4x RS Codeword Interleaving Proposal for 200 GbE and 400 GbE

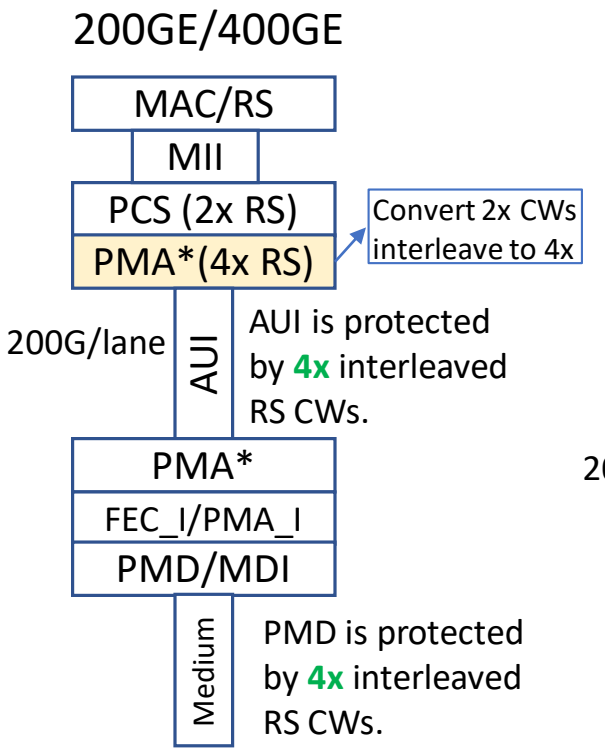
(Update July 2023)

Xiang He (Huawei), Matt Brown (Alphawave), Adee Ran(Cisco), Eugene Opsasnick (Broadcom)

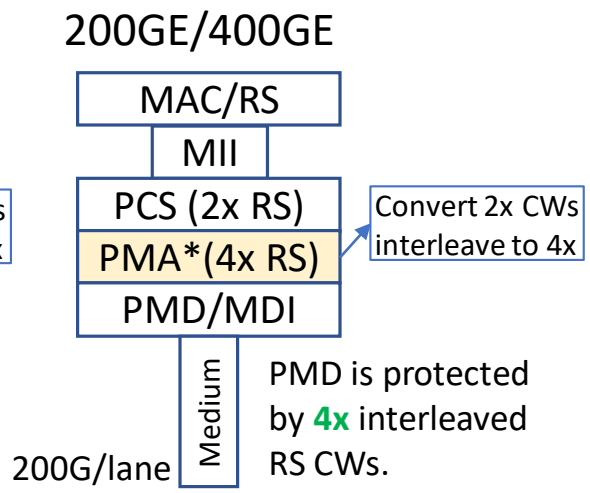
Introduction

- P802.3dj covers Ethernet rates for 200GE, 400GE, 800GE and 1.6TE.
  - Each rate needs to support 200Gb/s per lane AUIs and PMDs.
- All PCS designs for the above rates have been determined in the Task Force.
  - 200GE and 400GE both interleave 2x RS codewords in the PCS.
  - 800GE and 1.6TE both interleave 4x RS codewords in the PCS/PMA.
- Symbol-pair muxing PMA has been adopted for 200G/lane signaling.
  - Please see ran_3dj_01a_2303.
- This presentation analyzes performance differences due to interleaving depth of RS codewords for Type 1 and Type 2 PHY schemes, with recently adopted items in May interim:
  - RS(544,514) FEC has been adopted as the only FEC for 200G/lane CR/KR PMDs.
  - DER0 value of 2.67e-5 has been adopted as the total allocation for higher-loss AUIs within a PHY.
- Independent analysis in ran_3dj_elec_01_230420 and ran_3dj_logic_01_230629.pdf also showed performance difference of 2/4 RS codeword interleaving, with or without inner FEC.
- Straw poll in May interim showed good support for 4x RS codeword interleaving for all rates.

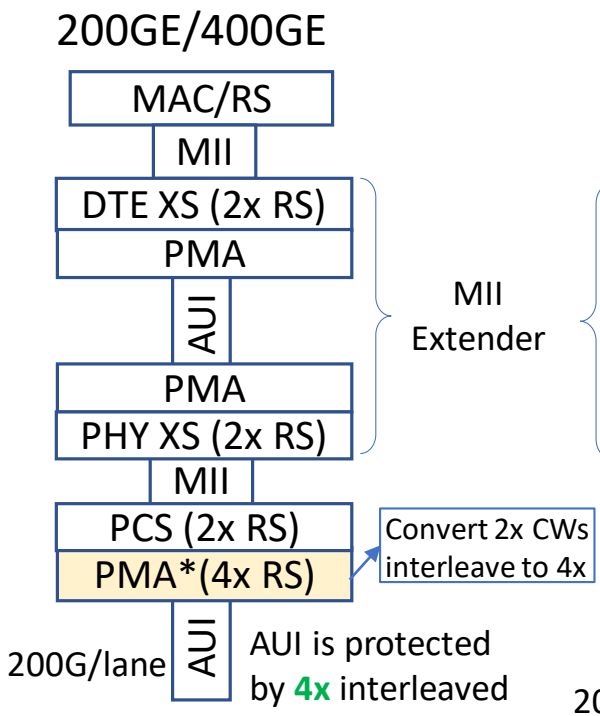
Architecture Overview



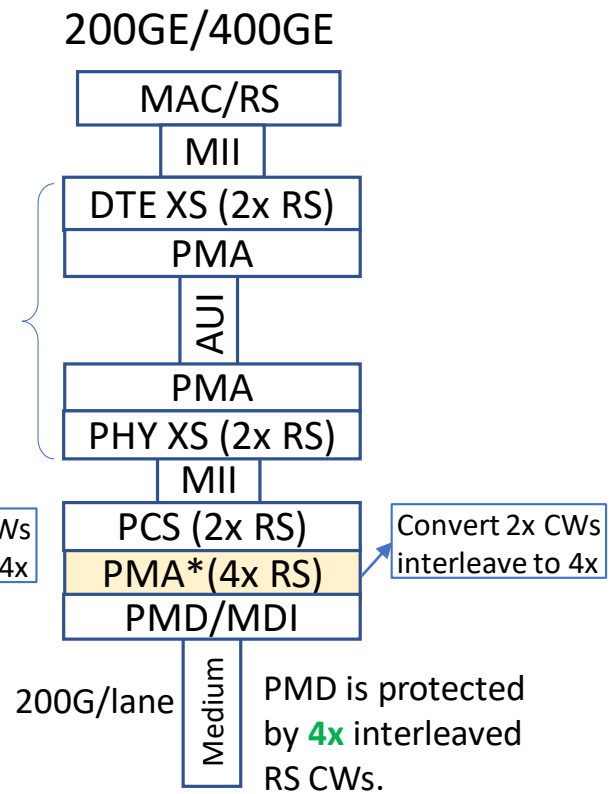
200G/Lane AUI + PMD



200G/Lane PMD w/o AUI



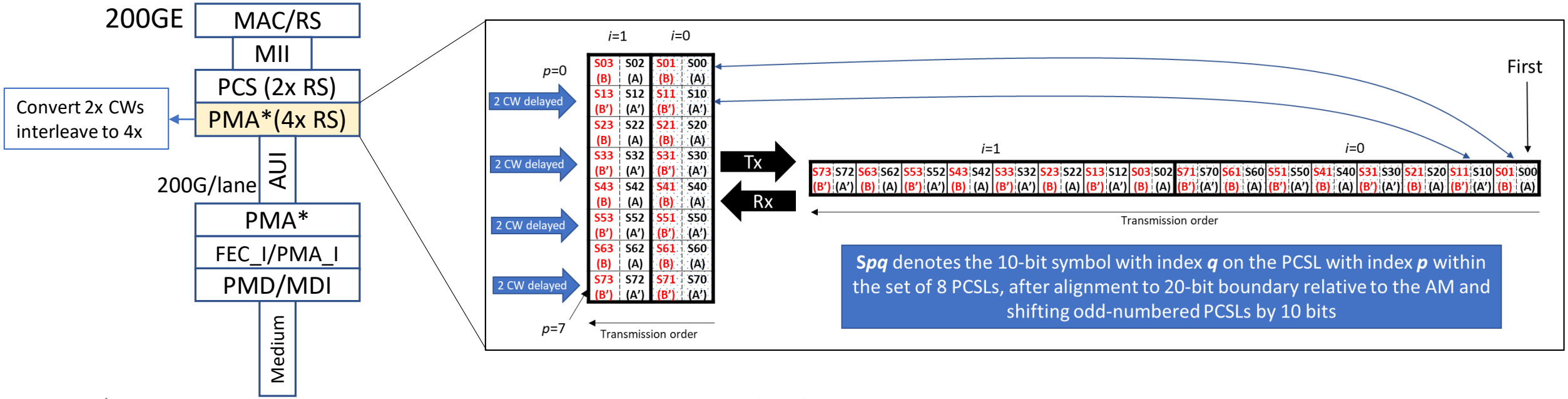
200G/Lane AUI + PMD + MII Extender



200G/Lane PMD w/o AUI + MII Extender

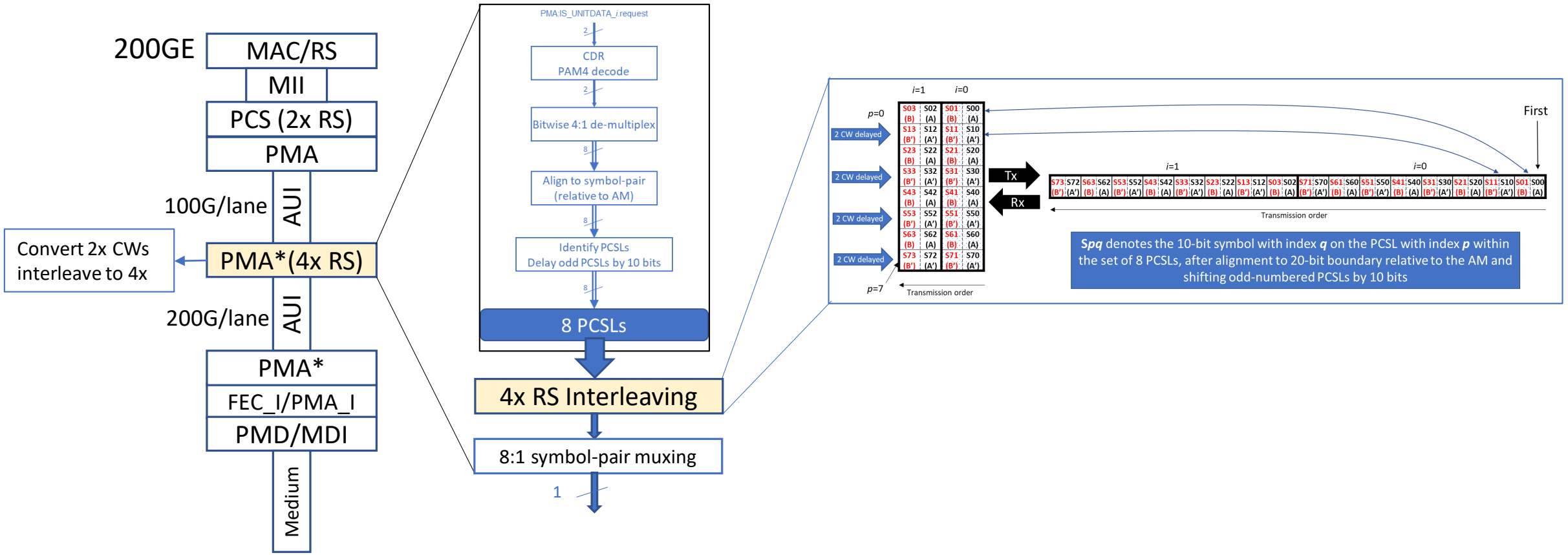
Proposed Method – Delay Half of PCS Lanes

- On TX side, PCS lanes with odd index ("*p*" as shown below) are delayed by 2 codewords.
  - For 200GE, each odd index PCS lane is delayed by 10880/8 = 1360 bits, on top of the 10-bit offset by symbol-pair mux.
  - For 400GE, each odd index PCS lane is delayed by 10880/16 = 680 bits, on top of the 10-bit offset by symbol-pair mux.
  - PMA performs symbol-pair muxing between even and odd lanes to form 4x RS CWs interleaving pattern (ABA'B' as shown below).
- In RX side PMA, PCS lanes with even index are delayed by 2 codewords (1360b for 200GE and 680b for 400GE).
  - If this reverse function is not done in the PMA, the RX PCS needs to tolerate this additional skew.
- Total delay penalty is 2 codewords for TX + RX.



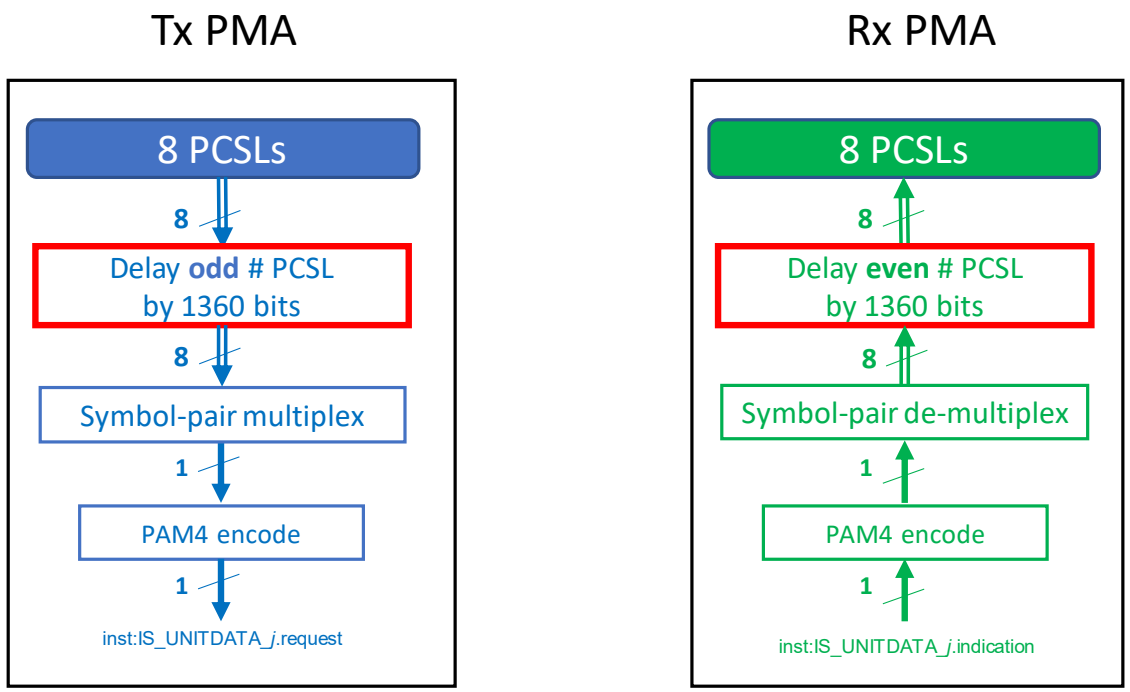
Proposed Method – Interop with 100G/Lane AUI

- When interoperating with 100G/lane AUIs, the 2x to 4x RS CWs conversion can be done within the PMA symbol-pair muxing function.
  - "PCS lane number aware" is required to ensure that TX PMA always delay odd lanes, and RX PMA always delay even lanes.
  - 8:1 muxing symbol-pair muxing must interleave even-odd-even-odd-... PCS lanes to get the 4x CWs striping.

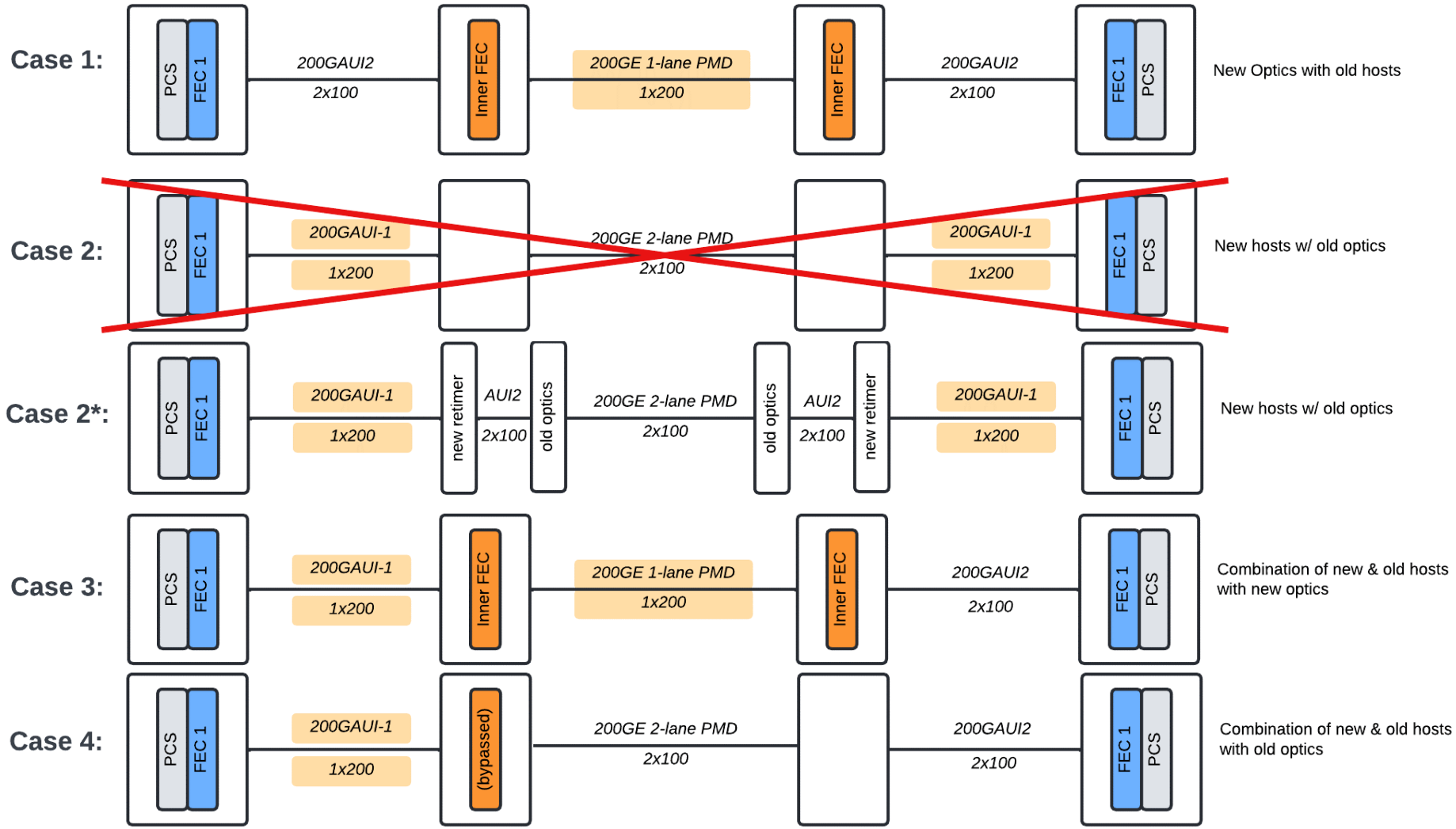


Tx and Rx Processing Illustration

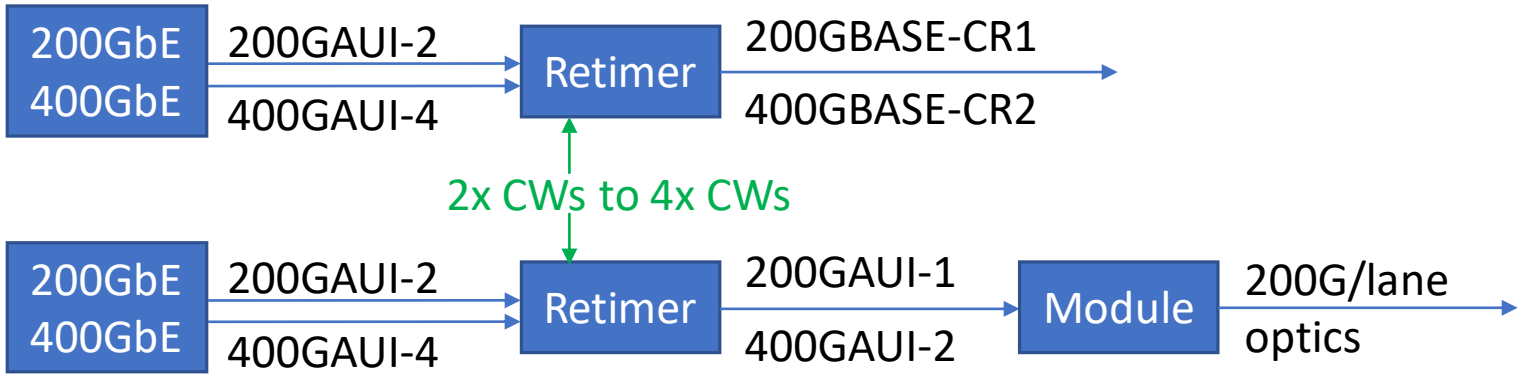
- The interleaving depth change should be converted back to 2x RS CWs before reaching PCS.
  - Reverse process is recommended to be performed in the PMA adjacent to PCS.
- Tx and Rx processing diagrams are shown below, using 200GE as an example.



Possible Interoperation Schemes with Optical PMDs



Skew Impact on FEC Performance



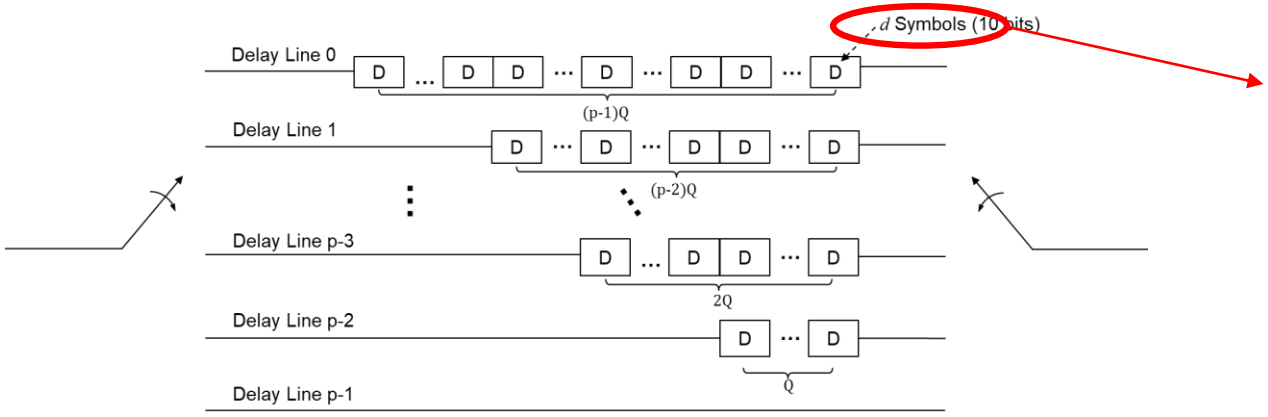
- Interop schemes between 100G/lane AUIs and 200G/lane AUIs could complicate things.
- Skew between the 100G AUI lanes could lead to <4 CWs interleaving, even back to 2xCWs.
  - Taking a 400GbE port as an example, assuming one AUI is from PCSL(0,2,4,6) and has +25.6ns skew comparing to its paring AUI from PCSL(1,3,5,7). Delaying the odd lanes by 10880/16 = 680 bits will lead to the original 2xCWs interleaving.
- If skew in actual implementation is much smaller than 25.6ns, the performance will be comparable to a true 4xCWs interleaving.

Impact on FEC_I Sublayer

- The total latency increased by this 4x RS CWs interleaver will lower the convolutional interleaver latency, resulting the same total end-to-end latency when FEC_I sublayer is used with CI.

Rate	RS CWs in PMA	d	P	Q	CI Depth	CI Latency, ns	PMA interleaver latency, ns	Total Latency, ns
400GE	4	4	3	96	12x RS	108.4	25.6	133.0
200GE	4	4	3	192	12x RS	216.8	51.2	268.0
400GE	2	2	6	48	12x RS	135.5	0	135.5
200GE	2	2	6	96	12x RS	271.1	0	271.1

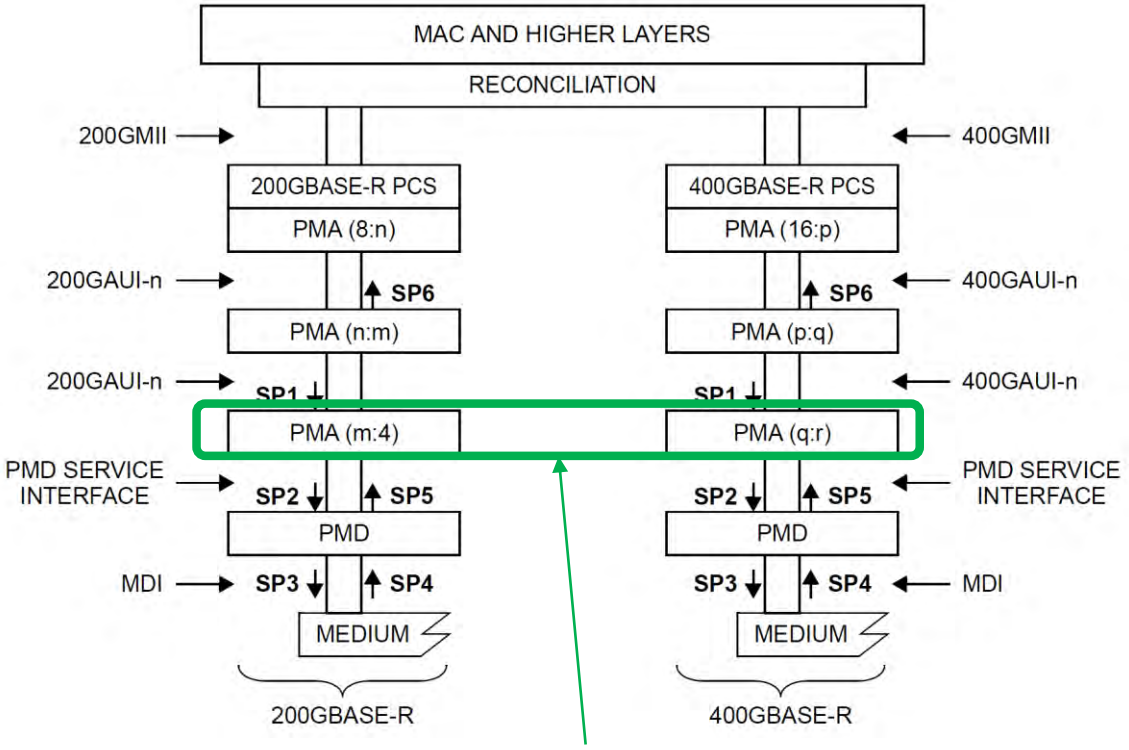
- When using 4xRS CW interleaver, 200GE and 400GE can have the same distribution method as 800GE and 1.6TE in FEC_I sublayer, in 40b blocks instead of 20b.



Rate	RS CWs in PMA	d
1.6TE	**4**	**4**
800GE	**4**	**4**
400GE	**4**	**4**
200GE	**4**	**4**
400GE	2	2
200GE	2	2

Impact on Skew Constraints

- The intentional skew offset added to the PMA sublayer could affect all skew constraints between SP2 and SP5 has very predictable behavior.
  - The artificially added skew can be treated differently from unpredictable skew from the physical interface or logic.
  - If both 200G/lane C2C and C2M are used, SP1 can also be affected.
- Proposed solutions:
  - Redefine the skew constraints in each sublayer to exclude the intentional skew offset.
  - Add a note to skew constraint tables for each sublayer:
    "Skew constraints do not include intentional skew offset to increase the RS codewords interleaving depth as described in <*subclause*>"
- The skew offset can be cancelled by the receiving side PMA, it will not impact the PCS skew constraints.



Intentional skew offset added on the Tx and cancelled on the Rx.

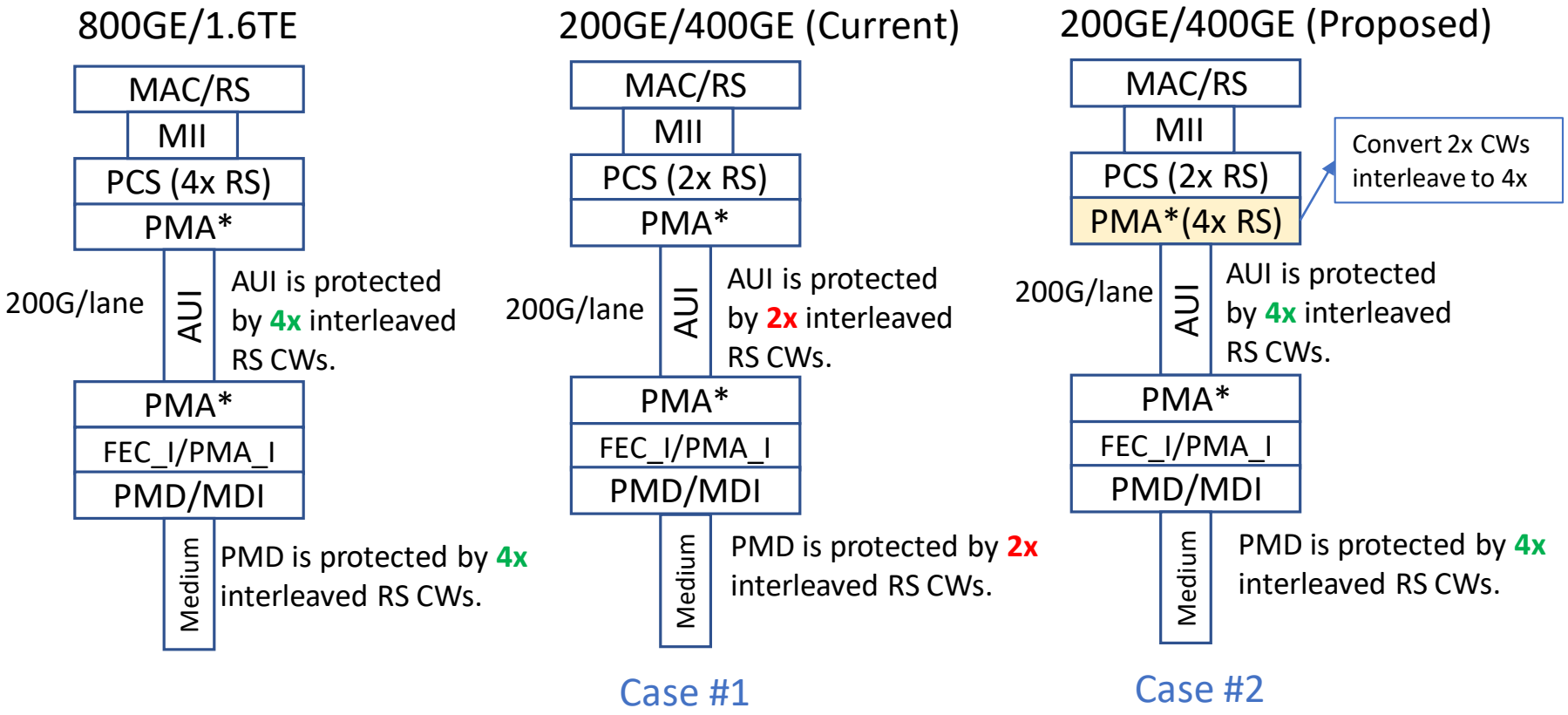
Summary

- Increasing interleaving depth from 2x codewords to 4x for 200GE and 400GE could improve overall coding gain by 0.4 – 0.55dB, if total AUI DER0 is 2.67E-5 on each side.
  - 2x codewords interleaving on worst case AUI leads to error floor around 1E-11 FLR.
  - Increasing the interleaving depth in the host (or the first PMA below the MII Extender) is preferred.
- The proposed method could provide 4x codewords interleaving in the PMA, with a total (Tx+Rx) added latency of 2 codewords.
  - Latency impact is 51.2ns for 200GE, and 25.6ns for 400GE.
- Skew constraints per sublayer may need to be redefined to exclude the intentional skew offset.

Thank you!

Back Up Slides

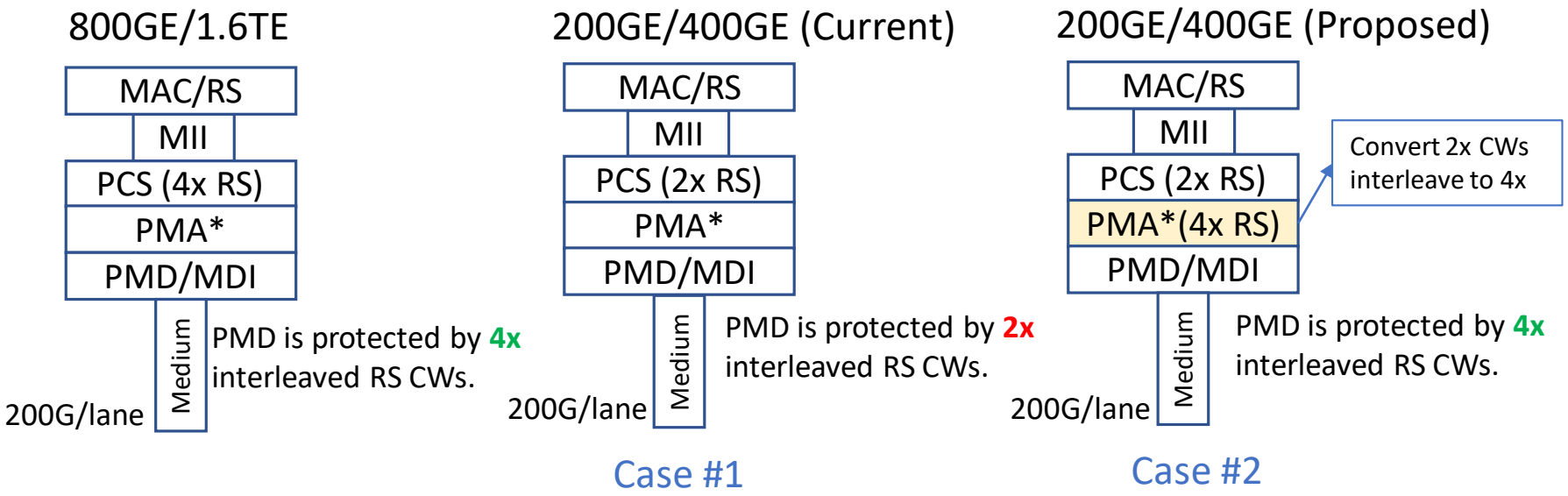
Interleaving Depth vs AUI Burst Protection



*Symbol-pair muxing PMA as in ran_3dj_01a_2303.

- Assuming same AUI BER threshold for all Ethernet rates using 200G/lane AUIs, the current 200GE/400GE will have lower performance due to 2x RS CWs interleaving compared to 4x of 800GE/1.6TE.

Interleaving Depth vs PMD Burst Protection (Type 1 PHY/FEC)

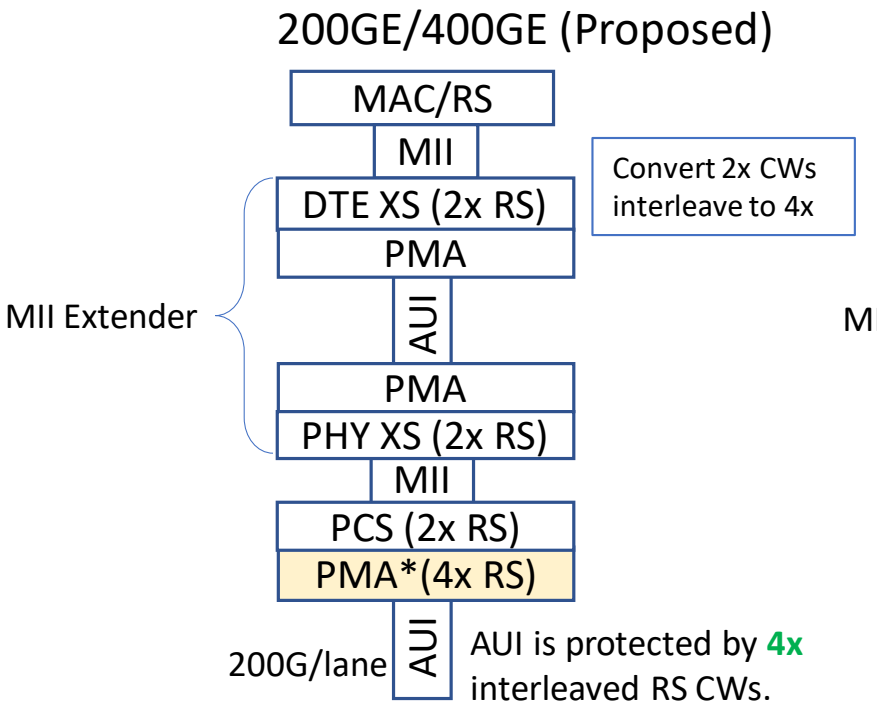


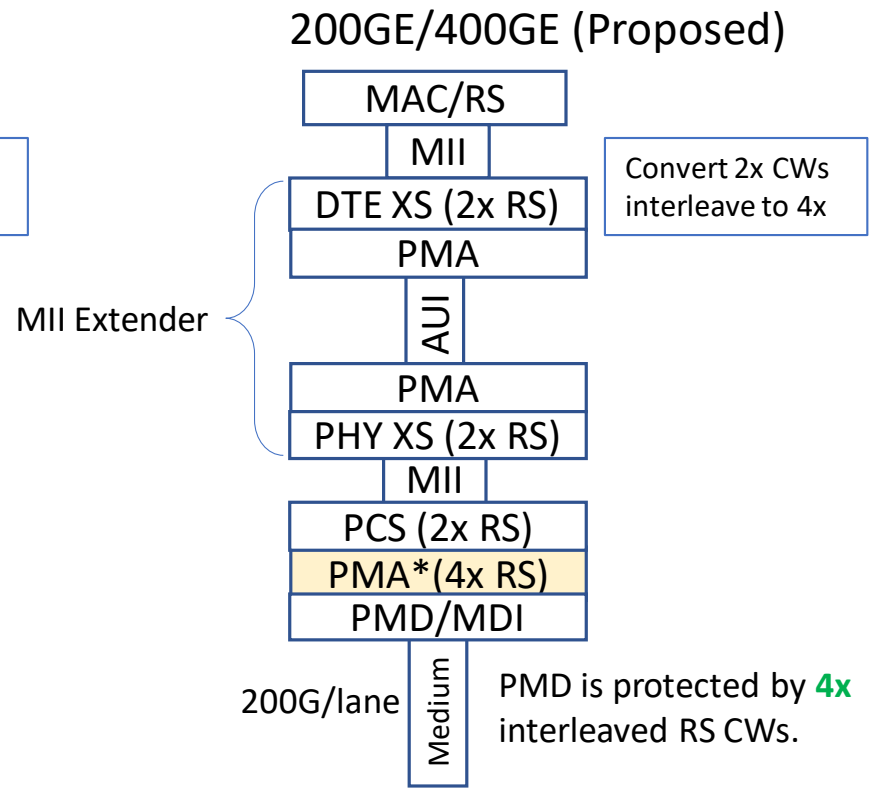
*Symbol-pair muxing PMA as in ran_3dj_01a_2303.

- For Type 1 PHY/FEC scheme, we used the simple Type 1 PHY with no AUI, but the simulation results could cover Type 1 PHY with 1 AUI to represent CR/KR applications with a C2C AUI.

Cases with MII Extenders (Type 2 PHY/FEC)

- MII Extenders terminate RS FEC and clears errors in the AUI above it.
- If the AUI or PMD directly below the MII Extender is 200G/lane, it is recommended to implement the 4x RS interleaver in the PMA below the MII Extender.

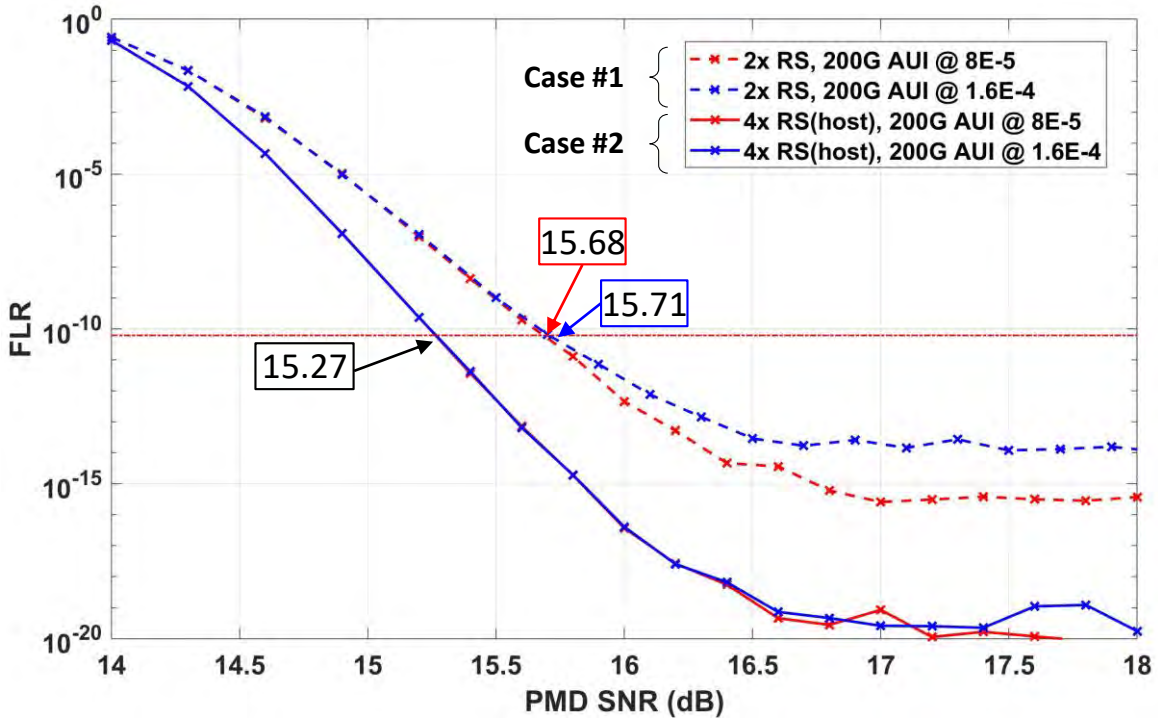




Two-part Link Simulation for 2x and 4x RS Codewords Interleave

- Case #1 (dotted lines): 2x RS
- Case #2 (solid lines): 4x RS(host)
- Increasing the interleaving to 4x RS codewords on the AUIs (Case # 2) is clearly beneficial.
  - The AUI electrical specs and/or the PMD's optical specs will be driven by the worst case.
  - It would be preferable if we can avoid burdening the AUI and PMD for the older rates.
- Latency will be increased but is reasonable as shown in brown_3dj_elec_01_230420.

Case #1 vs Case #2



- Simulation configuration:

AUI: 200G/lane, symbol-pair muxing. **DER0 = 2.67E-5 per PHY.**
Error propagation probability "a" = 0.75, pre-coding ON **and OFF**.
All BER values include~~s~~ additional errors due to bursts and precoding.

PMD: Pure AWGN.

FEC_I: Hamming(128,120), w/o convolutional interleaver.

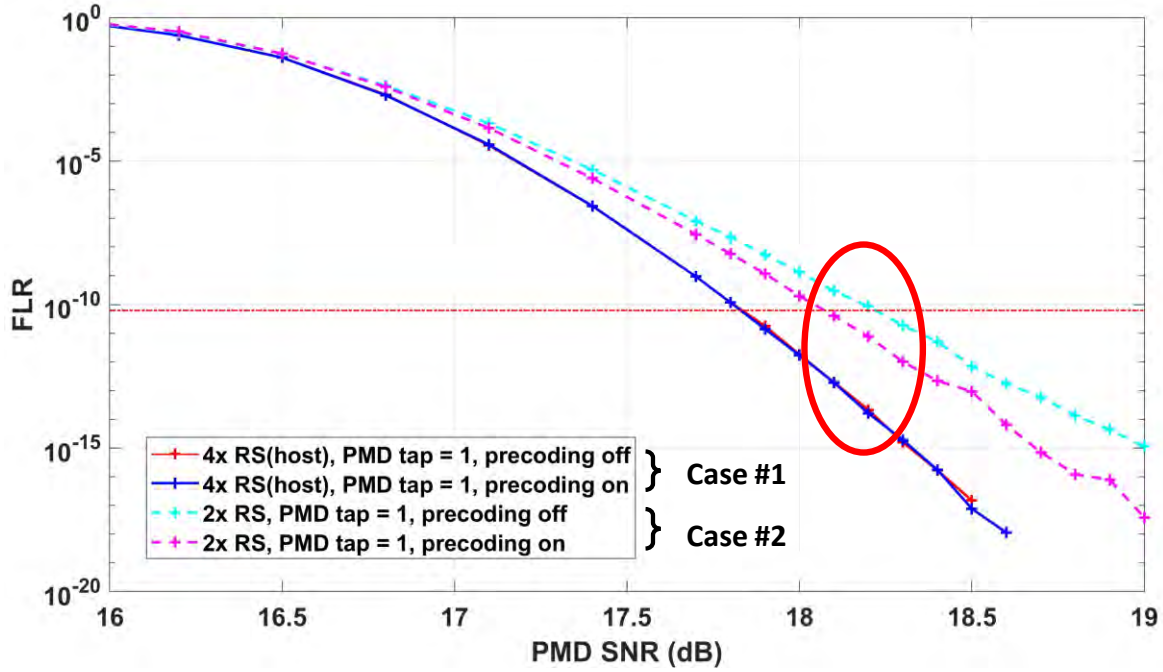
Precoding	AUI BER (per PHY)	Total AUI BER	PMD BER Threshold	
			Case #1	Case #2
ON	4E-5	8E-5	2.45E-3	3.55E-3
OFF	8E-5	1.6E-4	2.38E-3	3.55E-3

One-part Link Simulation for 2x and 4x RS Codewords Interleave

- Case #1 (dotted lines): 2x RS
- Case #2 (solid lines): 4x RS(host)

- Precoding is not needed if 4x RS codewords interleaving is used.
- SNR can be improved by ~0.4 dB comparing to 2x RS options.
  - FLR can be improved by **>5 orders** in the region of interest.

Case #1 vs Case #2



- Simulation configuration:

PMD: 200G/lane, symbol-pair muxing.
Error propagation probability "a" = 0.75, pre-coding ON and OFF.
All BER values include additional errors due to bursts and precoding.

FEC: Type 1 with RS(544,514) as the only FEC.

Precoding	PMD BER Threshold		PMD SNR Threshold	
	Case #1	Case #2	Case #1	Case #2
ON	2.53E-4	3.66E-4	18.08	17.84
OFF	3.97E-4	7.31E-4	18.23	