Architecture considerations for 800GbE and 1.6TbE

Yucun LU, Yan ZHUANG, Huawei Technologies

IEEE P802.3df Task Force

May 18 2022

Observations of adopted physical layer objectives

Ethernet Rate	Signaling Rate	Electrical			Optical					
		AUI	Backplane	Copper Cable	MMF 50m	MMF 100m	SMF 500m	SMF 2km	SMF 10km	SMF 40km
200Gbps	200Gbps	Over 1 lane 200GAUI-1	TBD * 200GBASE-KR1	Over 1 pair 200GBASE-CR1			Over 1 pair 200GBASE-DR1	Over 1 pair 200GBASE-FR1		
400Gbps	100Gbps							Over 4 pairs 400GBASE-DR4-2		
	200Gbps	Over 2 lanes 400GAUI-2	TBD * 400GBASE-KR2	Over 2 pairs 400GBASE-CR2			Over 2 pairs 400GBASE-DR2			
800Gbps	100Gbps	Over 8 lanes 800GAUI-8	Over 8 lanes 800GBASE-KR8	Over 8 pairs 800GBASE-CR8	Over 8 pairs 800GBASE-VR8	Over 8 pairs 800GBASE-SR8	Over 8 pairs 800GBASE-DR8	Over 8 pairs 800GBASE-DR8-2		
	200Gbps	Over 4 lanes 800GAUI-4	TBD * Over 4 pairs 800GBASE-KR4	Over 4 pairs 800GBASE-CR4			Over 4 pairs 800GBASE-DR4	Over 4 pairs 800GBASE-DR4-2 Over 4 lambdas 800GBASE-FR4	TBD	
	TBD								Over single SMF in each direction ?	Over single SMF in each direction ?
?1.6Tbps	100Gbps	Over 16 lanes 1.6TAUI-16								
	200Gbps	Over 8 lanes 1.6TAUI-8		Over 8 pairs 1.6TGBASE-CR8			Over 8 pairs 1.6TBASE-DR8	Over 8 pairs 1.6TBASE-DR8-2		

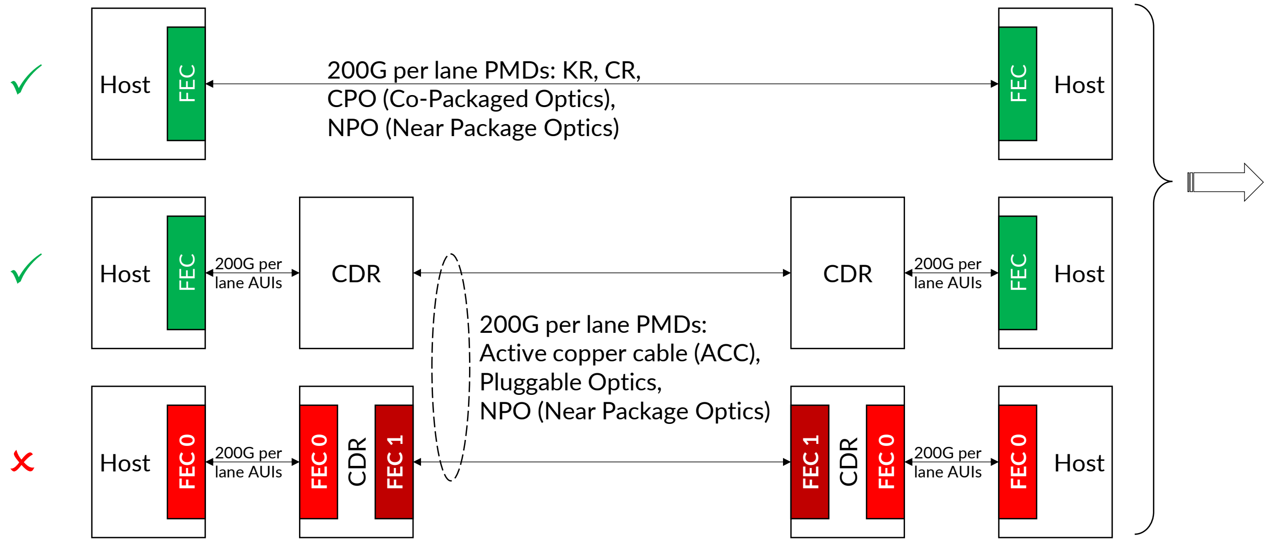
https://www.ieee802.org/3/df/proj_doc/objectives_P802d3df_220317.pdf

* Should be adopted as long as the signaling & modulation & insertion loss objectives for CR/KR channels are determined.

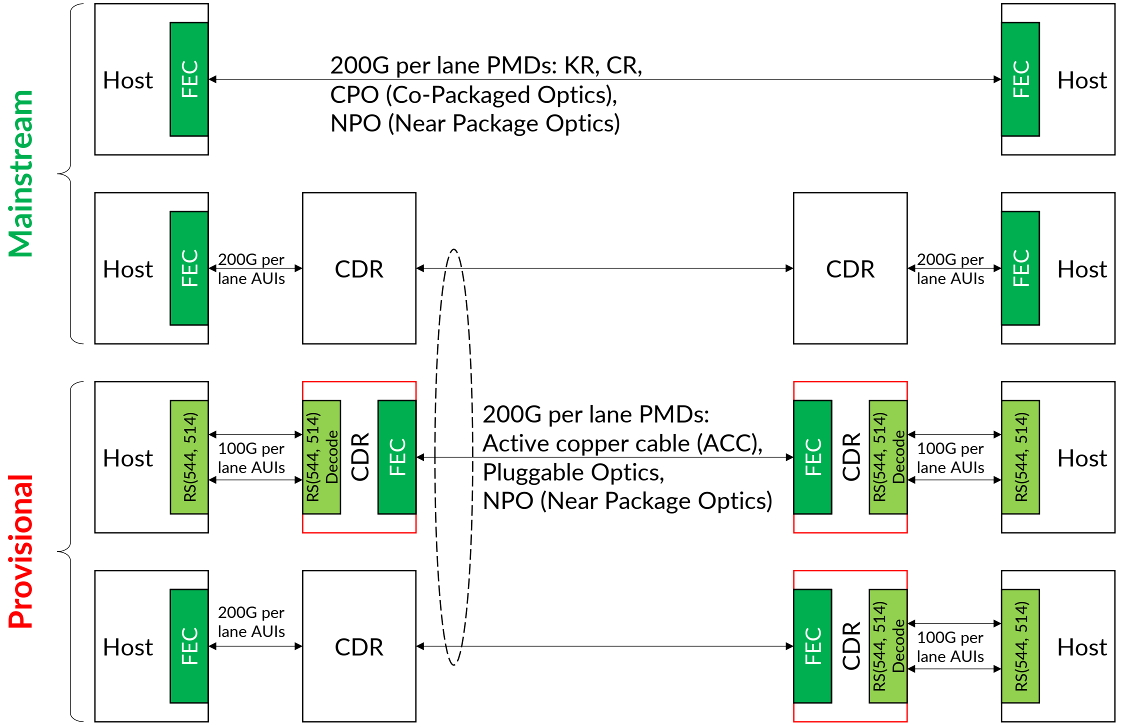
1. In mainstream applications, the number of AUI channels is the same as that of PMD channels.
2. Flexible breakout of the CDR is an essential requirement to support 1x, 2x, 4x and 8x 200Gbps Ethernet interfaces.

Start from 200G AUIs & PMDs and determine the optimization direction for the mainstream applications

lu_3df_logic_220425 page 2



lu_3df_logic_220425 page 3



For 200G AUIs and 200G PMDs:

- End-to-End FEC to support low latency, low power consumption and low cost;
- Bit-transparent CDR to support flexible breakout applications.

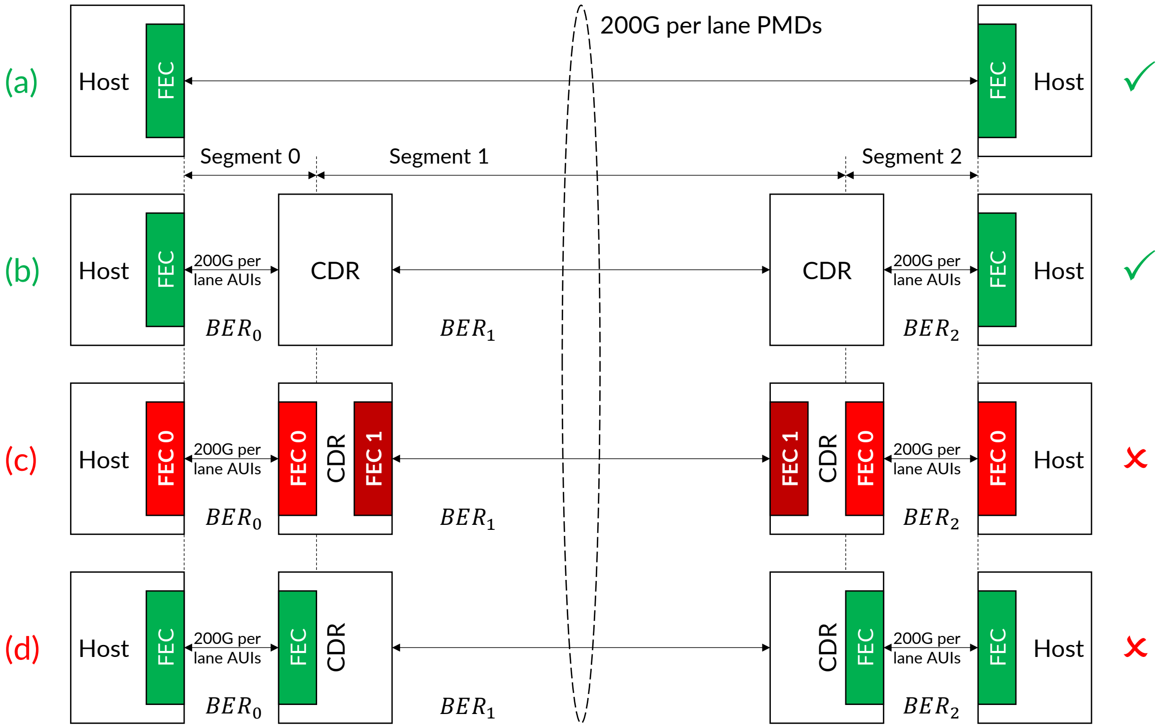
For 100G AUIs and 200G PMDs:

- Segmented FEC is natural, necessary and preferred.
  - Use the same FEC for 200G AUIs and 200G PMDs.
  - Use virtual lane aggregation to reduce the virtual lane number and the complexity.
  - Use 1:1 PMA to avoid error spreading and guarantee the FEC performance.
- End-to-end FEC with interleaved RS(544, 514) needs a lot of evidence to show that:
  - RS(544, 514) with bitmux PMA is sufficient for both 200G AUIs & PMDs.
    - Channel quality and DSP performance are good enough for RS(544, 514).
    - Error spreading of bitmux PMA does not hurt RS(544, 514) performance.

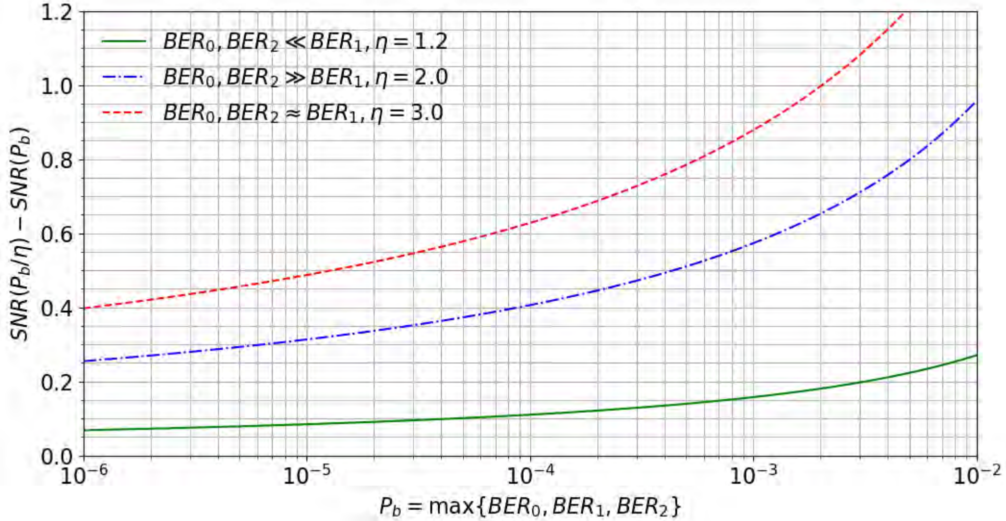
General design rules:

1. Remove unnecessary intermediate processing.
2. Simplify the CDR as much as possible and shift the necessary “complexity” to the host ASIC.
   (The Concatenated FEC does not simplify the CDR)

FEC processing in CDR has small SNR improvement, the gain is negligible if PMD is dominant in performance

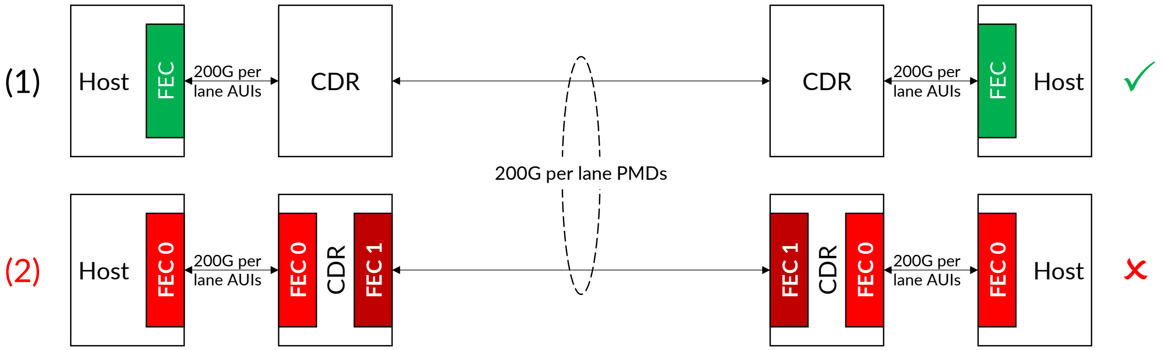


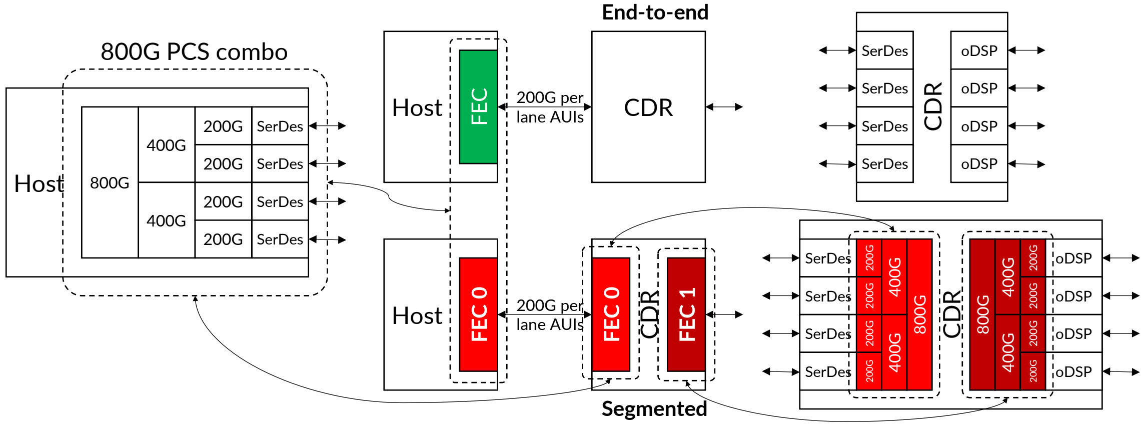
#	The end-to-end bit error rate: $\left[1-\prod_{i=0}^{2}(1-BER_i)\right] \approx \sum_{i=0}^{2} BER_i$.		
(i)	$BER_0, BER_2 \ll BER_1$	$BER \approx BER_1$	$\leq 1.2 \cdot \max\{BER_0, BER_1, BER_2\}$
(ii)	$BER_0, BER_2 \gg BER_1$	$BER \approx BER_0 + BER_2$	$\leq 2 \cdot \max\{BER_0, BER_1, BER_2\}$
(iii)	$BER_0, BER_2 \approx BER_1$	$BER \approx \sum_{i=0}^{2} BER_i$	$\leq 3 \cdot \max\{BER_0, BER_1, BER_2\}$

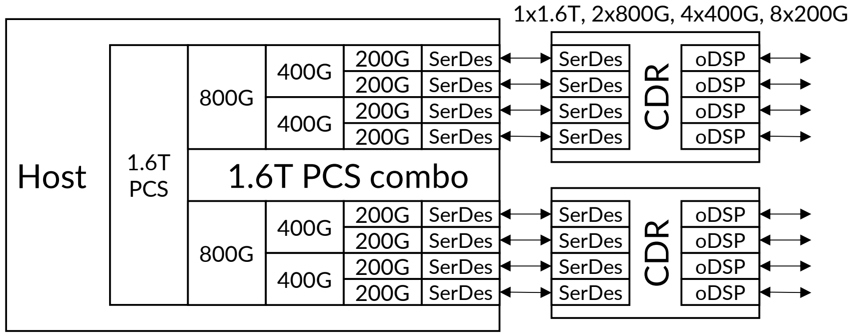


#	AUI BER_0	PMD BER_1	AUI BER_2	e2e BER	ΔSNR	Take-aways
1	1e-5	2.4e-4	1e-5	2.60e-4	0.056dB	✓ Segmented FEC has small SNR improvement compared with end-to-end FEC. ✓ If PMD is dominant, the segmented FEC has less than 0.2dB SNR gain. ✓ For other cases, the segmented FEC has less than 1dB SNR gain. ✓ Case 1 is “100G AUIs & 100G PMDs”; ✓ Case 2 is potential “100G AUIs & 200G PMDs”; ✓ Case 2 & 3 are potential “200G AUIs & 200G PMDs”. ✓ Case 4 ~ 6 are not reasonable, just for reference.
2	1e-5	2e-3	1e-5	2.02e-3	0.010dB	
3	2e-4	2e-3	2e-4	2.40e-3	0.186dB	
4	1e-3	2.4e-4	1e-3	2.24e-3	0.768dB	
5	2.4e-4	2.4e-4	2.4e-4	7.20e-4	0.831dB	
6	6.7e-4	6.7e-4	6.7e-4	2.01e-3	0.997dB	

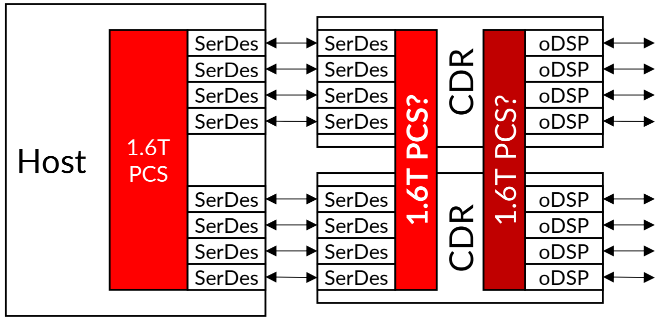
FEC processing in CDR is expensive and inflexible, it should be applied only when necessary (e.g. gearbox & new FEC)







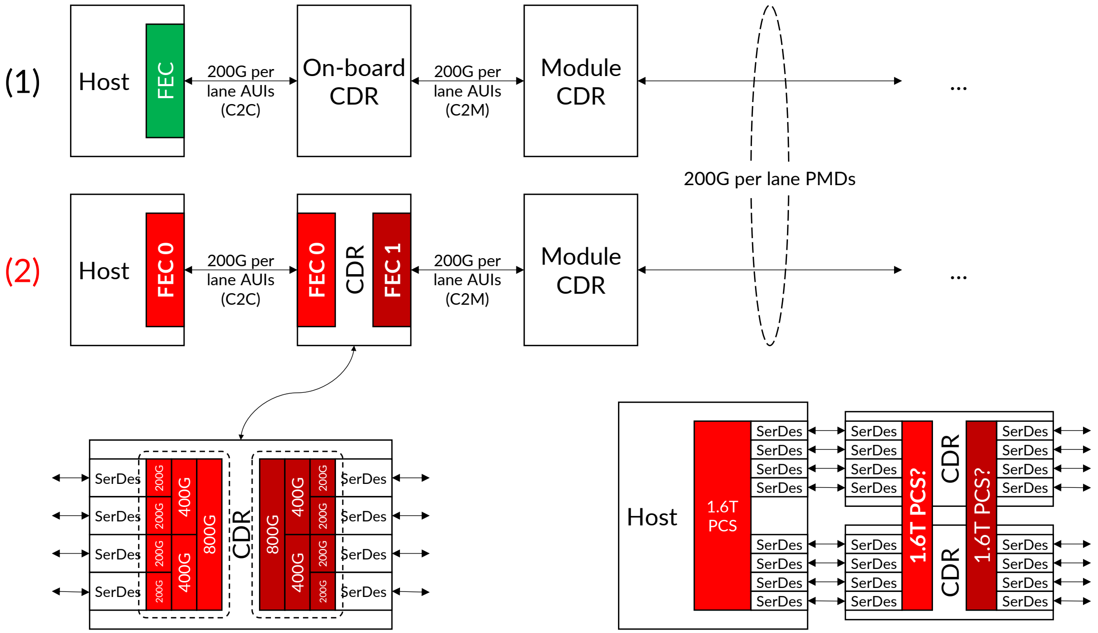
End-to-end FEC can support flexible break out of CDR.



How to achieve 1.6T PCS across two CDR chip?

- FEC processing inside CDR means integration of two back-to-back **full Ethernet PCS stacks inside the CDR**.
- If the rates of the client side and the line side are different, one **extra PLL** per direction is required for new frequency point.
- Ethernet PCS/FEC crosses multiple physical lanes, FEC termination **is difficult to support flexible CDR breakout**.

FEC processing in CDR is expensive and inflexible, choose the FEC according to the worst AUI link segment



Integration full Ethernet PCS stack inside the CDR is too expensive.

Inflexible for breakout applications. How to achieve 1.6T PCS across two CDR chip?

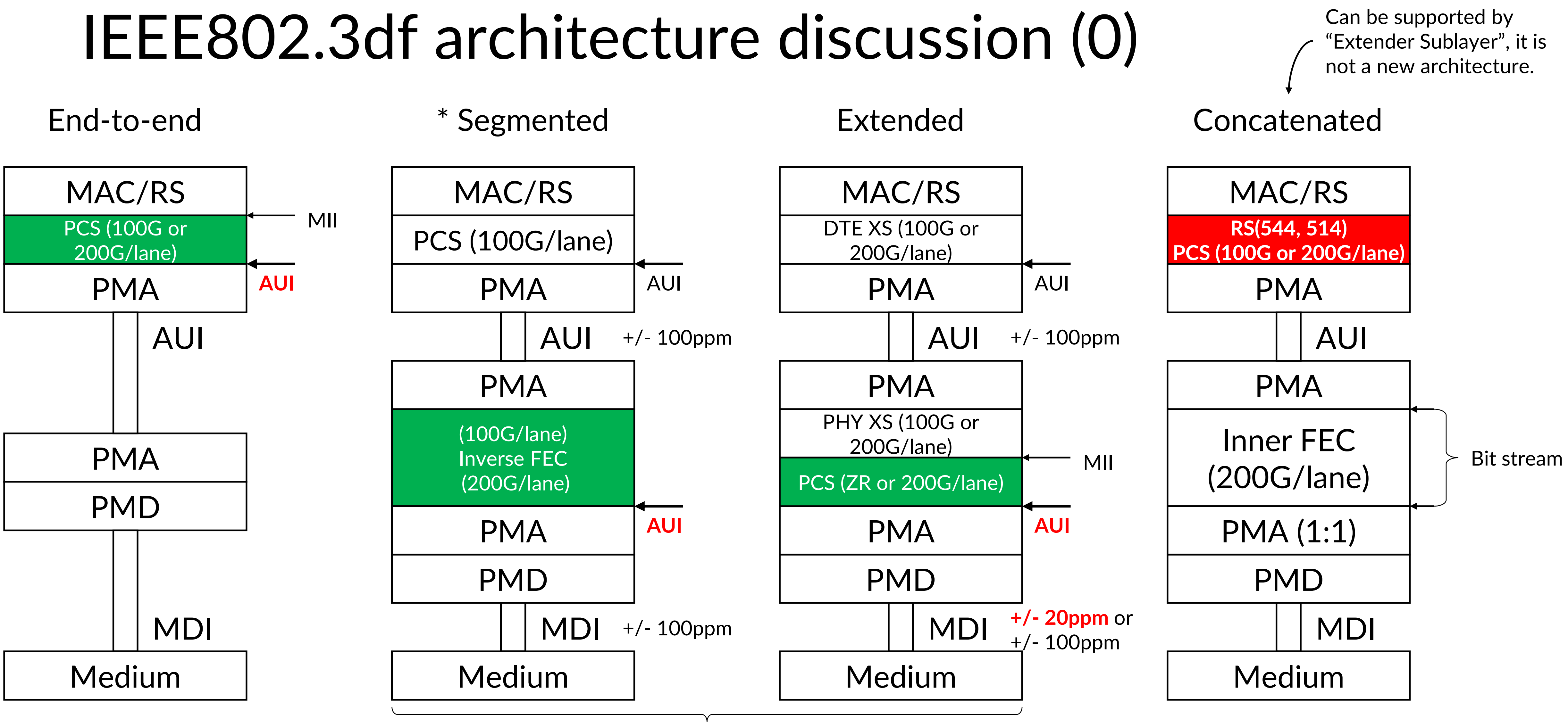
#	AUI-C2C *BER*	AUI-C2M *BER*	AUI E2E *BER*	ΔSNR
1	2e-4	1e-5	2.10e-4	**0.033dB**
2	1e-3	2e-4	1.20e-3	**0.163dB**
3	1e-4	1e-4	2.00e-4	0.445dB
4	2e-4	2e-4	4.00e-4	0.493dB
Case 1 is "100G AUIs"; Case 2 is potential "200G AUIs"; Case 3 & 4 are not reasonable, just for reference.				

- FEC termination in CDR has negligible SNR improvement, less than 0.2dB.
- FEC for termination for cascaded AUIs are too expensive and inflexible.
  - Two back-to-back full Ethernet PCS stacks inside the on-board CDR are required.
  - One extra PLL per direction is required to support new frequency point, if the rates of the client side and line side are different.
  - Cannot support flexible CDR breakout.

Design rules for on-board CDR (CDR for AUIs):

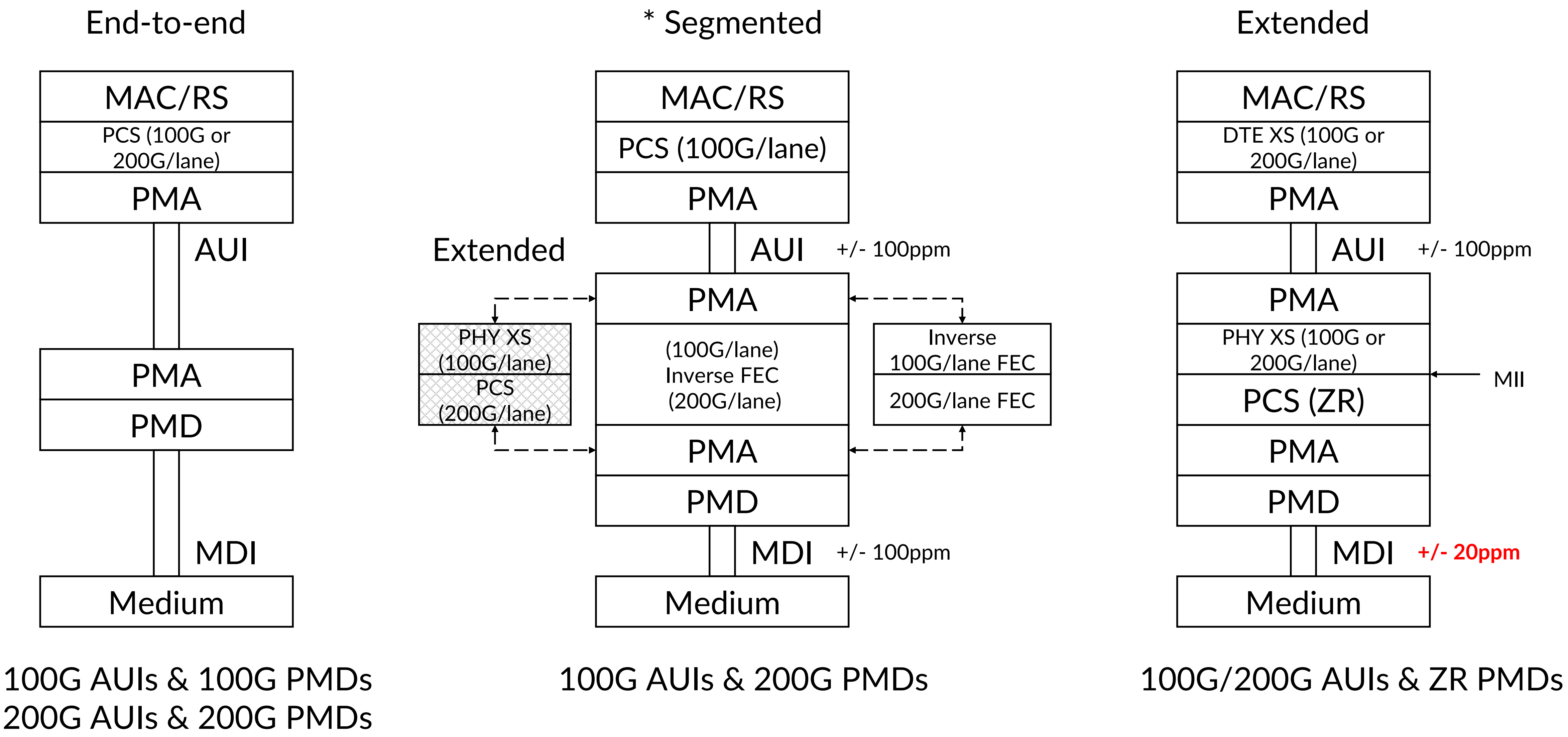
1. Simplify the on-board CDR as much as possible and shift the necessary "complexity" to the host ASIC. **Bit-transparent CDR is always optimal**.
2. Choose the end-to-end FEC according to the worst segment of the AUI links, i.e. AUI-C2C.

IEEE802.3df architecture discussion (0)



1. “Rate matching (supported by XS)” is mandatory for “ZR” applications due to the different requirement of clock drift in ppm.
2. It is not necessary to terminate the PCS, only the FEC needs to be terminated and new FEC is encoded/decoded.
3. Inverse RS-FEC Sublayer overview can be found in nicholl_3ck_02_0519.

IEEE802.3df architecture discussion (1)



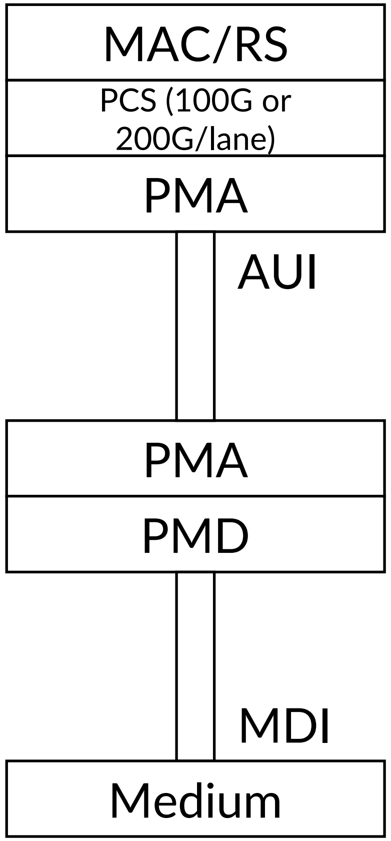
100G AUIs & 100G PMDs
200G AUIs & 200G PMDs

100G AUIs & 200G PMDs

100G/200G AUIs & ZR PMDs

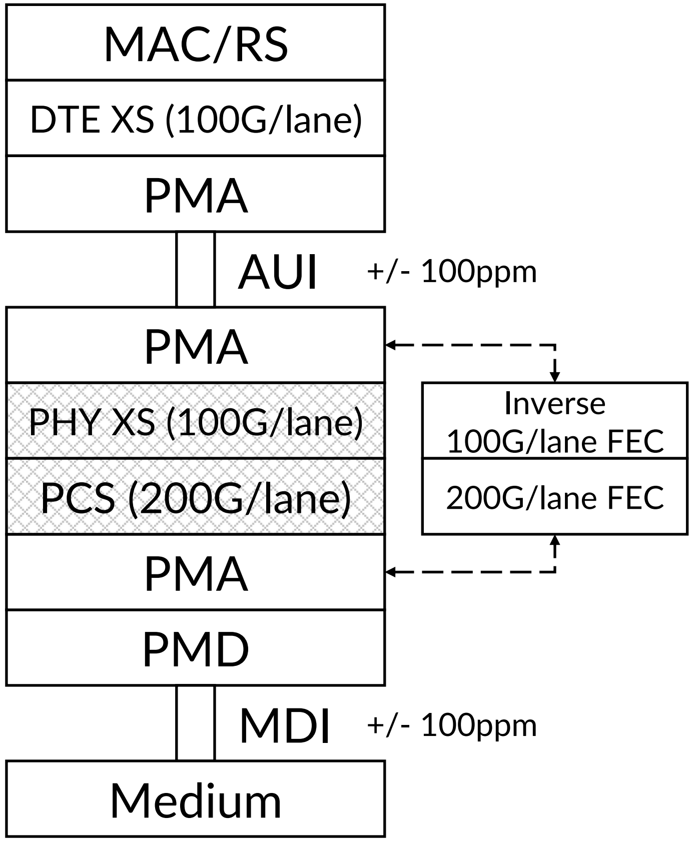
IEEE802.3df architecture discussion (2)

End-to-end



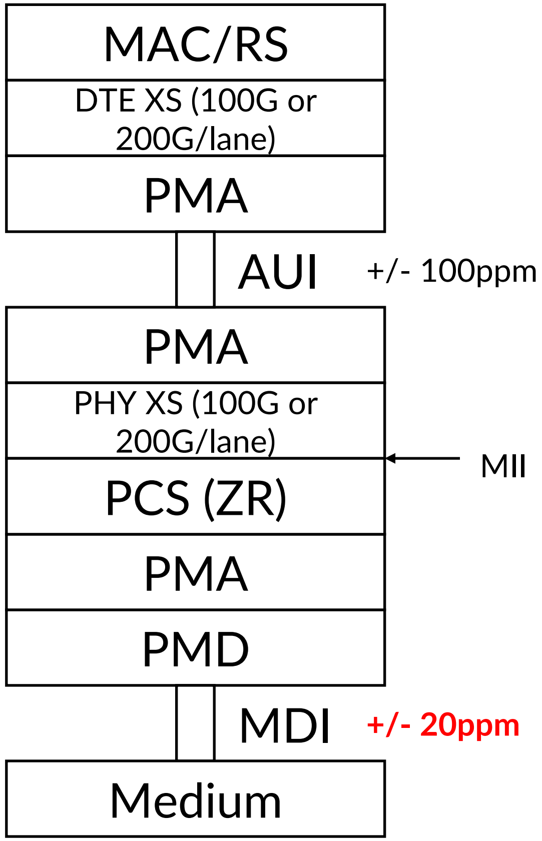
100G AUIs & 100G PMDs
200G AUIs & 200G PMDs

* Segmented FEC supported by XS



100G AUIs & 200G PMDs

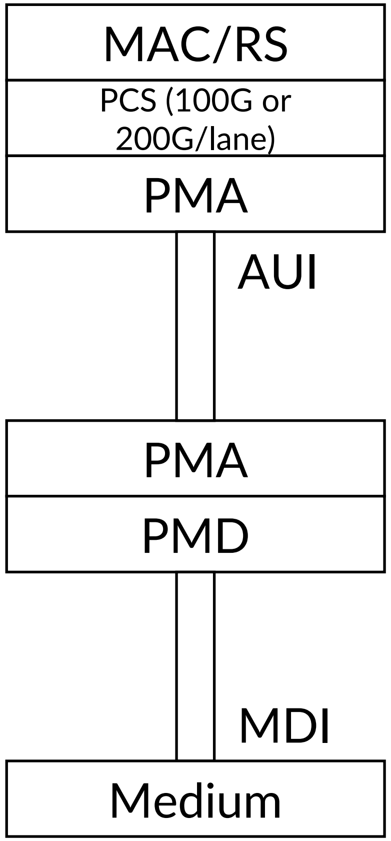
Extended



100G/200G AUIs & ZR PMDs

IEEE802.3df architecture discussion (3)

End-to-end



MAC/RS

PCS (100G or 200G/lane)

PMA

AUI

PMA

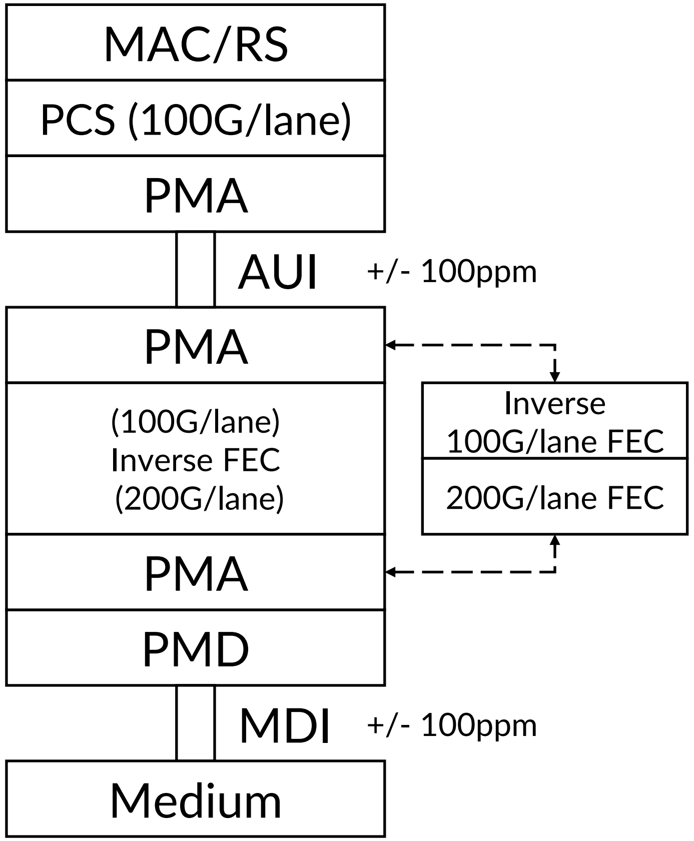
PMD

MDI

Medium

100G AUIs & 100G PMDs
200G AUIs & 200G PMDs

* Segmented FEC supported by FEC-Inv



MAC/RS

PCS (100G/lane)

PMA

AUI +/- 100ppm

PMA

(100G/lane)
Inverse FEC
(200G/lane)

Inverse 100G/lane FEC

200G/lane FEC

PMA

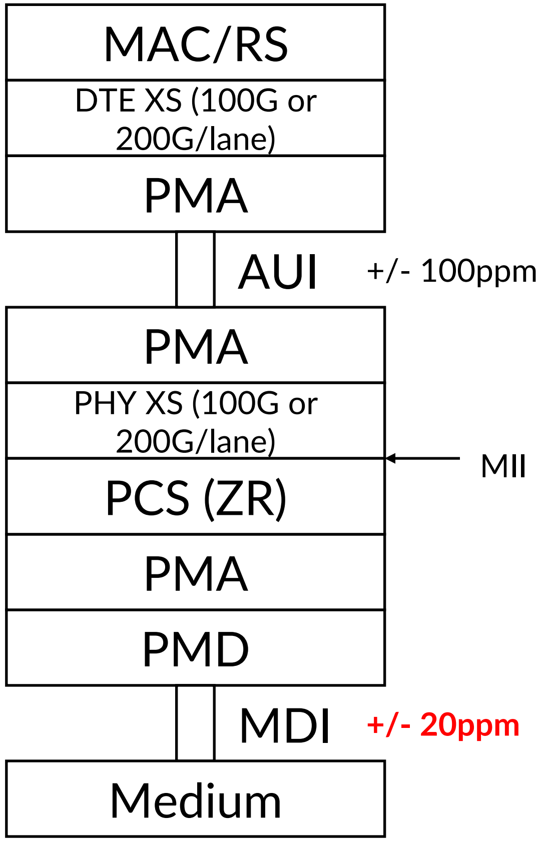
PMD

MDI +/- 100ppm

Medium

100G AUIs & 200G PMDs

Extended



MAC/RS

DTE XS (100G or 200G/lane)

PMA

AUI +/- 100ppm

PMA

PHY XS (100G or 200G/lane)

MII

PCS (ZR)

PMA

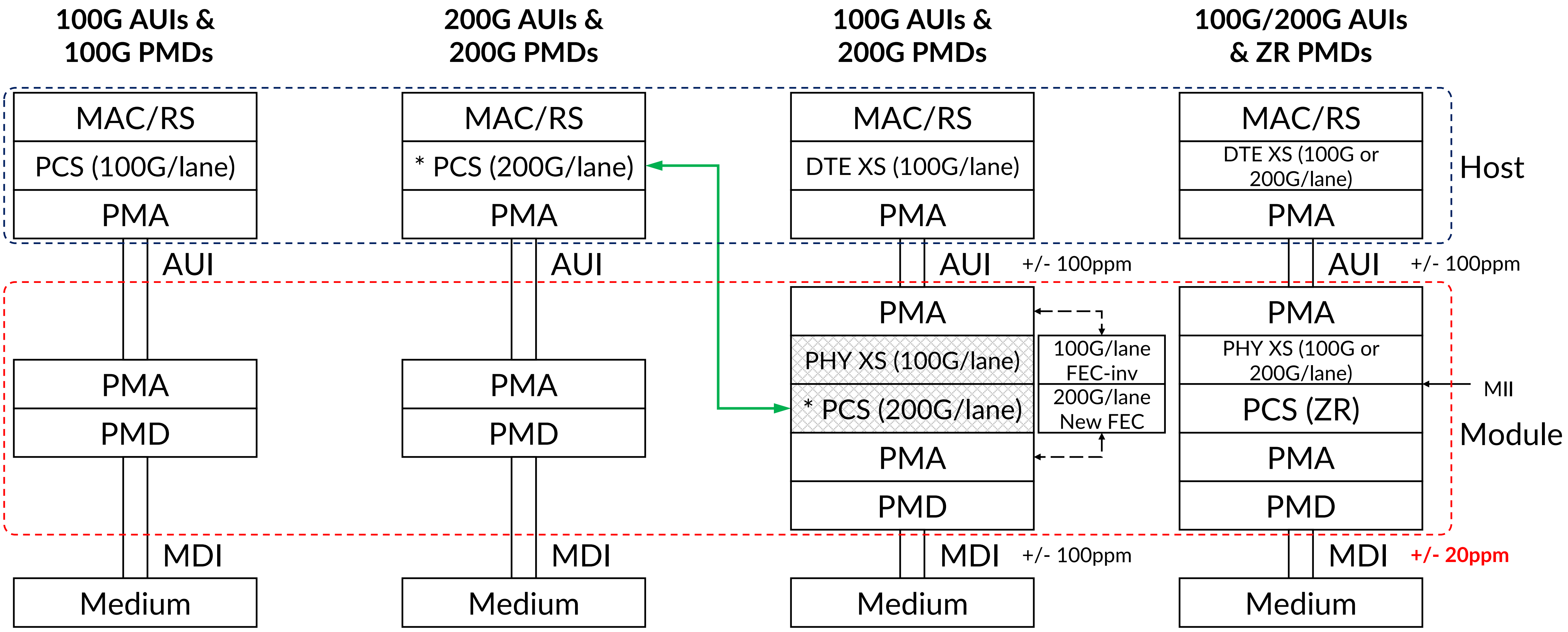
PMD

MDI **+/- 20ppm**

Medium

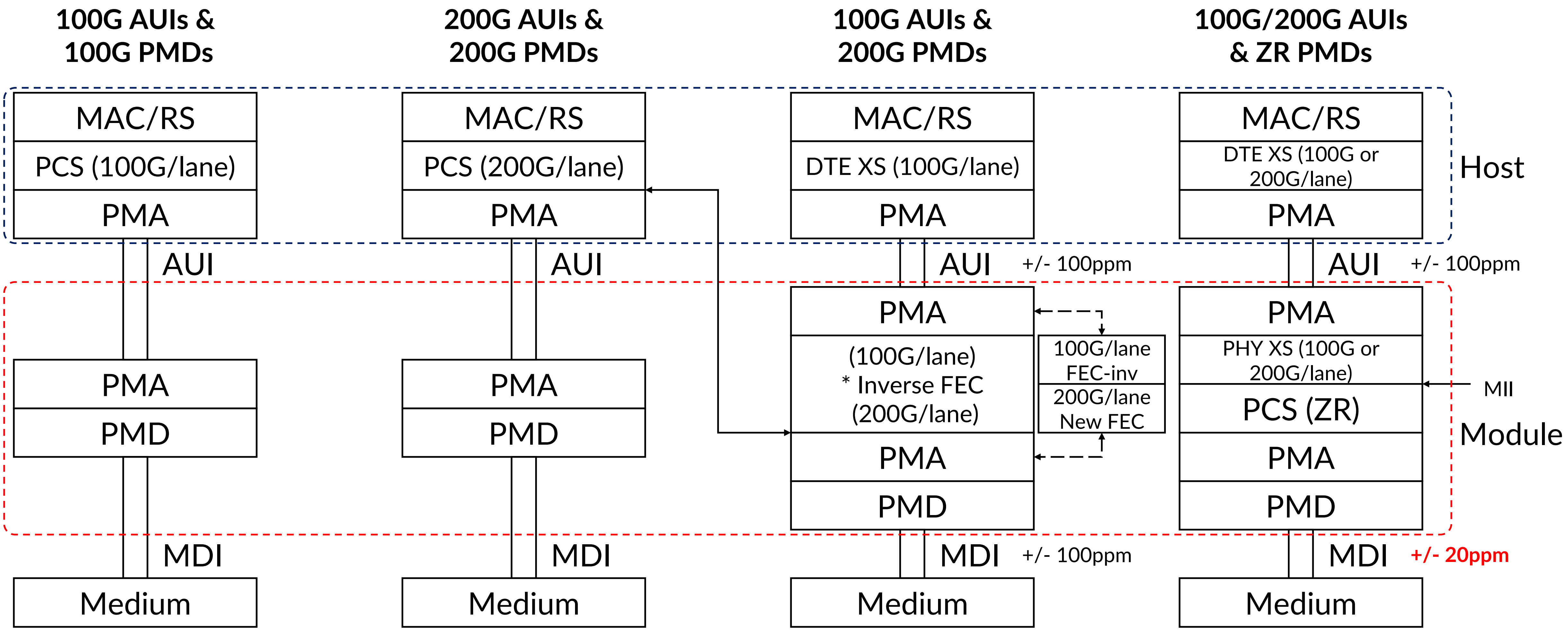
100G/200G AUIs & ZR PMDs

Proposed IEEE 802.3df architecture (0)



* Pros: Only editorial benefit, i.e. re-use “200G/lane PCS” clause for PHY XS; Cons: Lack of technical insights and details.

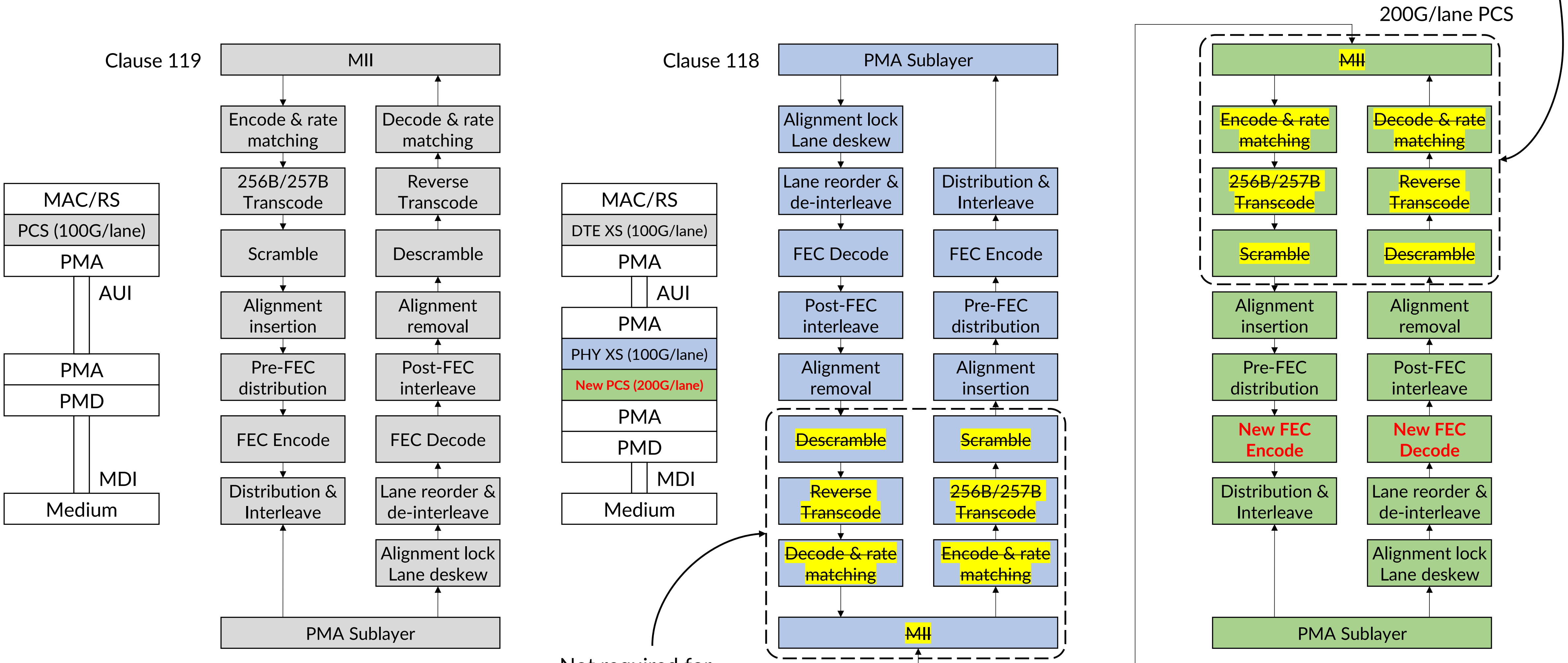
Proposed IEEE 802.3df architecture (1)



* Pros: Intuitive, simple and clear; Cons: Define the inverse FEC sublayer (Editorial).

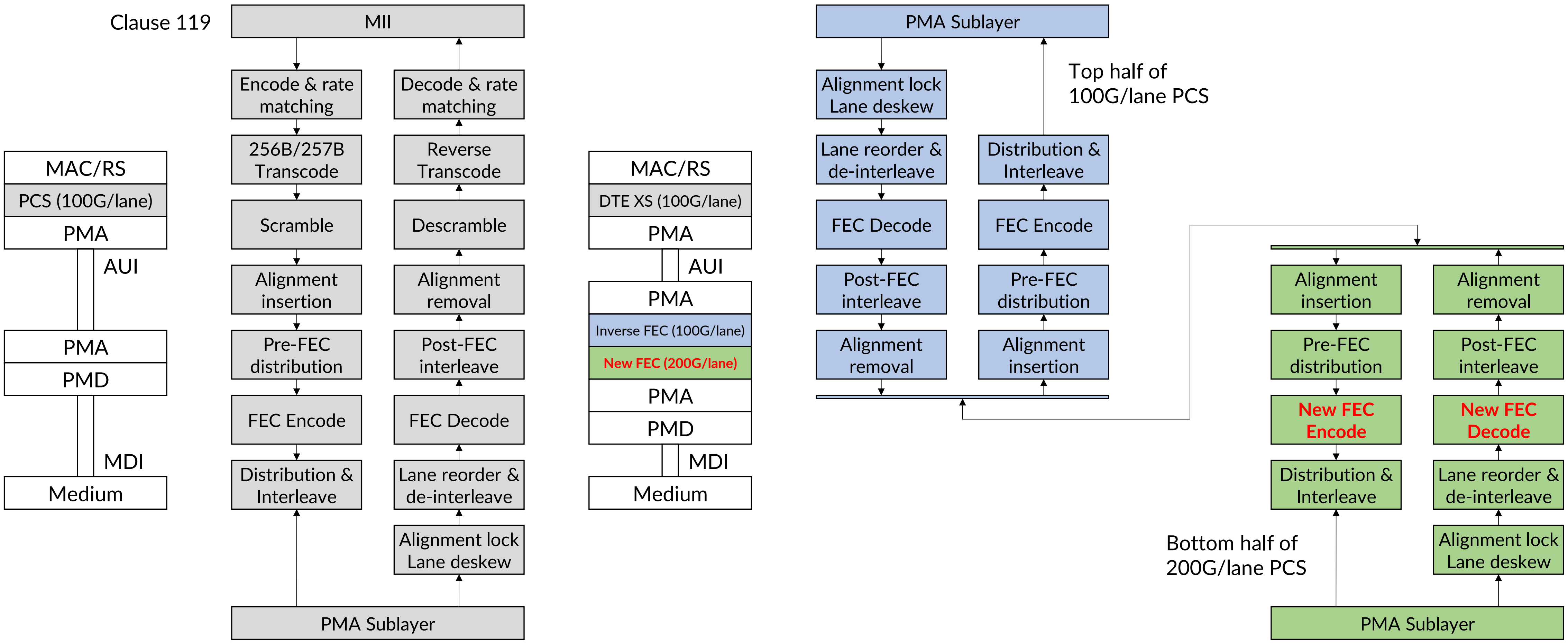
Extender Sublayer for the “new FEC”

Not required for
FEC conversion



Not required for
FEC conversion

Inverse FEC Sublayer for the “new FEC”



MAC/RS
PCS (100G/lane)
PMA
AUI
PMA
PMD
MDI
Medium

Clause 119

MII
Encode & rate matching
256B/257B Transcode
Scramble
Alignment insertion
Pre-FEC distribution
FEC Encode
Distribution & Interleave
Decode & rate matching
Reverse Transcode
Descramble
Alignment removal
Post-FEC interleave
FEC Decode
Lane reorder & de-interleave
Alignment lock Lane deskew
PMA Sublayer

MAC/RS
DTE XS (100G/lane)
PMA
AUI
PMA
Inverse FEC (100G/lane)
New FEC (200G/lane)
PMA
PMD
MDI
Medium

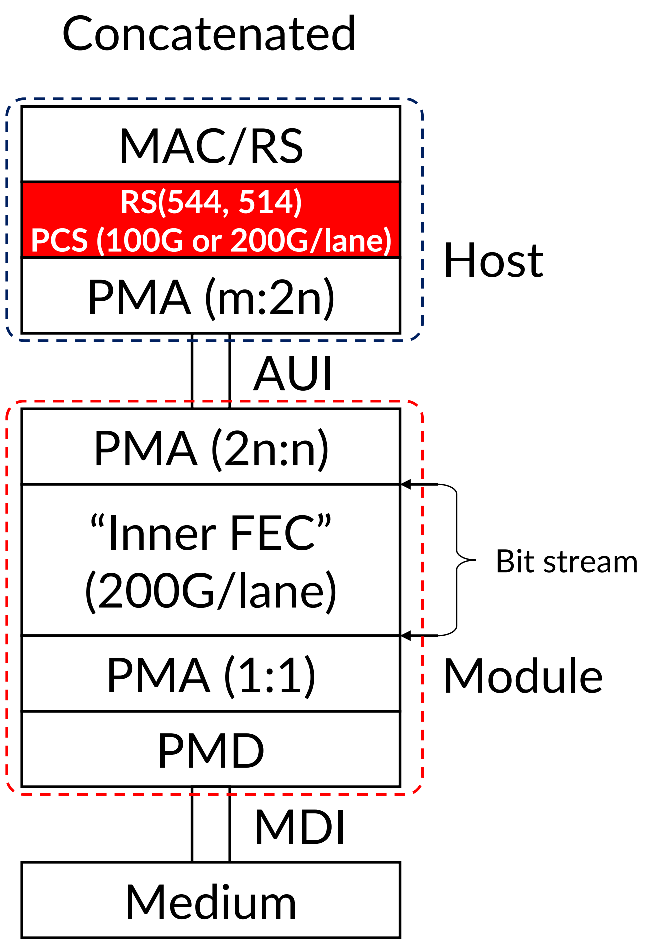
PMA Sublayer
Alignment lock Lane deskew
Lane reorder & de-interleave
FEC Decode
Post-FEC interleave
Alignment removal
Distribution & Interleave
FEC Encode
Pre-FEC distribution
Alignment insertion

Top half of 100G/lane PCS

Alignment insertion
Pre-FEC distribution
New FEC Encode
Distribution & Interleave
Alignment removal
Post-FEC interleave
New FEC Decode
Lane reorder & de-interleave
Alignment lock Lane deskew
PMA Sublayer

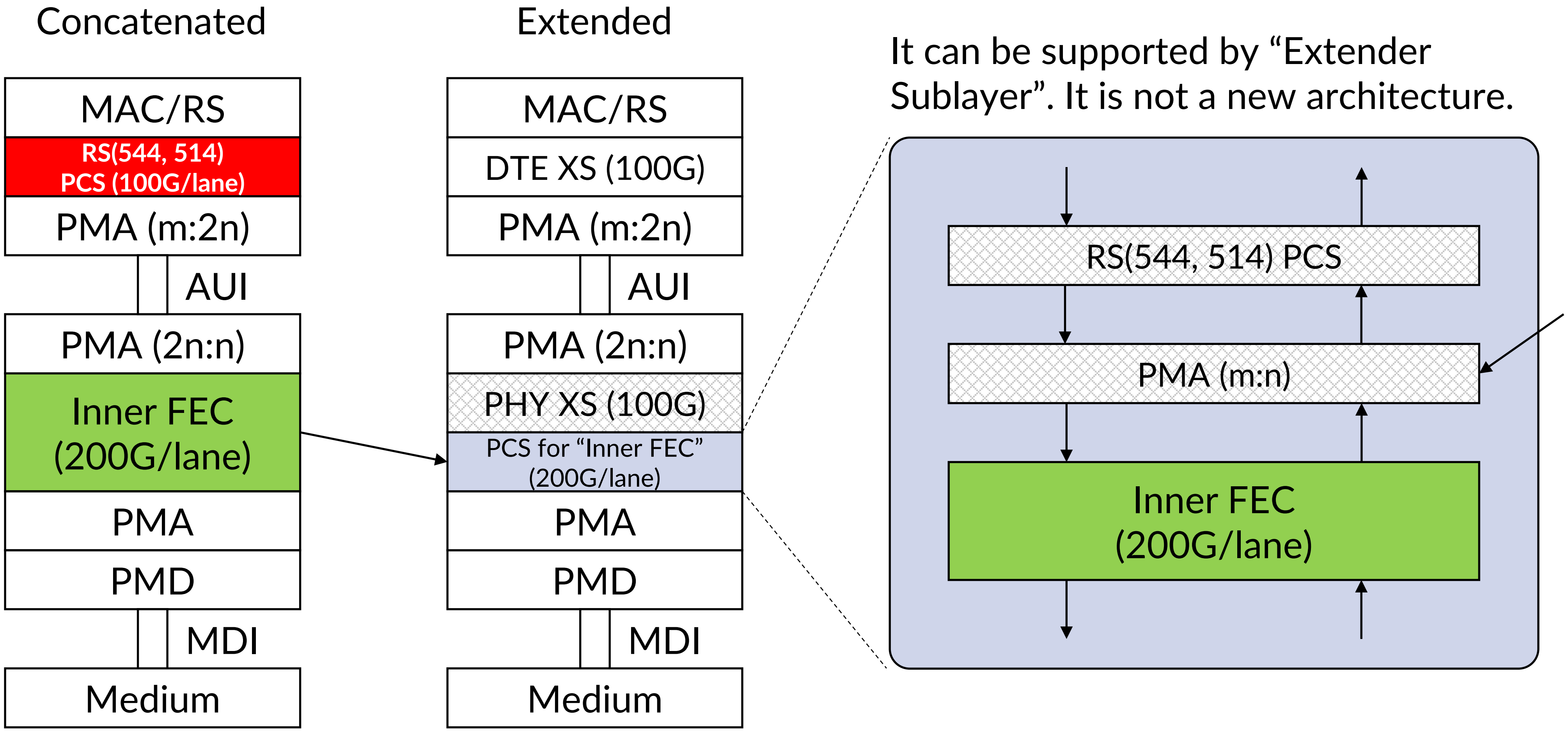
Bottom half of 200G/lane PCS

Questions for Concatenated FEC architecture



- It can be supported by "Extender Sublayer", It is not a new architecture.
- What is the detailed functionality of "Inner FEC" sublayer?
  - On the transmitter side, take the PMA bit stream per-lane and slice into data blocks of "inner FEC" payload size, and encoded as inner FEC code word. The "inner FEC" is per-PMD lane based.
  - On the receiver side, take the per-PMD lane based inner FEC stream, decode the inner FEC code, strip the inner FEC parity and recover the "bit-muxed" PMA bit stream lanes.
- Does it require RS(544, 514) FEC to cover all the 200G AUIs?
  - RS(544, 514) need to cover the 200G AUIs and PMDs end-to-end for 200G/lane native applications.
- What would be the instantiation on host PCS of this architecture?
  - Soft-decision FEC cannot be implemented in the host ASIC.
- How to address the performance concern?
  - The **net coding gain improvement is negligible or even negative** compared RS(544, 514) under burst channels (lu_3df_logic_220425, page 9).
  - Soft-decision FEC is not compatible with DFE & MLSE, it is ~~probably~~ the worst in performance (lu_3df_01b_220215, page 6~8).
- How to address the reliability concern (MTTFPA issue)?
  - The inner FEC decoding failure generates **burst errors** (lu_3df_logic_220425, page 12). The parity bits of "inner FEC" are wasted, the error detection capability is not improved. **Random error model for MTTFPA calculation is not available** and it is over optimistic, a new model account for burst errors is required.
  - **Reliable inner FEC frame alignment with "2e-3" channels, without the alignment markers is challenging**? The "bit-muxed" PMA streams do not have alignment markers.
- How to address the CDR complexity concern over the bit-transparent CDR scheme?
  - The rates of client side and line side are different, **it is overwhelmed to introduce one extra PLL per direction** to support new frequency point. It is far more complicated than the bit-transparent CDR.

Concatenated FEC can be supported by "Extender Sublayer", it is not a new architecture



It can be supported by "Extender Sublayer". It is not a new architecture.

It looks more reasonable if the PMA is removed.

Summary

- **End-to-end FEC architecture is optimal for 200G AUIs & 200G PMDs and cascaded multiple AUIs.**
  - FEC processing inside CDR means integration of two back-to-back full Ethernet PCS stacks inside the CDR. The rate difference between the client side and line side requires extra PLLs.
  - Ethernet PCS/FEC crosses multiple physical lanes. Only bit-transparent CDR can support low cost flexible breakout.
  - Choose the end-to-end FEC according to the worst segment of the AUI and PMD links.
- **Segmented FEC architecture is optimal for 100G AUIs & 200G PMDs.**
  - Can be supported by “Extended Sublayer” or “Inverse FEC Sublayer”.
  - “Extended Sublayer” is flexible, has editorial benefit, but lack of technical insights and details.
  - “Inverse FEC Sublayer” is intuitive, simple and clear.
- **Extended Sublayer is essential for 100G/200G AUIs & ZR PMDs**, due to the different clock drift of AUIs (+/-100ppm) and ZR PMDs (+/- 20ppm). It is different from the “Inverse FEC Sublayer” idea.

- **A lot of investigations and clarifications are required for “Concatenated FEC”.**

Recommendation

- Adopt “End-to-end FEC” as a baseline architecture (support bit-transparent CDR and flexible CDR breakout) for the mainstream applications, i.e. “100G AUIs & 100G PMDs” and “200G AUIs & 200G PMDs” to achieve low cost, low power consumption, low latency and flexibility (CDR breakout).

- Adopt the “Segmented FEC/Extended Sublayer” as baseline architecture for the provisional applications, i.e. “100G AUIs & 200G PMDs” . Adopt Extended Sublayer for ZR applications to compensate the clock drift difference between AUIs (+/-100ppm) and ZR PMDs (+/- 20ppm).
  - Support multiple FEC code schemes, e.g. RS(N, K), ZR FEC, etc.
  - The concatenated FEC can be supported by the “Segmented/Extended” architecture.

Q & A