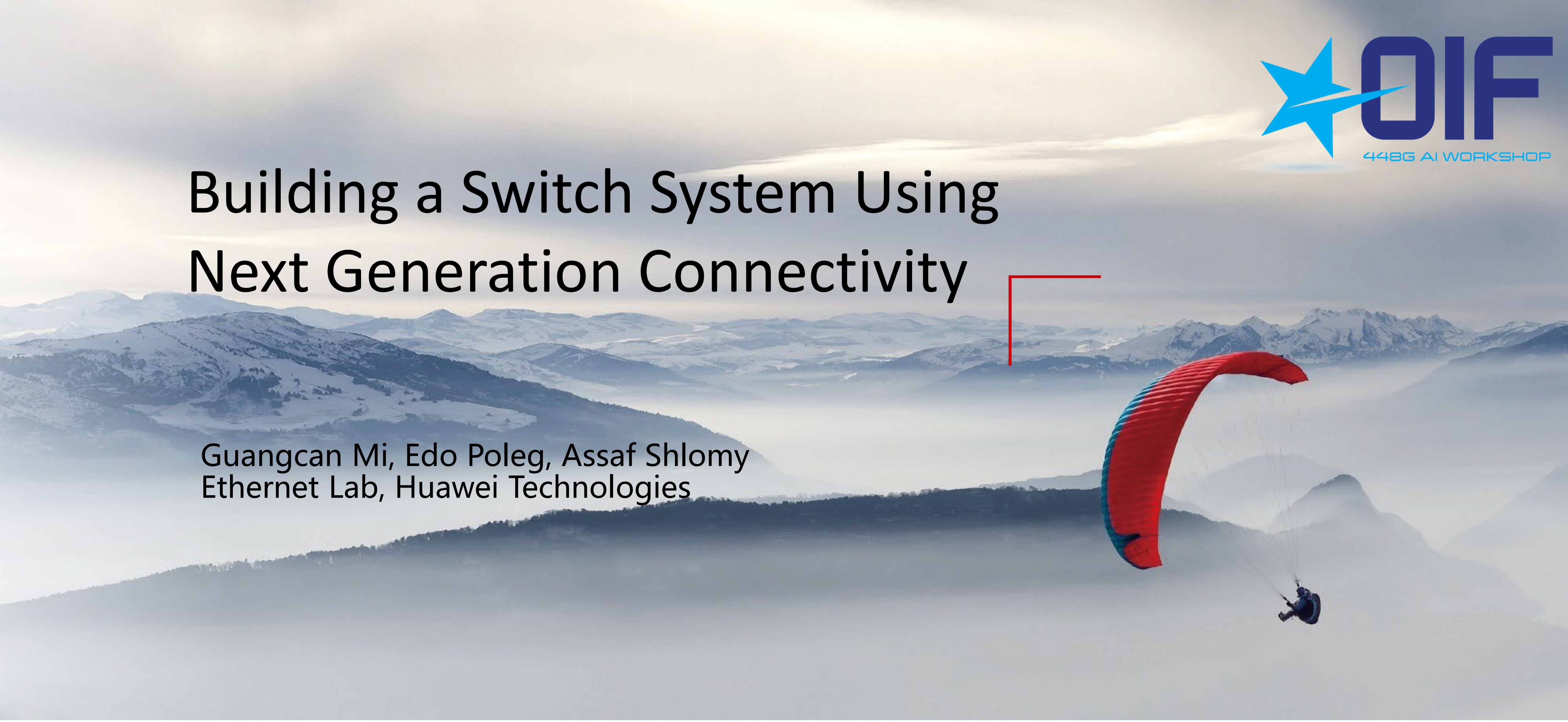


Building a Switch System Using Next Generation Connectivity

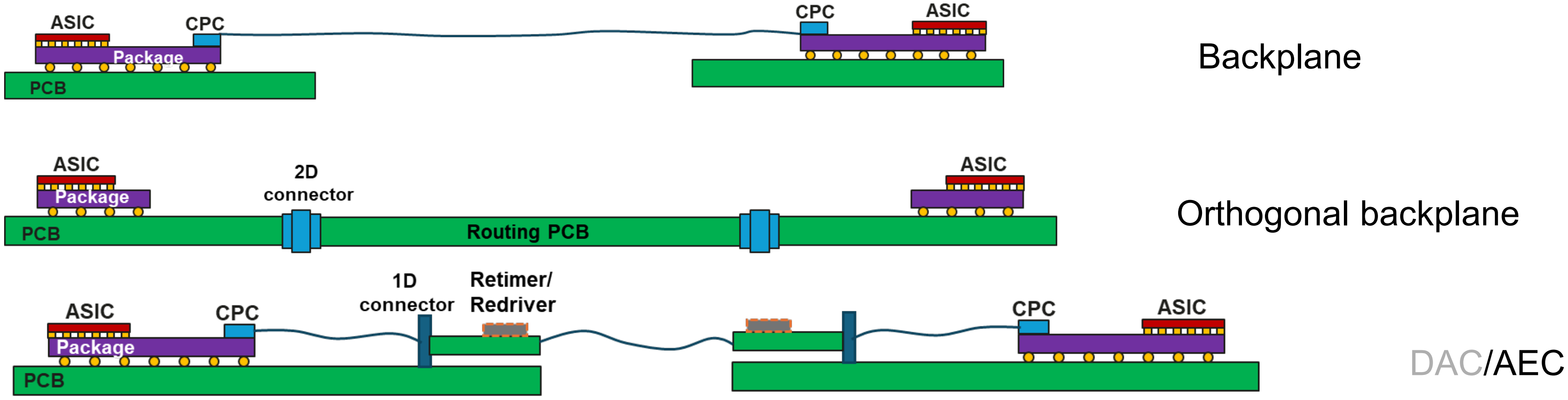
Guangcan Mi, Edo Poleg, Assaf Shlomy
Ethernet Lab, Huawei Technologies

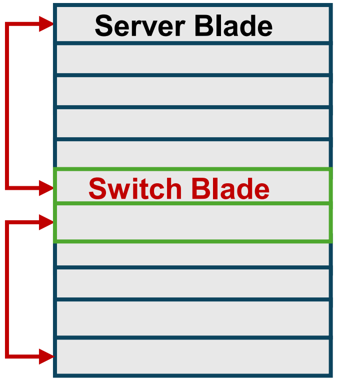
Introduction

- 6 months ago in EA's TEF:
  - Agreement on need of next generation interconnect driven by scale-out and scale-up network
  - Electrical signaling recognized as the main unknown factor
  - Exploration of various technologies: BIDI, SE-MIMO, PAM4/6/8
- 6 months later now
  - Industry showed demo of 1.6T coherent and 400G-PAM4 optics, both will be interfacing with 400G electrical signaling eventually
  - Still looking for more information regarding electrical link, which seems the biggest challenge here
  - Focus down PAM4 and PAM6, we will look at the common ground of optical and electrical interfaces, with the constraints from a switch system

Building the switch system：in the Rack

Possible architecture for electrical intra-rack links @ 448G/lane

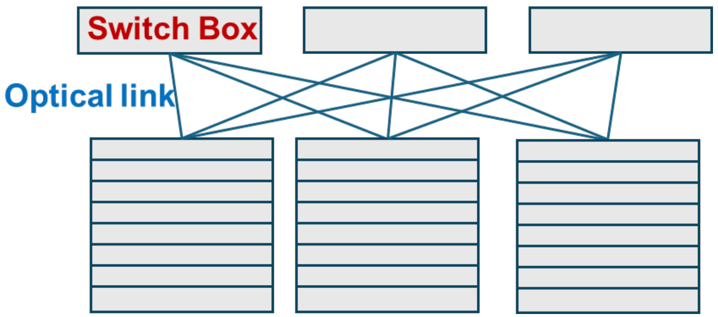




Lane Speed	backplane Link loss Per CEI-LR	Pluggable module Link Loss Per CEI-VSR	Preferred usage
50Gbps	30dB	10dB	4m for inter rack
100Gbps	28.5dB	16dB	4m for inter rack
200Gbps	40dB-IL_{dd} (equiv. 28~22dB ball to ball)	32dB-IL_{dd} With 28.2dB allocated from host up to connector	Up to 2m passive for intra rack AEC to extend to neighbouring racks
Next: 448Gbps	40? dB-IL_{dd} (< ? dB in ball to ball)	? dB-IL_{dd} (concerns on 1D connector bandwidth and loss)	Keep using electrical for intra-rack especially, scale-up

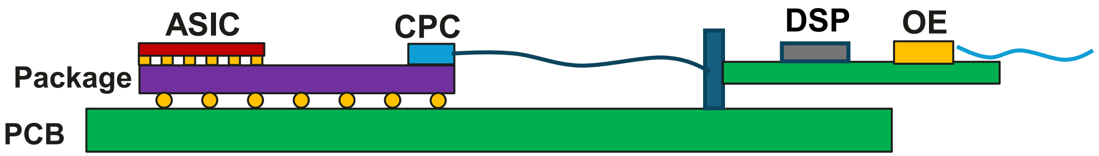
Building the switch system： Connecting the Racks

- The un-resolved concerns resolving:
  - Recent progress showed in OFC build confidence of 400G-PAM4 optics
  - Accommodate growing radix, 512→1024, for flat networking topology
  - High reliability as needed in both training and inference
  - The challenge points to passive electrical link again, and form factor



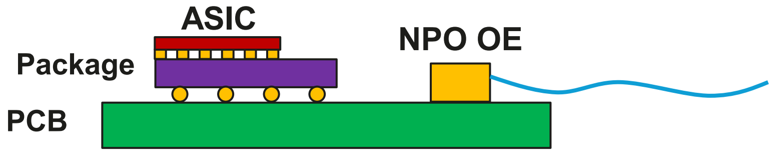
MTBF
First Power Down

Mean Time to Replace

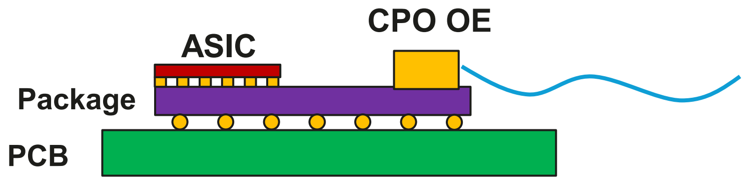


Module: $\text{MTBF}_{\text{lane}}$ accumulated by 8 lanes
Switch: Not triggered

Module: hot plug & play



Switch: $\text{MTBF}_{\text{lane}}$ accumulated by 512 lanes
If you want to change the components, power dwon the switch, all 512 lanes are affected.

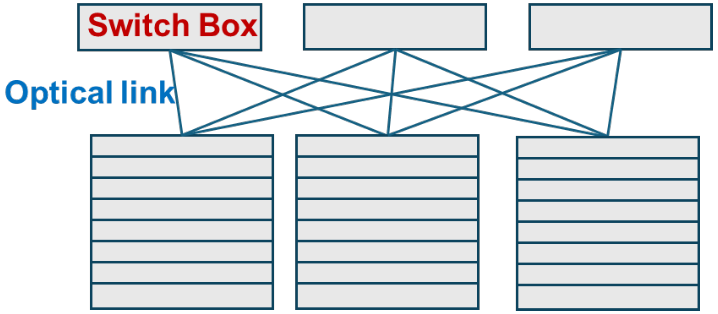


Switch: $\text{MTBF}_{\text{lane}}$ accumulated by 512 lanes

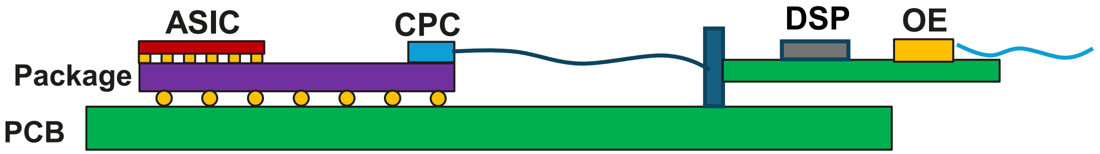
OE w/ socket: Some effort
OE solered: back to factory

Building the switch system：Connecting the Racks

- The un-resolved concerns resolving:
  - Recent progress showed in OFC build confidence of 400G-PAM4 optics
  - Accommodate growing radix, 512→1024, for flat networking topology
  - High reliability as needed in both training and inference
  - The challenge points to passive electrical link again, and form factor



Preferred. But huge challenge with VSR channel

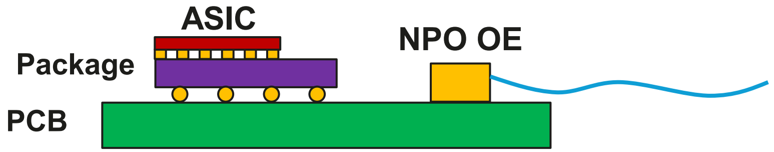


MTBF
First Power Down

Mean Time to Replace

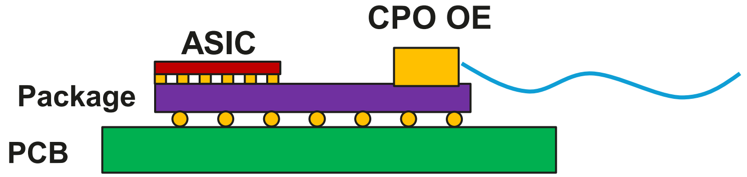
Module: $\text{MTBF}_{\text{lane}}$ accumulated by 8 lanes
Switch: Not triggered

Module: hot plug & play



Switch: $\text{MTBF}_{\text{lane}}$ accumulated by 512 lanes
If you want to change the components, power dwon the switch, all 512 lanes are affected.

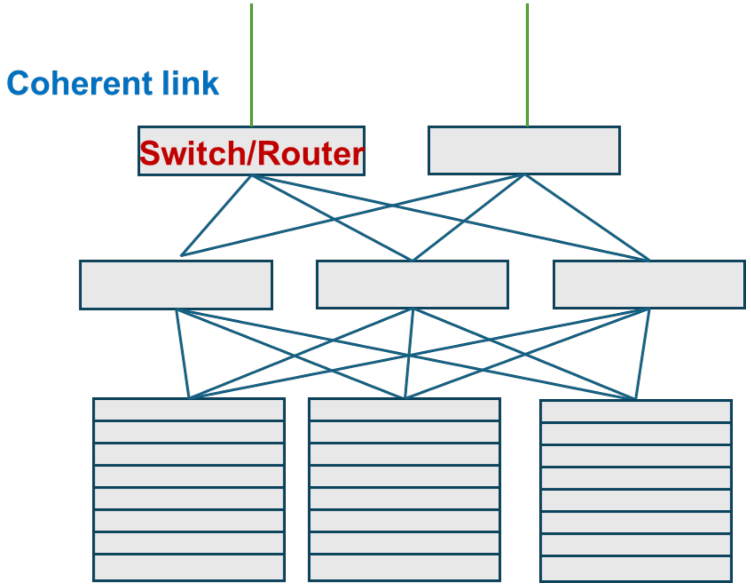
OE w/ socket: Some effort
OE solered: back to factory



Switch: $\text{MTBF}_{\text{lane}}$ accumulated by 512 lanes

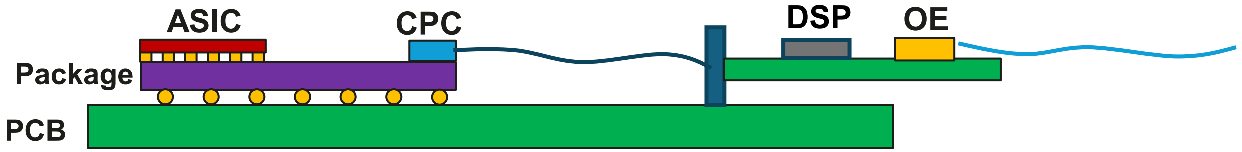
Building the switch system： Connecting the DCs

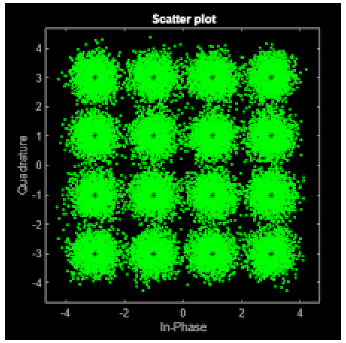
- 1600ZR/ZR+ being developed in OIF
- 1.6T-CL in OIF addressing the need of coherent for AI campus, i.e. short reach, point-to-point, low power



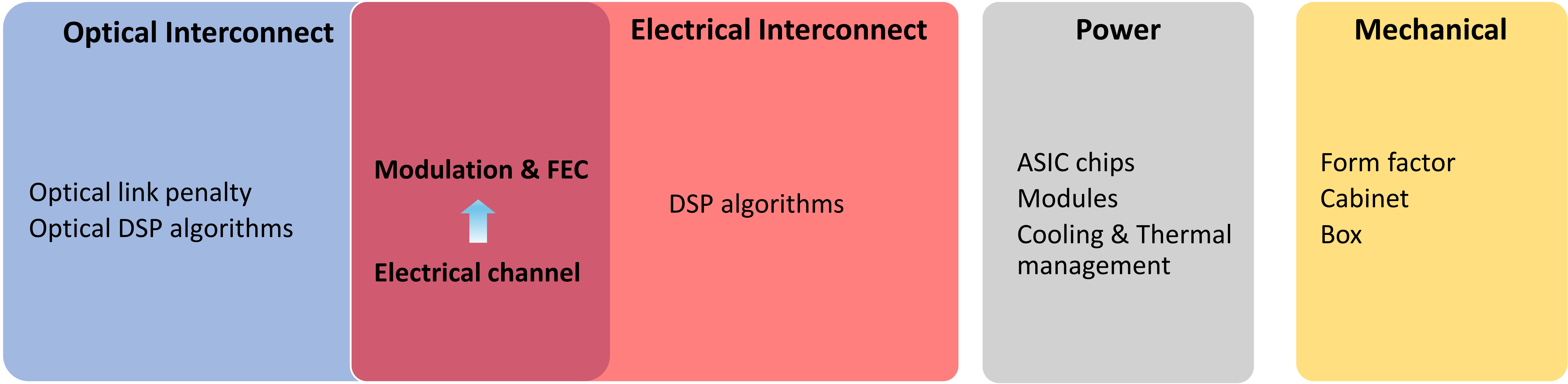
Simple question：

Do we want to roll back 15 years to on-board coherent?





Building the switch system



Optical Interconnect

Optical link penalty
Optical DSP algorithms

Modulation & FEC

Electrical channel

Electrical Interconnect

DSP algorithms

Power

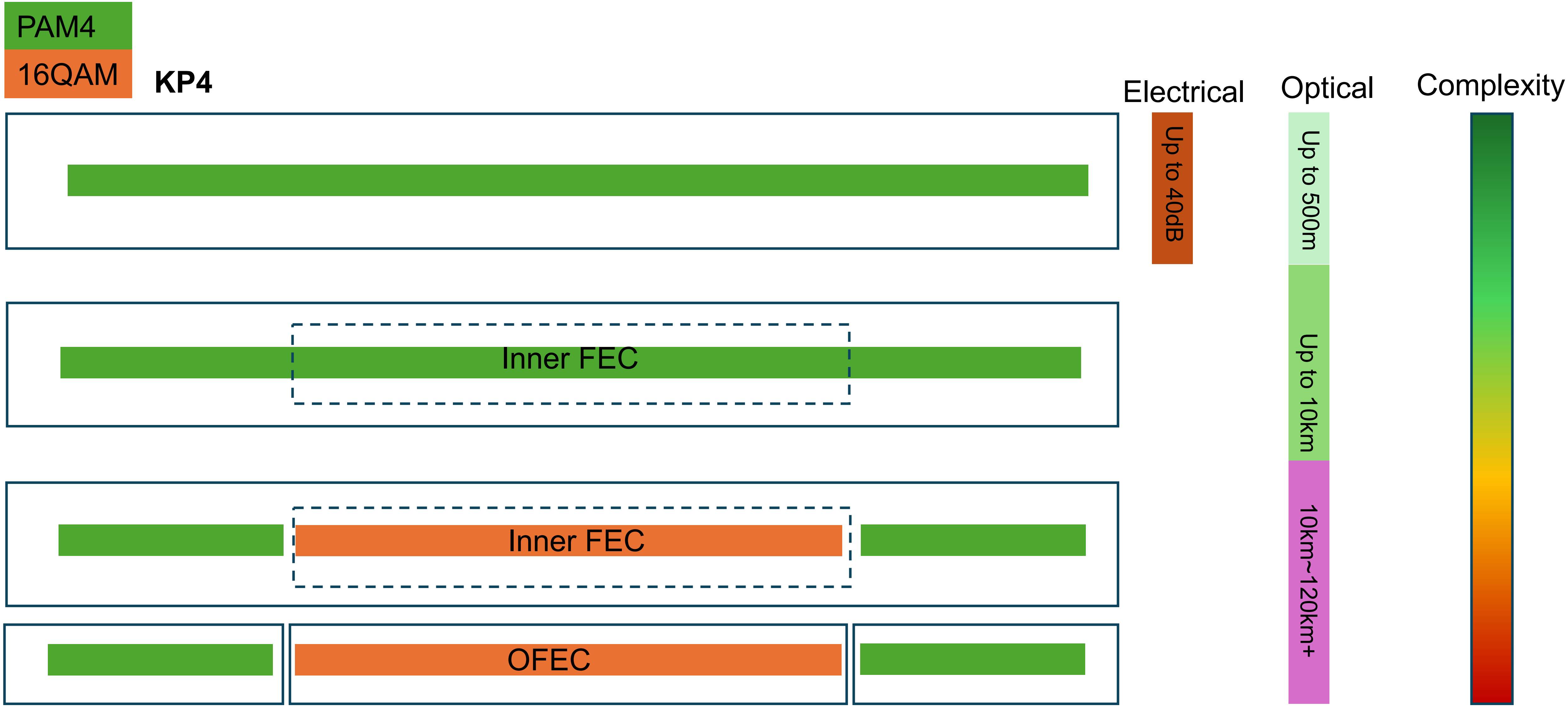
ASIC chips
Modules
Cooling & Thermal management

Mechanical

Form factor
Cabinet
Box

Conflicts of interest within the system
visit the common ground first

Modulation and FEC architecture – Now with 200Gbps



PAM4

16QAM

KP4

Electrical

Optical

Complexity

Inner FEC

Inner FEC

OFEC

Up to 40dB

Up to 500m

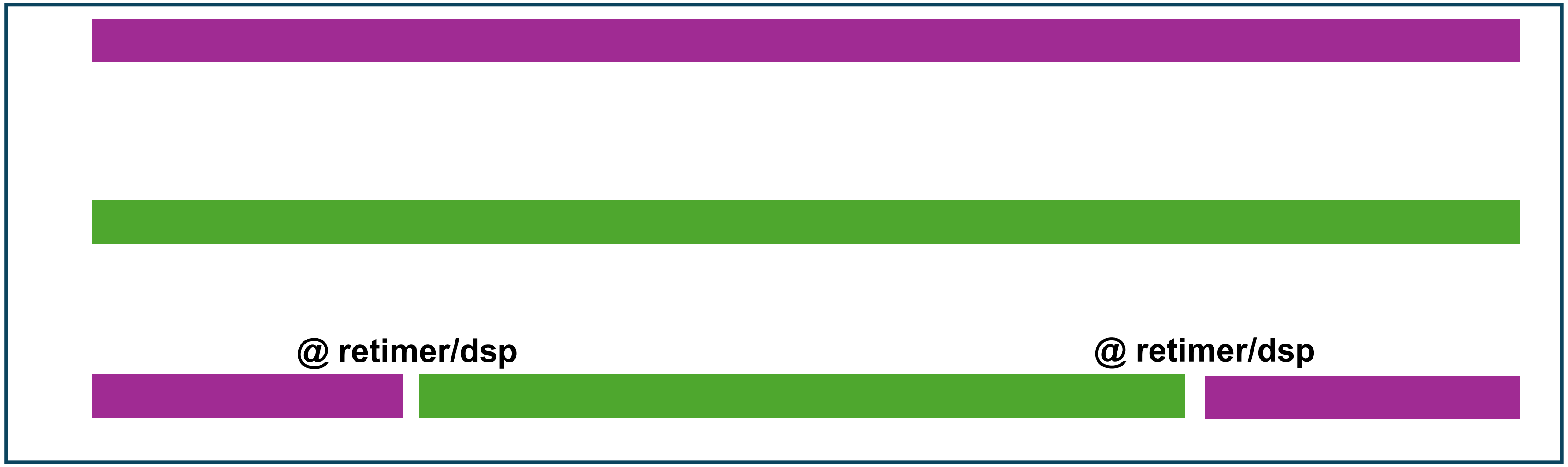
Up to 10km

10km~120km+

Modulation and FEC architecture – IMDD

🟪 PAM6 🟩 PAM4

FEC1 Preferred architecture: most simple, possibly most energy efficient



Technology Choices

Backplane DAC AEC	Not suitable for Optics
Backplane DAC AEC	Retimed Pluggable CPO
AEC	Retimed Pluggable

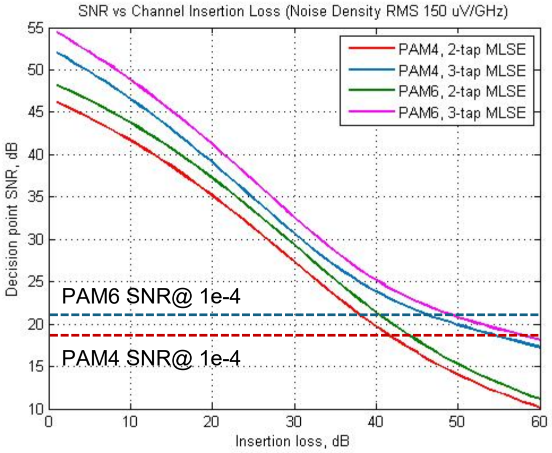
	PAM4	PAM6
Baud Rate (GBd)	212.5	170
BW Nyquist (GHz)	106.25	85
BW_{nq}x1.2 (GHz)	127.5	102
Required SNR @1e-4(dB)	18.2	21.8

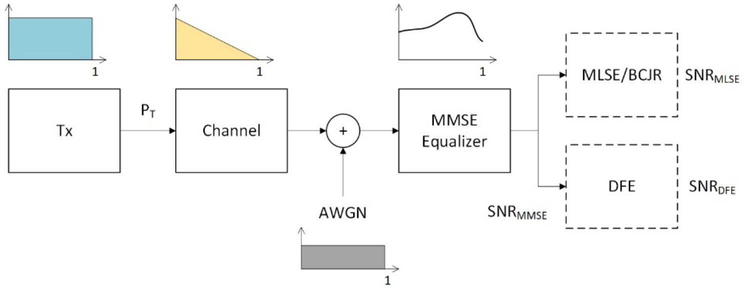
← Coding gain available from other choices of FEC1

Preliminary analysis of PAMn + IL + SNR

The case of electrical interconnect (backplane) :

- Assuming FEC1 = KP4
- Consider two partial equalization cases
- 80% BW limit & white noise
- Both PAM4 and PAM6 show potential for 40dB+ IL



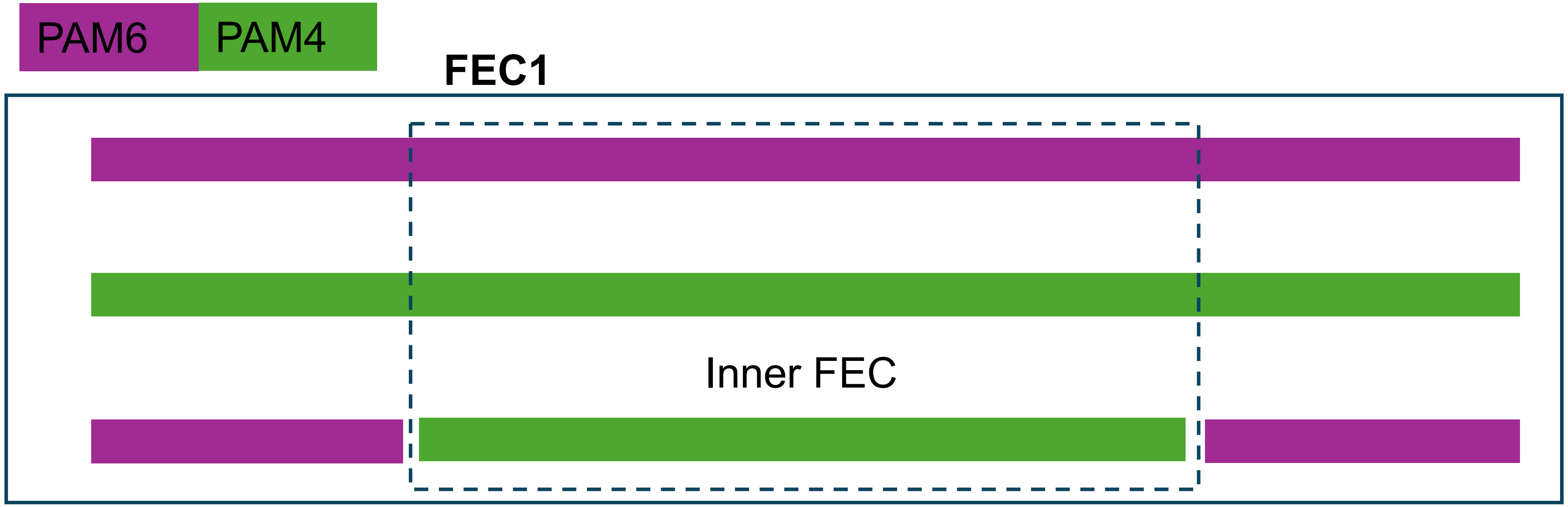


The case of pluggable modules (AUI-C2M/VSR):

- Assuming FEC1 = KP4
- interpolating the three channel data sets in kocsis_e4ai_01_250327 upto 115GHz
- Filter gain (TX FFE & CTLE) are considered
- AFE noise (AWGN) is included @ CTLE input

Modulation	Bandwidth/GHz	Required slicer SNR@1e-4(dB)	Salz margin (dB)		
			Channel A	Channel B	Channel C
PAM4	112	18.2	5.79	7.11√	9.63√
PAM6	89.6	21.8	6.42√	7.01√	8.24

Modulation and FEC architecture – IMDD



Extend Electrical Link budget

May Extend Optical reach

May benefit AUI-C2M &
Extend optical Reach

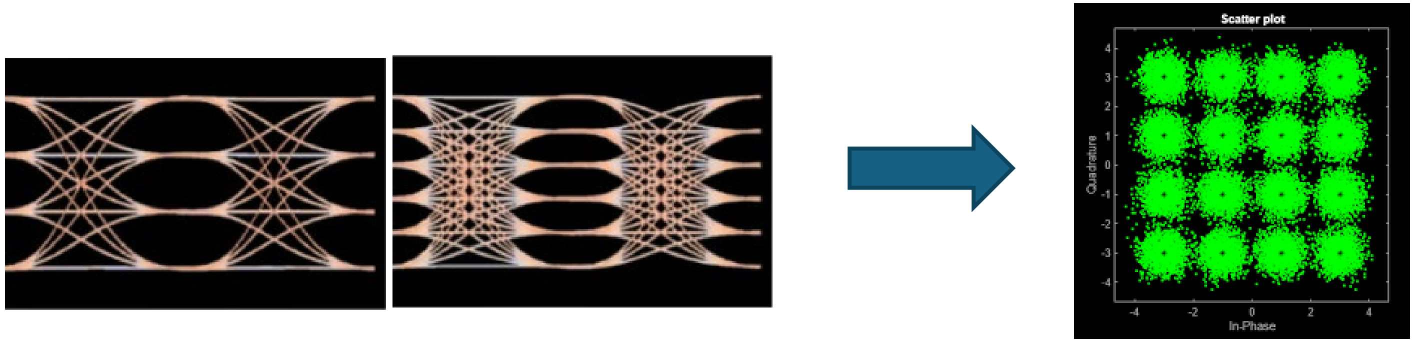
OH Same as 200G	PAM4	PAM4 Inner FEC	PAM6	PAM6 Inner FEC
Baud Rate (GBd)	212.5	226.875	170	181.5
BW Nyquist (GHz)	106.25	113.4375	85	90.75
BW_{nq}x1.2 (GHz)	127.5	136.125	102	108.9
Added e-channel loss	+ 2~8dB* and probably worse Cost and gain doesn’t sum up		+ 1.5~7dB** Something worthy of consideration	

Technology limited to:
AEC
Retimed pluggable optics

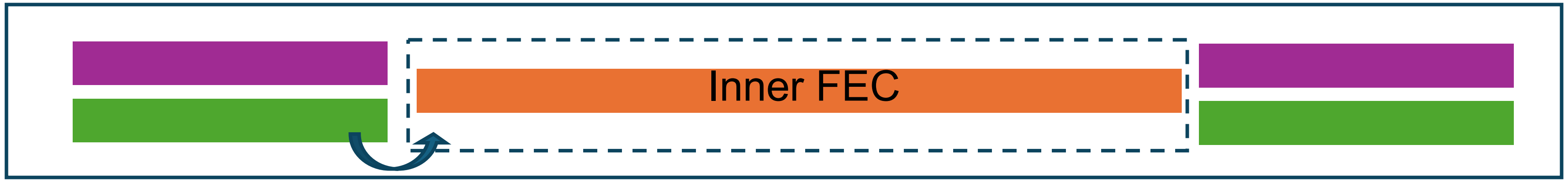
*Extrapolated from kocsis_e4ai_01_250327 A/B/C reference Channel; **estimated from kocsis_e4ai_01_250327 A/B/C reference Channel

Modulation and FEC architecture – Coherent

PAM6	PAM4	16QAM



FEC1

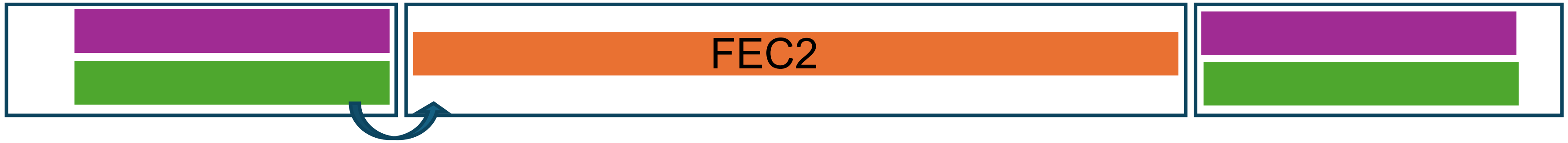


Inner FEC

Minimum difference with appropriate inner FEC design

Inner FEC Feasible proposal for 1600CL

In **Coherent optics** regime,
Changing electrical interface to PAM6 brings benefit:
Reduced challenge to AUI-C2M



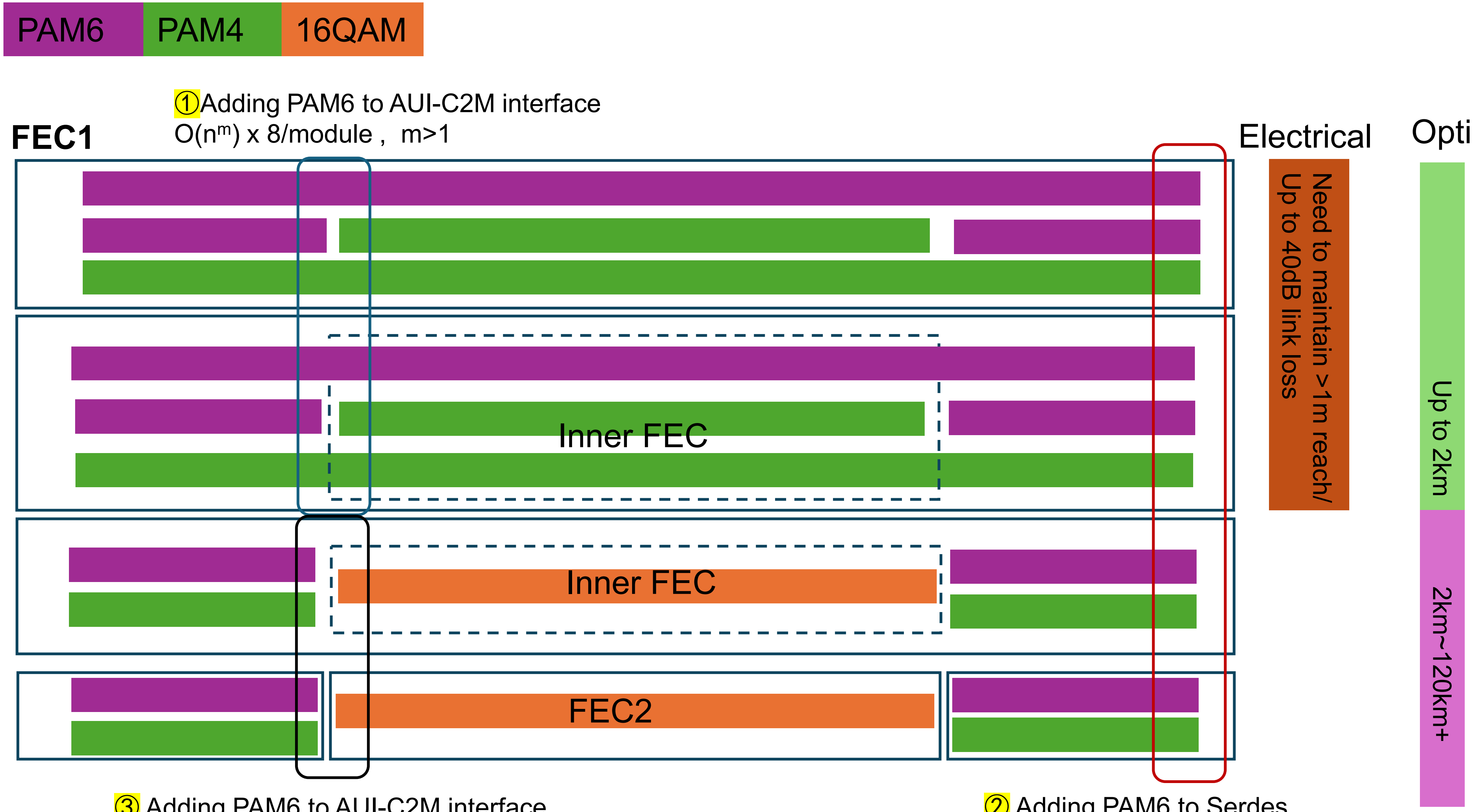
FEC2

No difference

PAM6 Benefit : easier AUI-C2M

Mandatory for ZR/ZR+ optics, currently at 1600ZR with next hop to 3.2T/6.4T not that far away

Modulation and FEC architecture



①Adding PAM6 to AUI-C2M interface
O(n^m) x 8/module , m>1

② Adding PAM6 to Serdes
O(*n*) x 512/switch

③ Adding PAM6 to AUI-C2M interface
No extra effort

Yet to explore:
Which **sets** could hit the target of 40dB and C2M

Surprise:
AUI-C2M/ pluggable module is mandatory to support pluggable coherent.

Technologies paving the way to a new balance

- **New connectors to maintain loss budget**
  - CPC connectors showing 100GHz+ BW, e.g. Luxshare showcased in 2024OCP
  - 2D connectors/twinax cable to front panel pluggable: HDC in EEI is the project addressing new need.
- **Advanced substrate and packaging technologies further reduces loss**
  - New materials such as glass substrate, ultra high density PCB, polymer infill
  - 2.5D and 3D packaging of OE chiplets, e.g. TSMC coupe, etc.
- **Advanced algorithms and FEC codesign to extend the life of current interconnect architecture**
  - Duobinary PAM4 offers high spectral efficiency, supporting higher data rates in bandwidth-limited scenarios.
  - Nonlinear compensation algorithm in Electrical and Optical channels with higher nonlinear impairments.
  - Low complexity MAP algorithm provides more reliable soft information to FEC
- **Asymmetrical architecture could be a safe net**
  - PAM6 electrical for intra-rack interconnect may help the reach
  - PAM4 electrical interfacing with IMDD help reduce power
  - Higher radix and lower rate(<400G) for intra-rack is also a feasible choice to support greater than 1m cable or backplane

Key Takeaways

- Electrical link awaits new technology on packaging and connector to close 1m cabling requirement
- Optical IMDD link may benefit from NPO and CPO for power savings, but impact to MTBF need to be considered.
- Coherent link becomes even more relevant as its application going towards 2km+
  - Maintaining Ethernet’s Plug-and-play is key
  - Lower power always needed
  - Creating the simple answer to the debate on supporting pluggable : Yes
- A solution space of modulation and FEC architecture was examined
- Building a switch system in the next AI era means: Finding new balance between E and O