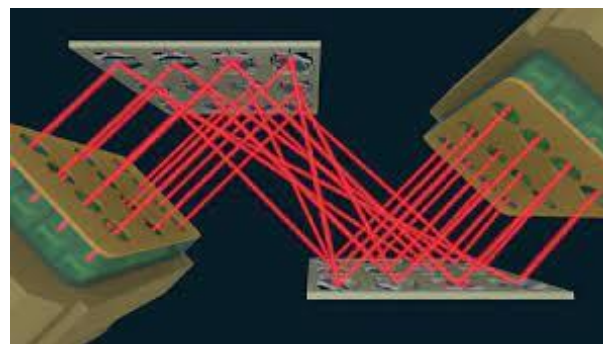
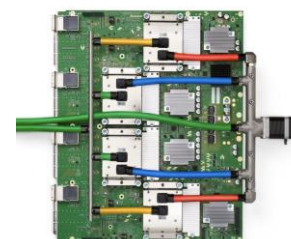
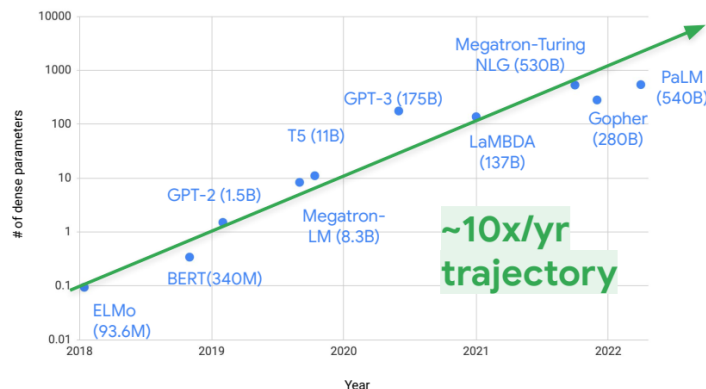
Introduction and proposed Ethernet Enhancements for AI

Uri Elzur, UEC Technical Advisory Committee Chair



This presentation has been developed within the Ultra Ethernet Consortium and is provided to the Ethernet Alliance, and is intended to educate and promote the exchange of information. Opinions expressed during this presentation are the views of the presenters, and should not be considered the views or positions of the Ethernet Alliance

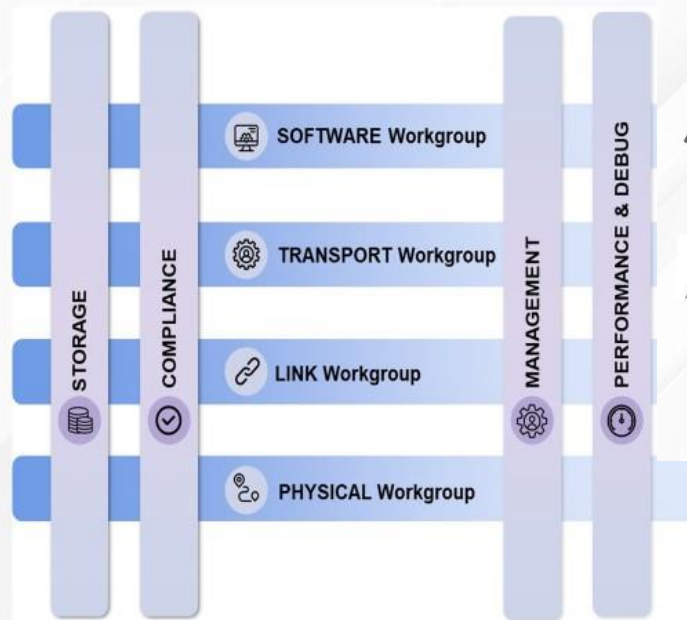
At a Crossroads or maybe a Perfect Storm...?



Ultra Ethernet Consortium

An Ethernet-based, open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale

AMD	ARISTA	BROADCOM
CISCO	EVIDEN an atos business	Hewlett Packard Enterprise
intel	Meta	Microsoft
	ORACLE	

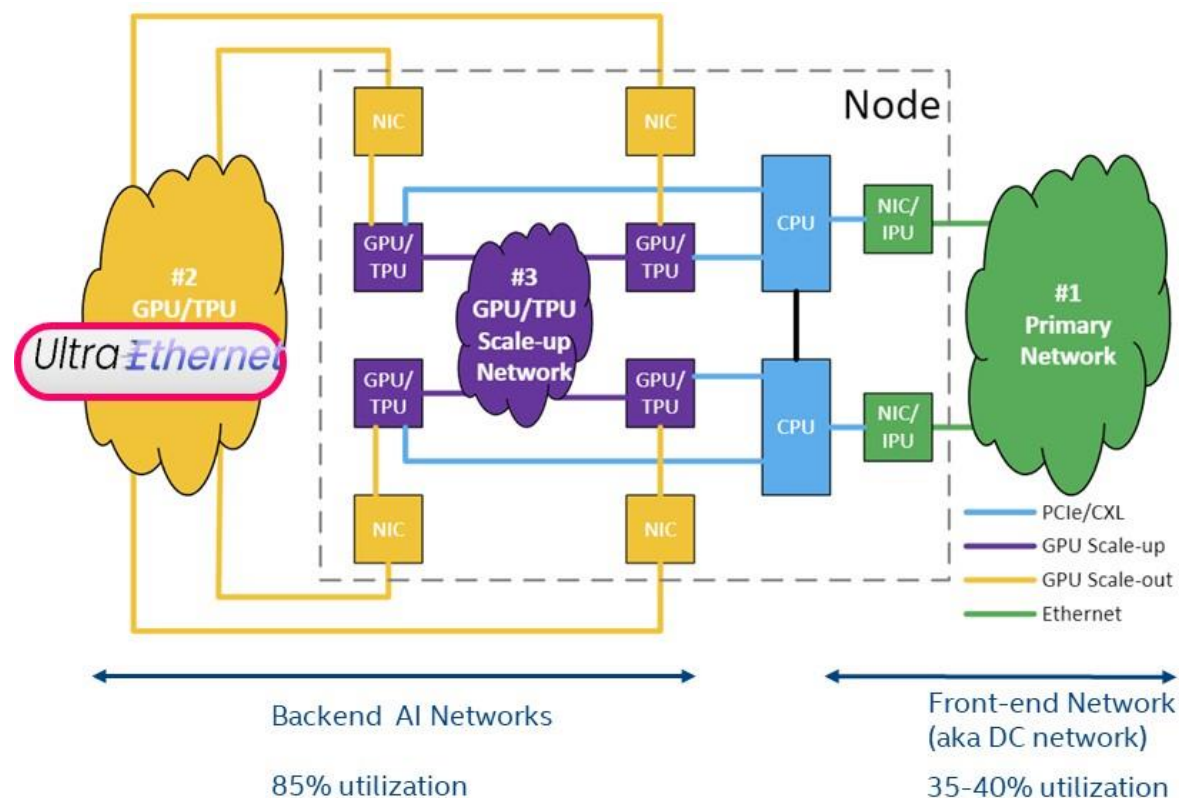


Ultra Ethernet Consortium



>90 member companies
>800 active members

AI/HPC Networks of Interest: Basic Characteristics



Some exemplary characteristics:

1. Primary DC network

- Used by all 3 deployment models
- Main network for some HPC At Scale
- Very large scale: up to 100K-1M Endpoints
- Distance: >150m ; RTT ~100 uS + ; BW/GPU ~10GB/S
- Storage attached e.g., over RoCE RDMA
- Network semantics

2. GPU/TPU Scale-Out Network

- DL/Inference Cluster -10k nodes and ↗
- Distance: <100m ; RTT <10 uS + ; BW ~100GB/S
- Main network for some HPC At Scale
- Network semantics

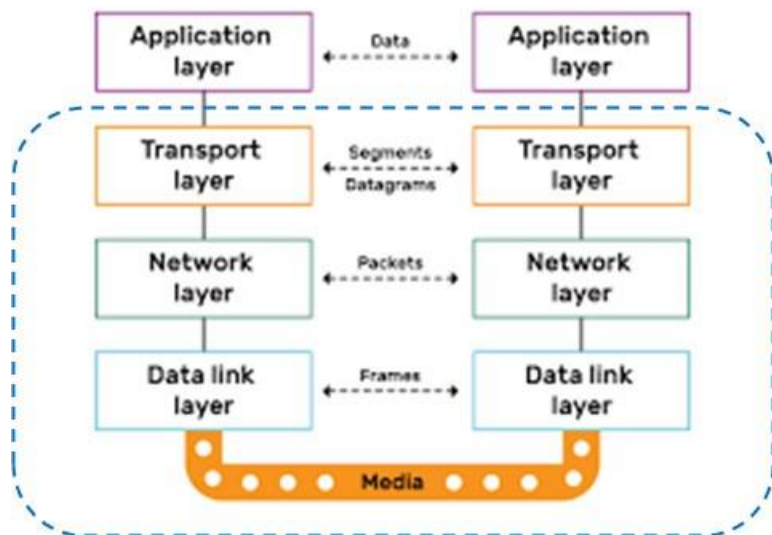
3. GPU/TPU Scale-Up Network

- Within a node; small scale e.g., 256 XPU?
- Distance: ~1m ; RTT ~1 uS + ; BW ~1200 GB/S ↗
- Direct connect and/or switched
- Memory and Network semantics

UltraEthernet V1.0 targets the Scale Out Network

- ❖ UEC v1.0 leverages the IEEE 802.3 2022 100G per lane signaling.
- ❖ UEC is highly interested in 200G per lane signaling for the high bandwidth density required. It will be incorporated into UEC next gen specs.
- ❖ Some optimizations/adaptation may be requested (see other presentations here at TEF)

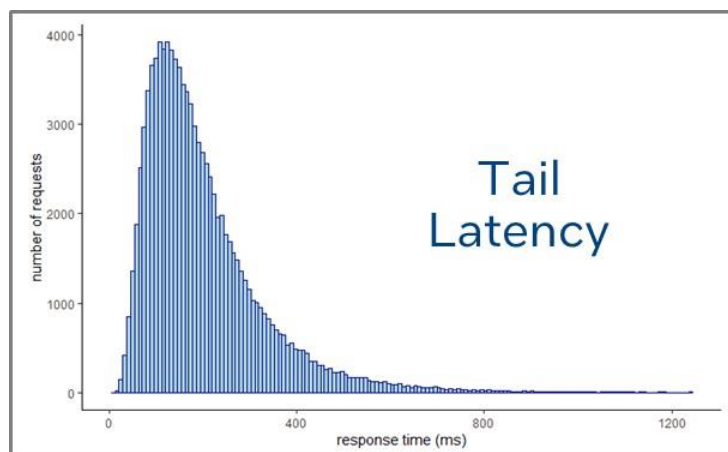
AI/HPC Common Requirements



- Transport primitives for
 - Large Scale
 - Multi pathing
 - Relaxed ordering
 - Modernized Congestion Control
 - Optimized RDMA
 - Performance – bandwidth, latency, tail latency, Packets/S
 - High network utilization
 - Stability and Reliability



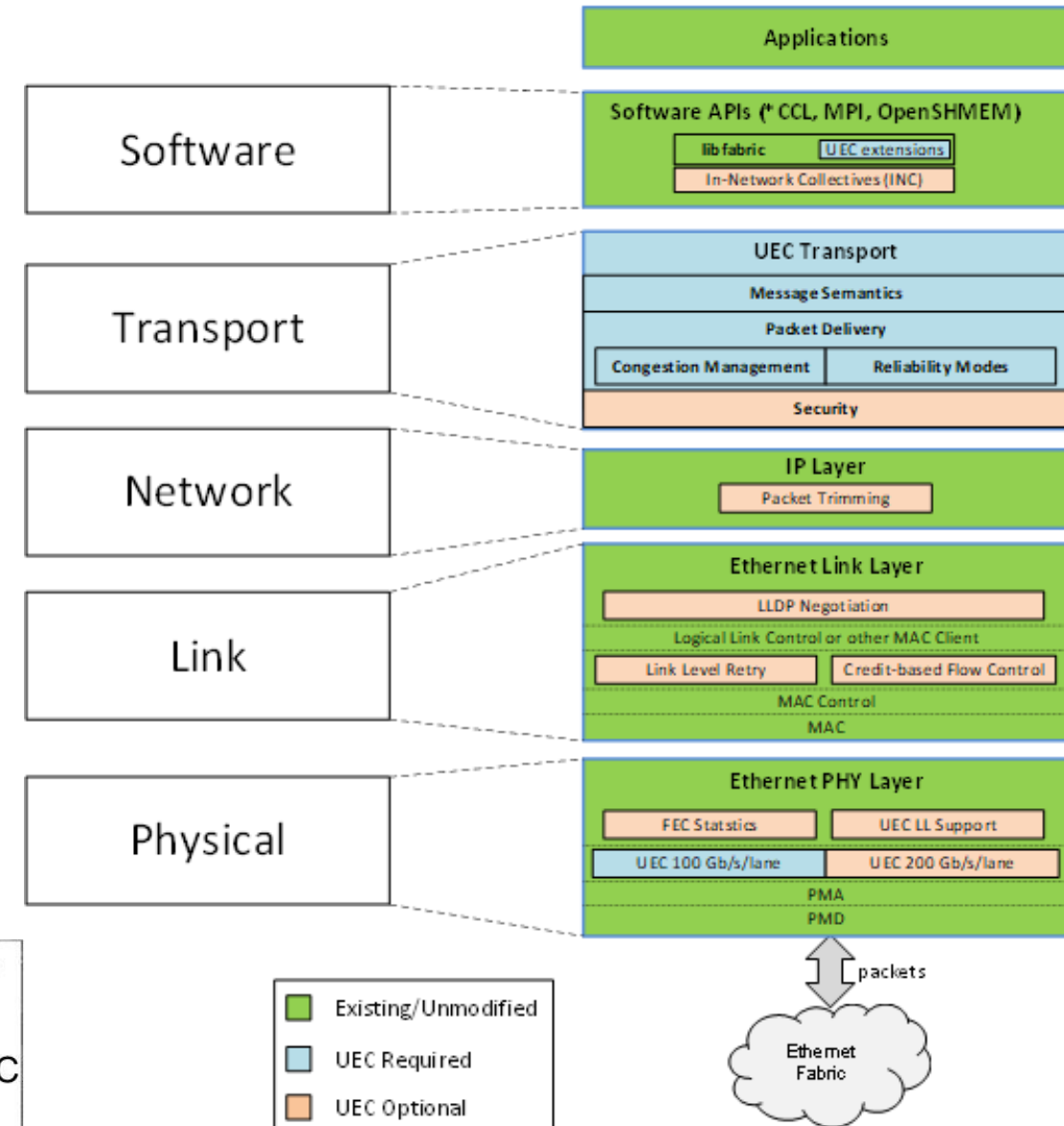
Key goals: high utilization & low tail latency!



UEC Stack Overview

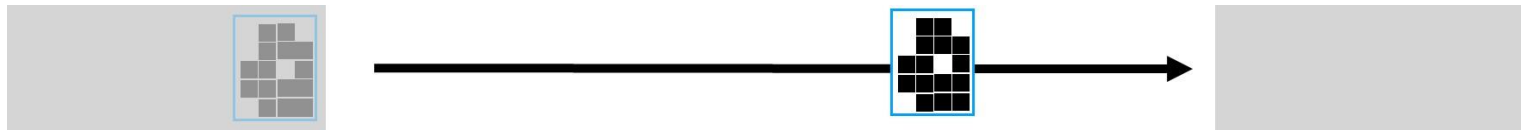
(partial feature list)

- Software API
 - Libfabrics 2.0 with extensions
- New Transport Layer
 - Multi-pathing
 - Packet spraying
 - Ordered (ROD) and un-ordered (RUD)
 - Lossy (no PFC) or Lossless
 - Congestion Control: Enhanced Tx and new Rx
 - Trimming
 - In Network Collective
- Network Layer
 - IP v4/v6
 - ECN
- Data Link Layer
 - Negotiation – LLDP
 - Link Level Retry - LLR
 - Credit Based Flow Control - CBFC
- Physical Layer
 - IEEE Compliant 100G Signaling (UEC v1.0)
- AI and HPC Profiles



Link-Layer Retransmission

...LLR, affectionately

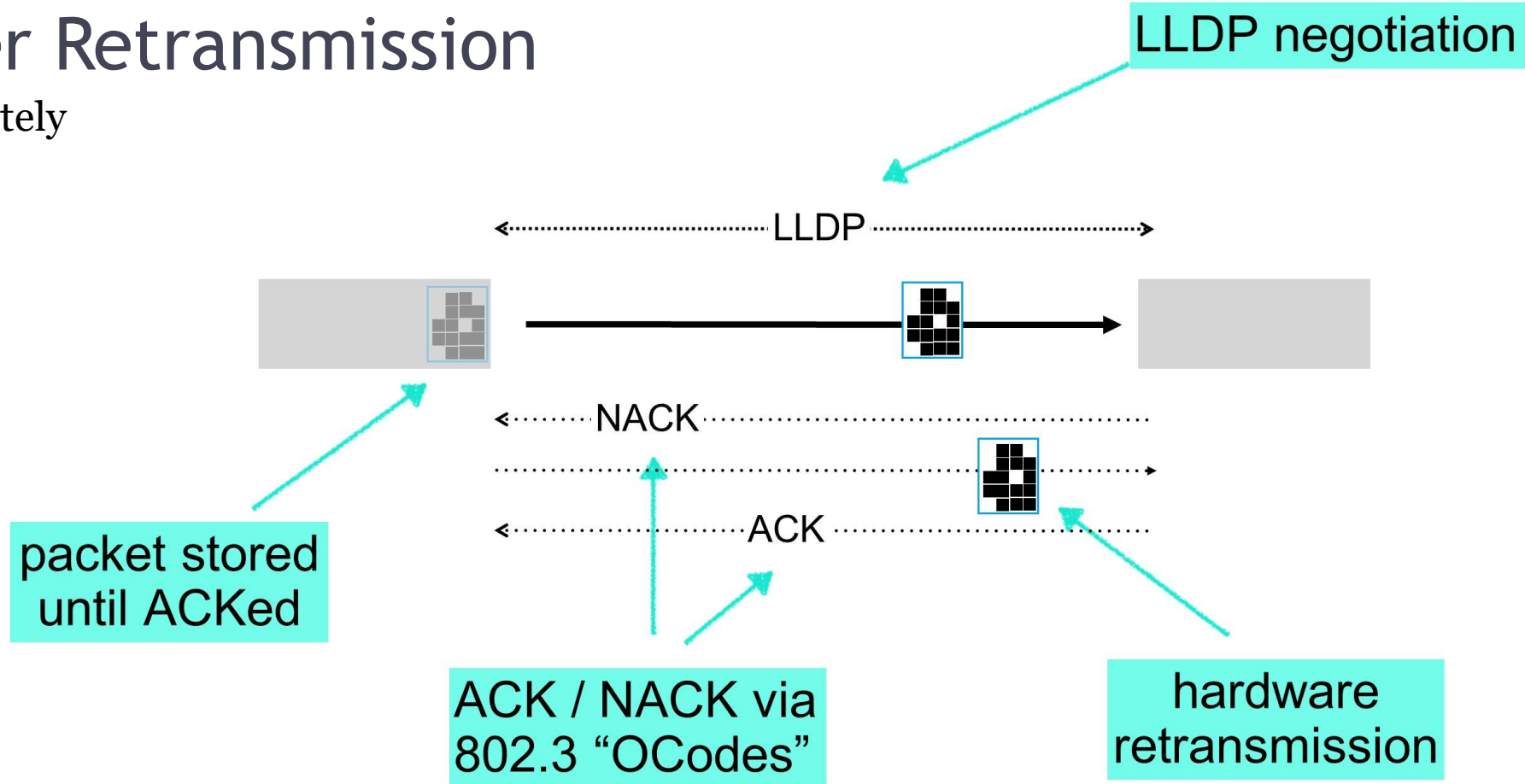


- Link and transceiver failures are a fact – and impact workloads
- An AI/HPC datacenter could have 256,000-512,000 transceivers
- Avoid poor tail latency, driven by a few bad links that trigger slow retransmits

local retransmission for flakey links

Link-Layer Retransmission

...LLR, affectionately



Fastest possible Retransmission – best remedy for potential Tail Latency due to physical layer issues

Modern Transport and RDMA Services Needs for AI and HPC

Requirement	UEC Transport	Legacy RDMA	UEC Advantage
Multi-Pathing	Packet spraying	Flow-level multi-pathing	Higher network utilization
Flexible Ordering	Out-of-order packet delivery with in-order message delivery	N/A	Matches application requirements, lower tail latency
AI and HPC Congestion Control	Workload-optimized, configuration free, lower latency, programmable	DCQCN: configuration required, brittle, signaling requires additional round trip	Incast reduction, faster response, future-proofing
In Network Collective	Built-In	NONE	Faster Collective operation, lower latency
Simplified RDMA	Streamlined API, native workload interaction, minimal endpoint state	Based on IBTA Verbs	App-level performance, lower cost implementation
Security	Scalable, 1 st class citizen	Not addressed, external to spec	High scale, modern security
Large Scale with Stability and Reliability	Targeting 1M endpoints	Typically, a few thousand simultaneous endpoints	Current and future-proof scale

UEC TARGET TIMELINES



- **Call to Action:** Help Optimize Ethernet for AI !
- **PHY:** expect UEC v1.1 to add 200G per lane signaling, 400G per lane signaling requirements (latency per network type, FEC, power), FEC statistics
- **Link Layer:** LLR, CBFC, LLDP updates



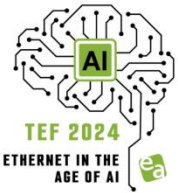
Migration of SFF connectors and modules

Tom Palkert, Samtec
Chair, SFF Transceivers Group

What is SFF?

- SFF (Small Form Factor) is part of SNIA (Storage Network Industry Alliance)
- SFF defines the front panel pluggable solutions for the industry (Transceiver, connector, electrical, pinouts)
 - SFP
 - QSFP
- SFF also defines storage device connectors

Join us!

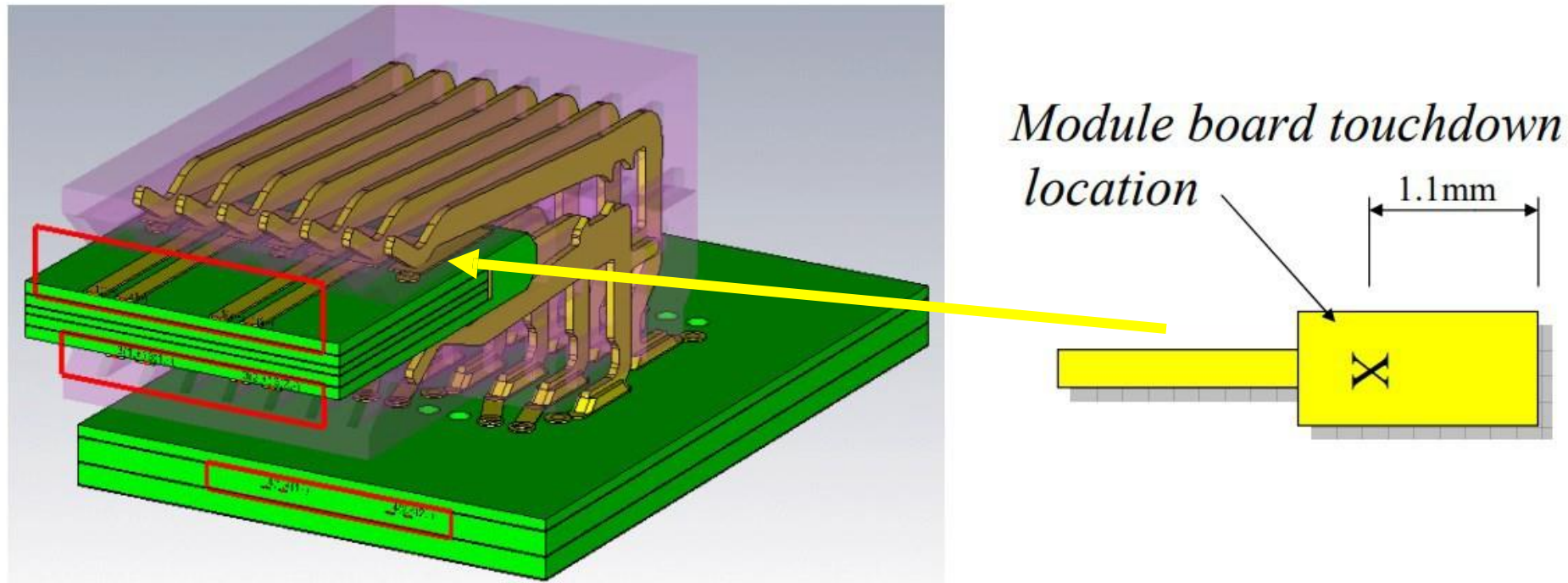


Migration to 400Gbps lane rate

- Pluggables will live on!
- Backwards compatibility may depend on modulation (PAM?)
- PCB routing will be challenging

Pluggable module connector basics

Module has pads made of 'hard' material that is capable of multiple insertions
Host connector has compression 'spring fingers' to contact pads



What limits the BW of a QSFP connector

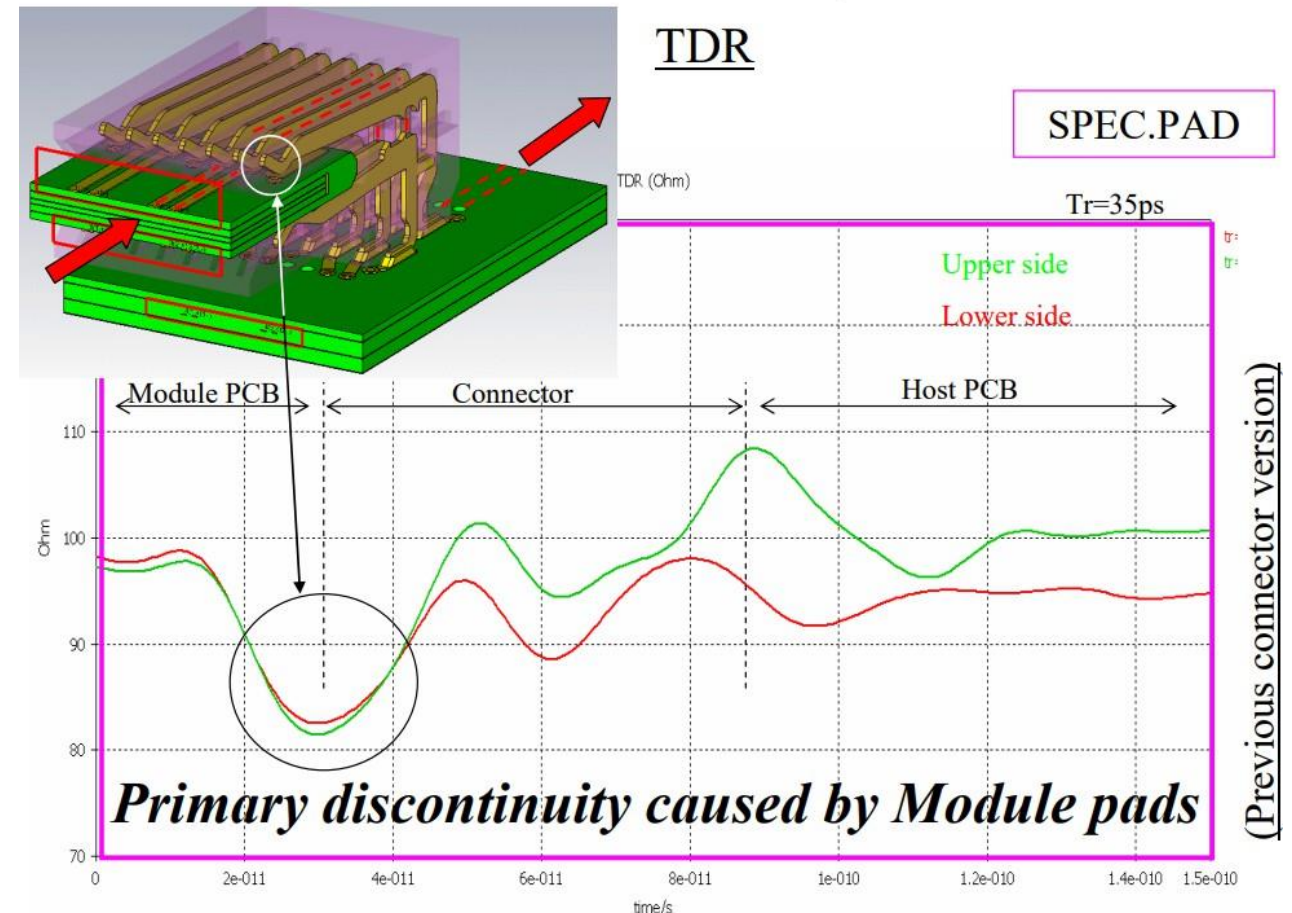
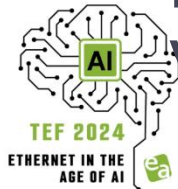
- nishimura_01_0710.pdf

Module pads are reflective due to:

pad wipe give stub
Spring finger has stub
Pad width gives reflection

Host solder pads are reflective due to:

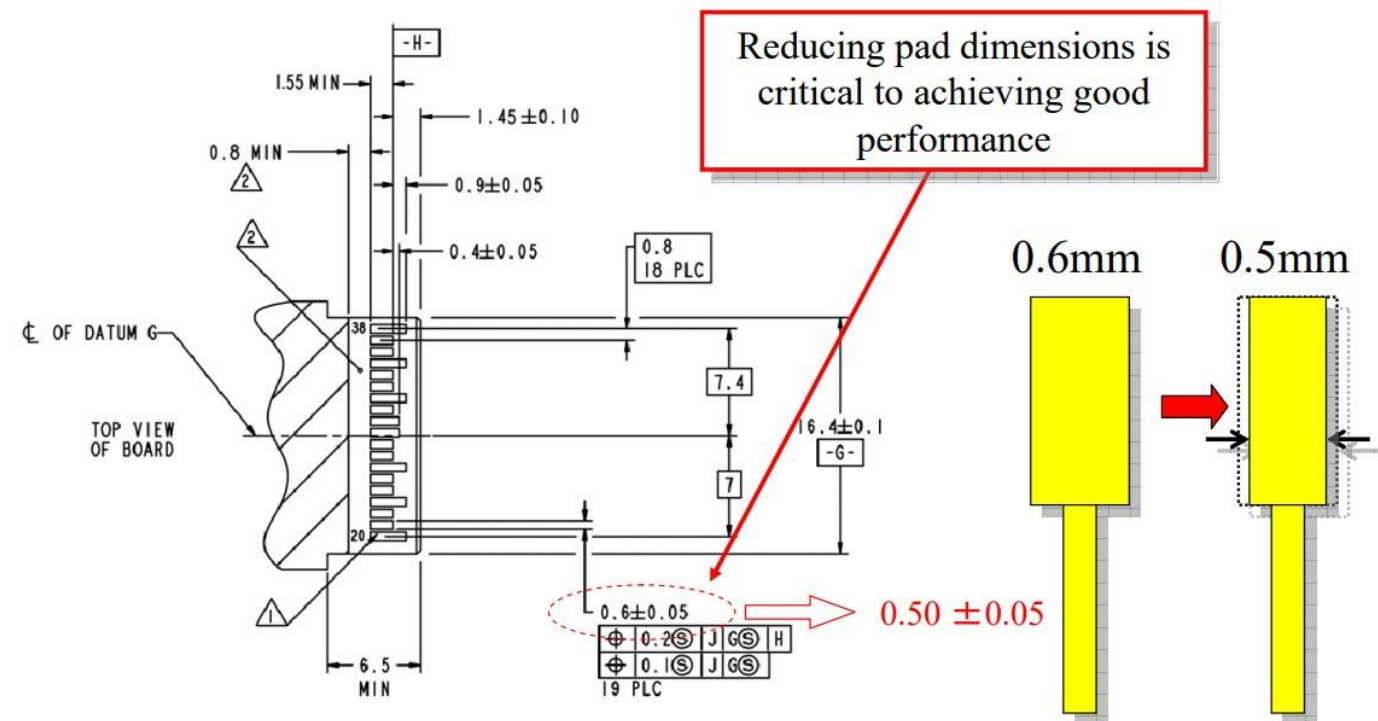
Pad size
Via stub



Moving from 10-25-50-100-200Gbps (Module Pad)

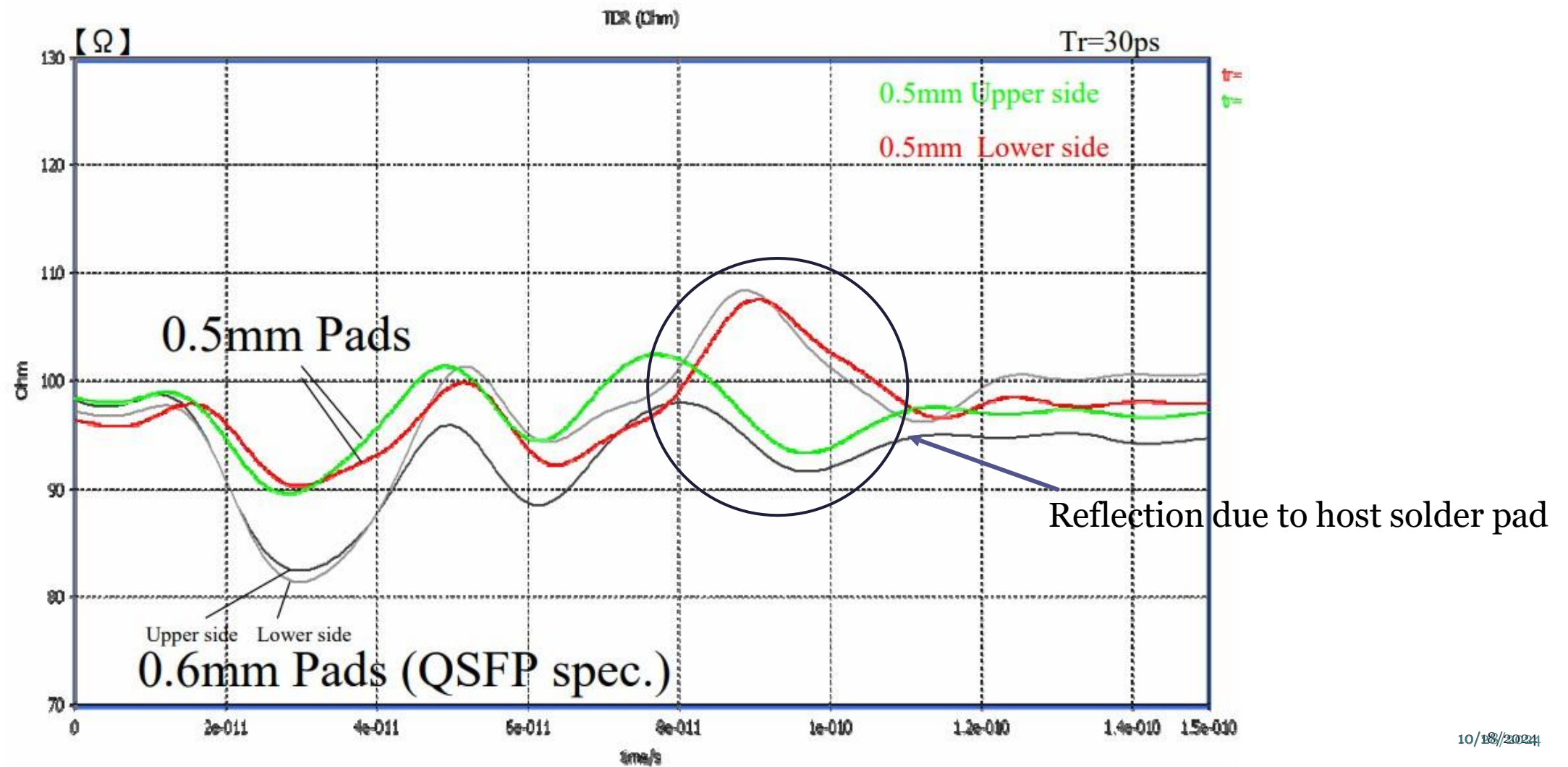
- Reduction in pad width,
- Isolation of wipe,
- Reduction in pad length

Reduced Module Pads Size



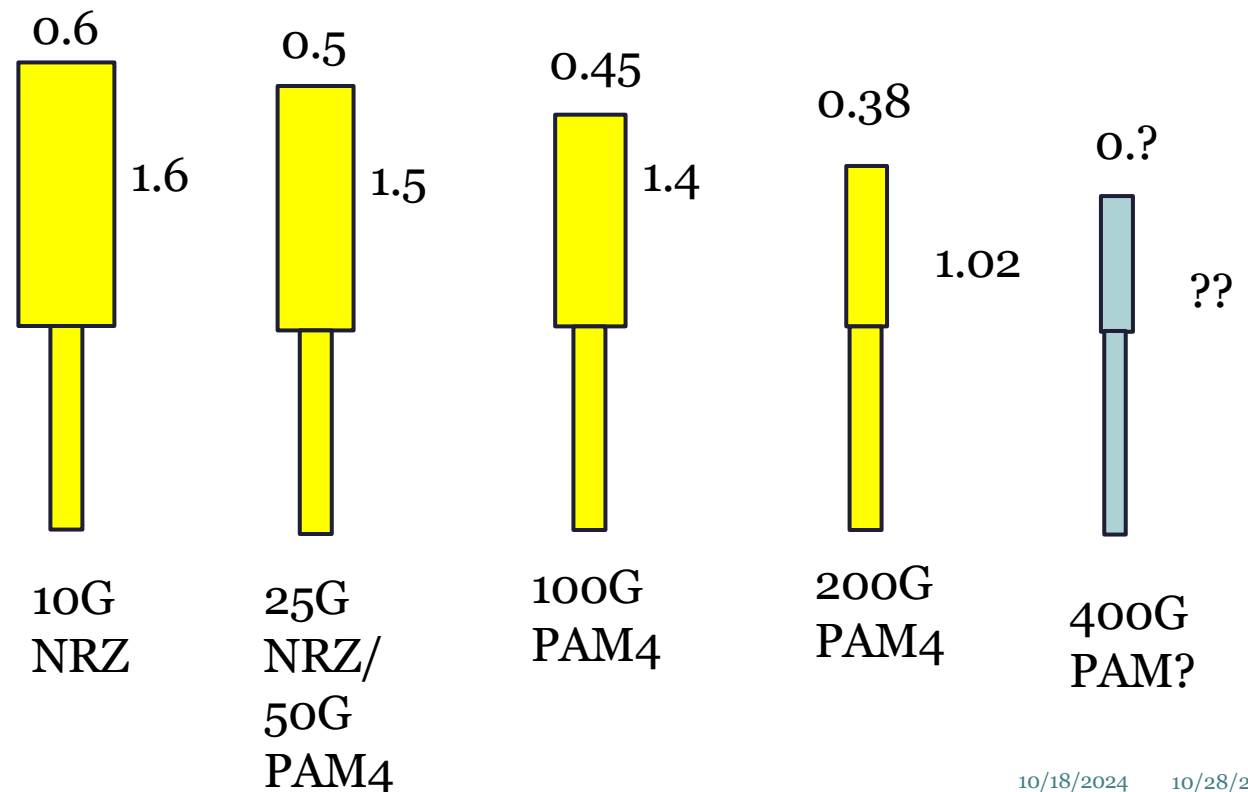
(Detailed explanation will be added at later updated presentation)

Moving from 10-25-50-100-200Gbps (Host Solder Pad)



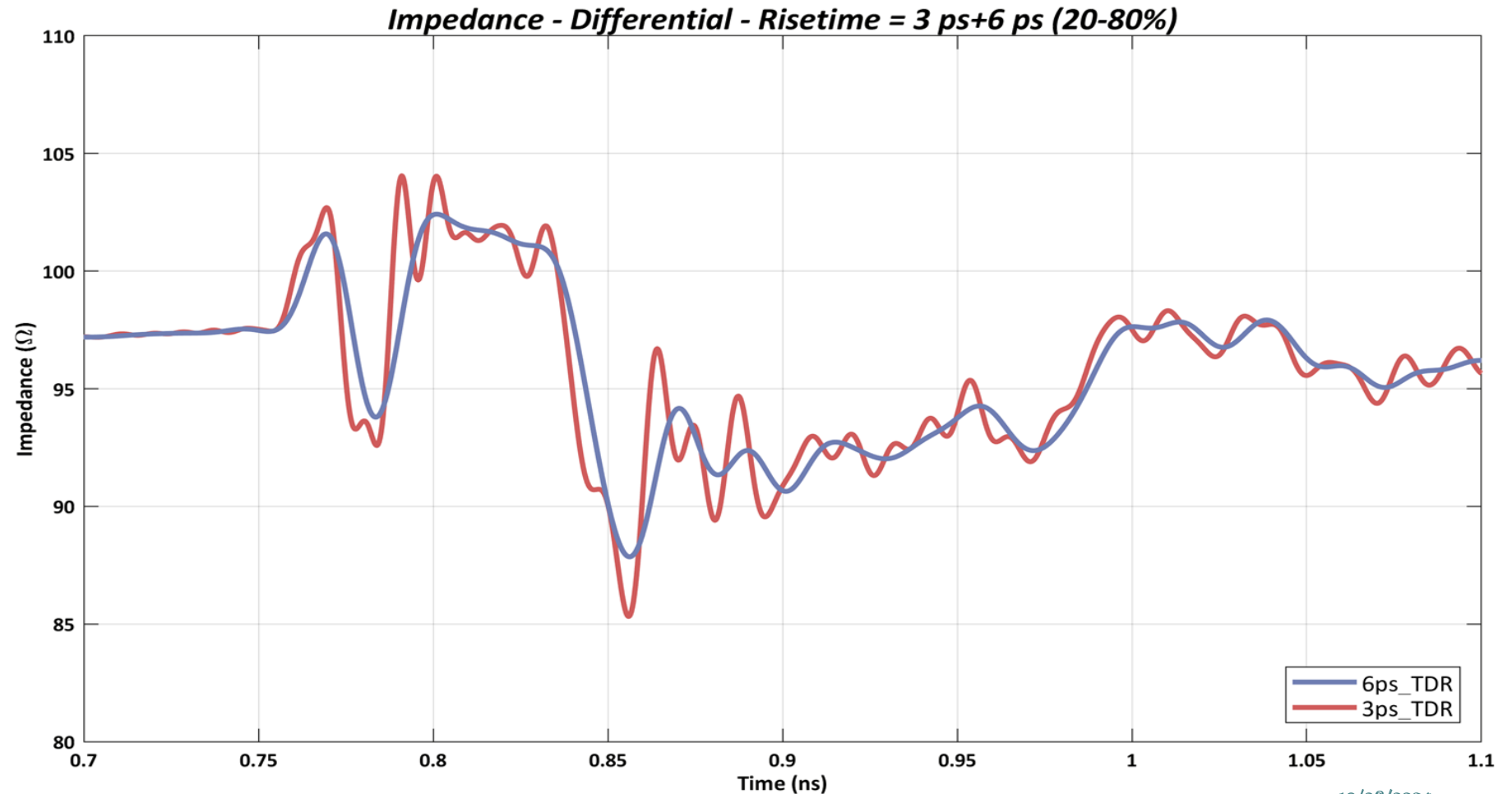
What can be done for backwards compatible 400G?

- Further reductions in pad dimensions
- Further reductions in host connector stub



TDR of 200G OSFP connector

- Module Connector impairments cover 50 UI for 400G PAM4



What about Flyover? (Cabled connector)

Eliminates Host solder pad and Via near module

Reduces module spring finger length (no right angle in spring finger)

Reduces crosstalk

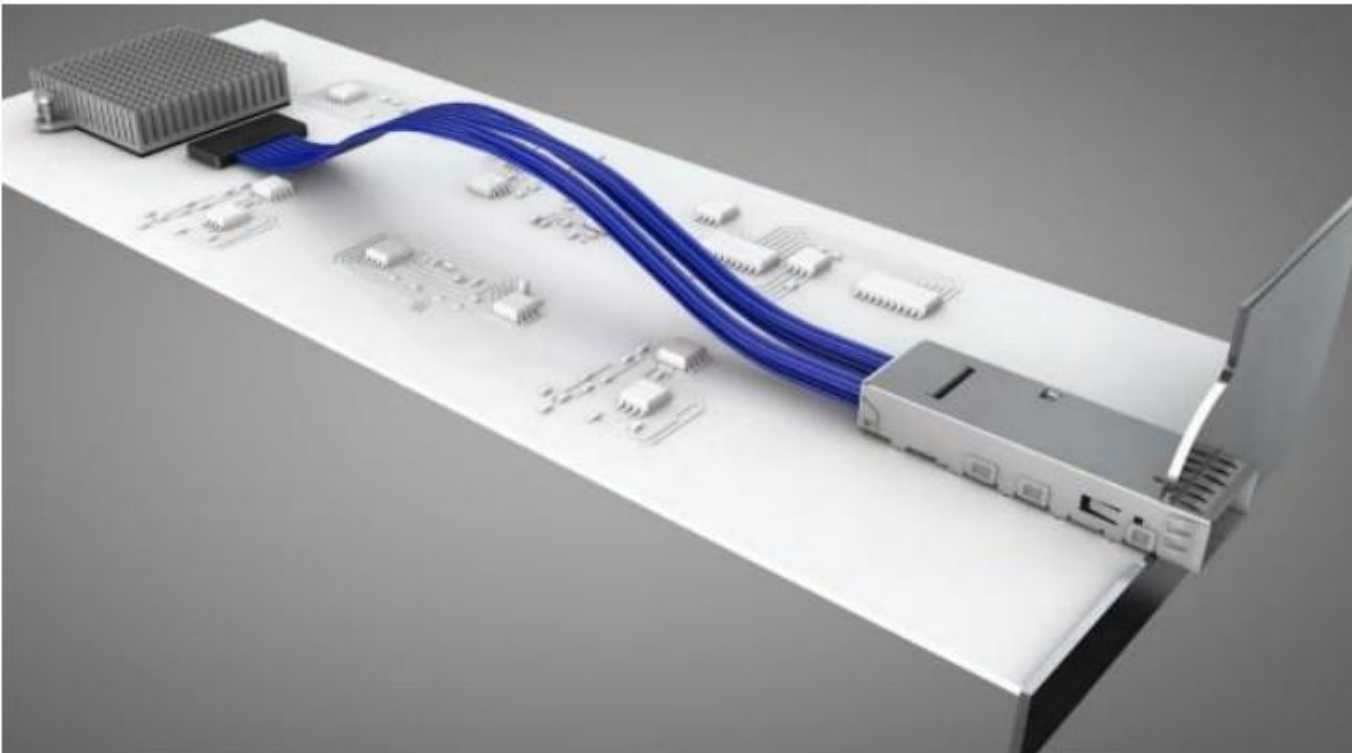


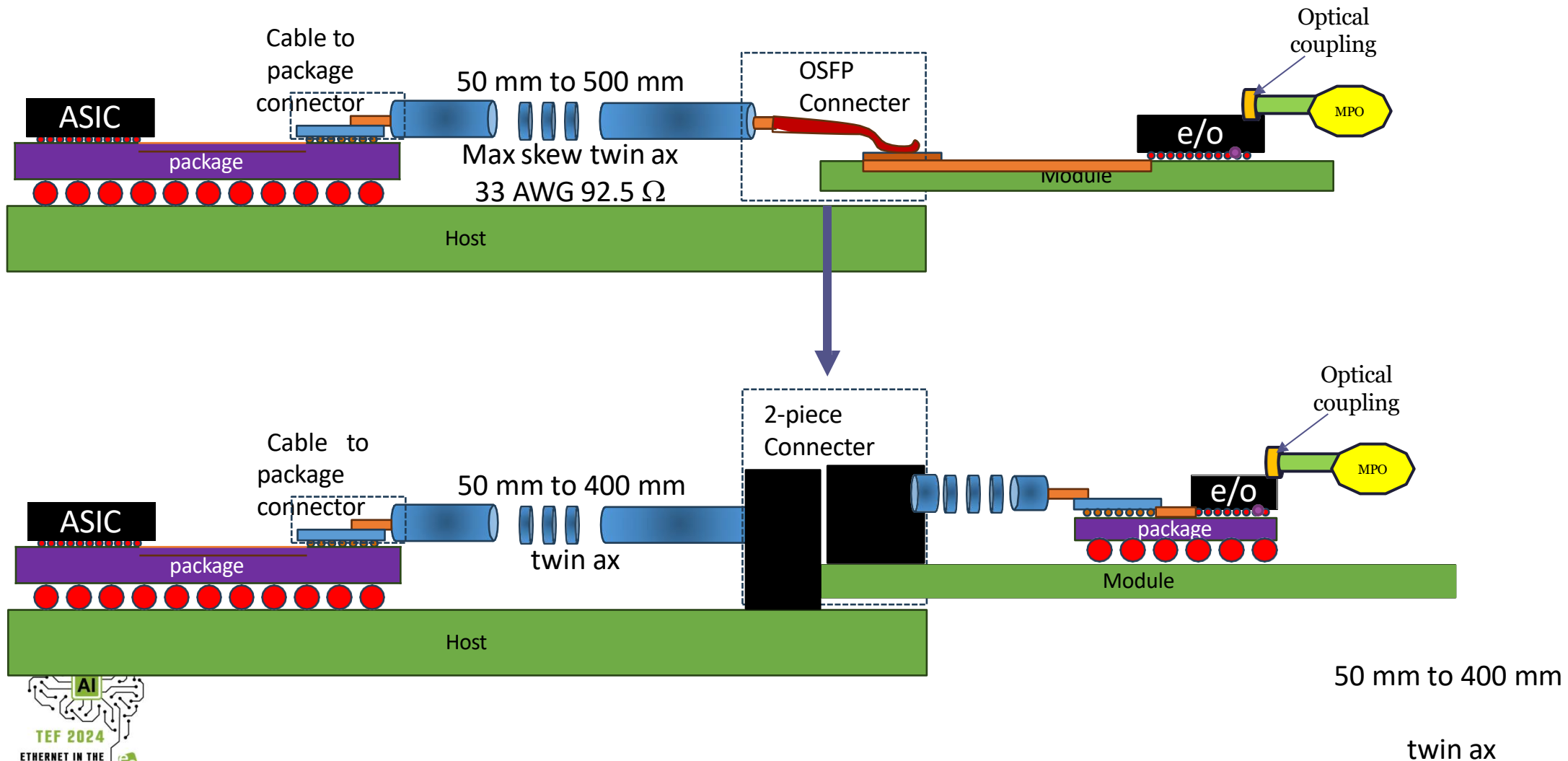
Figure 1. Implementation of Flyover QSFP on a PCB

www.ethernetalliance.org

Migration from 200G Spring finger/pad to 400G 2-piece connector

www.ethernetalliance.org

10/20/2024



400G connector Options comparisons

Reduce module pad dimensions

- Potentially backwards compatible
- Probably lower cost with no connector on module
- Requires improvements in pad to spring finger alignment mechanisms (could lead to higher cost)

2-piece backplane connector

- Improved Signal integrity
- Higher density
- Possible cost reduction vs Spring finger to pad alignment mechanical design

What about higher density? (Required for AI applications)

Reduce module pad dimensions

- Use 2nd row of pads (i.e. OSFP-XD)
- Use multiple PCB cards (i.e. CDFP)

2 piece backplane connector

- Backplane connectors can provide multiple rows
- Backplane connectors can have tighter pad pitch

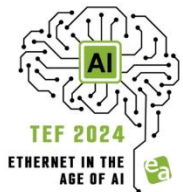


Conclusions

- Backwards compatible front panel pluggable may be possible but may not support PAM4 modulation
- 2 piece connector will provide higher BW, better Signal Integrity, higher density, but is not backwards compatible
- SFF will be addressing these issues

The Open Compute Project (OCP) Open Systems for AI Initiative: Towards Open Systems for AI

Bijan Nowroozi, CTO, Open Compute Project Foundation

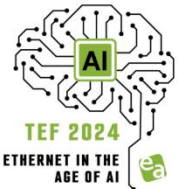


2024 OCP Global Summit

FROM IDEAS TO IMPACT



OCT 15-17, 2024
SAN JOSE, CA



Snapshot of Selected Tracks

Server	Networking	Data Center Facilities	Hardware Management
Storage	Cooling	Open Platform Firmware	Telco
Security	Rack & Power	Sustainability	Time Appliances

AI	Evenstar OpenRAN
Photonics	Market Impact
FTS	FTS: Quantum
FTI: AI HW/SW Codesign	FTI: Data Centric Computing
FTI: Short Reach Optics	

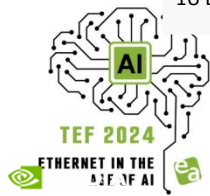
Chiplet Economy	Composable Memory Systems	HPC/NIC/OAI	MHS
-----------------	---------------------------	-------------	-----

Announcing GB200 NVL72 Design Contributions to OCP

Accelerating open innovation of AI data centers

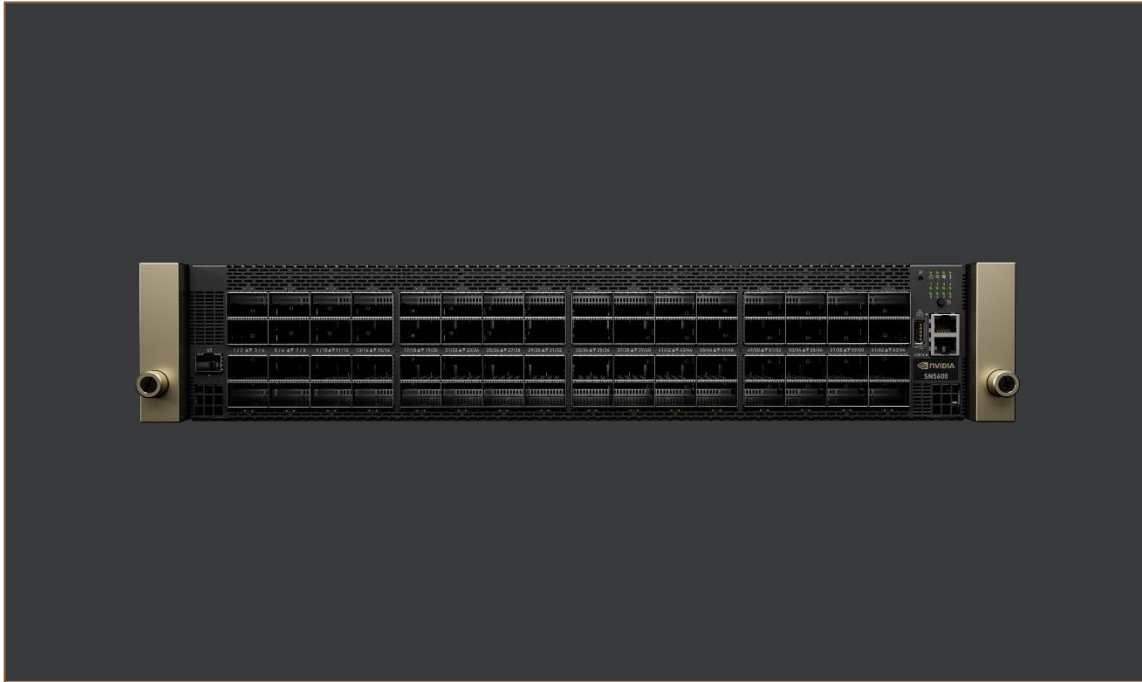
Key Contributions

NVLink Cable Cartridge 5,000 NVLink cables	Enhanced Bus Bar 1,400-amp capacity	Blind Mate Manifold 130 KW cooling capacity
Floating Blind Mate Reduces breakage	Blind Mate Slide Rails NVLink and LC connections	Latch Ejector Bar Support 6,000 lb. mating force
Rack Reinforcements +100lb. of steel support	Rear Rack Extenders Protects manifold fittings	Larger Rack Cable Tray Fits higher cable density
Compute Tray Form Factor 1U LC: 4 GPUs	Compute Tray DC-SCM Fits 1RU form factor	Compute Tray Bay System Highly configurable
Switch Tray Form Factor 1U LC: 550 NVLink Lanes	Switch Tray Bus Bar Connector Narrower 35mm width	Liquid Cool Flow Rates 5 liter/min



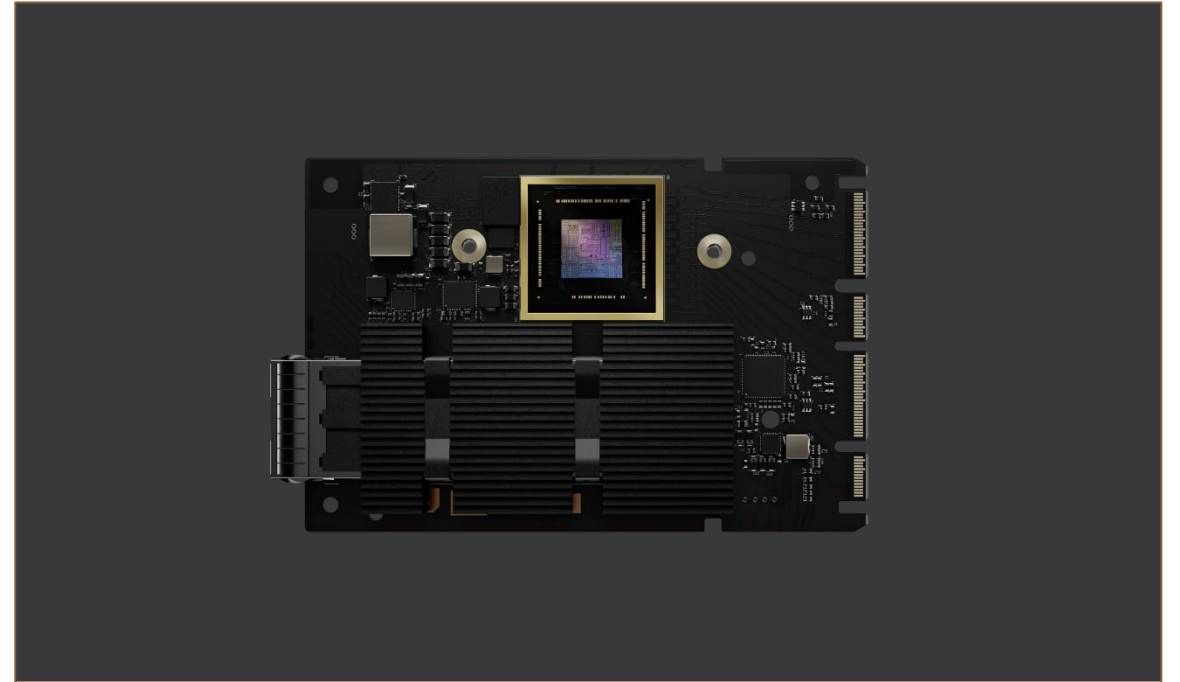
Announcing NVIDIA Spectrum-X Support for OCP Platforms

Re-imagining ethernet performance for massive-scale AI infrastructure



Spectrum-X800 Switch

800Gb/s per port | 51.2Tb/s aggregate bandwidth



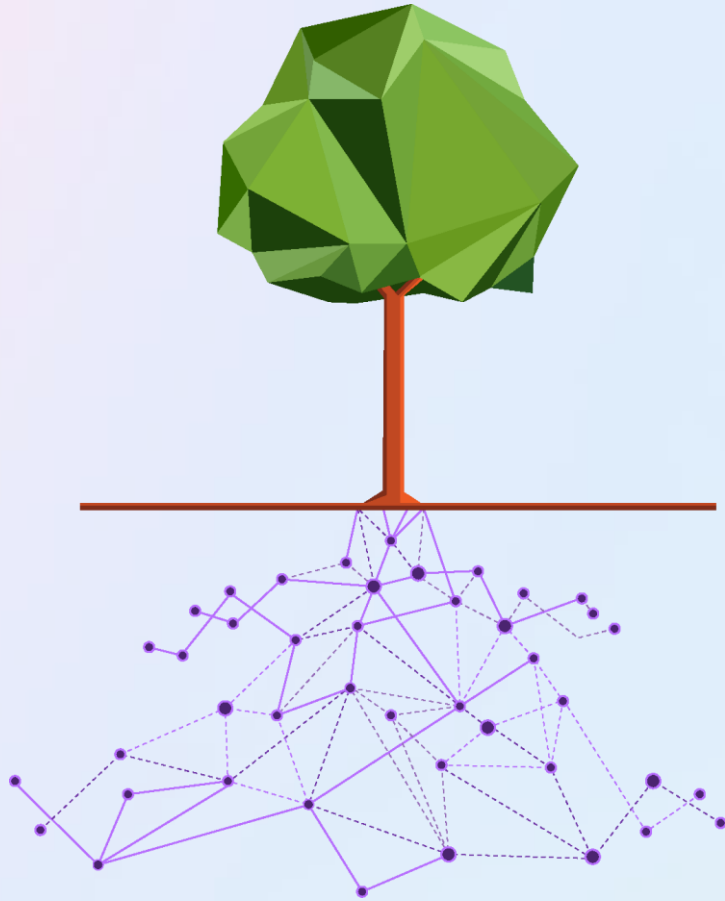
ConnectX-8 SuperNIC OCP 3.0

800Gb/s networking | PCIe 6

Full SAI / SONiC support for OCP-based AI factories
Accelerated networking with enhanced programmable I/O
Adaptive routing and telemetry-based congestion control

Blackwell in Full Production

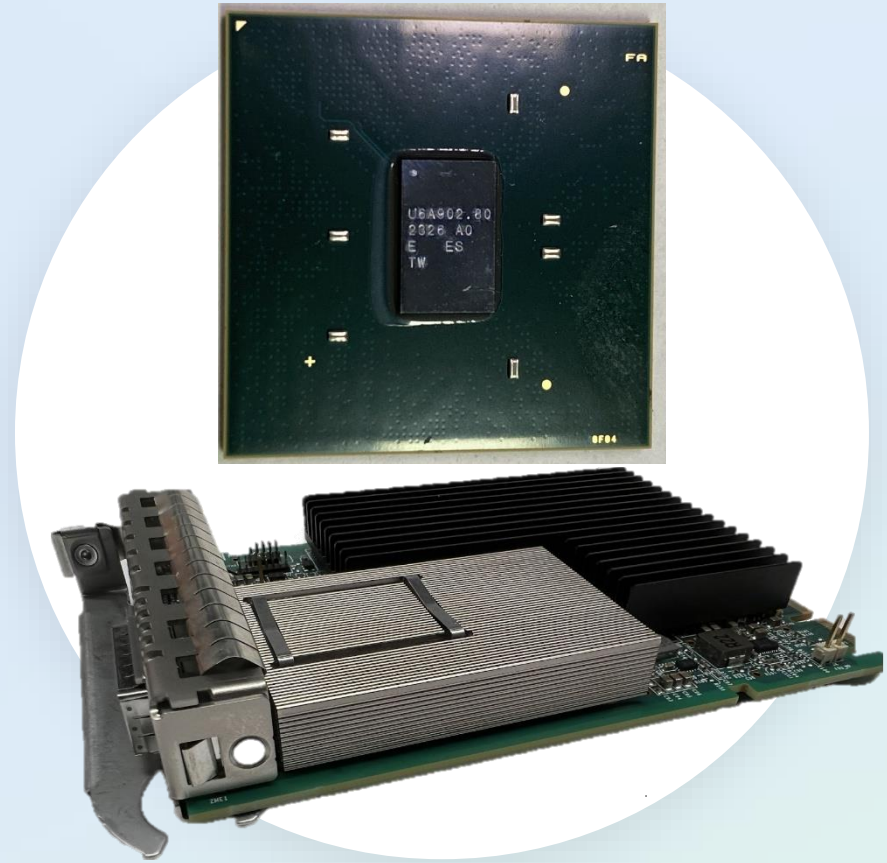




OPEN
Compute Project®

Our first Meta-designed NIC

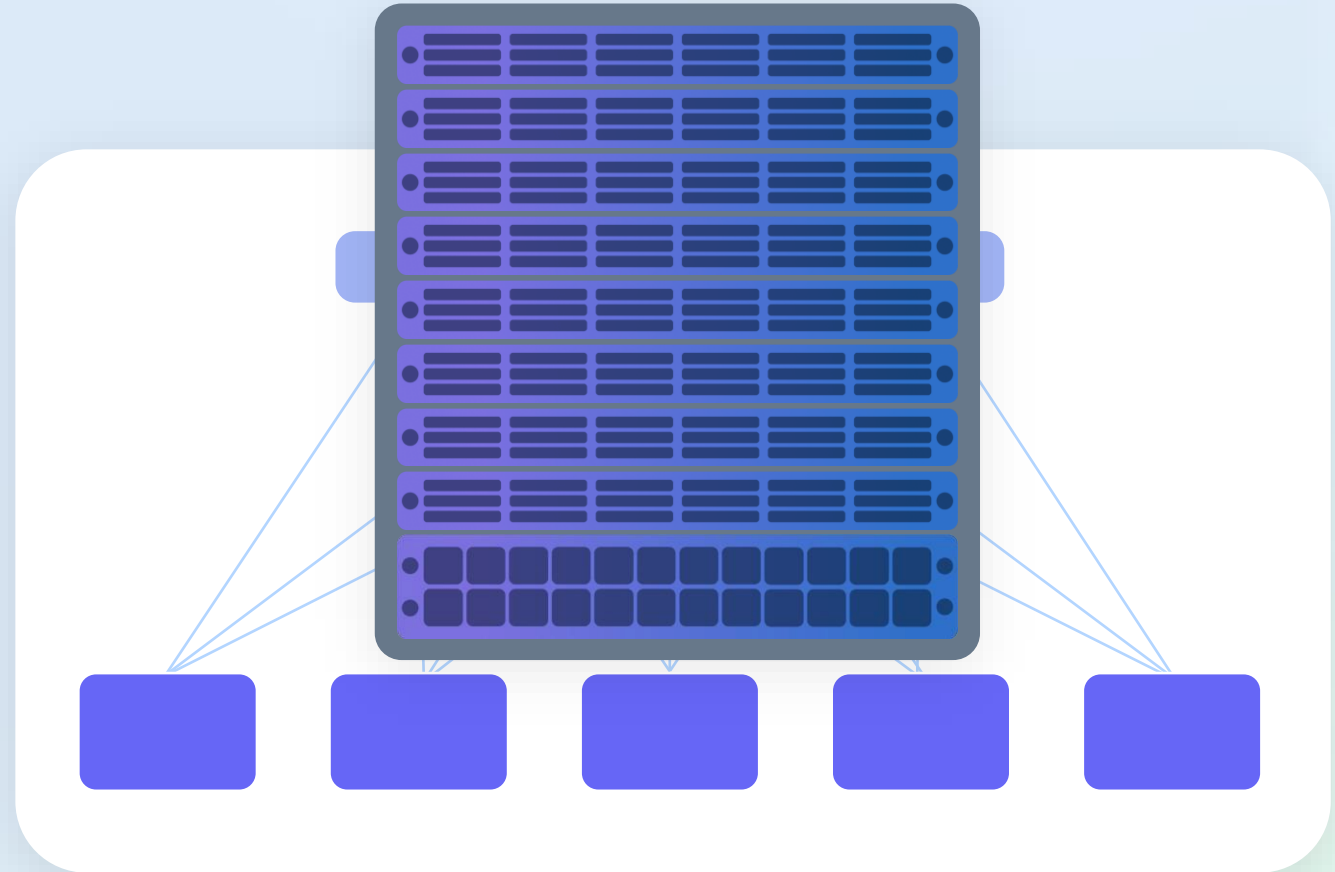
- 1st Meta-Designed NIC module and ASIC with Marvell
- Targeting compute platforms
- Open module with Marvell & up streamed Linux driver



Next-gen AI Fabric - DSF

- DSF: Disaggregated Scheduled Fabric
- Lossless/reliable fabric of switches
 - Like what we got from one chassis before in previous fabric
- Tuned for AI
- Multi-building scale
- Cross-building distances

Provides flexibility & speed across multiple generations and types of accelerators and NICs





Sustainable Silicon to Intelligent Clouds

Collaborating on the Compute Infrastructure of the Future



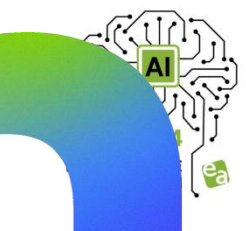
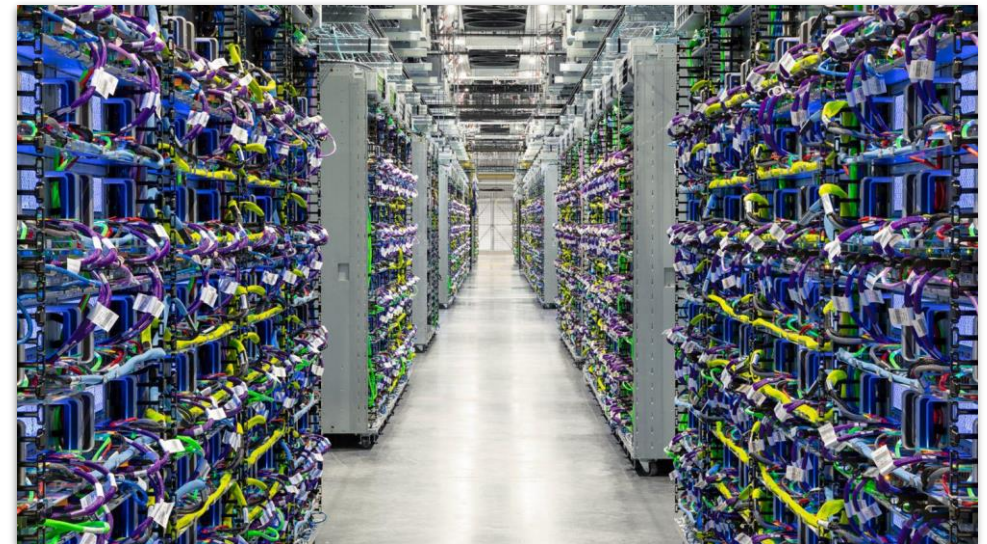
Partha Ranganathan
VP & Technical Fellow
Machine Learning, Systems & Cloud AI

Accelerator ecosystems

Standardized accelerator systems: Industry-aligned 1.0 spec: FW Update, Management Interfaces, RAS

Data formats for deep learning: OCP Floating Point 8 (OFP8) and OCP Microscaling Formats (MX)

Scale-out & Scale-up networking: Falcon, Ultra Ethernet w/ congestion control, Ultra Accelerator Link



From Bottlenecks to Breakthroughs Innovations in Data Center Infrastructure & Security

Saurabh Dighe

Corporate Vice President, Azure Strategic Planning and Architecture,
Microsoft

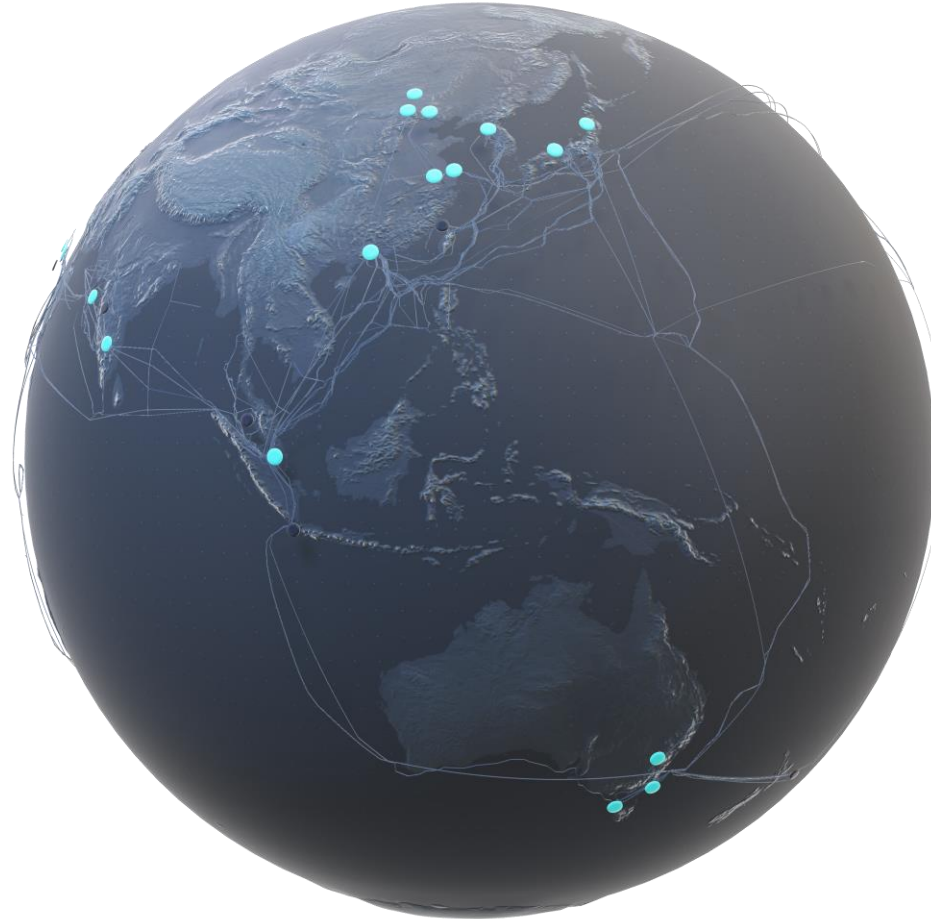
Azure

AI Supercomputer

Building the world's computer

60+

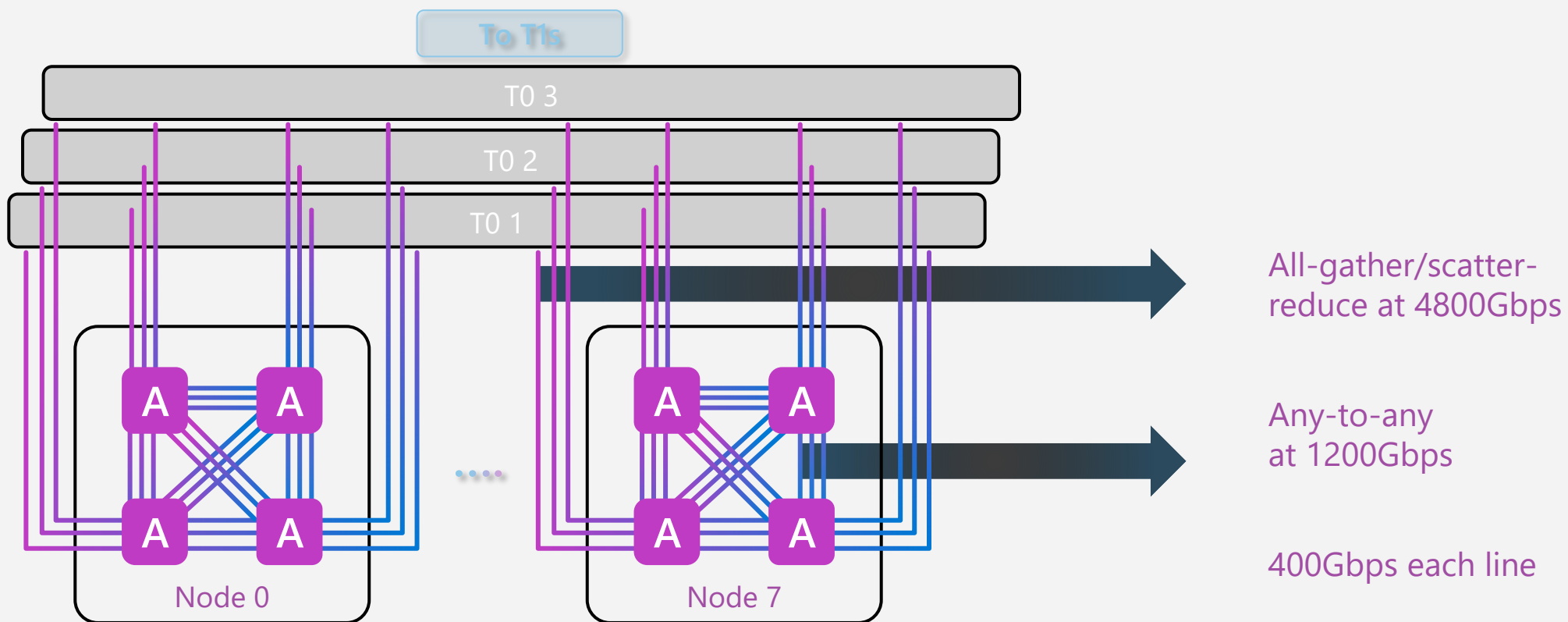
Azure regions



300+

Datacenters worldwide

Azure Maia Ethernet based networking



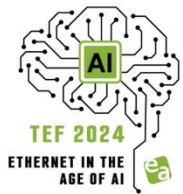
Ethernet

Ultra Ethernet Consortium

Open / multivendor	Switches, NICs, cables, optics, tools, software
Scalable	Addressing and routing for rack-, building-, DC- scale networks
Tools	Many tools for testing, operations, measurements
Cost	Economies of scale and competitive market
Supporting standards	Regular progress in IEEE, for many technologies, across layers

Tension in Design

Networking Trends Within OCP Projects



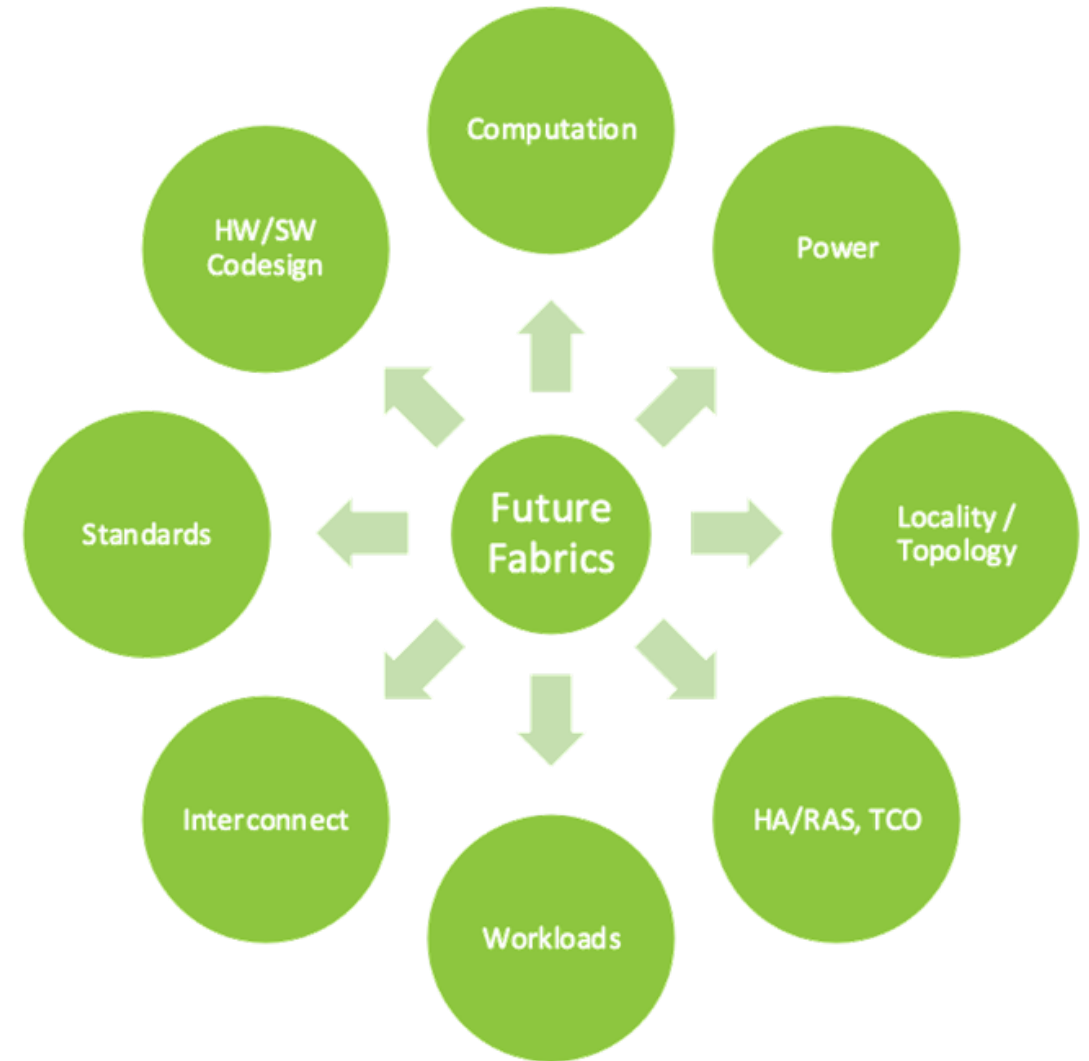
OCP Open Systems for AI Immediate Focus Areas

- Thermal Performance
- System Density
- Compute Throughput
- Scalability
 - (Max addressable compute)
- Maximize Commonality
- System Efficiency
- Resource / Facility



AI/HPC Future Fabrics

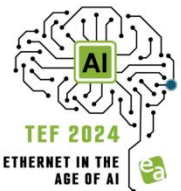
- Many factors influencing the fabric
 - Increasing complexity of applications, dimensions of cost
 - Technology: BW, Interconnect (Copper, optical)
 - The Drive to Parallelism, Ethernet
- Applications differ in demands
 - HPC: High BW, Low Latency, Scalable
 - AI: Both GenAI and Ranking / Recommendation
- Datacenter design, economics influencing fabrics
 - Power, cooling, geography
 - Workloads, tenants, resource management
 - HA/RAS, TCO, Data Services
 - HW/SW Codesign



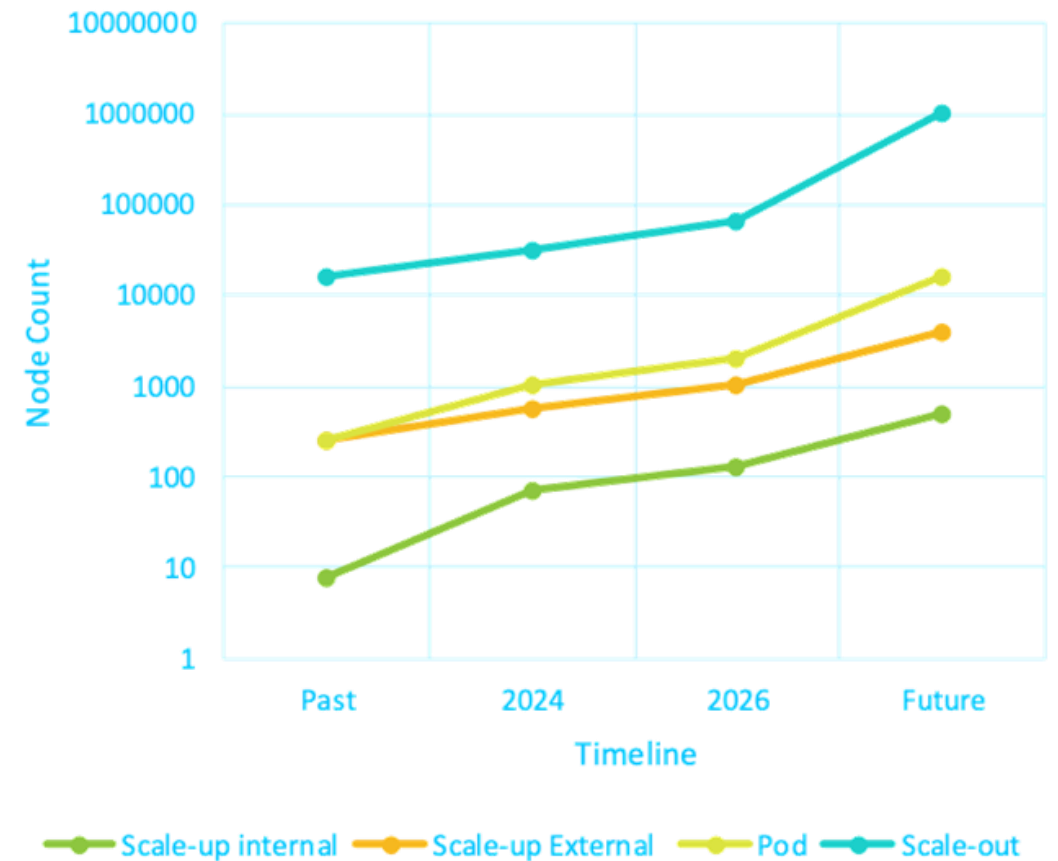
AI/HPC Future Fabrics

- Topology Doubling in Size, Speed every two years
 - Scale-up, Pod, Scale-out factors
 - Standards driven: HW/SW codesign
 - Adapting stack: Flow control, Traffic Engineering
- Scale-up limited by technology
 - BW, cost, application, physical packaging
 - High BW, low latency scale
 - xPU based interconnect
 - Parallel Path: interconnect, wavelengths
- Scale-out driven by standards, network engineering
 - Engineered to meet total application requirements, HA/RAS, Observability
- Pod driven by traffic engineering
 - Designed to bridge scale-up to scale-out
 - Enabled through high BW, parallel communication.

Rail optimized, many flows



Scaling Fabrics



AI System Design Driving SERDES Requirements

Design Roadmap: SERDES

56 Gbps

112 Gbps

224 Gbps

448 Gbps

~2028

- Systems moving from PCIe 5 towards PCIe 6/7
- Myriad PHY's increasing rates
- Interconnect
- Scale Up/Out

Design Roadmap: Ethernet

400 Gbps

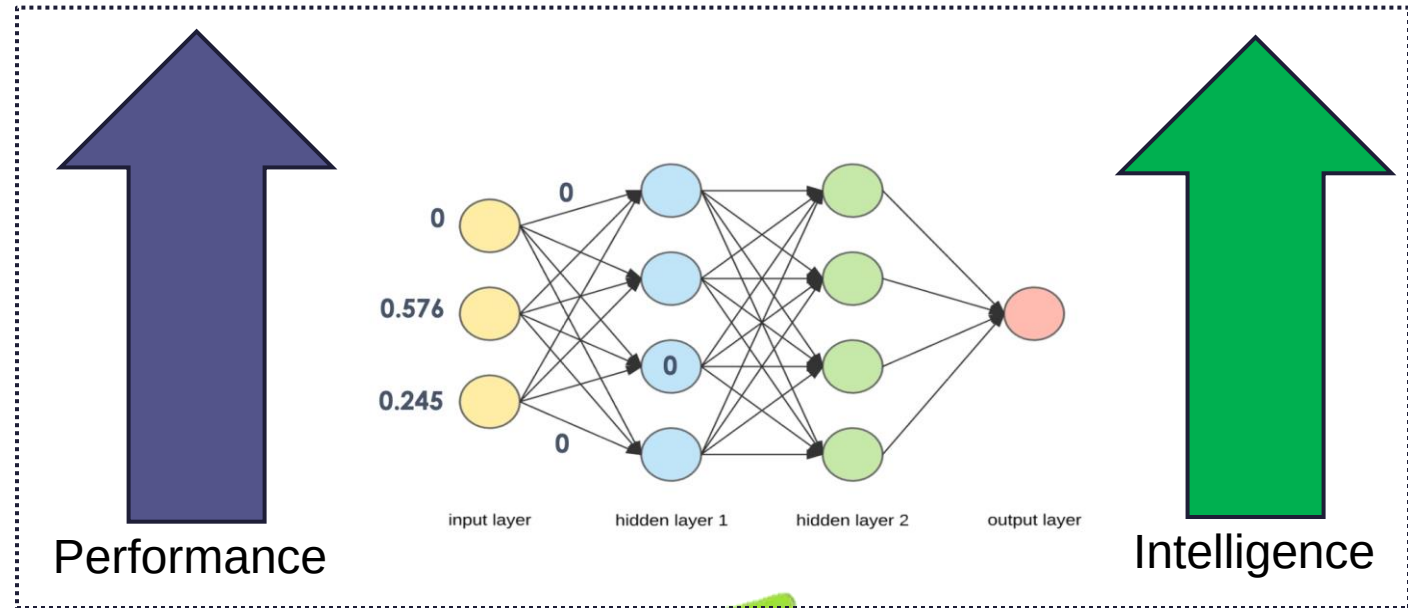
800 Gbps

1600 Gbps

X Tbps

~2028

What is the Tension?

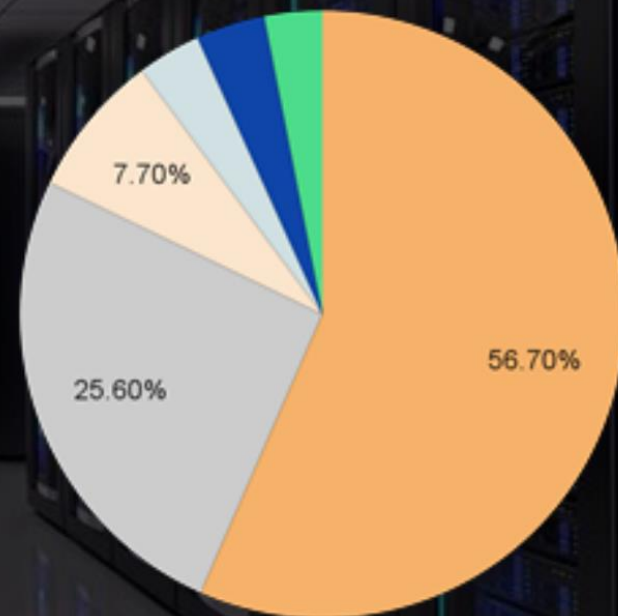


Efficiency



Electricity is already a huge *financial* and *environmental* burden on data centers

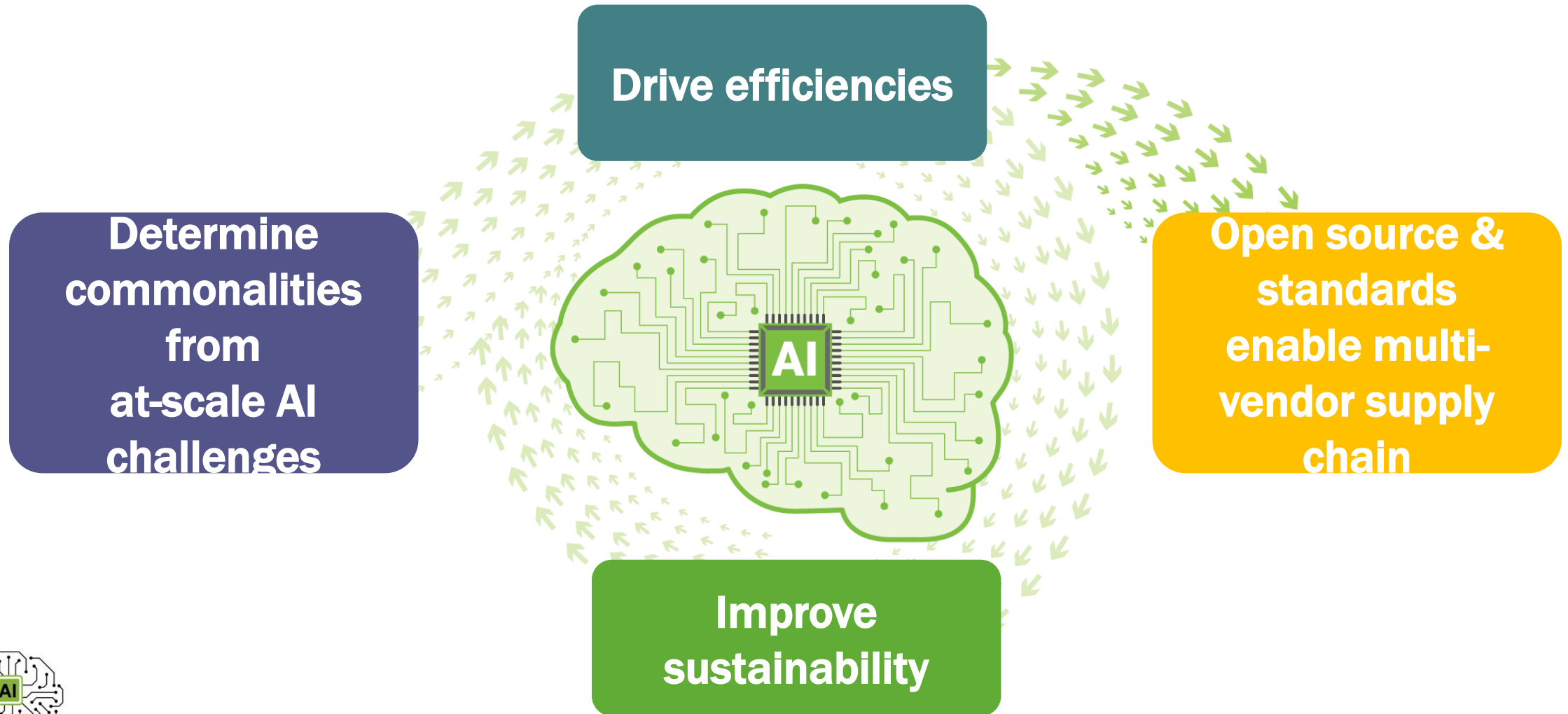
DC OPEX



Electricity Depreciation Building rent Equipment leasing Labour Others



OCP Open Systems for AI



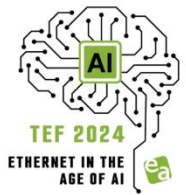
Related OCP Projects Activity

- **Server**
 - SERDES 112/224/448Gbs
 - Chiplet interconnects
 - Composable Systems
- **Networking**
 - NIC / AI Optimized
 - UEC/UAL
- **(FTI) Short Reach Optical Interconnects**
 - Interconnects to 10m
 - Opportunities ex: Thin-film Lithium Niobate (TFLN)
 - CPO/xPO

Other Noteworthy Discussion/Consideration:

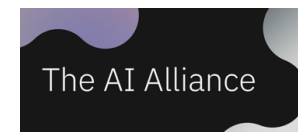
- Co-packaged Copper
- Active Electrical/Optical Cables
- Hollow Core FO
- MW + Fiber
- Waveguide Embedded Printed Circuit System (WPS)

BACKUP



Expanding Reach with Alliances

SYSTEMS MANAGEMENT



DC IT INFRASTRUCTURE



DC PHYSICAL INFRASTRUCTURE

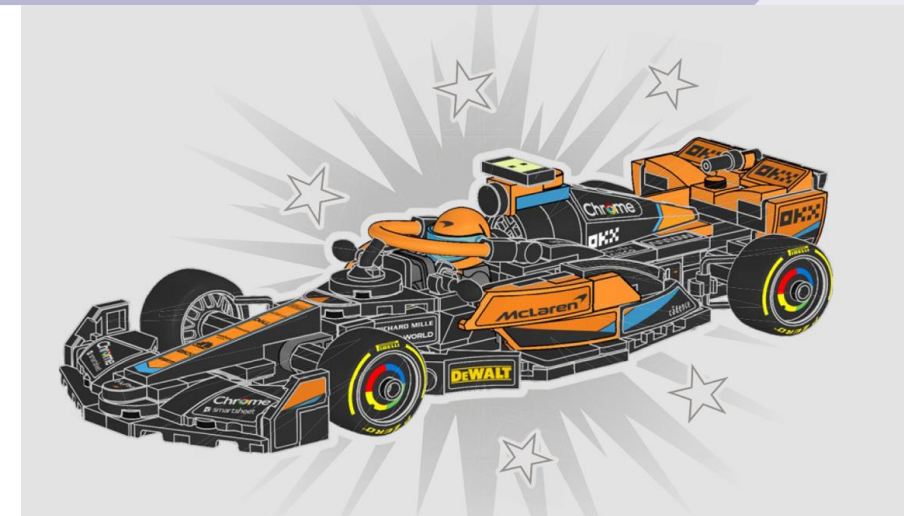
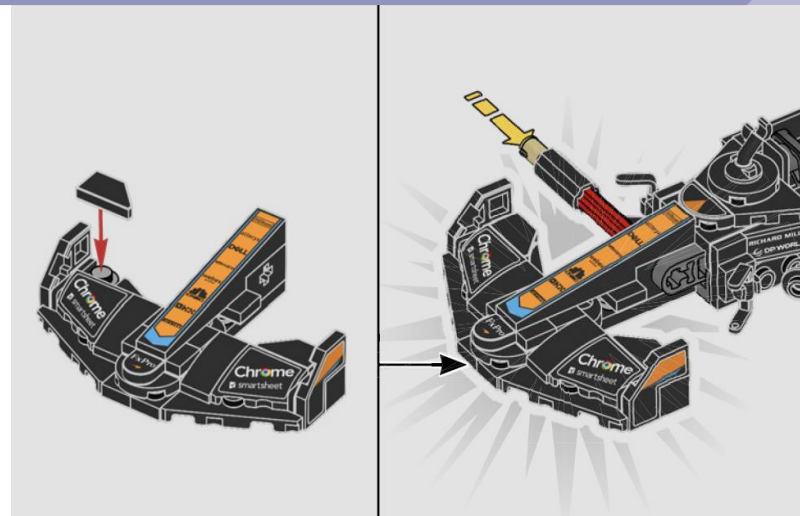
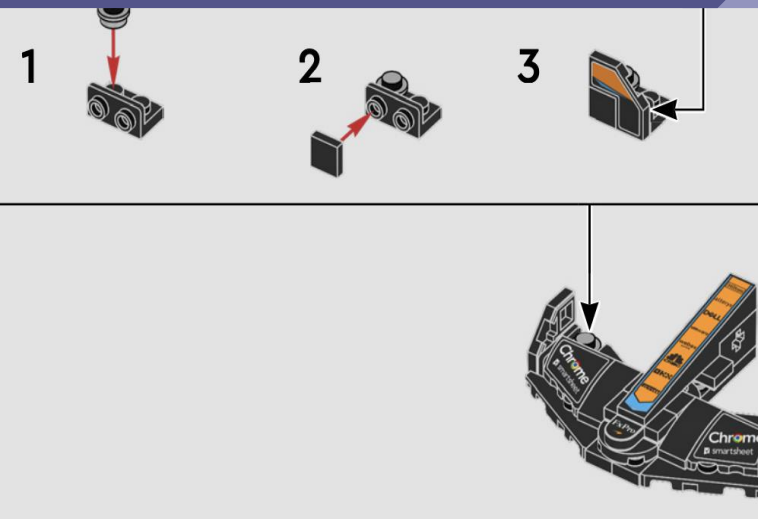


Assembling Open Systems of Systems

Project level

Cross-stack

Cluster level



Progressively Getting to Holistic AI Clusters Design with Open Source



OCP AI Community Work Efforts

Targets for Open Systems for AI	Progress to Date	Medium Term Plan (Actively Worked)	Identified Timeline TBD
Thermal Performance	•Targeted Flow Efficiency Heat Sink	•2 Phase Refrigerant based Direct Liquid Cooling Requirements	
System Density	•Modular Scalable Density Optimized HPM Form Factor (M-SDNO), •Storage, Catalina: PCIe Gen5 EDSFF NVMe SSD Reference Designs, Specification •Falcon Transport Protocol 1.0	•CSIG (w/UEC), •400+VDC Rack Requirements •Disaggregated Scheduled Fabric	•NG Open Rack (250-750kW) •UEC 2025 Specifications
Compute Throughput	•OCP Quantization/Micro-scaling Formats •Polymorphic Architecture Definition	•AI HW/SW Codesign Polymorphic Resource Architecture •Quantization Improvements •BoW V3.0, Modular form Factors, Design Kits	•OAI G3, •UALink (Networking) •AI Hardware System Benchmarking, •Short Reach Optical Interconnect •224G+ SERDES
Scalability (Max addressable compute)	•MGX Accelerated Computing Rack and Tray Specification •Meta Catalina Rack Specification	•CMS Memory Pooling •UEC backend Networking	
Maximize Commonality	•Blind Mate Quick Connector Specification •Open Rack V3 Blind Mate Manifold •HPM Common Circuit Type 1 Design Specification Rev1.0 •Catalina: PCIe Gen5 EDSFF NVMe SSD Reference Designs • GPU Management Redfish Interoperability Profile	•ORv3x Common High Power Rack (NVIDIA/OTHER) •DCF Technology Cooling System Pipe Distribution	
System Efficiency	•GPU & Accelerator RAS Requirements •Server Component Resilience SDC	•RAS (multiple layers)	•Sustainability Metrics
Resource / Facility		•OCP AI Data Center Ready Checklist	•Cogeneration, •Heat Reuse 2G

QUESTIONS?

