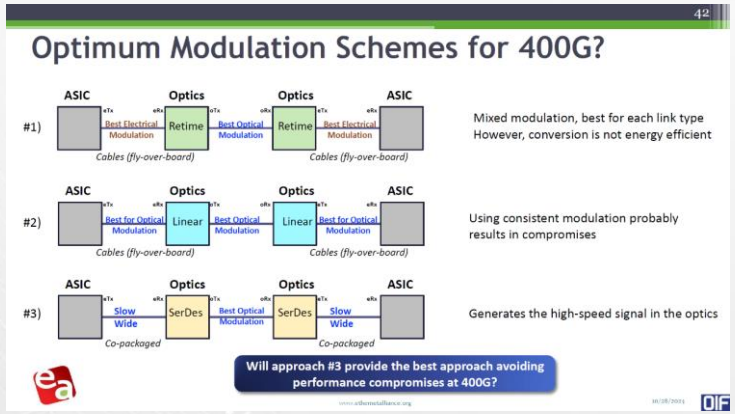
448Gb/s native O/E modulation format for AI compute networks: PAM4 vs. PAM6

Maxim Kuschnerov, Balazs Matuz, Tom Wettlin, Nebojsa Stojanovic, Stefano Calabro

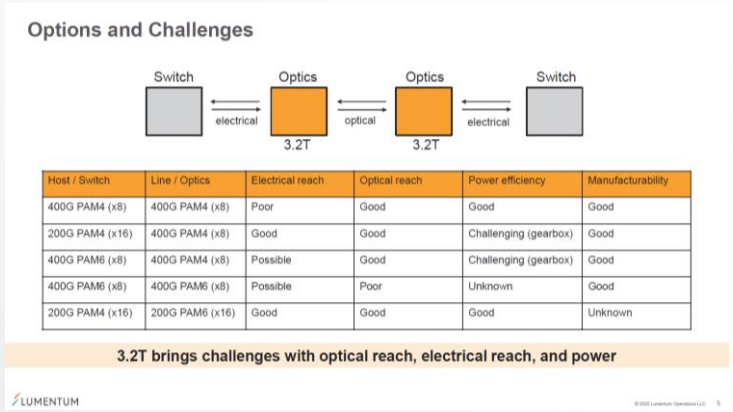
Our vision and mission is to bring digital to every person, home and organization for a fully connected, intelligent world.

OIF 448Gbps Signaling for AI Workshop - April 15-16, 2025

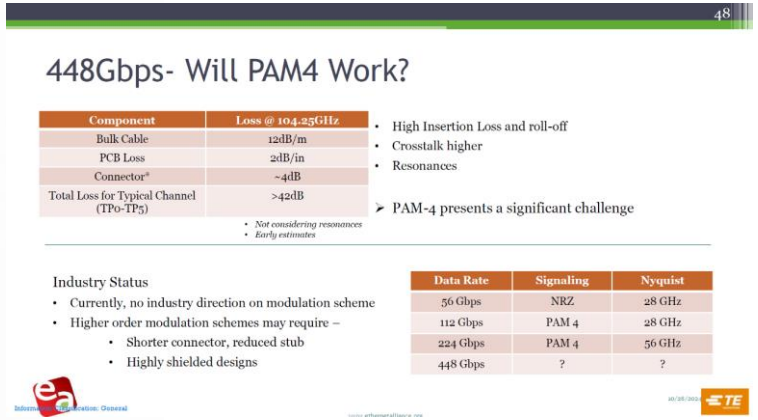
Industry on 448G O/E modulation (PAM4/6/8)



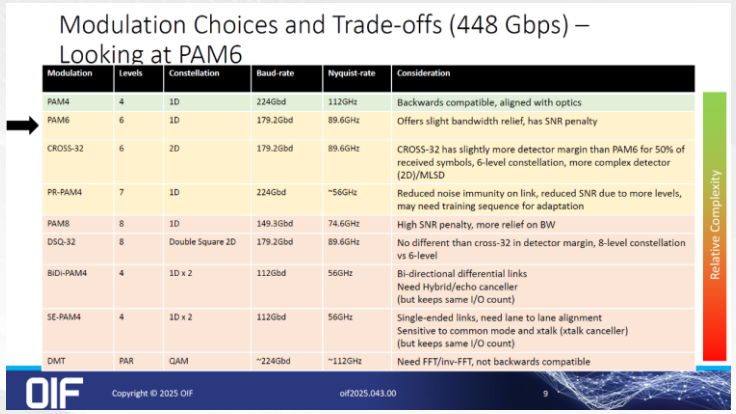
Jeff Hutchins, Ranovus



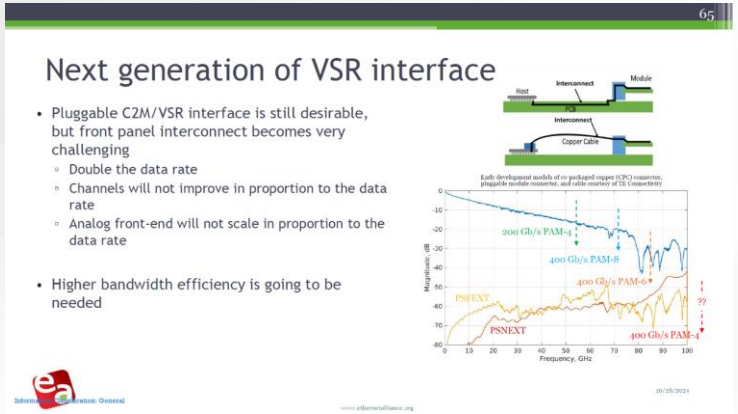
Lumentum



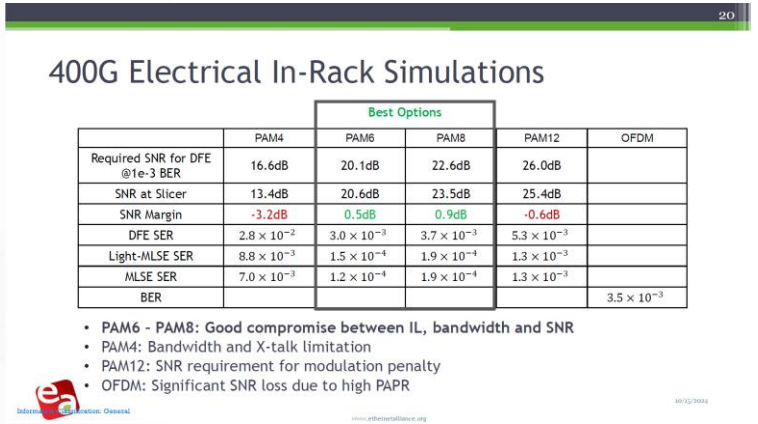
Ashika Shaji, Nathan Tracy, TE



Ken Lusted, Synopsis



Cathy Liu, Broadcom



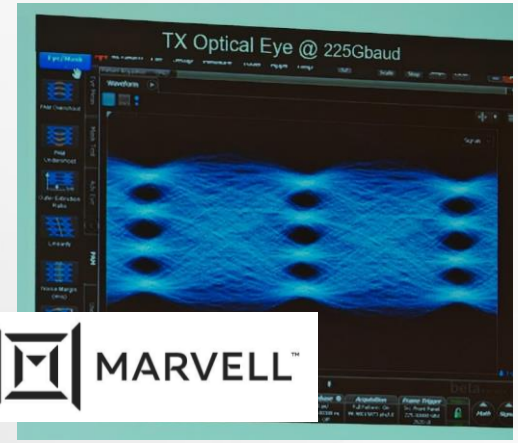
Halil Cirit, Meta

Bandwidth limitations on the electrical channel side are dominating the modulation discussion

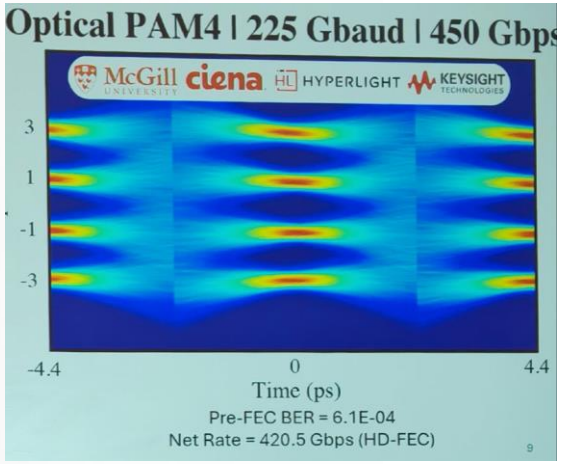
Optical PAM4 demonstrations at OFC 2025

- OFC 2025 has seen several demonstrations of 400G/lane optical feasibility
- TFLN achieves a higher bandwidth overall, with the highest EML baud rate shown by Lumentum
- SiP has been limited to 160-175Gbaud demonstrations
- First products will include gear-boxed solutions with 224G SerDes

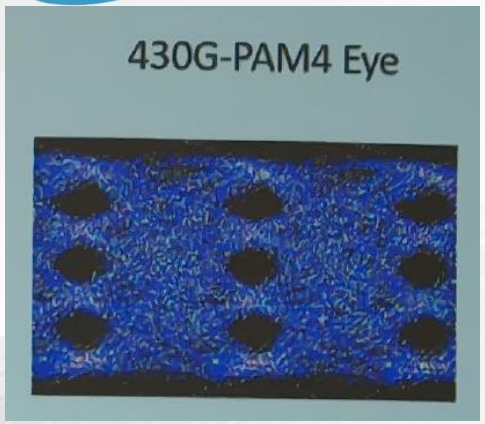
TFLN



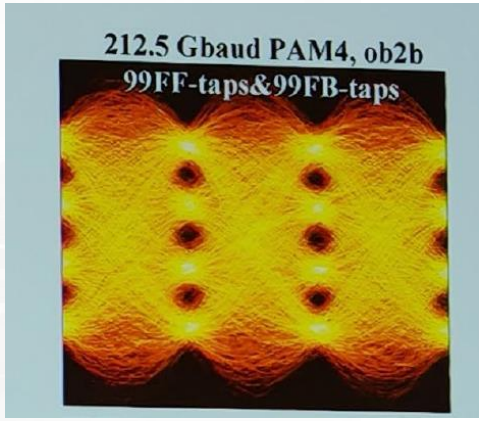
MARVELL



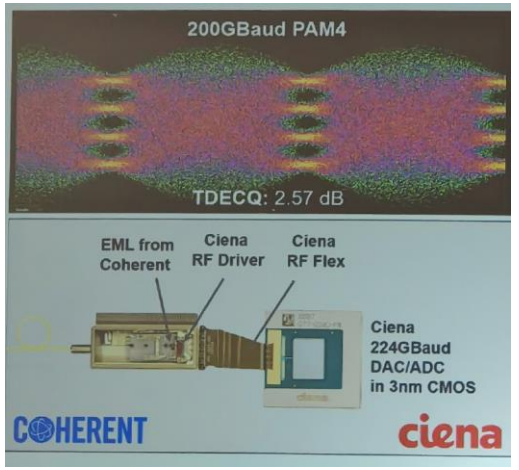
INNO LIGHT



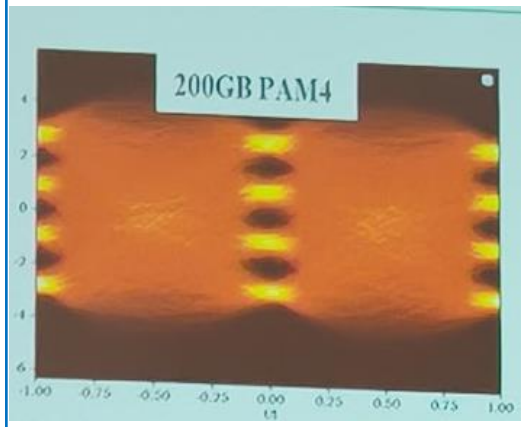
Riga Technical University



EML



HUAWEI



HUAWEI

Native signaling for various architectures

Native modulation needed

- First 448G/lane optical modules will be based on gear-boxed 224G SerDes
- However, a native modulation scheme supporting both electrical and optical channels is the ideal choice for future Ethernet

Support all architectures

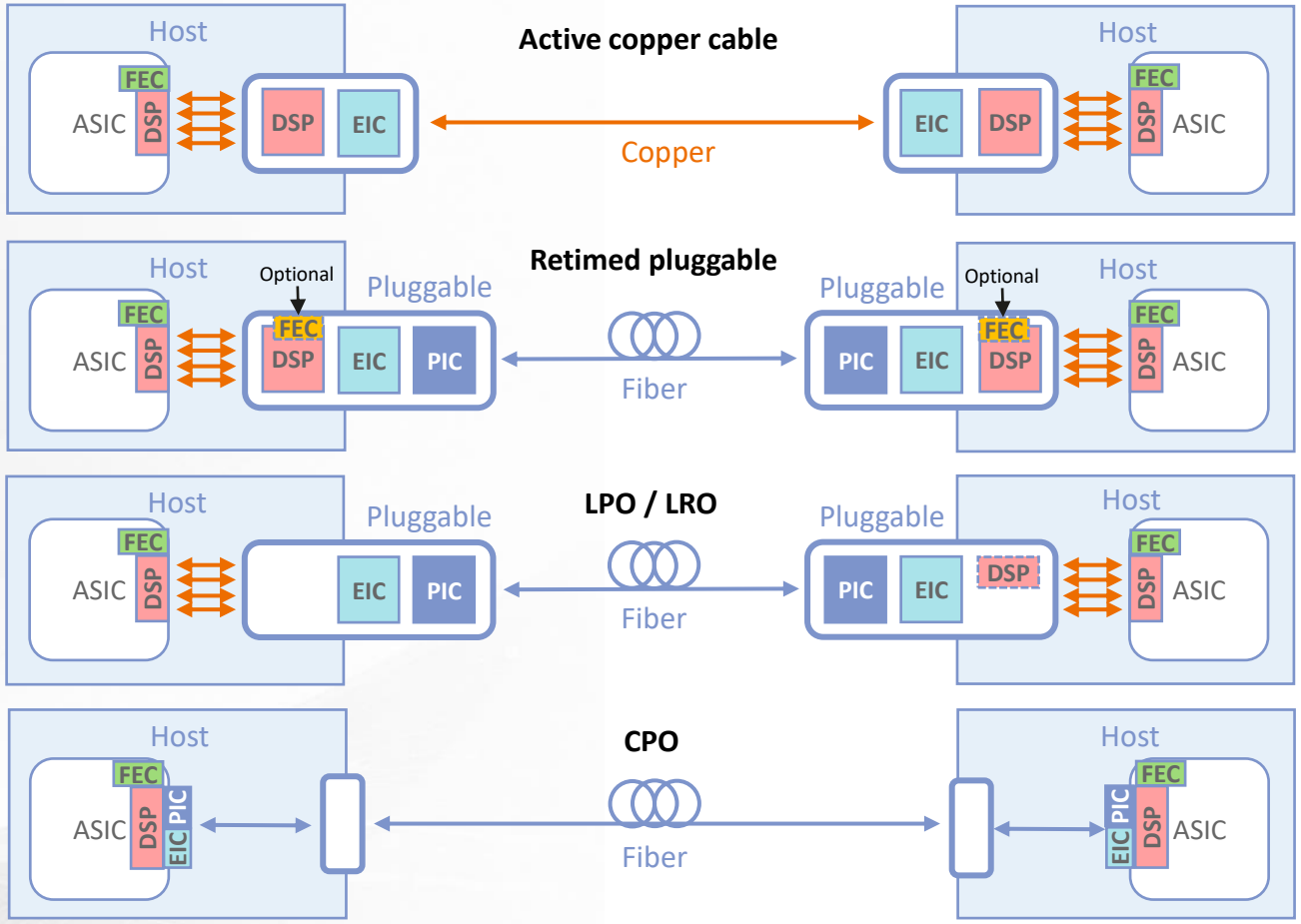
- E2E low latency FEC architecture support needed for AECs, retimed pluggables, LPO, LRO, NPO, CPO transceivers

Inner FEC not primary use case

- Better inner FEC in the pluggable module is an extended use case, but should not guide the modulation format choice

PAM6 better for electrical channels

- 448Gb/s PAM6 performs better over current electrical channel models
- Can PAM6 also be a competitive format for optics or is PAM4 the best native modulation?



HUAWEI

DSP power consumption PAM4 vs. PAM6

DAC and ADC: **Advantage for PAM6**

- Benefit from the 20% lower symbol rate of PAM6
- No increase in resolution for PAM6 with respect to PAM4 needed

FFE: **Slight advantage for PAM4**

- Time domain implementation assumed to reduce latency in the SerDes
- It benefits from the 20% lower symbol rate in terms of both operations per sec. and number of required taps
- It suffers from the increased number of levels of PAM6 with respect to PAM4

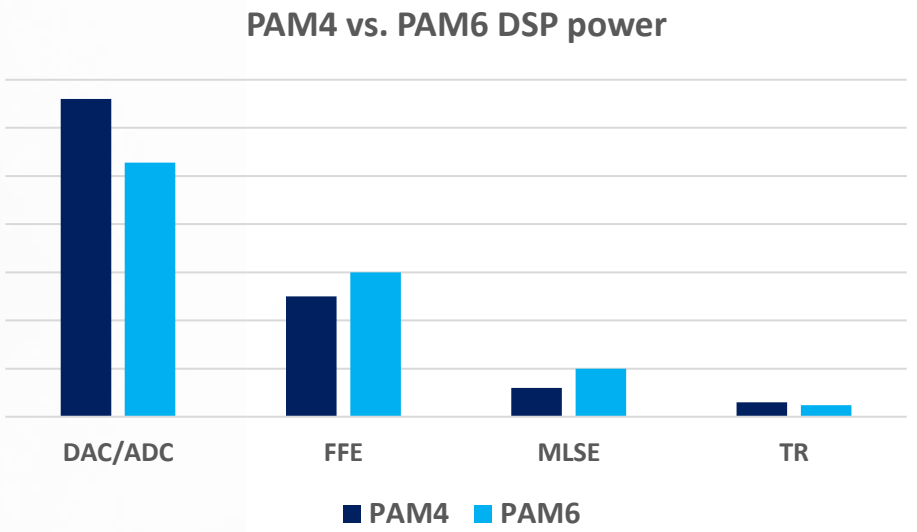
MLSE: **Advantage for PAM4**

- Can be simplified through state reduction. It benefits from the 20% lower symbol rate for PAM6
- It suffers from the increased number of levels and from the 2D nature of the PAM6 constellation

Overall: **Slight advantage for PAM6**

- Our preliminary estimate results in a slight power advantage for PAM6

For the same power consumption, once can e.g. assume a slightly higher overhead FEC for PAM6



Assumptions

- Symbol rate PAM6 = 80% symbol rate of PAM4
- State-reduced maximum likelihood sequence detection (MLSD) and time-domain feed forward equalizer (FFE)
- FFE complexity for PAM6 is assumed to be 50% higher than for PAM4 (excluding symbol rate impact). This accounts for larger constellation (more complexity) and less stringent bandwidth limitations (less complexity)
- MLSE complexity for PAM6 is assumed to be 100% higher than for PAM4 (excluding symbol rate impact)
- Timing recovery (TR) with similar assumptions, although easier to implement for PAM6 if there is less bandwidth limitation

FEC assumption

KP4 for PAM4

- 200G PAM4 legacy mode will require **KP4** by definition
- KP4 is the best initial assumption for 448G PAM4 in the host

Better FEC for PAM6

- We assume a **higher overhead FEC** for PAM6 to achieve a fairer comparison to the higher baud rate / power PAM4
- HD-FEC to support all retimed architectures

Lower overall risk

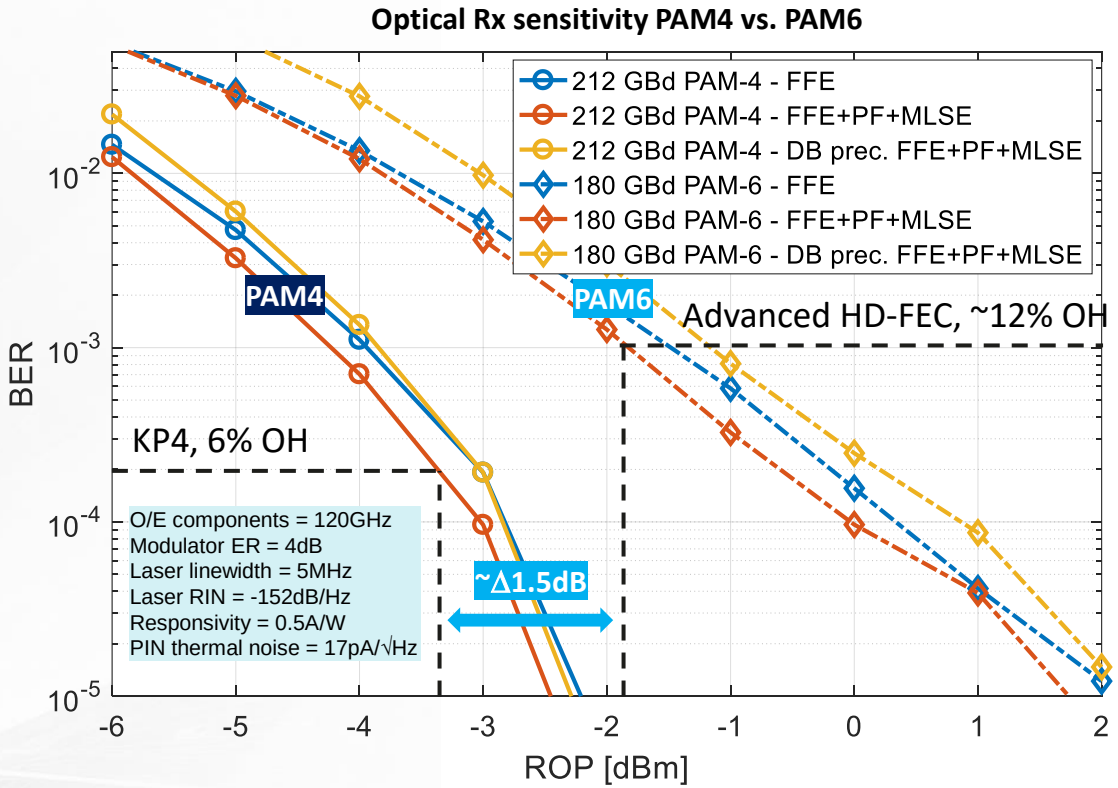
- Technological risk of 180Gbaud PAM6 SerDes with 12% FEC is still lower than 212Gbaud PAM4 with KP4

Reduces SNR gap

- SNR gap can be reduced from ~3dB → ~1.5dB, which is relevant for optical channels to limit laser output power

Error floor margin

- Better FEC for PAM6 is needed also to improve the error floor margin



HUAWEI

Chromatic dispersion & wavelength plan

MLSE for higher CD

- MLSE is already part of 224G AUI and will be part of 448G SerDes DSP to increase the CD tolerance

PAM4 vs. PAM6

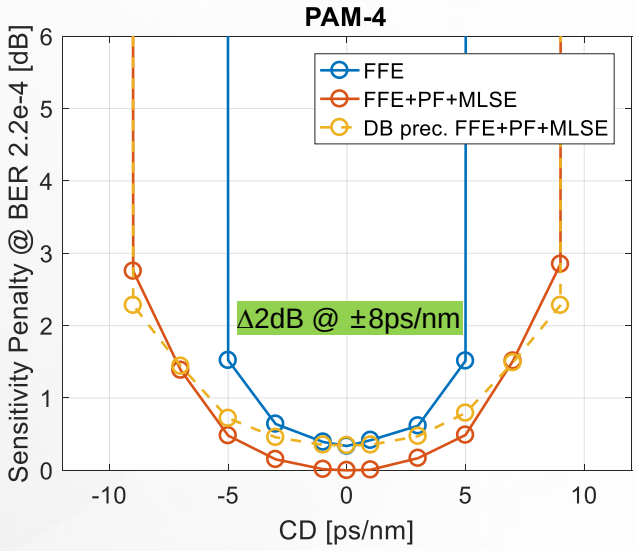
- No substantial advantage for PAM6 enabling a new applications

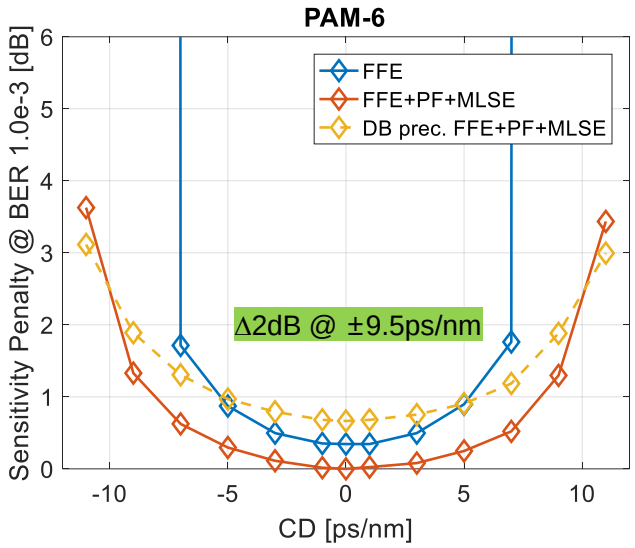
1.6T FR4

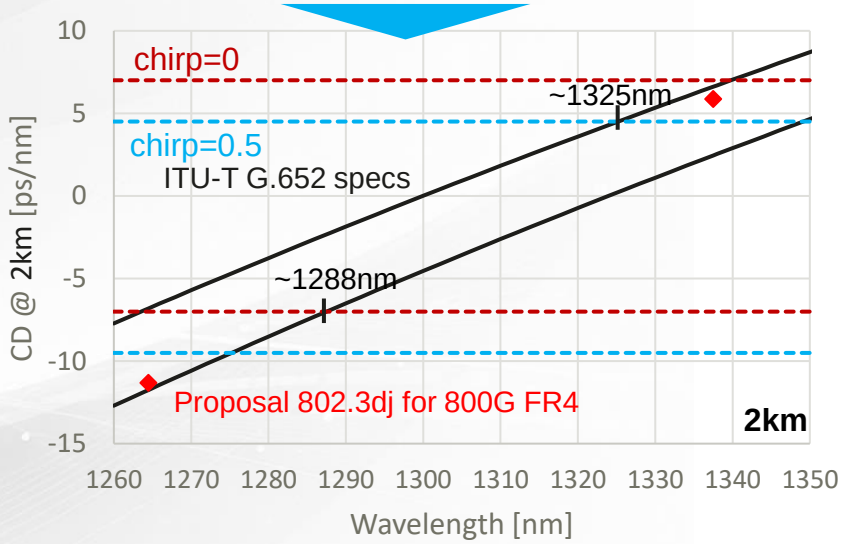
- Accomodating for transmitters with different chirp, a 1.6T FR4 interface with PAM4/PAM6 looks feasible
- Uncooled FR4-2km with 10nm spacing possible
- Chirp managed FR4-2km with 20nm spacing possible

3.2T FR8

- On paper possible on a LAN-WDM grid
- LAN-WDM would require a tighter laser accuracy of +/-0.5nm compared to today‘s cooled lasers with +/-1nm, which would further increase costs







1.6T FR4-2km

Uncooled

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

Chirp management

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

MPI

Networking interruption

- Networking failures in GPU training clusters have a significant effect on cluster performance and amount of GPU sparing
- **~80%** of all optical transceiver failures come from link contamination (link flaps)

More stringent MPI spec

- New data centers with initially more dust in the air
- Legacy Ethernet MPI spec is **-35dB**, but should be increased for future scenarios
- Linear drive (LPO/LRO/CPO) use cases will lead to more reflections in the analog signal path
- **PAM4** has a higher inherent MPI tolerance employing receiver sided compensation techniques of up to **-25dB**

Improving PAM6

- Better PAM6 performance would require **additional signal overhead** (e.g. 1.5-2%) and more effort with standardization of the equalization scheme

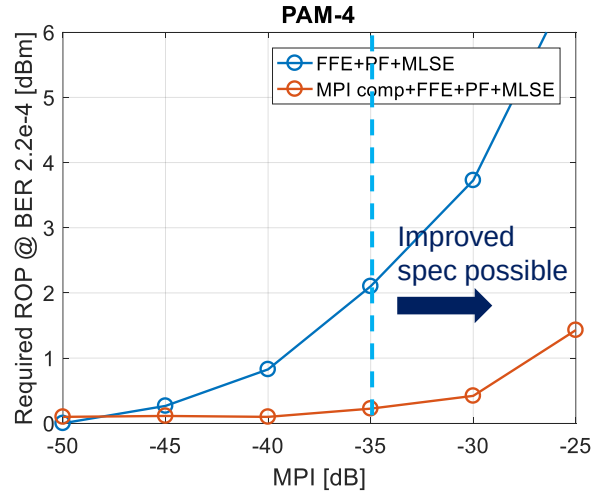
Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

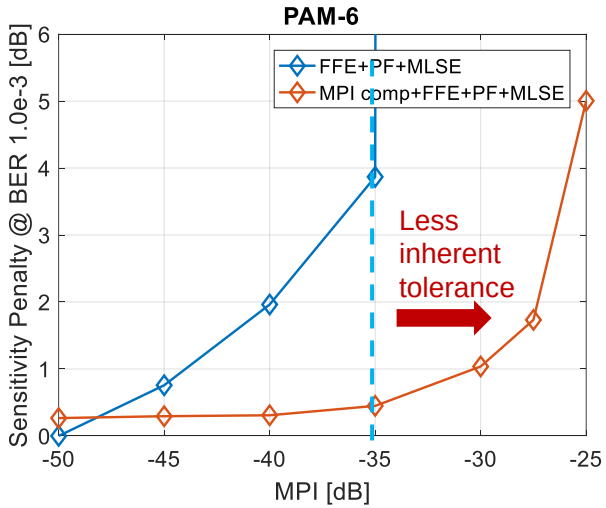
Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

[Meta] https://arxiv.org/abs/2407.21783

Estimated Time to First Job Failure (Minutes)				
Mean Time to Failure Per Link	3 years	4 years	5 years	10 years
Number of GPUs				
10,000	157.7	210.2	262.8	525.6
20,000	78.8	105.1	131.4	262.8
50,000	31.5	42.0	52.6	105.1
100,000	15.8	21.0	26.3	52.6

[SemiAnalysis]





Improving FEC latency & power

E2E FEC

- Soft decoding is not an option for the host FEC due to retimed interfaces

PAM4 FEC

- Better performance could improve electrical channel performance
- Increased OH for PAM4 is generally not desired in the host
- KP4 FEC will be part for the SerDes for the 200G interop mode and should be ideally reused
- **MLC** with different Reed-Solomon FECs can achieve same performance as KP4 at lower overhead and power consumption (5.8% → 4.1%)

PAM6 FEC

- MLC can also deliver optimized codes for PAM6 and provide better performance than BICM decoders

Code (Hard decoded)	OH	BER @ 1e-15	NCG	Complexity
RS(544,514)	5.8%	2.2e-4	6.9dB	1x
MLC RS(554,514) + RS(544,542)	4.9%	2.3e-4	7.0dB	1.05x
RS(560,514)	8.9%	6.1e-4	7.4dB	2.86x
RS(576,514)	12.1%	1.1e-3	7.8dB	5.79x
RS(544,514) + BCH(128,120)	12.9%	1.4e-3	8.0dB	1x
MLC RS(544,514) + BCH(128,120)	7.4%	7.33e-4	7.8dB	0.5x

Selection of basic FEC options

Conclusions

- Electrical and optical domains require an identical modulation format similar to previous Ethernet standards to avoid gearboxes for every architecture using 448G SerDes
- PAM4 has an obvious advantage in the optical domain
- PAM4 limitation in the electrical domain largely come from connectors
- PAM6 could potentially overcome the drawbacks in the optical domain with more effort (higher overhead)
- Other optical effects, like DGD or FWM, not critical at 2km for PAM4/6 for FR4
- Next to retimed architectures, linear drive optics and copper cabled designs will dominate the decision finding

Thank you

HUAWEI