Architectural Considerations and Managing PMDs Timeline

Ali Ghiasi, Ghiasi Quantum LLC

Jamal Riani, Marvell

802.3df Task Force Meeting

Virtual Meeting

Feb 15, 2022

Contributors

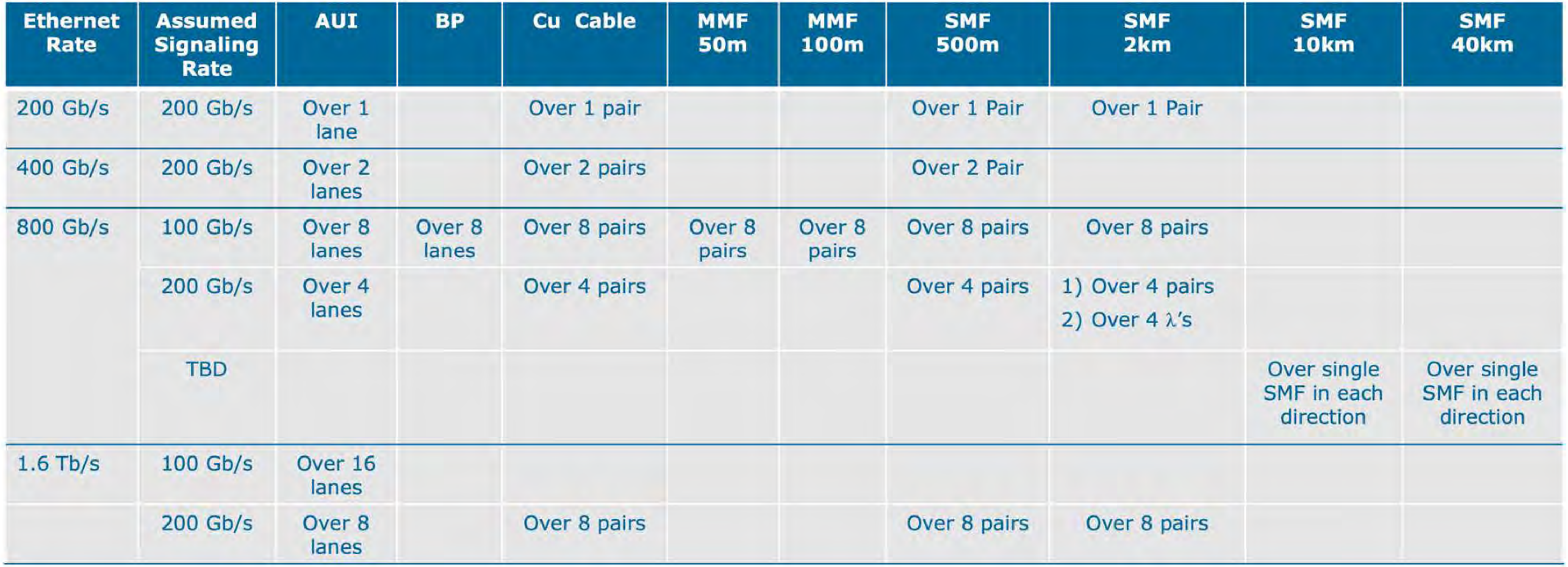
- **Lenin Patra – Marvell**
- **Arash Farhoodfar – Marvell**

Overview

- ❑ 802.3df PMD landscape
- ❑ ETC PCS/FEC
- ❑ 802.3bs FEC/PCS architecture
- ❑ Potential optics/Cu FEC architecture
- ❑ AUIs, PPI, PMDs mode of driving
- ❑ Potential 802.3df FEC architecture
- ❑ How to potentially partition 802.3df into 3 task forces
- ❑ Summary

Adopted Objectives

- 13 optical PMDs
- 6 Cu PMDs
- 6 AUIs.



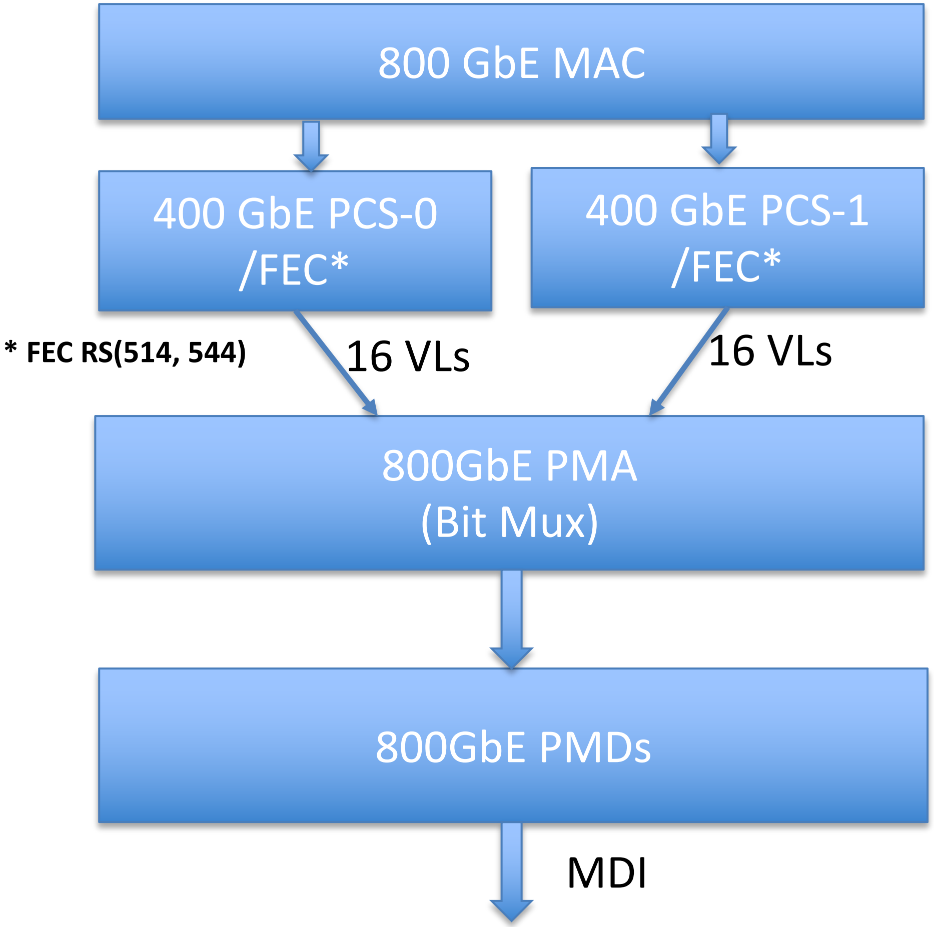
Ethernet Rate	Assumed Signaling Rate	AUI	BP	Cu Cable	MMF 50m	MMF 100m	SMF 500m	SMF 2km	SMF 10km	SMF 40km
200 Gb/s	200 Gb/s	Over 1 lane		Over 1 pair			Over 1 Pair	Over 1 Pair		
400 Gb/s	200 Gb/s	Over 2 lanes		Over 2 pairs			Over 2 Pair			
800 Gb/s	100 Gb/s	Over 8 lanes	Over 8 lanes	Over 8 pairs	Over 8 pairs	Over 8 pairs	Over 8 pairs	Over 8 pairs		
	200 Gb/s	Over 4 lanes		Over 4 pairs			Over 4 pairs	1) Over 4 pairs 2) Over 4 λ's		
	TBD								Over single SMF in each direction	Over single SMF in each direction
1.6 Tb/s	100 Gb/s	Over 16 lanes								
	200 Gb/s	Over 8 lanes		Over 8 pairs			Over 8 pairs	Over 8 pairs		

Underlaying Assumptions

- 1st 800 GbE deployment will be based on 8 lanes PMDs with 8 lanes AUI
- 1st 1.6 TbE deployment will be based on 8 lanes PMDs with 16 lanes AUI
- 1st 200G/lane deployment will occur on optical PMDs
- 1st expected switches with 800 GbE MAC will be 51.2T (512x100G) expected sample time mid-2023 and likely will have ETC MAC
- 1st 8x100G optical PMDs are being currently deployed in conjunction with 25.6T switches operating as 2x400 GbE
- 1st 200G/lane optical PMDs will be deployed on 51.2T switches
- 1st 200 GbE and 400 GbE optics based on 200G/lane will be deployed in conjunction with 51.2T
- 1st 200G electrical IO will be in conjunction with 102.4T switches.

800 GbE PCS/FEC

- **The decision regarding 800 GbE MAC-PCS/FEC is urgent given some of the potential ASICs in flight**
  - We either need to adopt Ethernet Technology Consortium (ETC) proposed 800 GbE and if we define a new 800 GbE PCS/FEC can be disruptive to product in flight
    - PCS/FEC based on clause 119 but with unique identifier to indicate PCS-0/1
  - ETC 800 GbE implementation is based on
    - Dual 400 GbE instance of PCS/FEC
    - With 32 virtual lanes instead of 16
    - With additional set of markers to allow interleaving odd/evens codewords
- **800 GbE PCS/FEC decision is urgent and whether to to stay with ETC 800 GbE PCS/FEC or not**
  - ETC 800 GbE PCS/FEC should meet 100G/lanes AUIs and PMDs.

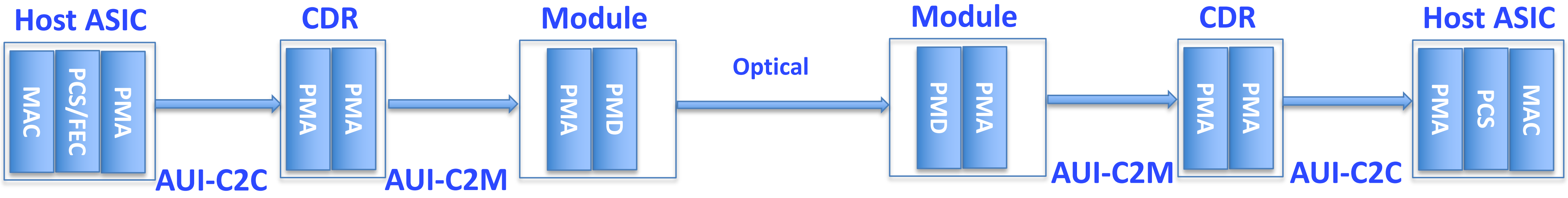


* FEC RS(514, 544)

https://ethernettechnologyconsortium.org/wp-content/uploads/2020/03/800G-Specification_r1.0.pdf

802.3bs FEC Architecture

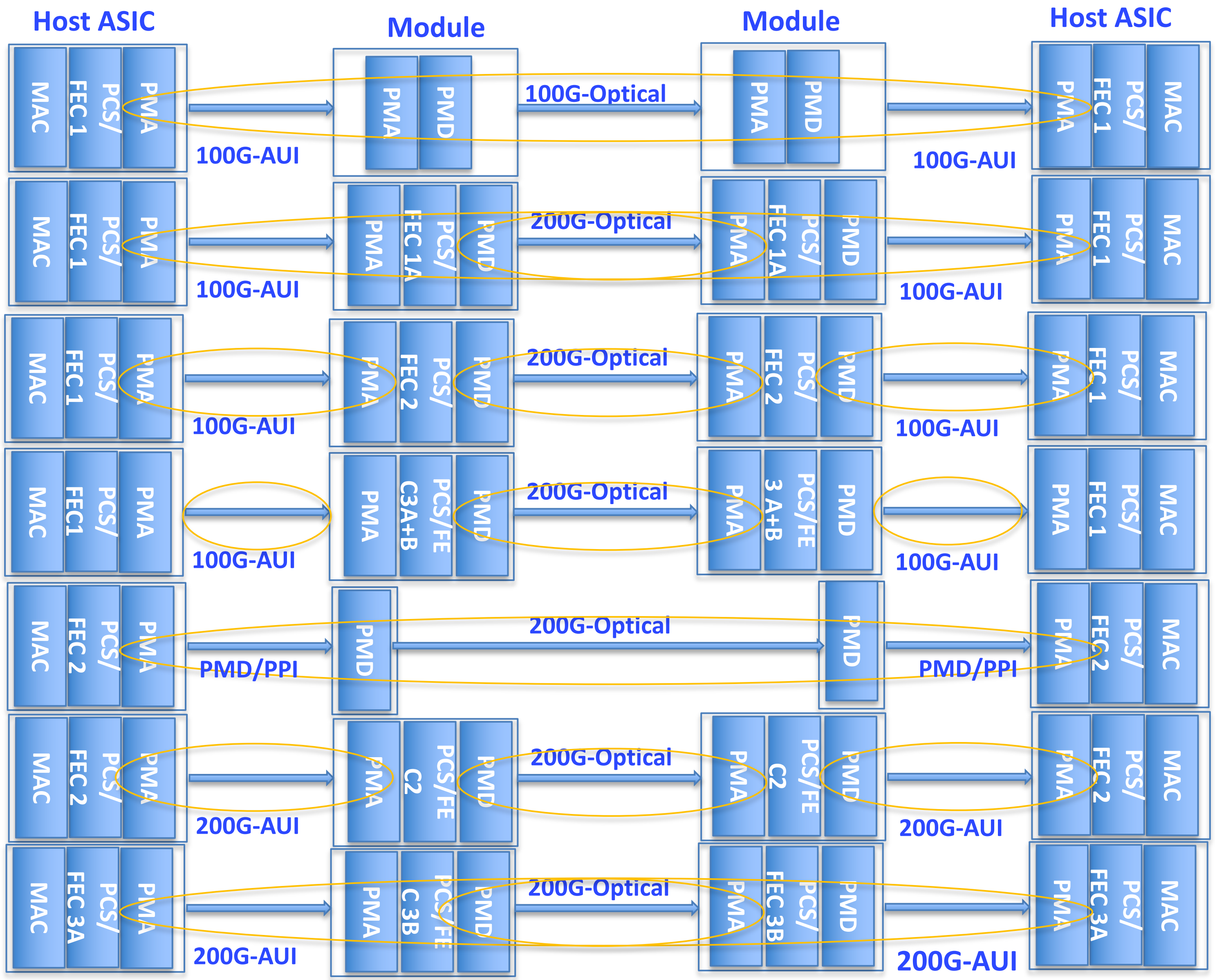
- **802.3bs contribution from Anslow supports 4 AUI sub-links as shown below by stealing 0.1 dBo of optical budget to allow operation with one end-end FEC**
  - Single end-end FEC architecture unlikely to support 200G/lane Cu-Optical links!



	RS(544,514) FLR = 6.2E-11			
	Electrical		Optical	
1:2 Same FEC, a = 0.75 worst skew	Burst	1.4E-6*	Random	2.4E-4
1:2 Same FEC, a = 0.75 worst skew	Burst	2.9E-6*	Random	2E-4
a = 0.75 misaligned	Burst	5.2E-6*	Random	2.4E-4
Random errors	Random	8.2E-5	Random	2.4E-4

Note – these values are the BER **including** the additional errors due to the bursts. To account for burst errors, the values marked with "*" have been multiplied by 4 when a = 0.75.

Optical 800/1600 GbE FEC Options



I. **End-end FEC1 RS (514, 544)**
   - 1st deployment of 800 GbE/1600 GbE

II. **Concatenated RS(514,544)+ SFEC (soft decision) on top of FEC1**
   - 1st 200G optics deployment
   - SFEC can have 1.61-2.7 dB additional NCG

III. **Segmented RS(514,544)+~8.5 dB optics FEC2**
   - 1st 200G optics deployment

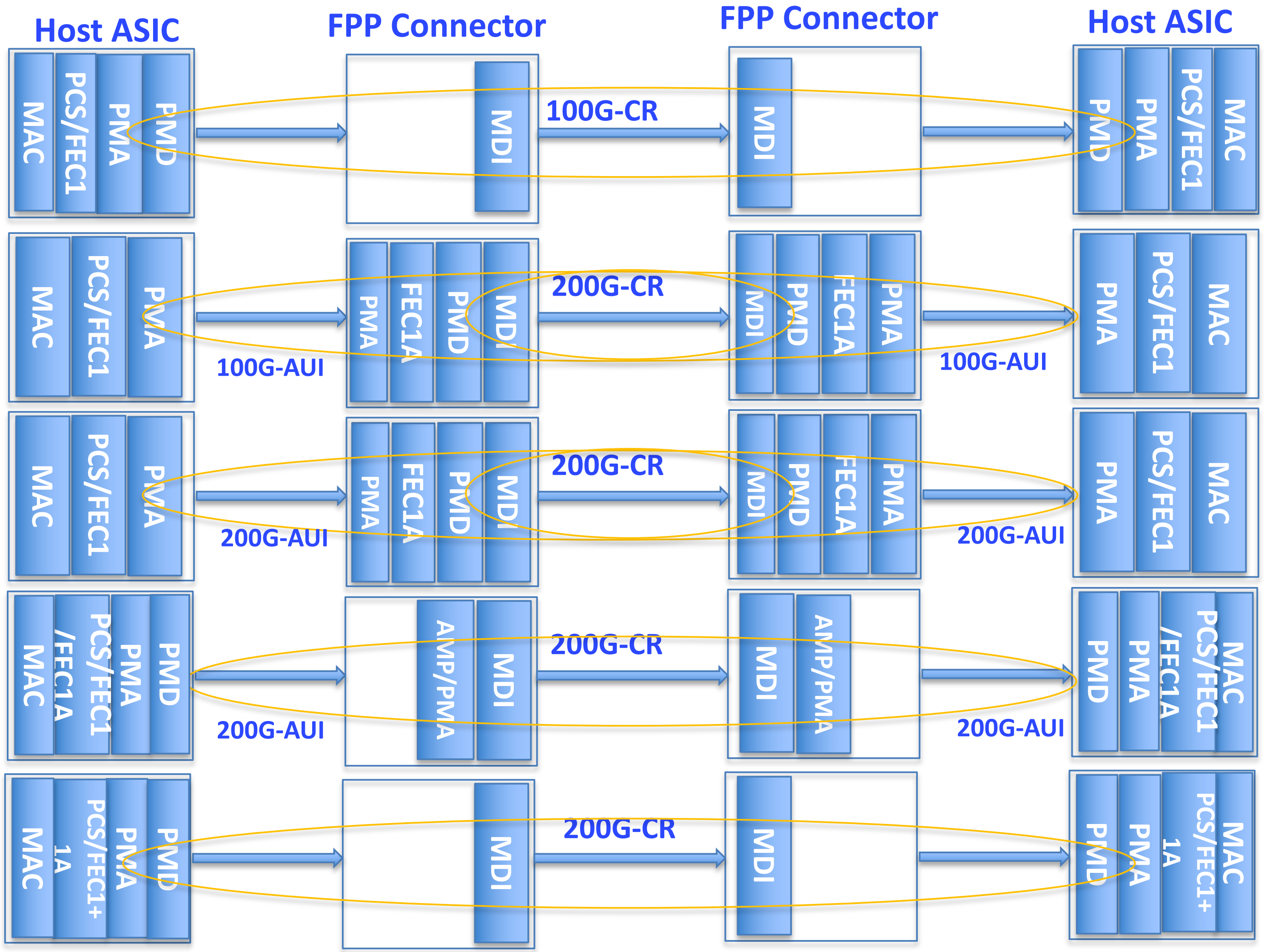
IV. **New concatenated stronger RS+SFEC FEC 3A+3B**
   - Hard to justify at this point
   - This is what is being used in 400ZR

V. **End-end FEC from III will support PMD/PPI**
   - Stronger RS FEC will have about the same or less NCG than III

VI. **Segmented FEC will use optics FEC2 from III**
   - Costly but some AUI variant may require segmented FEC

VII. **Concatenated strong FEC from IV**
   - Too complex for mainstream optics.

Cu CR 800/1600 GbE FEC Options



VII. End-end FEC1 RS (514, 544)

- 1st deployment of 800 GbE/1600 GbE CR

VIII. Same FEC as in II for 200G optics RS(514,544)+SFEC

- SFEC gain 1.61-2.7 dB

IX. Active 200G-CR with SFEC in the module + FEC1

- SFEC in the ASIC unlikely to be compatible with PMA in the cable

X. Active 200G-CR/ACC with FEC1+SFEC

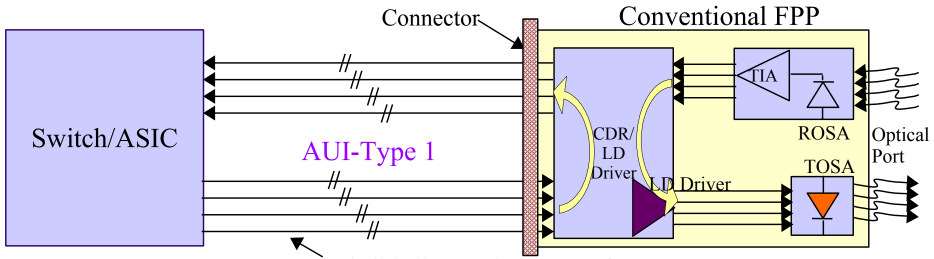
- SFEC in the ASIC unlikely compatible

XI. Passive 200G-CR will use end-end FEC1+SFEC in the ASIC

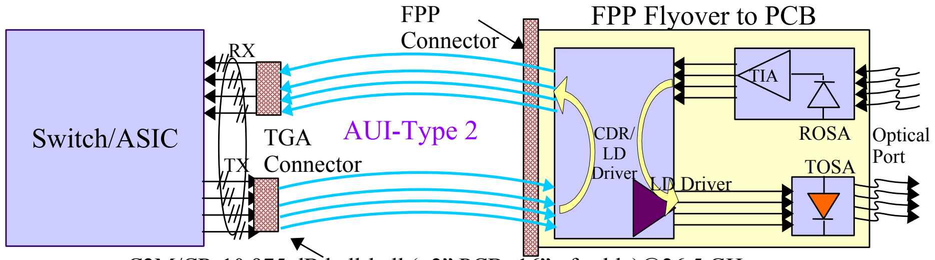
Note: FEC2 and FEC3 not considered for CR as justifying segmented FEC defeat's purpose of low-cost Cu!

Various PMA/PMD/PPI Driving Modes

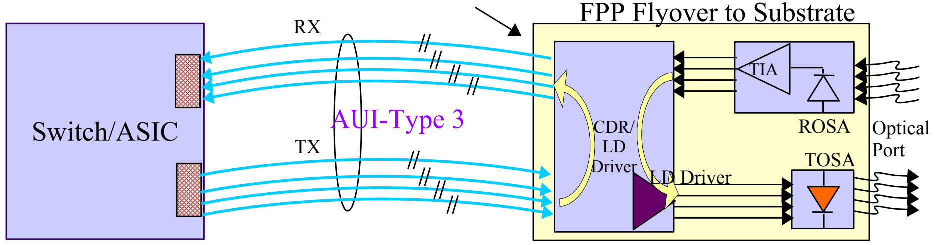
- **With switch radix and data rate increase switch IO power exceed >40% total power**
  - There are potentially 3-4 types of AUIs and several possible AUI/PPI interfaces not all shown here
  - Figure below does not include potential parallel or lower speed buses that can be utilized in conjunction with high density packages.



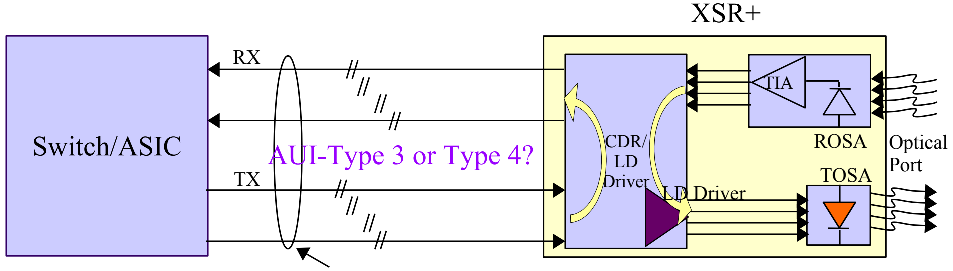
C2M AUI- 16 dB ball-ball (~11” host PCB) @26.5 GHz



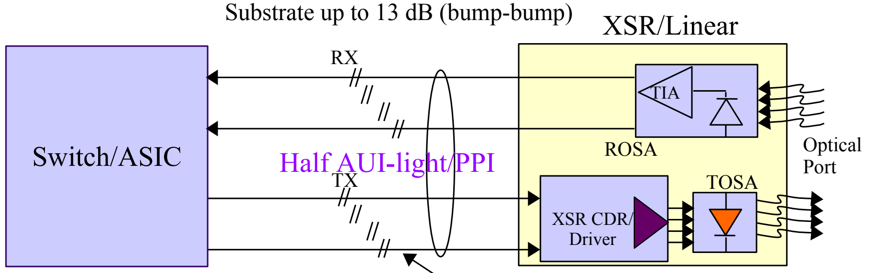
C2M/CR-10.975 dB ball-ball (~2” PCB+16” of cable)@26.5 GHz



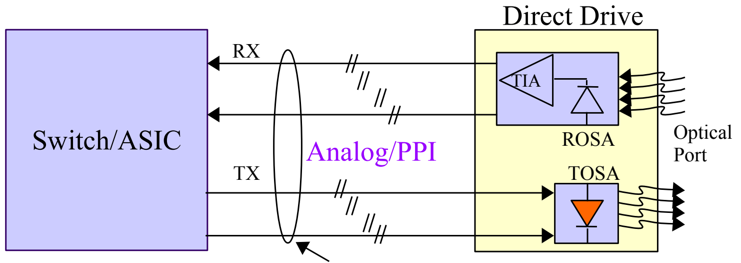
Bump-Bump Loss <13 dB (~16” of cable)@26.5 GHz
Loss inline with XSR+ but may require CTLE+ RX



Substrate up to 13 dB (bump-bump)



Substrate up to 6 dB (bump-bump)



Substrate up to 3.5 dB (bump-bump)

Popular Parallels and 112G-XSR Interfaces

Parameters	AIB*	LIPINCON	IF	HBM2	BOW**	BOW(Turbo)**	112G-XSR
Company	Intel	TSMC	AMD	JEDEC JESD235B	ODSA	ODSA	OIF
Technology	14 nm	7 nm	14 nm	7, 10, 14, 16 nm	7 nm	7 nm	7 nm
Bitrate/pin	2 Gb/s	8 Gb/s	5.3 Gb/s	2.4 Gb/s	16 Gb/s	32 Gb/s	100 Gb/s diff
Architecture	Clock forward Uni-directional	PLL/DLL Uni-directional	Uni-directional With DLL	Clock forward Bi-directional	Clock forward/ DLL Uni-directional	Clock forward/ DLL Bi-directional	CDR Uni-directional
Packaging	EMIB	CoWoS	Organic	CoWoS/Organic	Organic	Organic	Organic
Channel Length	1 mm	4 mm	~10 mm	~5 mm	~10 mm	~10 mm	50 mm
Termination	Unterminated	Unterminated	Terminated	Unterminated	Terminated***	Terminated***	Double Terminated
Chiplet Bump Pitch	50 μm	40 μm	150 μm	40 μm-150 μm	130 μm	130 μm	0.6 mm LGA Socket
PHY Power	0.85 pJ/bit	0.56 pJ/bit	2 pJ/bit	Depend on process	~0.5 pJ/bit	~0.6 pJ/bit	~1.5 pJ/bit
Bandwidth Density	1.2 Tb/s/mm^2	1.6 Tb/s/mm^2	~0.22 Tb/s/mm^2	Depend on bump	0.64 Tb/s/mm^2 (Include ECC)	1.28 Tb/s/mm^2 (Include ECC)	~0.6 Tb/s/mm^2 (Include RS FEC)

* https://github.com/intel/aib-phy-hardware/blob/master/docs/AIB_Intel_Specification%201_2%20.pdf

** http://files.opencompute.org/oc/public.php?service=files&t=6bfc2493f2f3e0a1d1a14a3314062bdd&download

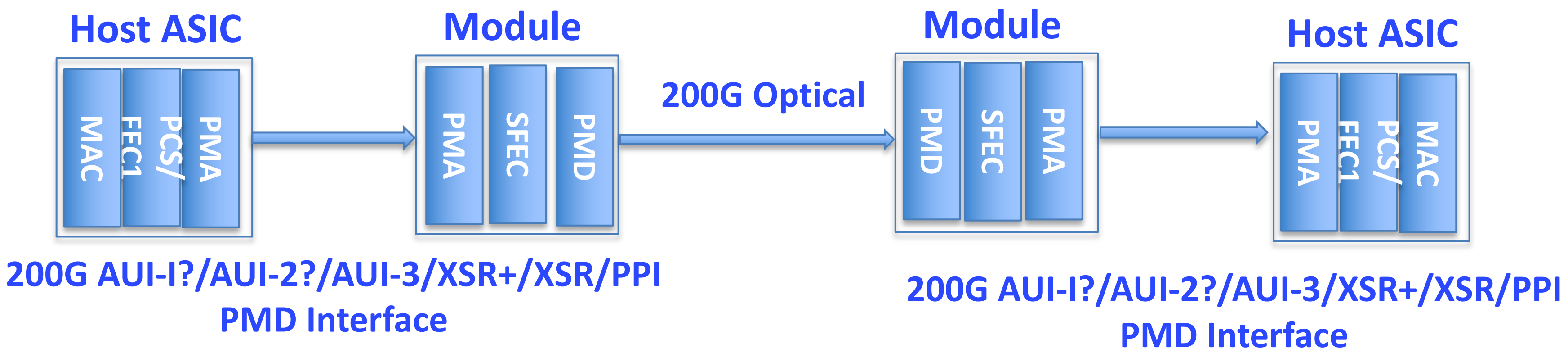
*** Can operate unterminated for trace up to 1 mm up to 5 Gb/s.

How to Define 200G/lane Optical PMDs Prior to 200G/lane AUI

- **802.3bs successfully defined an architecture that operated with an end-end FEC by allocating 0.1 dBo to 4 AUI sub-links and prior to defining 100G-AUI**
  - There are potentially 3-4 types of AUIs some expect to operate with end-end FEC with 0.1-0.2 dBo allocation to the electrical sublinks
  - With emergence of optics/Cu co-packaging there are more implementation options than traditional AUIs
  - Some of the optics co-packaging may use low speed parallel buses, PPIs, or even PMD interfaces
  - It is plausible that future 200G system may not have any conventional PCB based AUIs
- **Some variant of 200G/lane AUI expect to be have substantially higher loss, ILD, and reflections**
  - 802.3df should not tax everyone for implementation that may not get used broadly
  - Segmented FEC is a fairer option in such cases
  - Some of the 200G AUI that are more challenging now over time could improve and segmented FEC could then go away.

Following 802.3bs FEC Architecture

- **802.3df task force need to define a new 200G/lane optics FEC with 0.1-0.2 dBo reserved for PMA/PMD/PPI sub-links as shown below**
  - SFEC+RS(514,544) allow seamless upgrade of 100G-AUIs to 200G/lane optics without rate increase on the 802.3ck interfaces
  - It is also expected the end-end SFEC+RS(514,544) to support a range of AUIs, PPIs, and PMD interfaces
    - But AUI-1 or AUI-2 implementations initially may require segmented FEC due to high loss and high reflections!


Host ASIC

MAC

PCS/ FEC1

PMA

Module

PMA

SFEC

PMD

200G Optical

Module

PMD

SFEC

PMA

Host ASIC

PMA

FEC1 PCS/

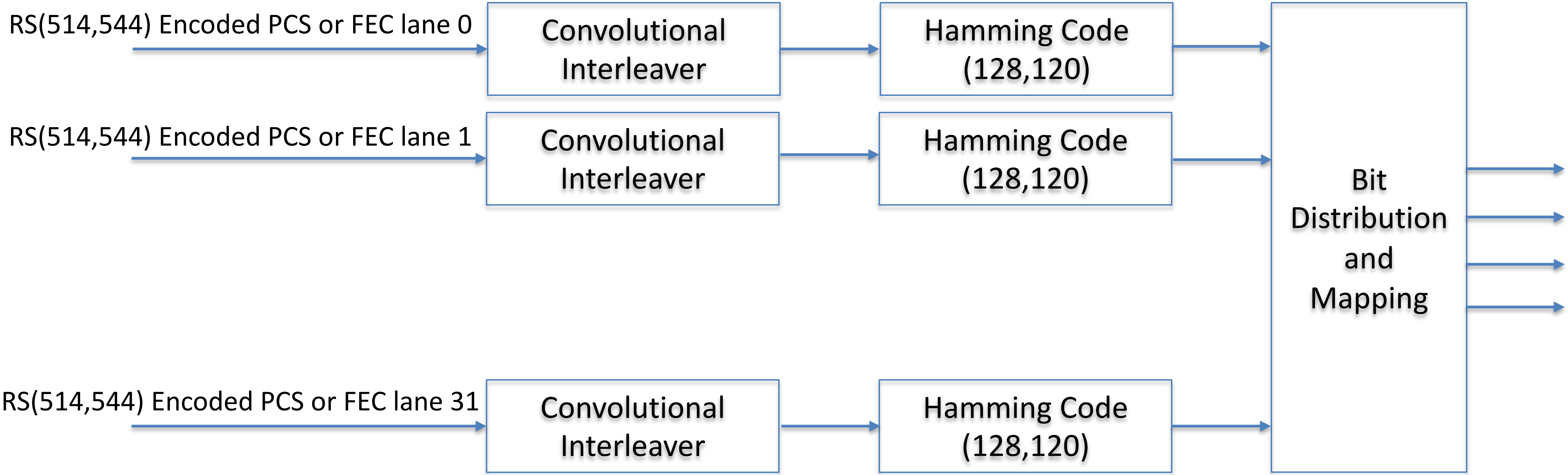
MAC

200G AUI-I?/AUI-2?/AUI-3/XSR+/XSR/PPI
PMD Interface

200G AUI-I?/AUI-2?/AUI-3/XSR+/XSR/PPI
PMD Interface

Concatenated KP FEC + Hamming ”SFEC”

- **SFEC200 is concatenation of RS(514,544) with (128,120) Hamming code**
  - SFEC200 Hamming gain is +2.7 dB compared to lyubomirsky_nea_01_200914 +1.61 dB gain
  - Code overhead =12.89%, Net Coding Gain (NCG)=9.5 dB, and BER limit of 4.8E-3
  - KP FEC + enhanced SFEC200 also being considered for 800G-LR coherent.


RS(514,544) Encoded PCS or FEC lane 0
Convolutional Interleaver
Hamming Code (128,120)
RS(514,544) Encoded PCS or FEC lane 1
Convolutional Interleaver
Hamming Code (128,120)
RS(514,544) Encoded PCS or FEC lane 31
Convolutional Interleaver
Hamming Code (128,120)
Bit Distribution and Mapping

Breaking B400G PMDs Sets Potentially into 3 Taskforces

- □ **802.3df taskforce**
  - – Consider adopting 800G Ethernet Tech. Con. MAC/PCS
  - – Define 800G-DR8, 800G-SR8, 800G-FR8
  - – 800G-AUI8, 800G-CR8/KR8
- □ **2nd taskforce starts ~ Nov 2022**
  - – 200G/lane SMF optics PMDs
  - – 800G-ZR
  - – 1600 GbE MAC/PCS
- □ **3rd taskforce starts ~ March 2023**
  - – 200G-AUI/C2C (let the MSA continue improving the connector as OIF investigates)
  - – Other optical PMDs including more efficient MMF PMDs.

Summary

- **Given that next generation switches are <18 months away the industry needs direction from 802.3df task force no later than November 2022 by publishing D2.0**
  - In this time frame all the 100 Gb/s/lane PMD should also be defined
- **The most important decision for the task force is how to architect the FEC for legacy compatibility (802.3bs, cu, ck, and db), support 200G optics, and support some variants of AUIs/PPIs/PMDs without the need for segmented FEC**
- **KP RS(514,544) FEC + SFEC200 offers legacy compatibility and supports 200G PMDs**
  - KP FEC+SFEC200 allow extending the 802.3bs architecture to most 802.3df PMDs
  - Directly will leverage 802.3ck 100 Gb/s/lanes AUIs without any Baudrate increase
  - 800ZR will use segmented FEC and the more conventional AUIs may also use segmented FEC
  - KP FEC + SFEC200 expect to support some variants of AUIs sub-links, PPI, and PMD links
  - KP FEC + SFEC200 generally is compatible with Cu/CR but some ACC Cu variant may not be compatible
- **The flexibility of KP FEC + SFEC200 allow to defining 800 GbE/1600 GbE FEC/PCS architecture now**
- **Given 802.3df is a condensed project but with some PMDs requiring longer development should consider spinning out new task forces as needed.**