



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 Role of Optics in Future AI Applications

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 Alliance

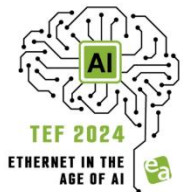
MODERATOR & PANELISTS

- Moderator - Vlad Kozlov, LightCounting
- Panelists
 - Maxim Kuschnerov, Huawei
 - "Feasibility analysis of optical 400G/lane IOs for IMDD"
 - David Rodgers, EXFO
 - Navigating the AI-driven Evolution: Next-Gen Transceiver and Interconnect Testing Challenges
 - Tony Chan Carusone, Alphawave Semi
 - "Advancing AI Scalability with New Optical Connectivity Technologies"
 - Sam Johnson, Intel
 - "Interoperability in an Open Ethernet Ecosystem - Lessons from EA's Interoperability Events"

Feasibility analysis of optical 400G/lane IOs for IMDD

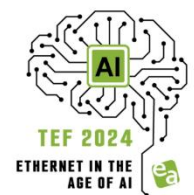
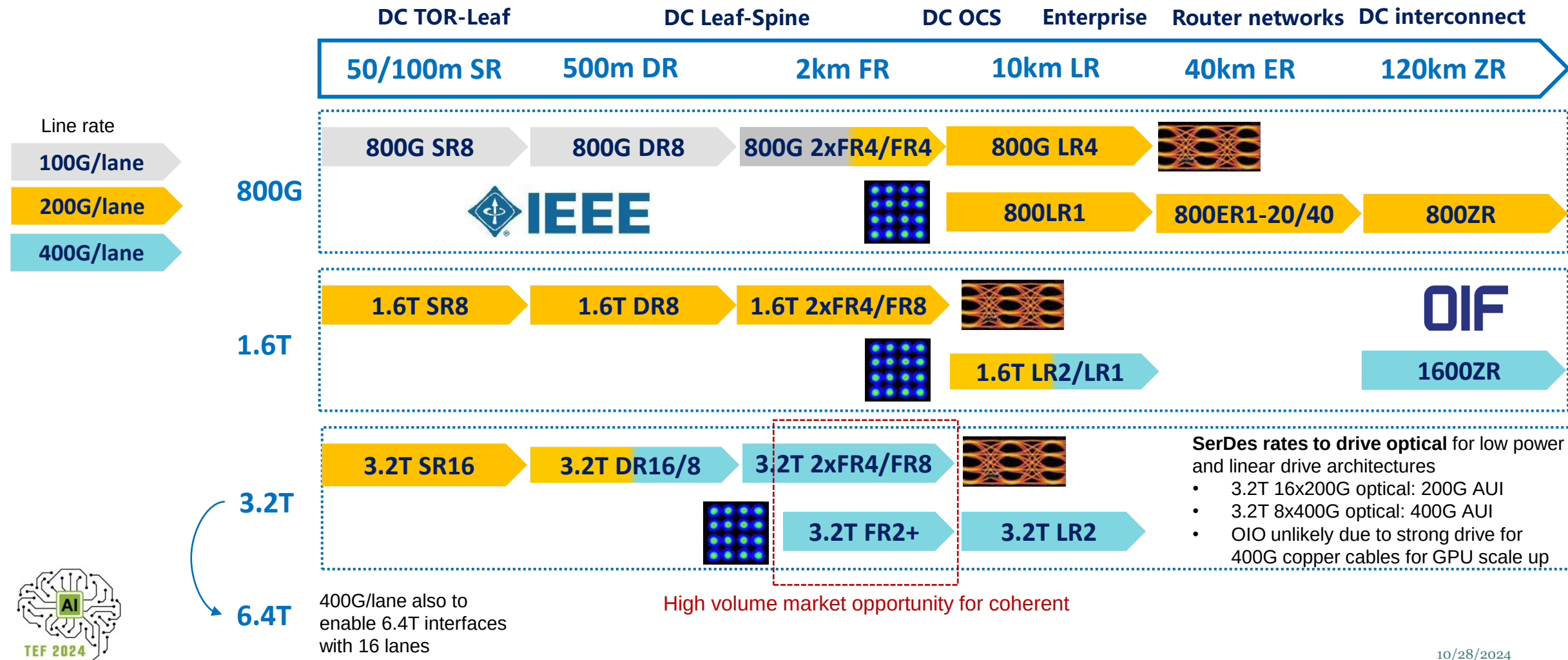
Maxim Kuschnerov, Director R&D, Huawei Technologies

Contributors: Tom Wettlin, Stefano Calabro, Lin Youxi, Nebojsa Stojanovic



10/28/2024

Optical Ethernet transceiver overview



400G/lane optical component feasibility

Various modulator platforms demonstrated for 400G/lane (TFLN, InP, SiP, BTO, plasmonic)

MMF (SR8/16)

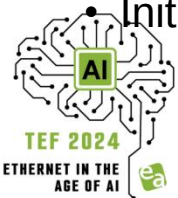
- 200G VCSELs lagging behind the market
- 400G VCSELs not demonstrated yet

PSM (DR8/16)

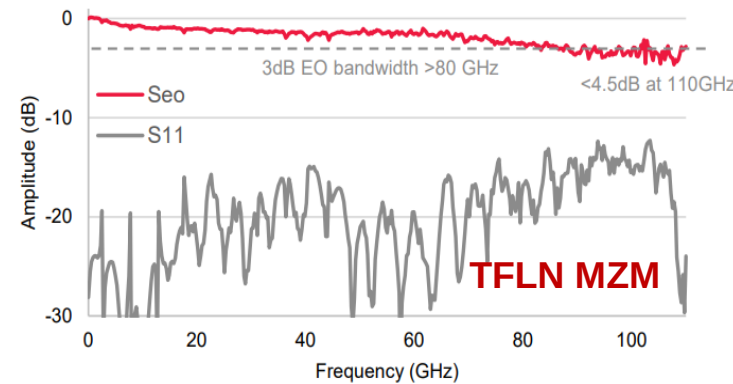
- TFLN MZM as a likely candidate for PSM with shared high power laser source(s)
- SiP MZM with initial feasibility
- BTO without a broader eco-system

WDM (FR4/8)

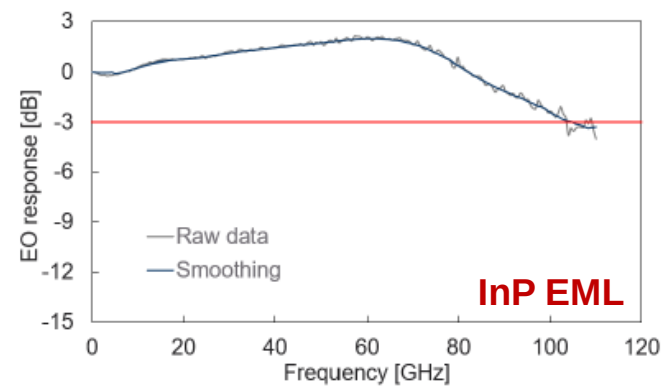
- Multiple companies demonstrating InP EML technology
- Initial results on plasmonic ring resonators



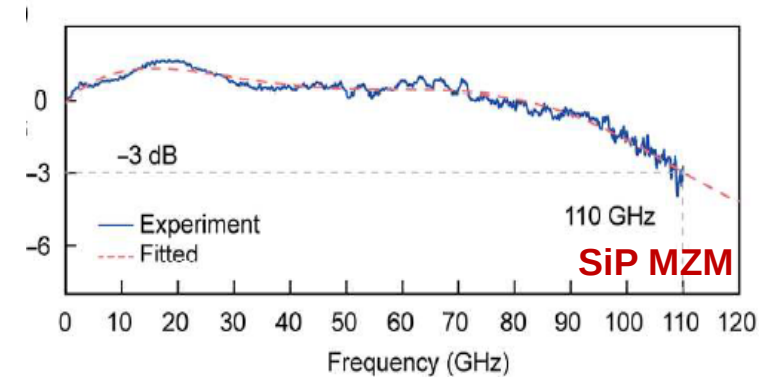
Typical E/O response and RF return loss



© Hyperlight 2023

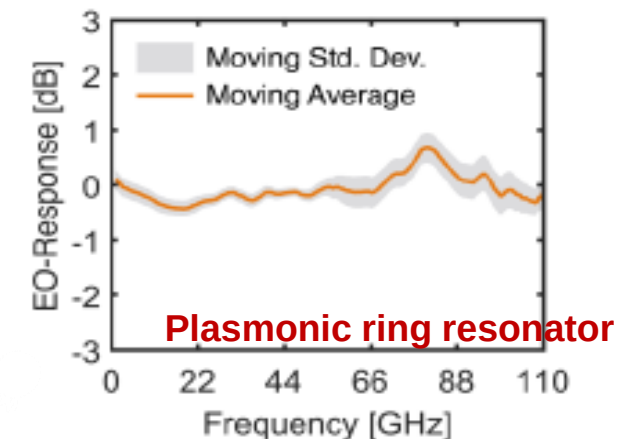


ECOC 2024 Lumentum, Asakura et al.



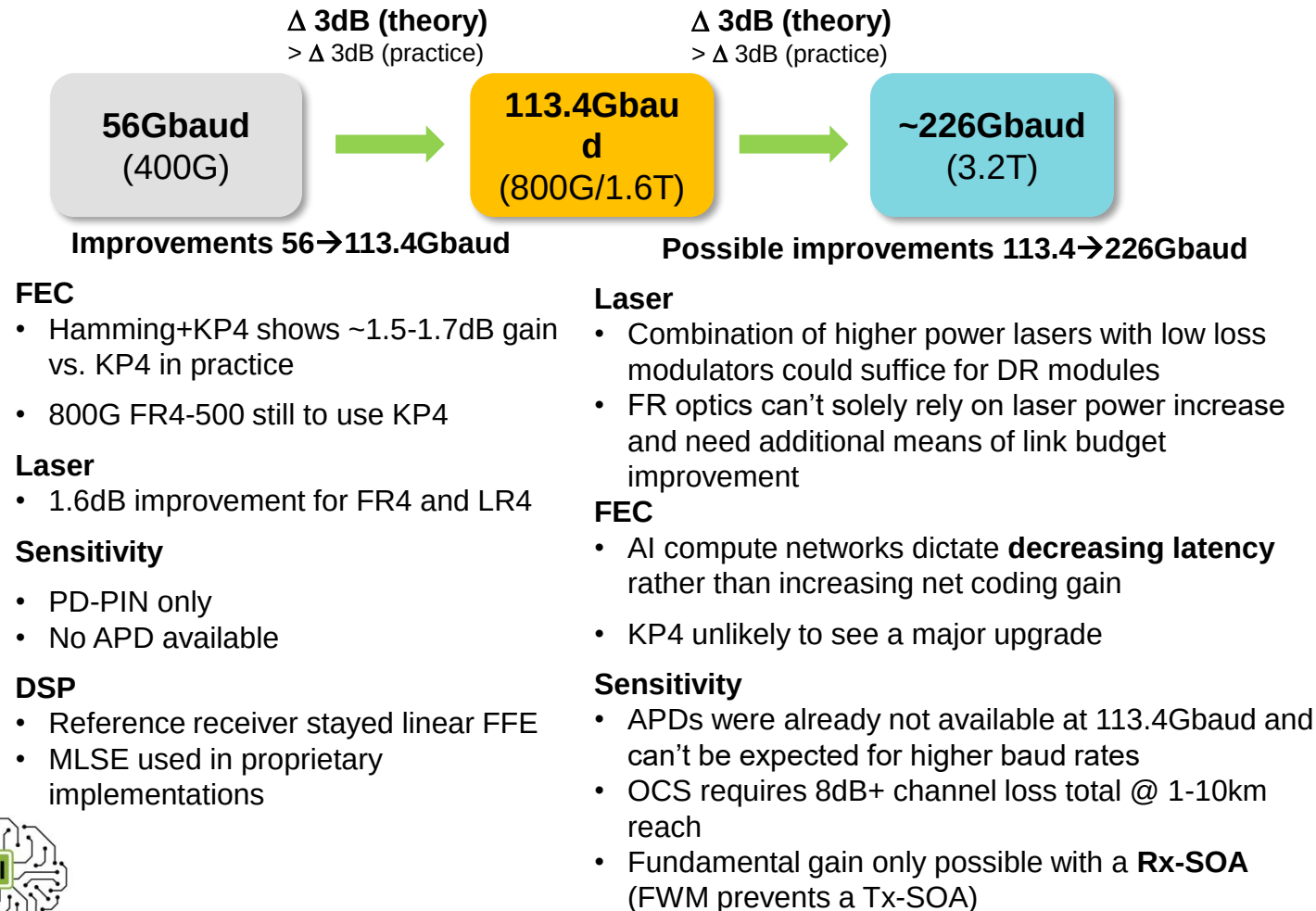
Slow-light silicon modulator with 110-GHz bandwidth

Han et al., *Sci. Adv.* **9**, eadi5339 (2023) 20 October 2023

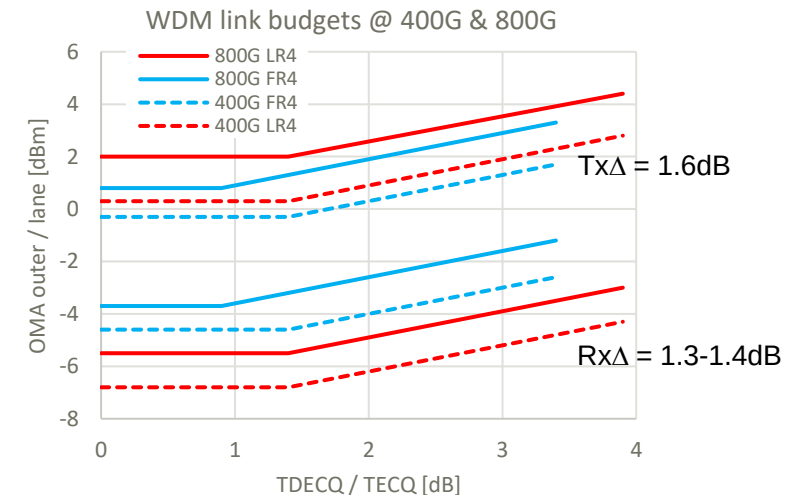


10/28/2024

Link budget challenges

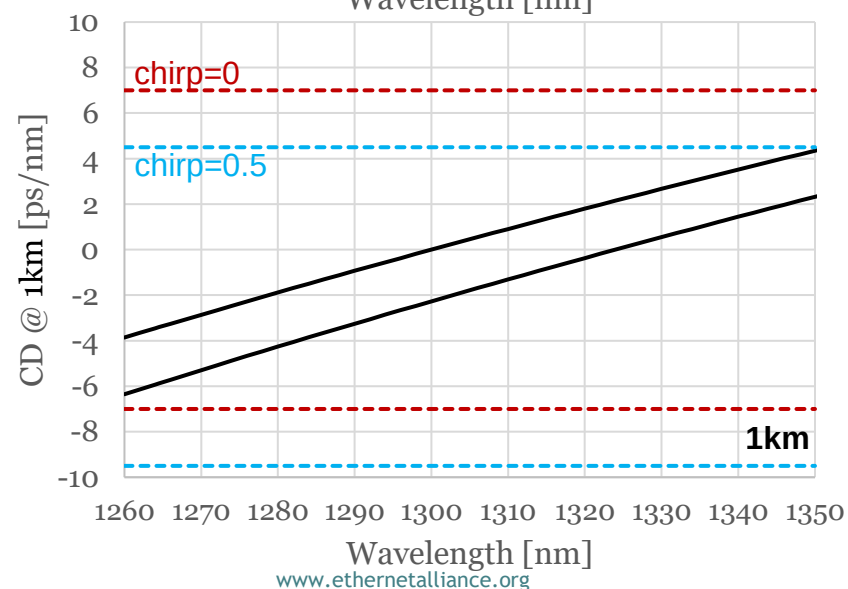
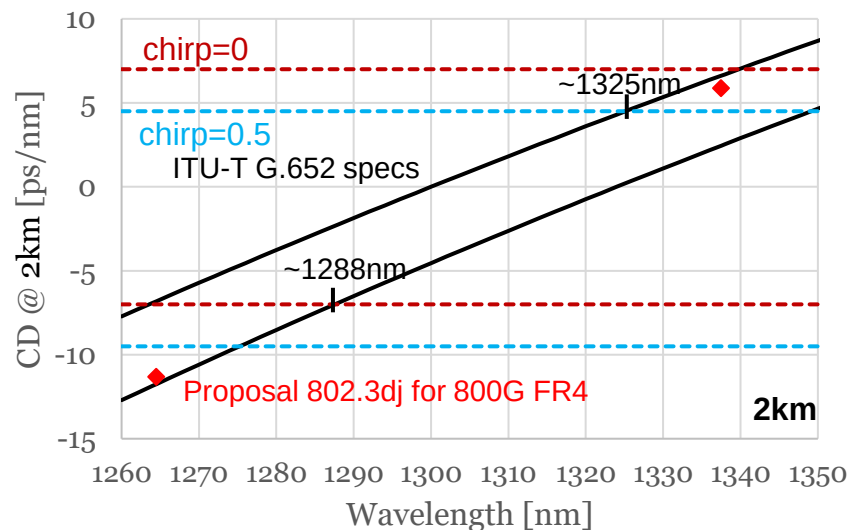
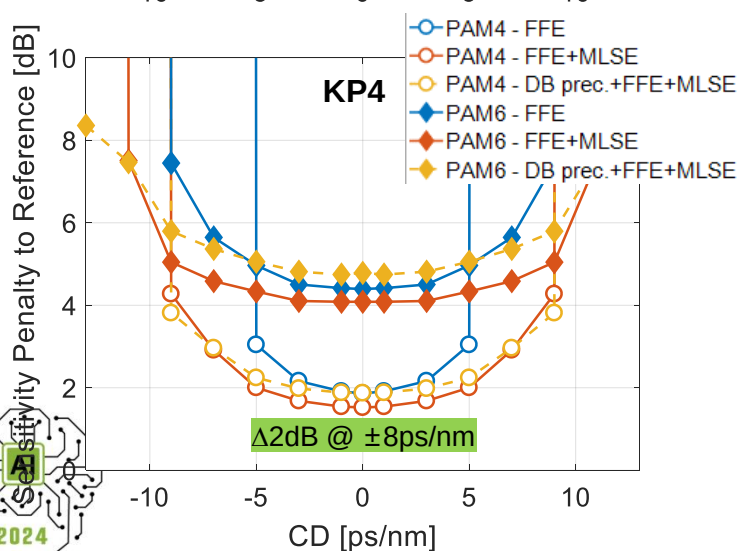
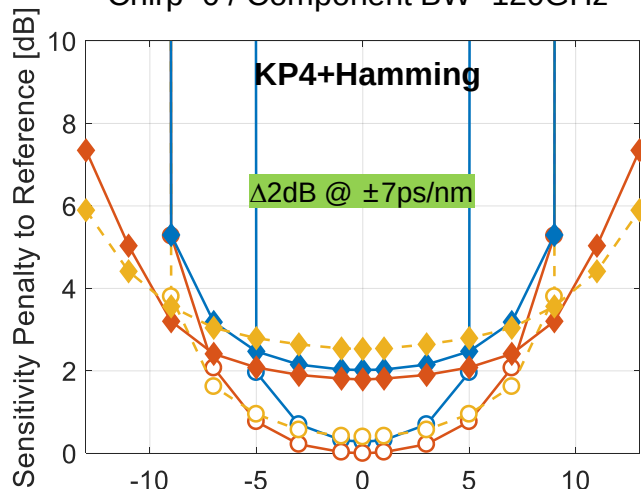


Class	Fiber	Reach	Type	Band	IL [dB]
DR	PSM	500m	PAM4	O	3dB
FR	duplex	500m-2km	PAM4	O	4dB
LR	duplex	10km	PAM4	O	6.3dB
LR	duplex/bidi	10km	16QAM	O/C	6.3-8dB+
ER-20	duplex	20km	16QAM	C	6.5dB
ER	duplex	40km	16QAM	C	12dB



Chromatic Dispersion

Chirp=0 / Component BW=120GHz



Different grid
Cooling needed

3.2T 2xFR4-2km
Uncooled

#	Wavelength [nm]
1	1291
2	1301
3	1311
4	1321

Chirp management

#	Wavelength [nm]
1	1271
2	1291
3	1311
4	1331

3.2T FR8-2km
LAN-WDM grid

#	Wavelength [nm]
1	1291.1
2	1295.56
3	1300.06
4	1304.59
5	1309.14
6	1313.73
7	1318.32
8	1322.97

3.2T FR8-1km
Uncooled grid

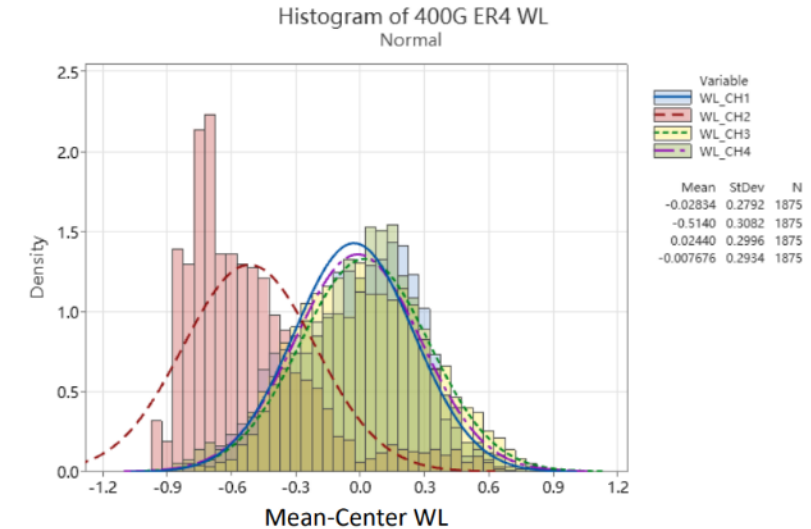
#	Wavelength [nm]
1	1271
2	1281
3	1291
4	1301
5	1311
6	1321
7	1331
8	1341

No cooling
Leverage FR4 lasers

3.2T FR8-2km grid

- Optical switching benefits from bidi fiber solutions to reduce necessary port count (see 1.6T FR8)
- **3.2T FR8 not possible in 10nm** spacing and requires LAN-WDM
- Production modules show a +/-1nm accuracy of fixed LAN-WDM lasers (see top right)
- 400G PAM4 spectrum is around 1.3-1.4nm depending on FEC choice and signal roll-off
- >3.4nm passband of the filters needed to support fixed cooled lasers
- Passband of LAN-WDM filters is ~2.3-3nm at best (see bottom right)
- ➔ FR8 would require a tighter laser accuracy of +/-0.5nm
- **Coherent optics** become a challenger for a duplex 3.2T FR 2km and can cover a 1km-10km/8dB+ application space easier than IMDD

LAN-WDM laser accuracy



[yu 3dj optx 01a 240829](#)

LAN-WDM filter specs

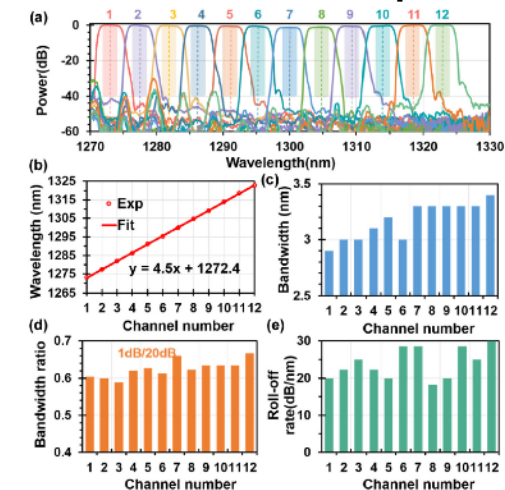
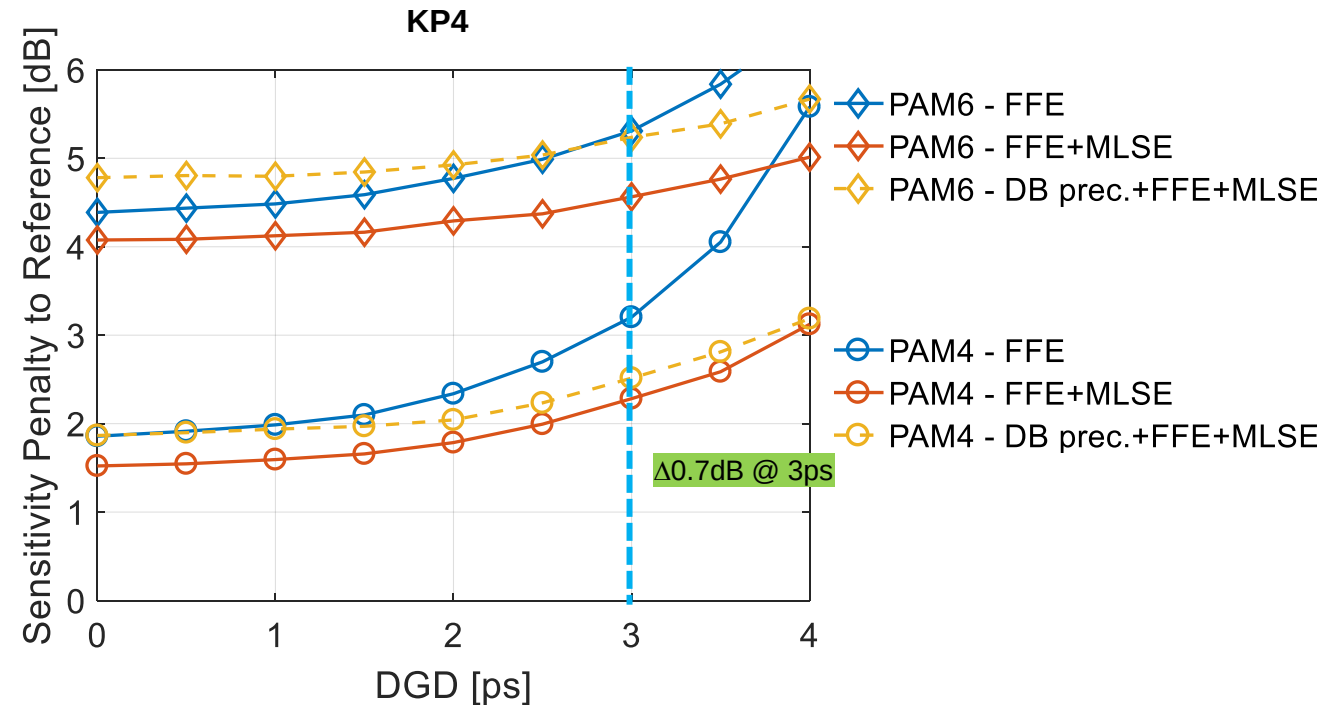
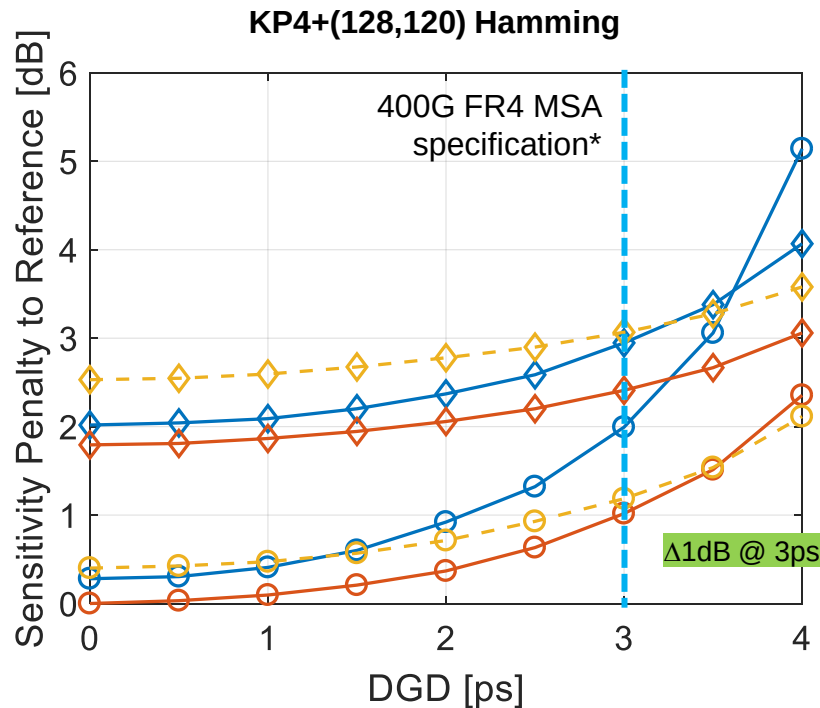


Figure 5: Measured results of the proposed twelve-channel LWDM filter. (a) Measured spectral responses at the drop port of the fabricated 12 channel LWDM filter; the center wavelength (b) and the 3-dB bandwidth (c) of all twelve channels.

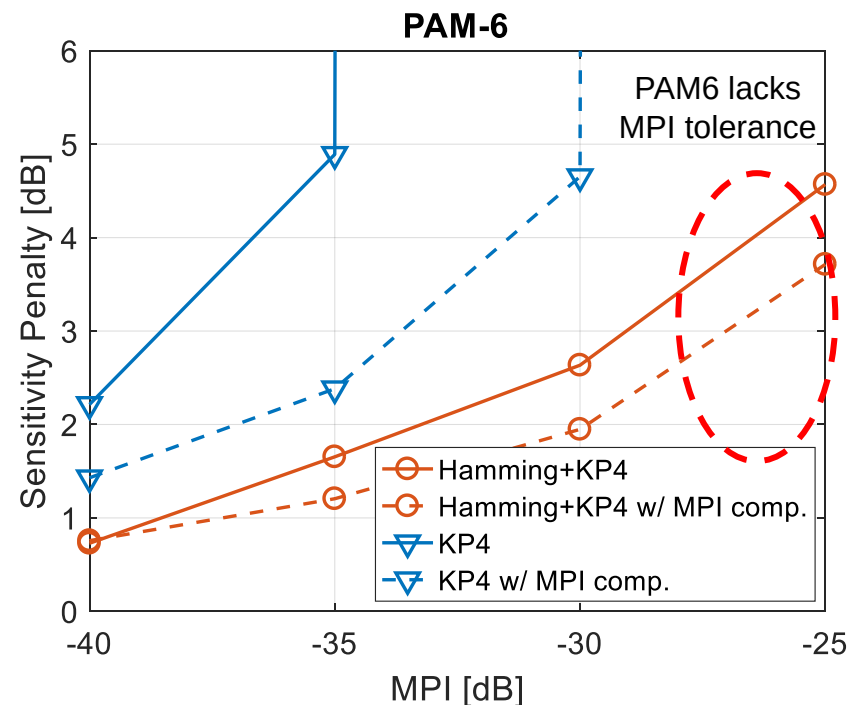
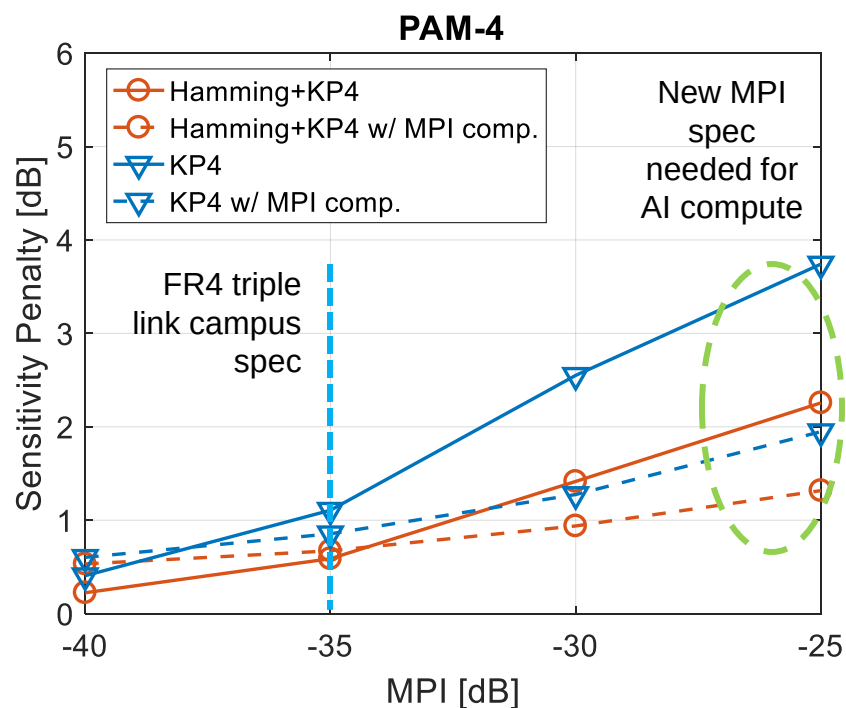
<https://doi.org/10.1515/nanoph-2023-0665>

Differential group delay (DGD)



- Historically up to 0.7dB penalty were allocated to the maximum DGD
- 400G/lane PAM4 with MLSE can cover the legacy 400G FR4 MSA specification of $DGD_{max} = 3ps$
- [ferretti 3dj optx 01b 240829](https://www.100glambda.com/specifications/category/2-specifications) discussed a further possible reduction to the max $DGD_{max} = 2.28ps$, which can be supported by a FFE-only with both legacy FECs

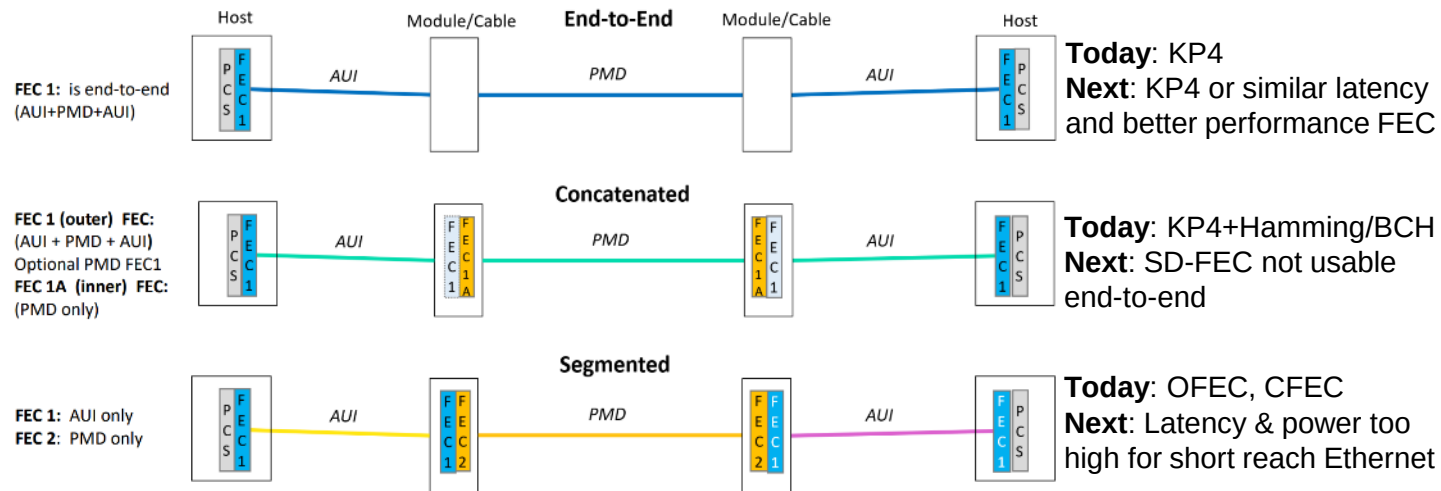
MPI tolerance



Tougher MPI spec needed

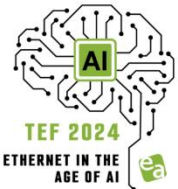
- New AI data center deployments happen rapidly and the air is more dusty with MPI > -35dB
- Linear drive use cases will lead to more reflections in the analog signal path
- PAM4 is the preferred modulation format due to higher intrinsic MPI tolerance for KP4-only

FEC performance and latency



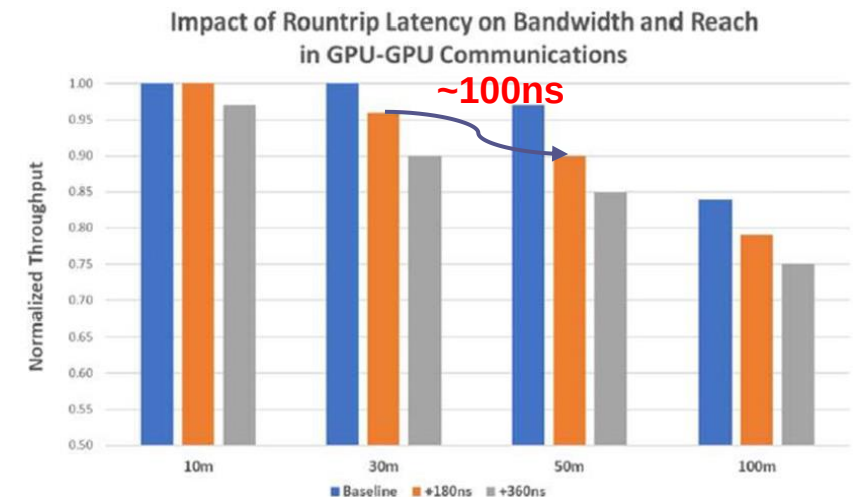
Next generation Ethernet FEC requirements:

- In practice, KP4+Hamming only gives ~1.7dB gain vs. KP4
- **SD-FEC not useful** in most scenarios (retimers, DSPs), since the soft information from the channel is not preserved until the decoder
- Impact of added **latency** quantified for GPU to GPU communications and is an optimization criterion (see bottom right)
- **FEC overhead** larger than KP4 increases ADDA power consumption and relative channel penalties (CD, DGD at 2km)
- **KP4** will already be in the host chip due to interop with 200G AUI applications



Ethernet FEC at 800GbE and 1.6TbE

Codes	Overhead	BER threshold	Latency @ 1.6TbE
KP4 RS(544,514)	5.8%	2.2e-4	49.8ns
KP4 RS(544,514) +Hamming(128,120)	12.9%	4.8e-3	90.4ns
KP4 RS(544,514) +BCH(126,110)	21.2%	1.2e-2	96.8ns



© NVIDIA

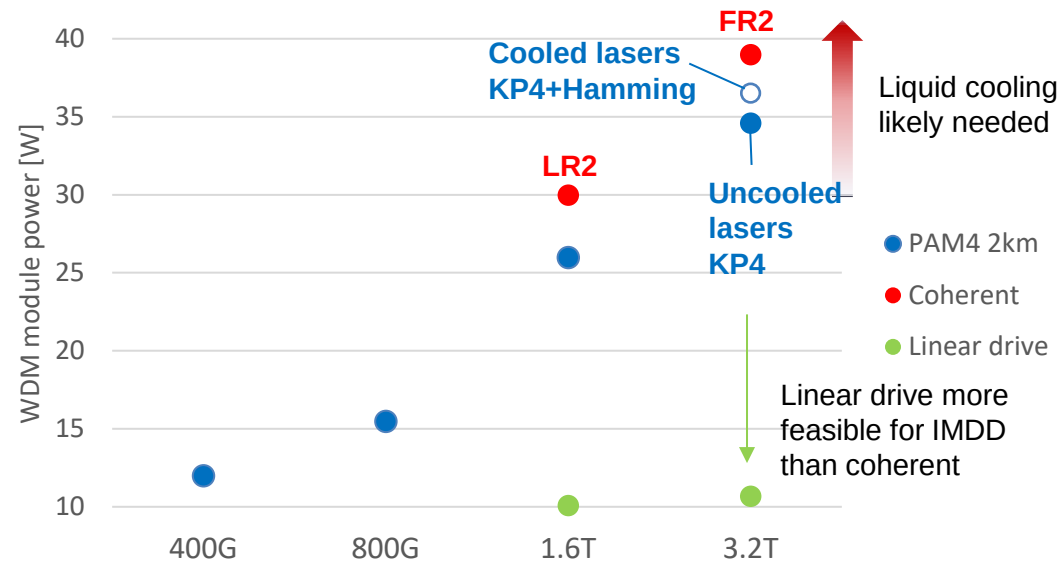
10/28/2024

Coherent optics architecture for 2km Ethernet

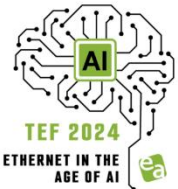
Component	IMDD vs. Coherent	Future power/cost penalty of coherent
DAC	DACs for ZR coherent today might have more fidelity and power than IMDD, but going forward a single DAC architecture for lower distances can address both. However, coherent can't compete with KP4-based ADDA for PAM4	Rate/power to remain higher
ADC	Today, coherent ADCs use more power due to higher oversampling of the signal. 1sps ADCs however are possible for coherent (Issue with DGD penalty & optimum sampling phase needs to be solved). FEC overhead to define the power consumption	Rate/power to remain higher
Driver/TIA	Not fundamentally different with likely higher swing for coherent drivers depending on the used modulator (EML vs. MZM) and the material (InP, TFLN, etc.)	Similarity depends on the used optics
Modulator	At 400Gb/s MZMs could become a joint platform for both IMDD and coherent. Coherent has additional components (phase shifters and PBC) and fundamentally higher insertion loss for the modulators, which impact the laser output power requirement. TFLN MZM would need more laser power vs. EMLs.	DP-IQ-MZM vs. EML cost difference for WDM
Laser	@ FR4: 4 uncooled fixed lasers for IMDD vs. 1 high power tunable laser for coherent. Tunability will always induce a power gap vs. IMDD, but it can be reduced to a smaller level for gray links. Optical power of coherent lasers must always be higher than IMDD due to higher insertion losses in Tx/Rx	Inherent power and cost gap
Rx	Coherent has higher insertion loss and complexity	ICR a bit more complex than PD
Equalization	Coherent needs 4x4 MIMO equalizers + CPE, while IMDD needs 4x1 filters only	Inherent power gap
FEC	Coherent KP4+BCH FEC (OIF) has a high overhead and leads to a higher ADC sampling rate and power. Ideally lower rate FEC should be used for coherent (KP4+Hamming). But: KP4-only not feasible, due to inherently higher error floor of coherent	Inherent power gap

10/28/2024

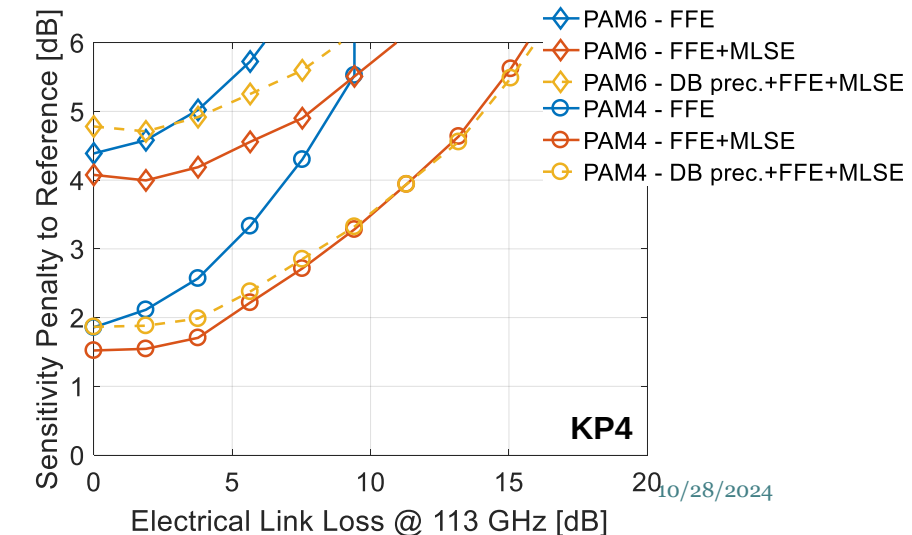
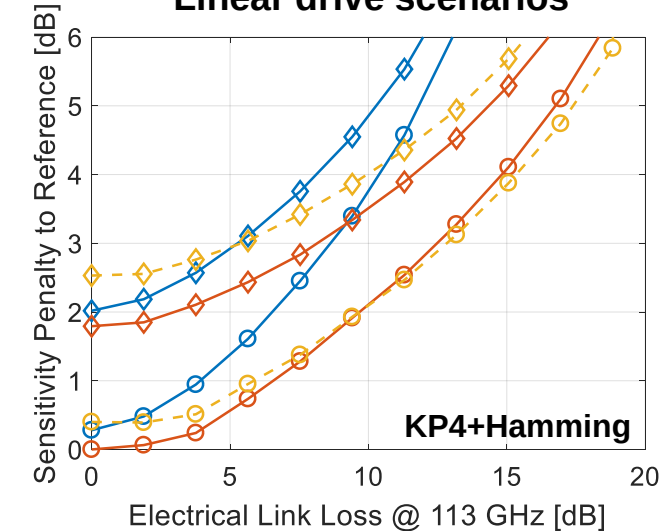
Optical module power consumption



- Majority of module power consumption is in the DSP ASIC (**70%-75%**)
- Most of the DSP ASIC power is allocated to SerDes and ADDA (**75-80%**)
- Native 400G/lane transmission doesn't support classical C2M PCB based transmission and would require cabling
- 400G AUI could lead to a stronger move towards **linear drive** based networking systems to drastically reduce OPEX and simplify switch design (no cabling, no liquid cooling)



Linear drive scenarios



Conclusions

400G/lane PAM4 vs. PAM6

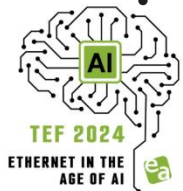
- PAM6 has certain benefits for electrical links, but is not suitable for optical links due to high MPI penalties
- Linear drive optics to play a crucial role at 400G/lane (requires same modulation for electrical & optical, host FEC only)
- 400G copper cables for intra-rack GPU interconnect will have to be active to cover similar reaches to 200G DACs. Support of PAM6 could be a design choice in the AECs.

FEC

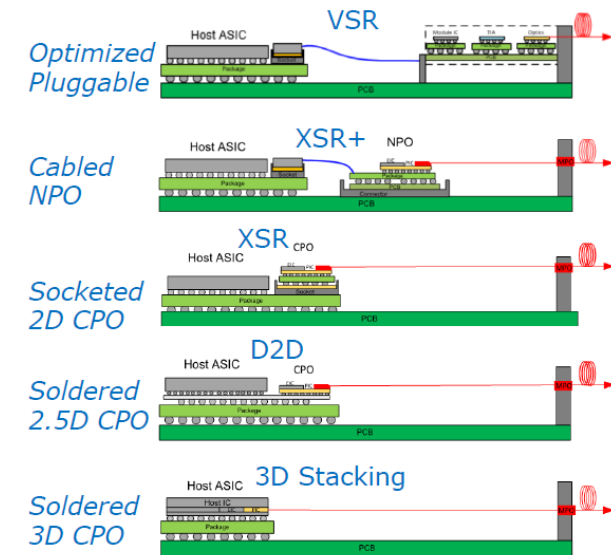
- KP4 will already be part of future hosts due to 200G legacy interoper
- Having a higher power/latency FEC is undesirable

3.2T IMDD vs. coherent

- 3.2Tb/s IMDD FR8 duplex/bidi will be limited by CD
- 3.2Tb/s FR8-1km could be feasible in a 10nm grid with uncooled lasers enabling EMLs and MZMs
- 3.2Tb/s coherent can cover a flexible range of 1-10km for all inter-building interconnects for future DC campuses but unlikely to support linear drive architectures due to higher power of ADDA & DSP/FEC

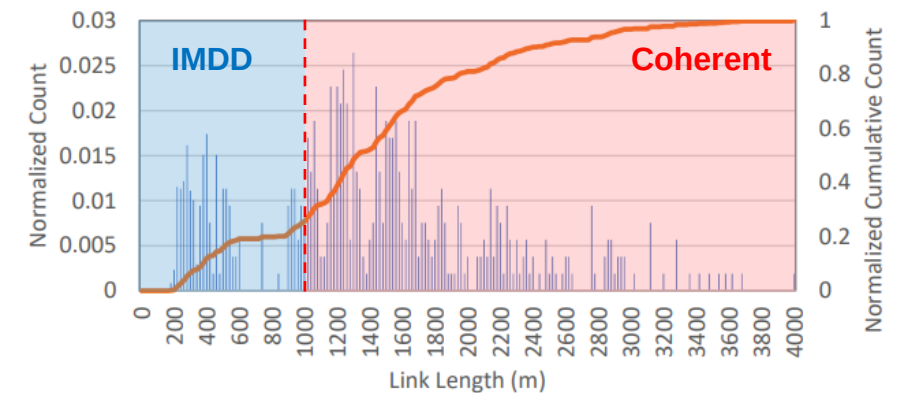


400G/lane architecture support



3.2T duplex / bidi

Outside Hall Fiber Link Lengths

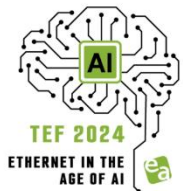


[welch_3dj_02a_2309.pdf](#)

Navigating the AI-driven Evolution: Next-Gen Transceiver and Interconnect Testing Challenges:

Consideration of test tools and practices for successful 400G deployments.

David J. Rodgers,
Sr. Business Development Manager - EXFO
Events Chair and BoD member - Ethernet Alliance



Abstract

- The landscape of high-speed optical communication is undergoing rapid evolution, fueled by the surge in artificial intelligence (AI) applications. This surge has led to an increased demand for higher data rates, reduced latency, and more efficient network architectures. Next-generation transceivers are at the forefront of this transformation, incorporating innovative approaches to address these challenges.
- These advanced transceivers, including the 800G-ZR and 200G/lane optical modules, have transcended their traditional role as mere transmitter-receiver pairs. They now function as intelligent modules, incorporating features like forward error correction (FEC) capabilities. As we step into exciting new times, these transceivers also introduce new testing challenges. They revolutionize the way we test transceivers during development, manufacturing, and deployment.
- Cutting-edge test solutions that closely mimic real-world operating scenarios play a pivotal role in accelerating development timelines and mitigating interoperability risks in the field.

Test Tools

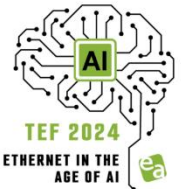
Key Considerations

What will 400G look like?

How will Signal integrity and Protocol test paradigms need to evolve?

Can test tools adequately scale out to support capabilities roadmap and customer timelines?

Will 'interoperability' mandate conformance certification?



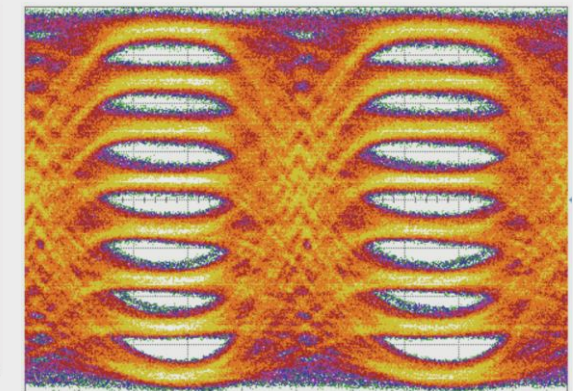
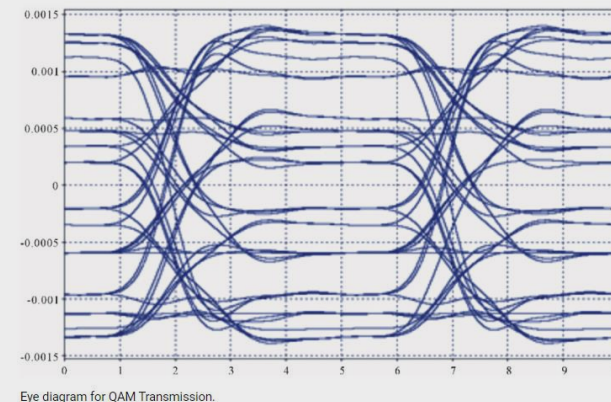
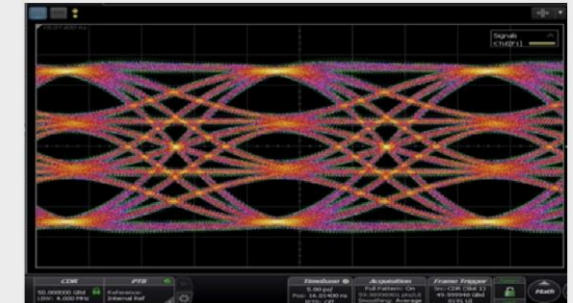
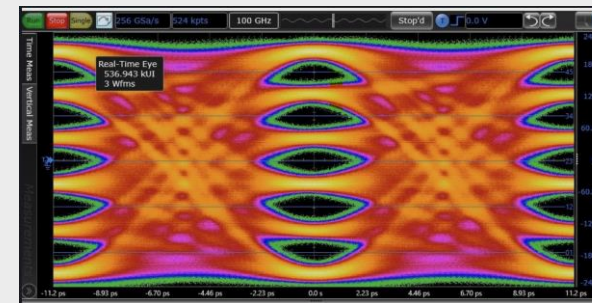
What will 400G look like?

Modulation scheme (PAM4 or PAM8, QAM) will dramatically impact test tools and schemes.

- More bits/clock or higher baud rates will impact test tools/requirements.

How much of the 200G test bench can we leverage for 400G?

Will electrical and optical signaling align?



How will SI and Protocol test evolve?

Baud, Bit/clock will drive T&M roadmap and product evolution

- Protocol tools becoming more SI aware.
 - Can we adapt existing tools?
- Certainly, closer coordination/integration will be required.
- “Competitors” will also become “Collaborators”.



Scaling out test tools for capabilities roadmap?

There must be synergy for supporting specification influences and adaptations to meet customer/market timelines.

- Influencers include SIGs, multiple MSA(s), and unique requirements.



Is ‘interoperability’ conformance required?

Will interoperability be sacrificed for the sake of speed?

- Is interoperability a requirement?

Will there be bifurcated “interoperability”?

- Lower speed breakout, specific modules/interconnects.

Would the industry support conformance certification?

- Building blocks, sub-assemblies, finished products.



Test Tools

Key Considerations

What will 400G look like?

How will Signal integrity and Protocol test paradigms need to evolve?

Can test tools adequately scale out to support capabilities roadmap?

Will 'interoperability' require conformance certification?

Greater coordination between electrical and optical test tools.

Leveraging 200G test enhancements for 400G applications.

200G moved the FEC needle to improve error tolerance and mitigate latency. Will 400G require architectural changes?

How will AN and LT be mandatory for both Electrical and Optical performance and interoperability?

Modulation scheme (PAM4 or PAM8, QAM) will dramatically impact test tools and schemes.

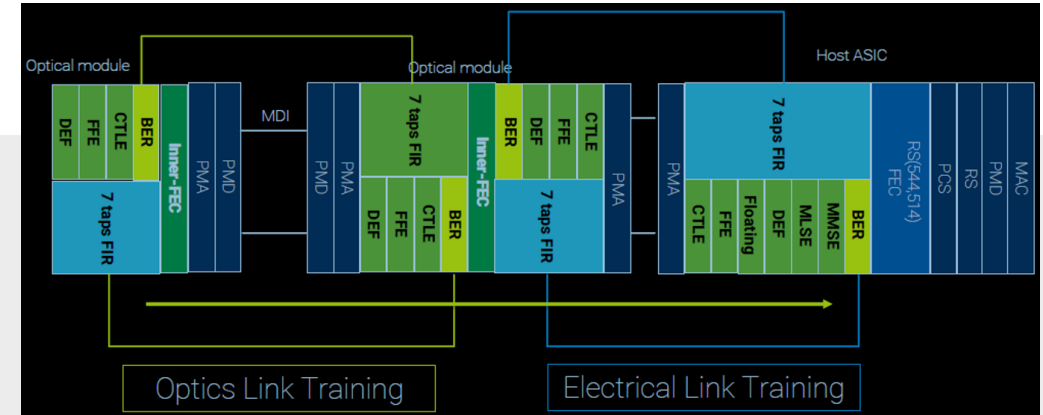
Test requirements now include intelligent modules/cables.

Adapted tools will be required early on for the 400G journey. Close collaboration btw SerDes and T&M vendors a must.

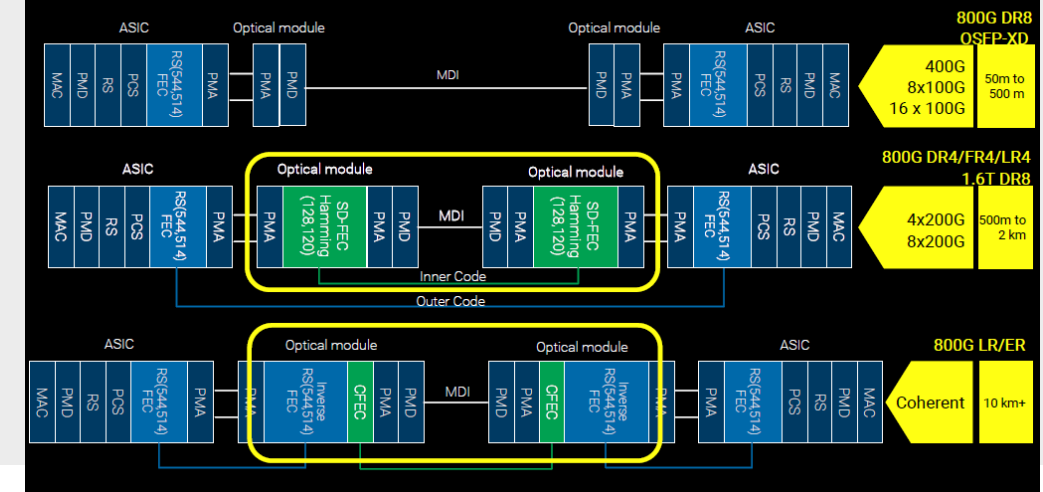
Auto-Negotiation and Link Training

Recap of 802.3dj AN & LT

- 200GbE proposal adapts CR/KR style AN & LT for C2M
- Leverages existing LT to include FECi on/off control in Optics
 - More intelligence interconnect options
 - Better interoperability



Next Gen Optics (Module Types)



Testing for Optical Link Training?

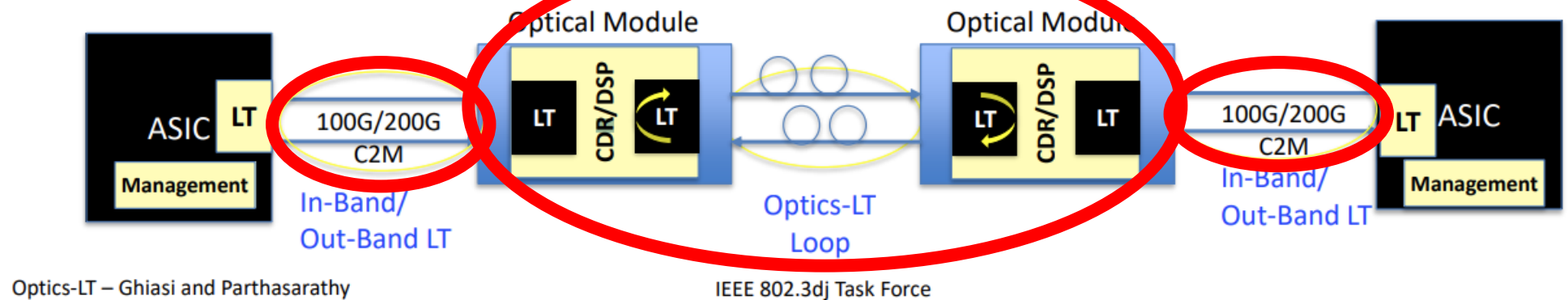
C2M LT, then module to module, simultaneous?

- Module specific switch ports?

Optical link-up is challenging

Mitigate SI/Burst Error conditions

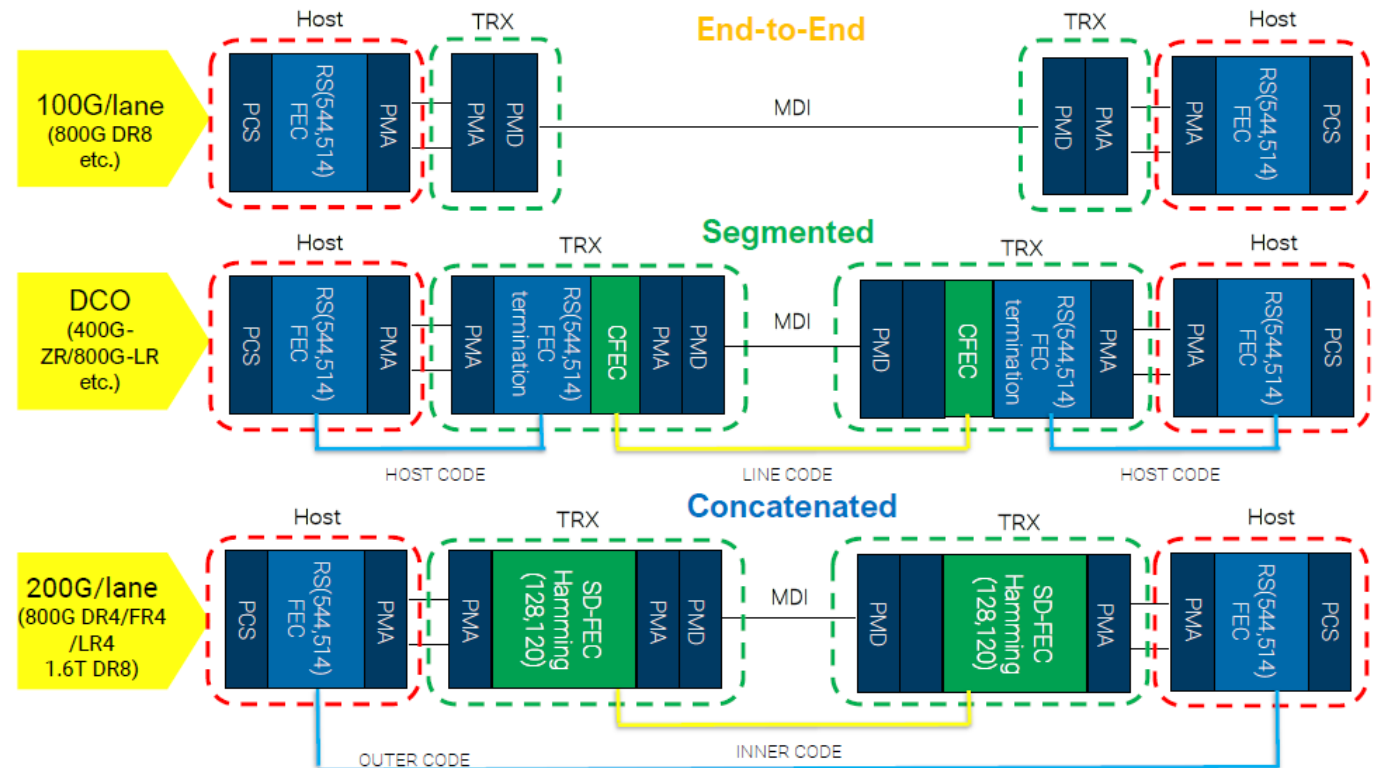
Adjusting the TX based on the RX



Factoring in FEC

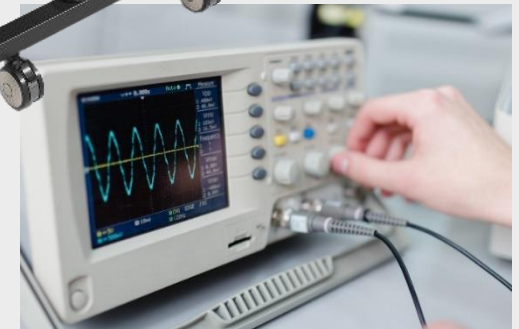
200G moves the FEC needle to improve burst error tolerance and reduce latency.

Will 400G require architectural changes?



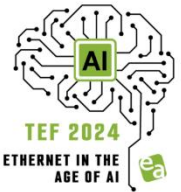
Call to action

- Close coordination between IP and Test community
 - Chip suppliers to engage test community early
 - Robust test ecosystem increases successful adoption
- Convergence of & Cooperation between test tools
 - Component and advanced system testing required



Scaling AI with New Optical Connectivity Technologies

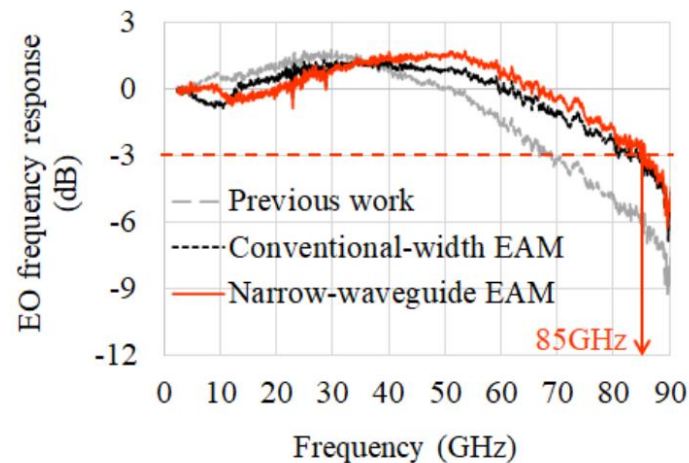
Tony Chan Carusone
Chief Technology Officer
Alphawave Semi



400G Optics - EML

- > 90GHz EMLs already being demonstrated
- Further enhancements on the way
- However, chirp + CD promises to impose reach limitations

OFC24 - Asami Uchiyama [Mitsubishi]



ECOC'24 - [Lumentum]

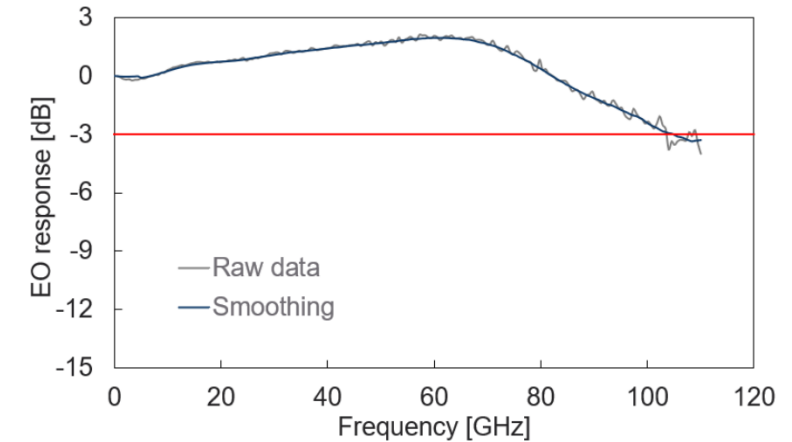
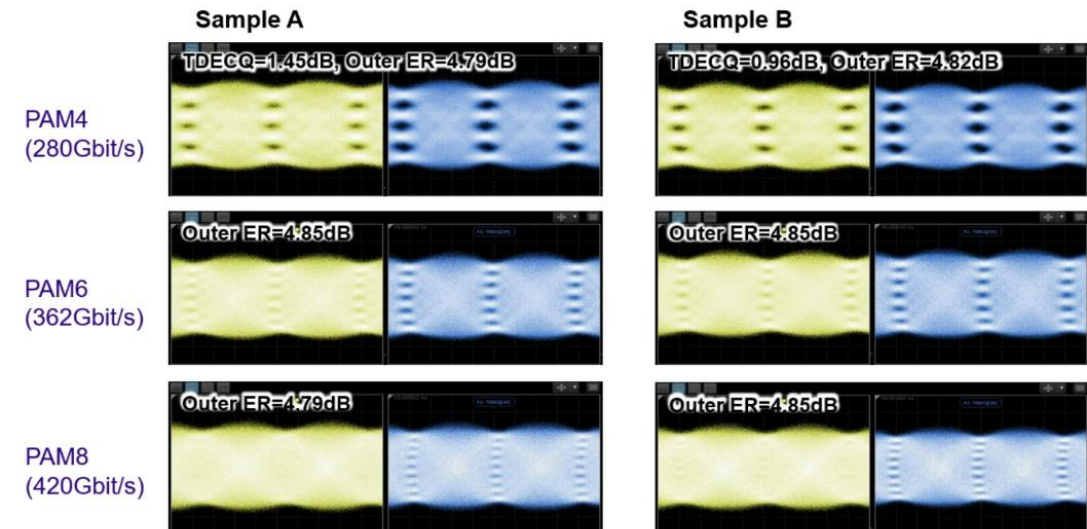


Fig. 4: EO-response at 25°C.

OFC24 - [Sumitomo]

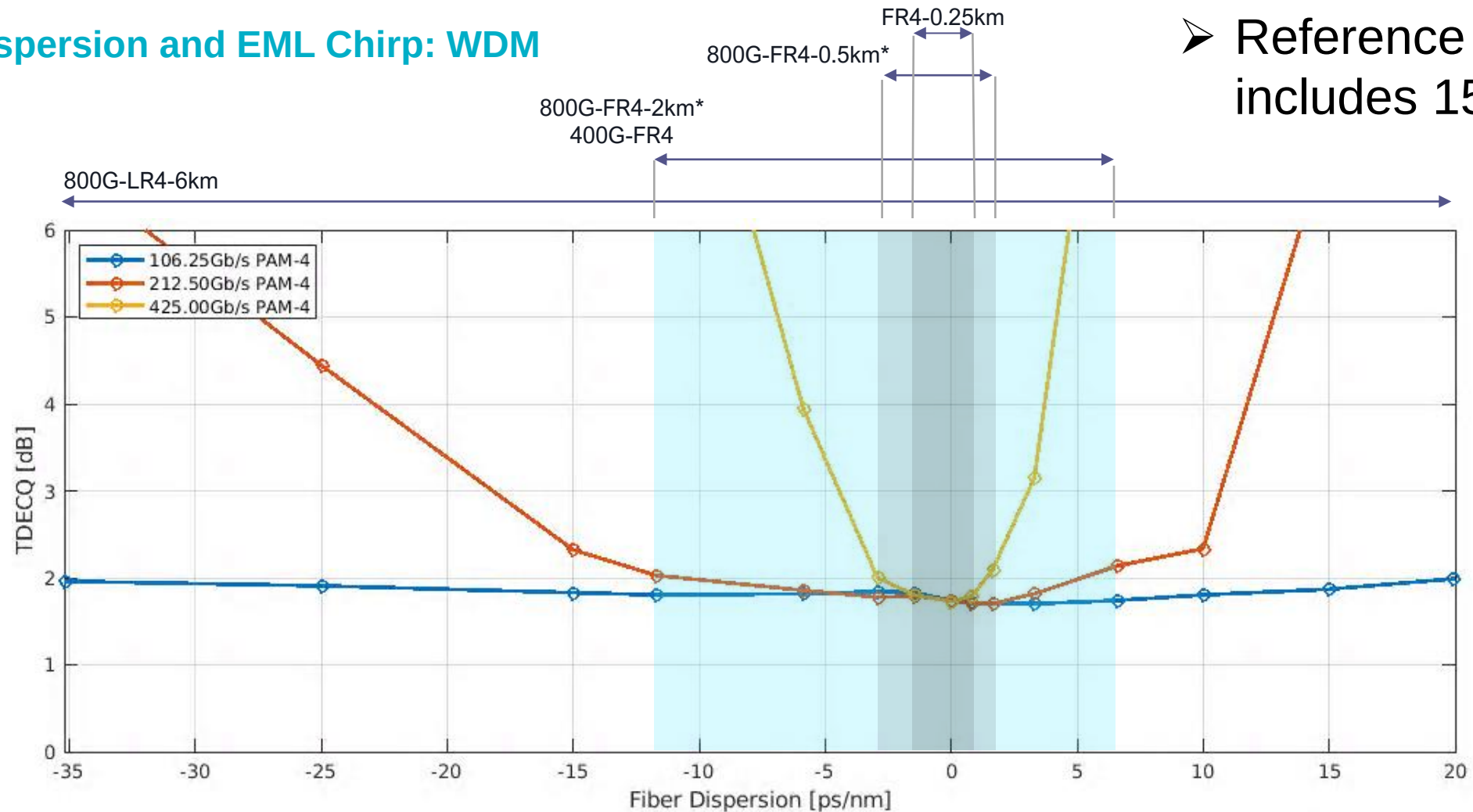
140Gbaud eye, PAM4/PAM6/PAM8



400G Optics Impairments

Fiber Dispersion and EML Chirp: WDM

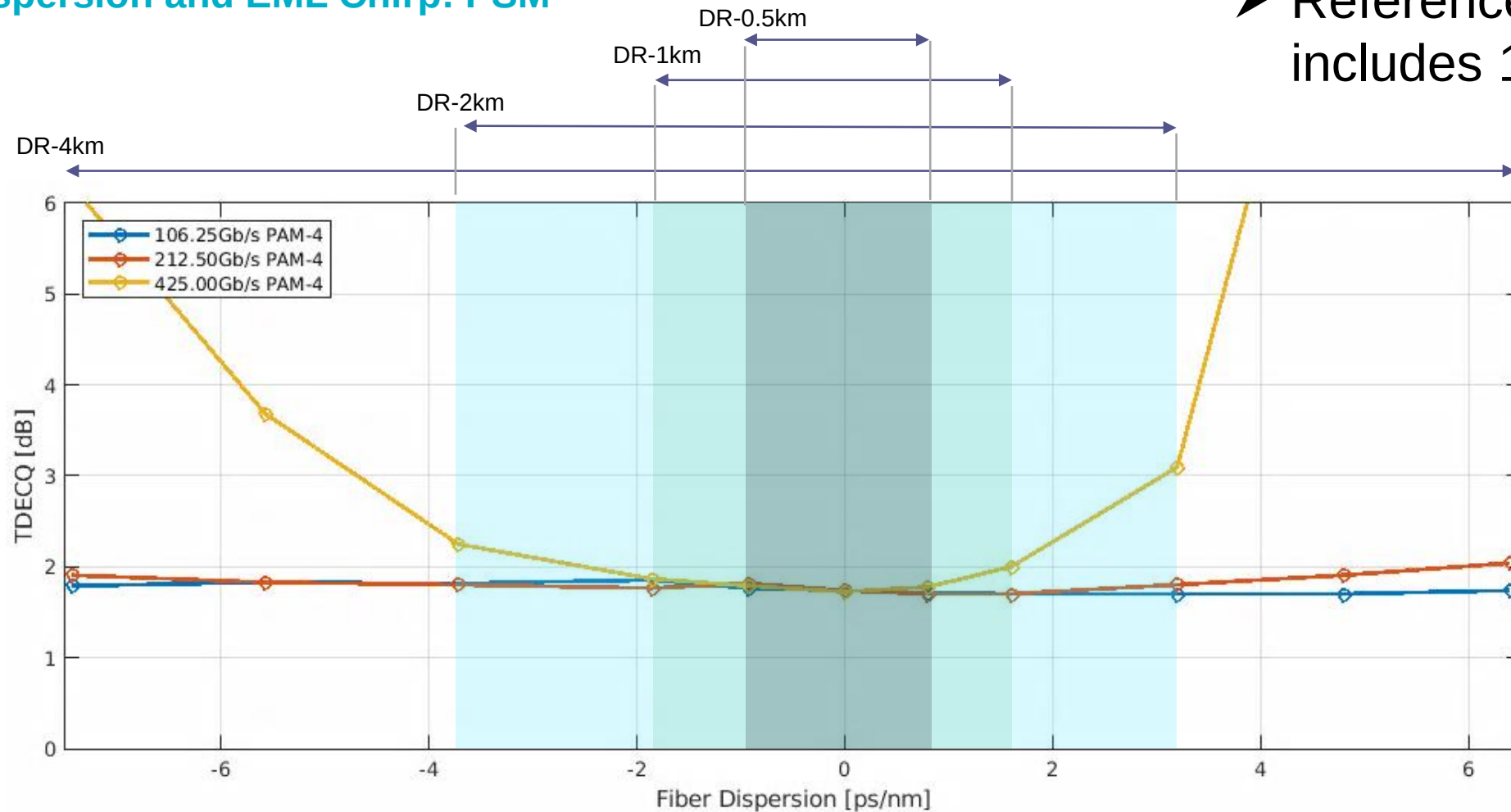
➤ Reference receiver includes 15-tap FFE



400G Optics Impairments

Fiber Dispersion and EML Chirp: PSM

➤ Reference receiver includes 15-tap FFE

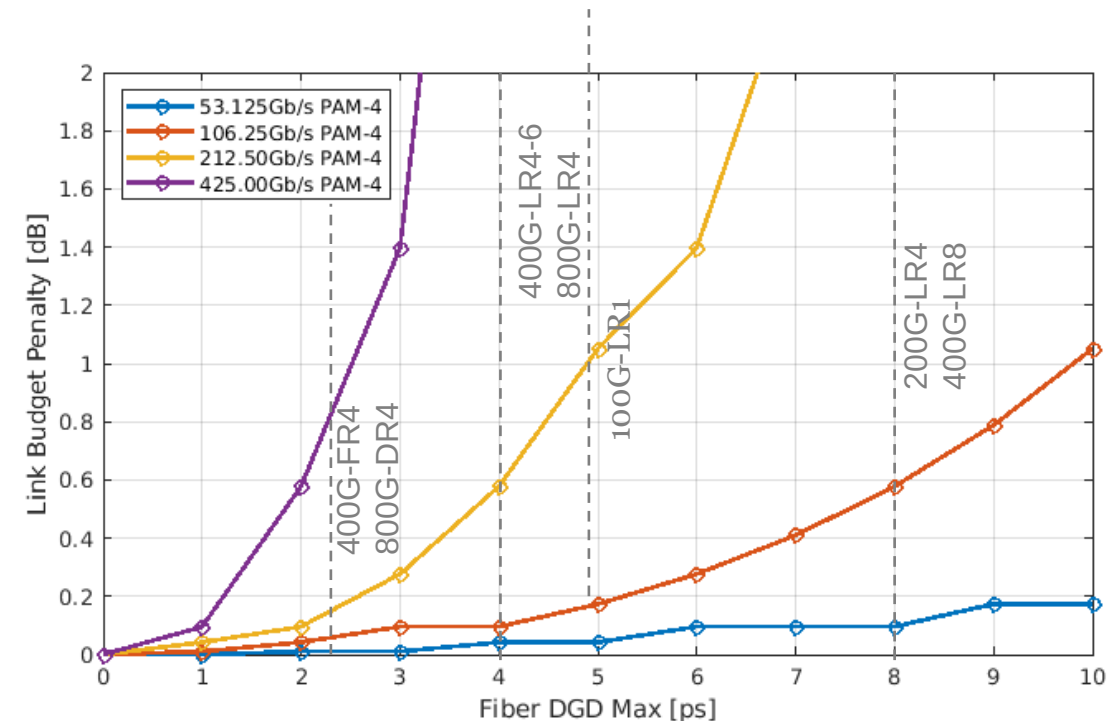


400G Optical Feasibility

Fiber Differential Group Delay

- Link penalty due to DGD has been negligible over short reaches
- At 400G/lane, DGD penalty becomes significant over FR4/DR4 distances
- Will fiber specs need to be tightened or the added link penalty accommodated?

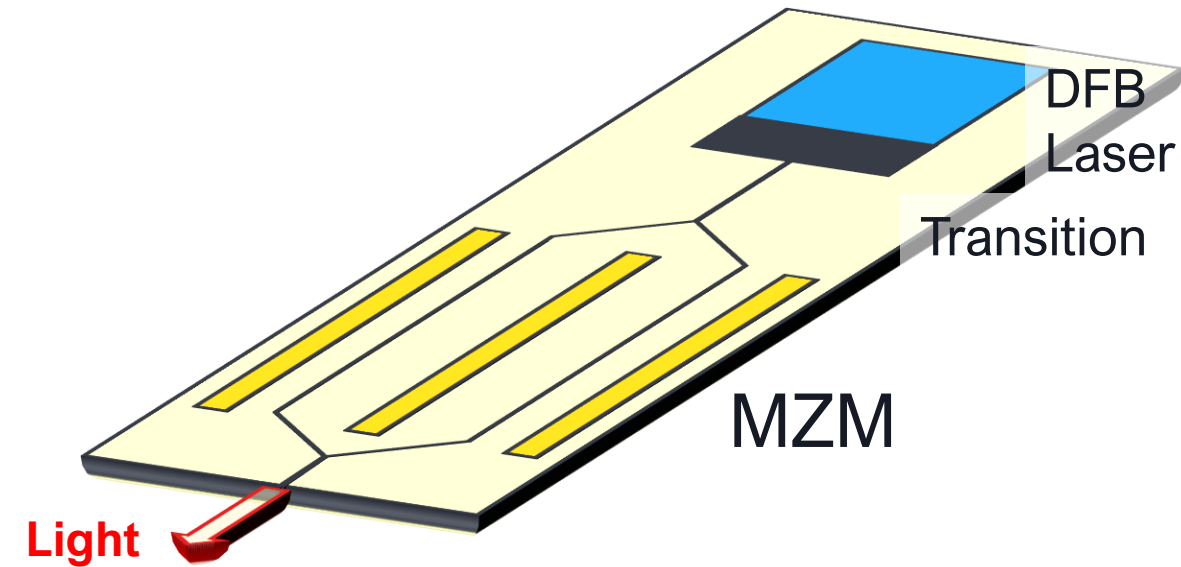
➤ Reference receiver includes 15-tap FFE



*Link penalty is calculated based on OMA-TDECQ in the transmitter

400G Optics - New Material Systems

- Silicon MZMs do not currently show potential for 400G 4-PAM or 6-PAM signaling
 - May necessitate use of III-V MZMs, or other material systems
 - ⇒ higher cost and lower maturity
- Integrated laser + MZM on an InP platform
 - e.g. [S. Lange et al., "100 GBd Intensity Modulation and Direct Detection With an InP-Based Monolithic DFB Laser Mach-Zehnder Modulator," JLT 2018]
 - Compatible with differential drive
- TFLN on Silicon

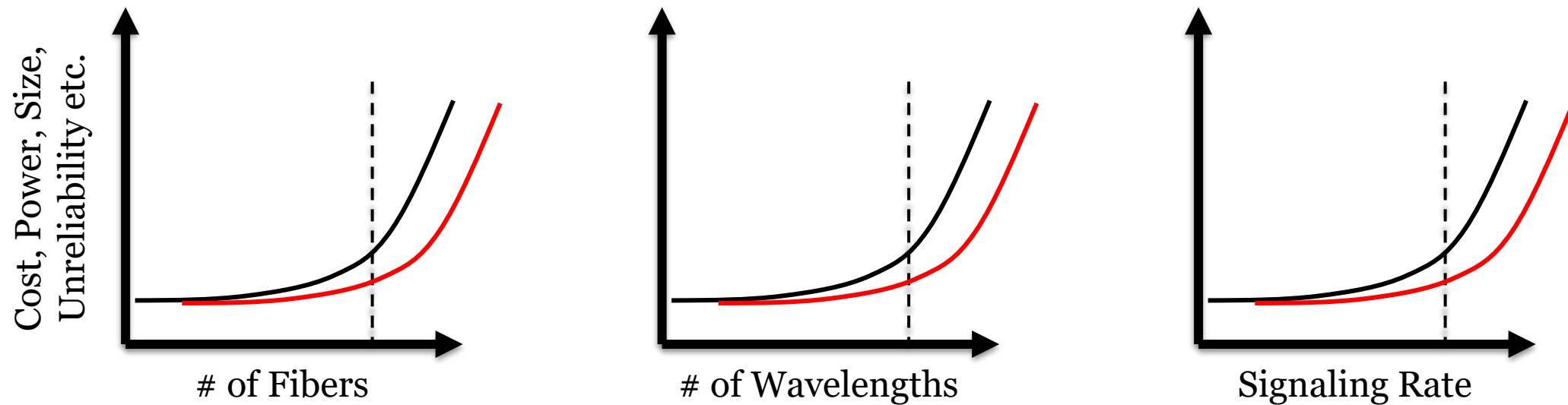


- Expect higher cost and lower maturity than EML technology

400G Optics Summary

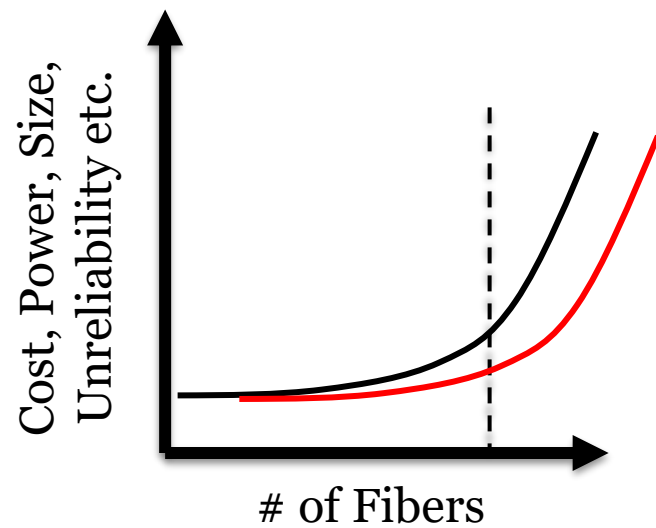
- Chirp + Dispersion & DGD threaten to limit the reach of 400Gbps PAM signaling
 - MZMs in other material systems (e.g. InP or TFLN) may have less chirp but introduce added cost, lower maturity
- Rising interest in alternative optical connectivity technologies

Maximizing Optical Connectivity for AI

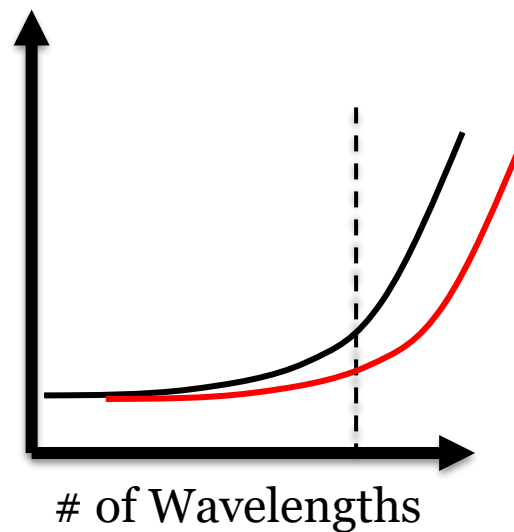


Operate links at the inflection point
R&D pushes hockey-sticks to the right

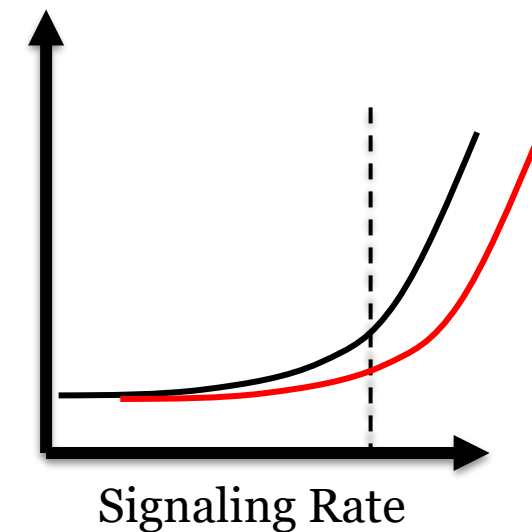
Maximizing Optical Connectivity for AI



- Facilitates breakout, but complicates manufacturing, assembly, reliability



- Link budget penalties due to optical losses of the wavelength mux/demux & additional CD



- Analog front-end circuit limitations
- Additional DSP complexity

Trends in Optics for AI

Summary of 400G PAM (IMDD) Optics for AI

- New analog bandwidth requirements
- IMDD FEC continues to swell, approaching coherent FEC in terms of power consumption, gate count (silicon cost), and latency
- Optical components becoming more expensive
 - Pure-silicon modulators do not show potential for sufficient bandwidth
 - EML chirp, in combination with CD, introduces significant link penalties
- DSP power, circuit area and/or implementation penalties likely to increase
- Optical link loss budgets also becoming challenging
 - AI scale-out, even across multiple sites will necessitate long reaches
 - Multiple product announcements on optical circuit switching (OCS) for AI networking may challenge optical link budgets

800G/λ Coherent-lite

- Analog bandwidth requirements comparable to 200G PAM
- FEC complexity comparable to 400G IMDD
- Optical component requirements comparable (or even easier) than 400G IMDD:
 - A single gray laser per 800G λ without wavelength-locking
 - Pure-silicon modulators already demonstrated
- Use of O-Band mitigates CD compensation power, gate count and latency compared to previous coherent DSPs
- Link budget affords additional margin over 400G IMDD to accommodate the longer reaches of scale-out networks and/or OCS losses

Coherent vs. IMDD for 1.6T and beyond

	IMDD		Coherent	
	Discrete Optics (e.g. EML)	Integrated Optics (SiPho, hybrid,...)	ZR	Tailored to Intra-Datacenter
Tx DSP Relative Complexity	1.0	1.0	1.2	1.0
Rx DSP Relative Complexity	1.0	1.0	1.5	1.1 – 1.2
DSP Relative Latency	1.0	1.0	2.0	1.0
Laser Comparison: Relative no. of lasers TEC required Total Laser Power Linewidth Wavelength Locking	4 Maybe High 1MHz No	2 (more if CWDM) Maybe High 1MHz No	1 Yes High 300kHz Etalon	1 Yes Low 1MHz No (DSP feedback)
Reach Limitations	CD + chirp, DGD	CD + chirp, DGD	No fund. limits	No fund. limits
Scale to 3.2+Tbps	Difficult	Difficult	Yes (high power)	Yes

Comparison of Coherent Rx DSP Architectures

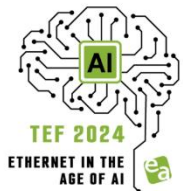
Block	Feature	Synchronous Baud-Rate Sampling	M/N x Oversampling
ADC	ADC sampling rate	f_{baud}	$f_{baud} \cdot \left(\frac{M}{N}\right)$
	ADC ENOB	5.5	6
Timing loop	Timing phase correction	Analog clock phase shifter	Digital Polyphase Re-sampler
	Timing error detection	Detector @1sps	Detector @1 or 2 sps
Data path	CD compensation	@1sps	@M/N sps
	Pol. MIMO	@1sps	@1 sps
	Phase recovery	@1sps	@1 sps

Summary

- Optics are demonstrating sufficient bandwidth to accommodate 200+GBaud signalling
- Depending on reach targets, optical impairments may challenge link budgets
- If IMDD solutions face higher cost, less mature optics, and higher power consumption in both the optics and DSP, coherent-lite solutions can play an increased role over shorter reaches

Interoperability in an Open Ethernet Ecosystem

Sam Johnson
Engineering Manager, Intel Corp



Interoperability Definition

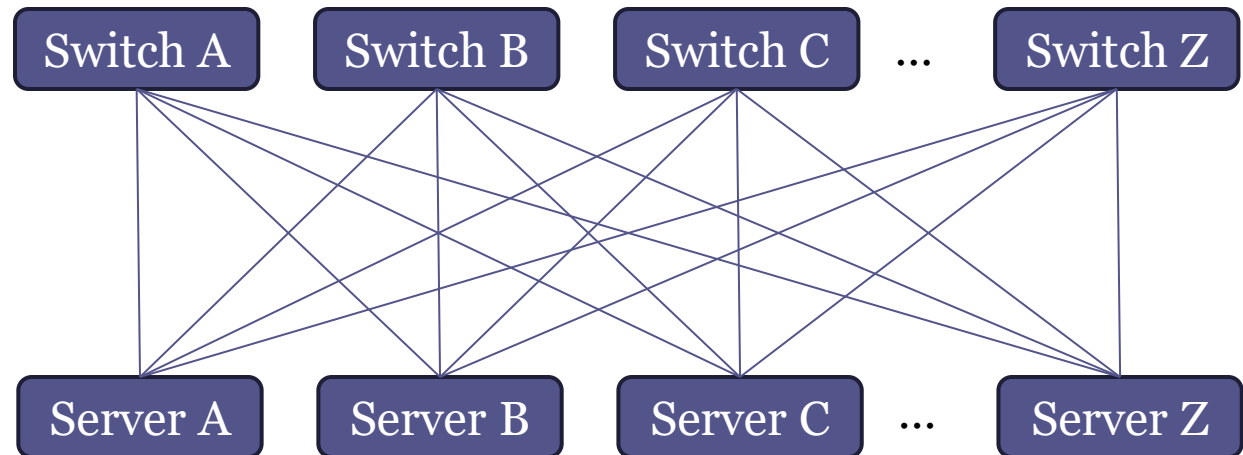
- **Interoperability:** The ability of devices to establish a reliable data communication connection for passing Ethernet traffic
 - L1 PHY connectivity: “*Link up*”
- **Open Interoperability:** The ability of an Ethernet device to work seamlessly with any other device
 - Requires compatible configuration of host, link partner, media
 - Consistent standards support
 - Device/vendor agnostic



Historical Interop Expectations

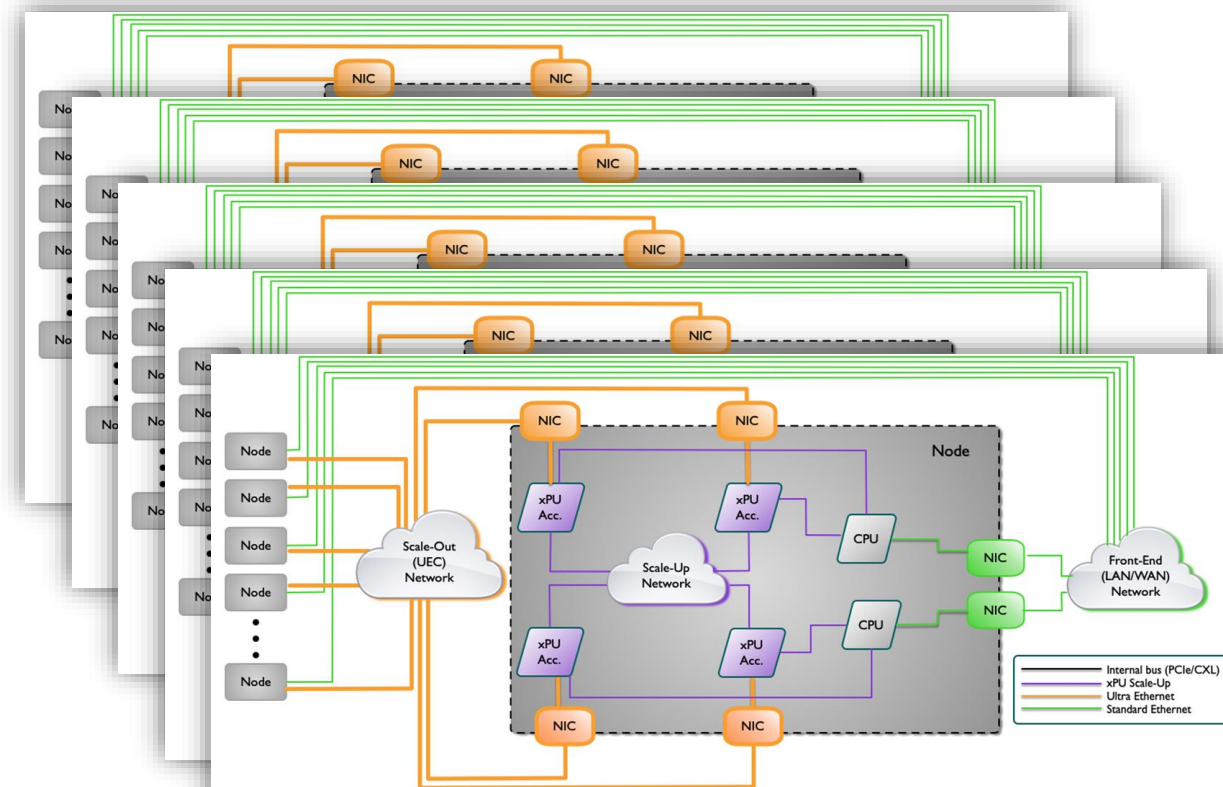
- Use Case: Any to any, plug and play
- Success Criteria:
 - Spec compliance + basic link and BER testing across ecosystem
- Interconnect: DAC and retimed optics
- Open spec/standards: IEEE, ETC, SFF

*“If the cables fit,
you must link with it”*



Interop Requirements for AI

- Use Case:
 - Large scale deployment of consistent configurations
 - Limited devices, few to few
 - Speed is king
- Success Criteria:
 - Flawless reliability, per use case
 - Greater than spec data integrity
- Interconnect:
 - DAC(?), retimed/linear/half-retimed/TX retimed optics/cables
- Open spec/standards:
 - IEEE, ETC, SFF, OIF, UEC, LPO MSA, UALink



Source: Ultra Ethernet Consortium

Ecosystem Shifts Impacting Interop

Shift	Impact
Narrowing interop needs, per system	Reduced ROI to extensive interop testing
Diversity of Ethernet applications New specs, interconnect types	Configuration variety, increasing complexity
Pushing signal integrity boundaries Increased performance expectations	Advanced technology development (FEC, coherent, ILT, etc.) Host capabilities design dependent Link health metrics and margin testing
Inconsistent system capabilities, configuration variety	Need smarter hosts, end users Manual system configuration
Advanced technology implementation	Debuggability, protocol compliance validation
IP time to market	Early device specification compliance



Importance of Open Interoperability

- The promise of Ethernet: Compliant device functionality
- Open standards-based interop == Ecosystem health
- Value:
 - Ease of new product integration
 - Reduces released product support burden
 - Enables economy of scale
 - Allows extended applications and adoptions

The Fight for Open Interop

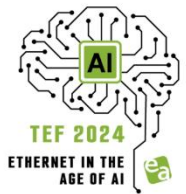
- Open interop remains critical for AI ecosystem development
- Beware of standards divergence, non-compliant functionality
- Maintain ability to integrate and expand applications
- Develop advanced debug tools and diagnostics
- Engage with EA interop initiatives

Ethernet Alliance Interop Mission

- Establish and demonstrate multi-vendor interoperability
 - Plugfest hosting
 - Interop demonstrations at industry events
 - Test and measurement methodology correlation
- Open standards support
 - Technology and standards incubation
 - Industry consensus building
 - Supporting emergent Ethernet-based applications



QUESTIONS?



10/28/2024