800GbE PCS/FEC/PMA Baseline Proposal for PHYs using 8 x 100G PMD lanes - *Update*

Kapil Shrikhande (Marvell), Eugene Opsasnick (Broadcom), Gary Nicholl (Cisco),
David Ofelt (Juniper), Eric Maniloff (Ciena), Shawn Nicholl (AMD), Jeff Slavick (Broadcom)

July 12, 2022

IEEE 802.3df Plenary meeting, July 2022

Supporters

- Rob Stone, Meta
- Brad Booth, Microsoft
- Kent Lusted, Intel
- Brian Welch, Cisco
- Lenin Patra, Marvell
- Venu Balasubramonian, Marvell
- Piers Dawe, Nvidia
- Bill Simms, Nvidia
- Arthur Marris, Cadence
- Liav Ben Artsi, Marvell
- Jerry Pepper, Keysight
- Chris Cole, Quintessent
- Ted Sprague, Infinera
- Dave Estes, Spirent
- Adee Ran, Cisco
- Chris DiMinico, PHY-SI/SenTekse
- Ben Jones, AMD
- Jeffery Maki, Juniper Networks
- Ali Ghiasi, Ghiasi Quantum LLC
- Paul Brooks, Viavi Solutions
- Nathan Tracy, TE Connectivity

This Talk

- Review updates to the Baseline
- Summary of work since May’22 interim

Outline

- **Introduction**
- PCS/FEC/PMA Baseline proposal
- Implementation considerations
- Architecture considerations
- Summary of work since May’22 interim
- Conclusions

Goals

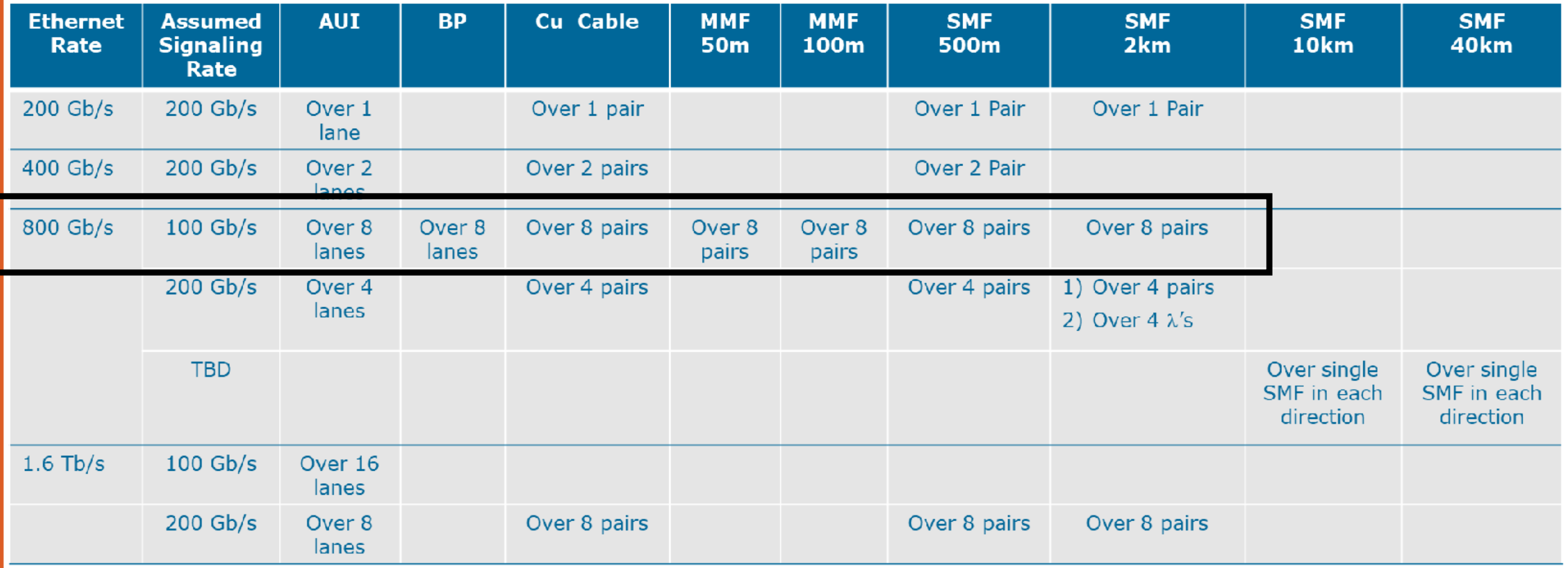
- Fast time to an 800GbE PCS/FEC/PMA specification for PMDs using 100G/lane
  - Re-use 400GbE PCS/FEC (CL119) as much as possible
  - Support 800GbE with simple modification to the 400GbE PCS/FEC
  - Leverage 802.3bs Cl120 PMA; leverage 802.3ck 100G/lane PMA and AUI specifications

- Maximize the re-use of existing logic sub-blocks used in 400GbE PCS/FEC
  - Leverage industry investment in 400GbE technology

- Enable systems using current 8-lane 800G connectors (OSFP / QSFP-DD) to also support 800GbE
  - E.g. 8-lane C2M AUIs used as : 8 x 100GAUI-1 / 4 x 200GAUI-2 / 2 x 400GAUI-4 <u>and</u> 1 x 800GAUI-8

Scope

802.3df Adopted PHY Objectives*



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		

Making it all work together

Technology Reuse

Leverage existing or work-in-progress 100 Gb/s per lane (e.g. 3cu, 3ck, 3db) to higher lane counts

Develop 200 Gb/s per lane electrical signaling for 1/2/4/8 lane variants of AUIs and electrical PMDs

Develop 200 Gb/s per optical fiber for 1/2/4/8 fiber based optical PMDs and 4 lambda WDM optical PMD

Potential for either direct detect and / or coherent signaling technology

Scope of this Baseline : 800GbE PCS/FEC/PMA for all PHY objectives that use 8 x 100G PMDs and AUIs

* *Table from https://www.ieee802.org/3/B400G/public/21_1028/B400G_overview_c_211028.pdf*

AUI and PMD assumptions

- 802.3df Task Force has adopted 800GbE 8-lane AUI baseline proposals leveraging existing 100G/lane AUI specs, drafts
  - https://www.ieee802.org/3/df/public/22_03/lusted_3df_01a_220315.pdf

- 802.3df Task Force has adopted 800GbE 8-lane PMD baseline proposals leveraging existing 100G/lane PMD specs, drafts
  - https://www.ieee802.org/3/df/public/22_03/lusted_3df_01a_220315.pdf
  - https://www.ieee802.org/3/df/public/22_02/welch_3df_01a_220222.pdf
  - https://www.ieee802.org/3/df/public/22_03/murty_3df_01a_220315.pdf

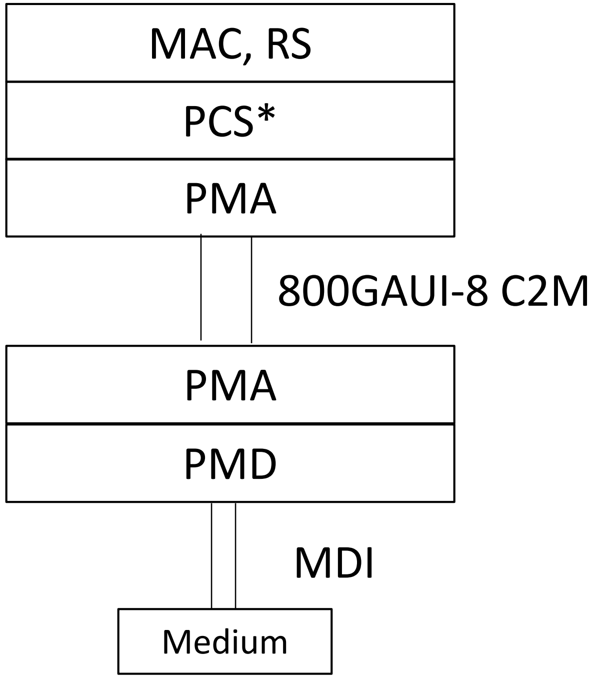
- 802.3bs CL119 PCS works for all 100G/lane AUIs and PMDs for 400GbE

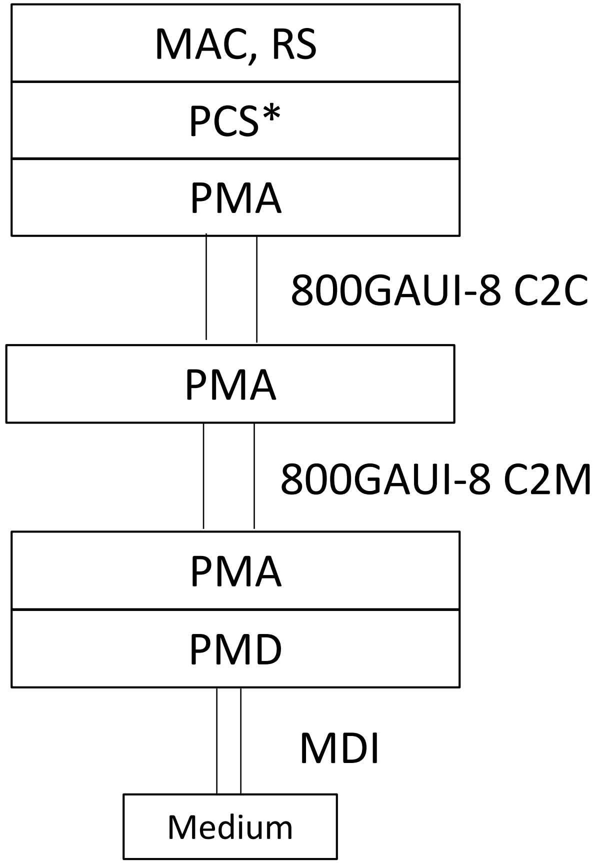
- Similarly, this PCS/FEC Baseline (leveraging CL119) works for all adopted 800GbE 8-lane AUIs and PMDs

Outline

- Introduction
- **PCS/FEC/PMA Baseline proposal**
- Implementation considerations
- Architecture considerations
- Summary of work since May’22 interim
- Conclusions

Architecture

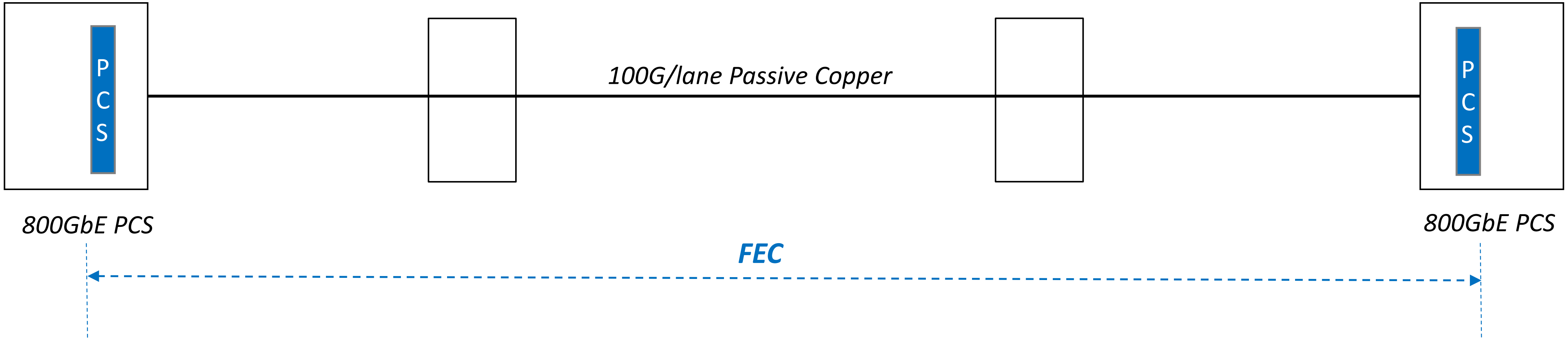


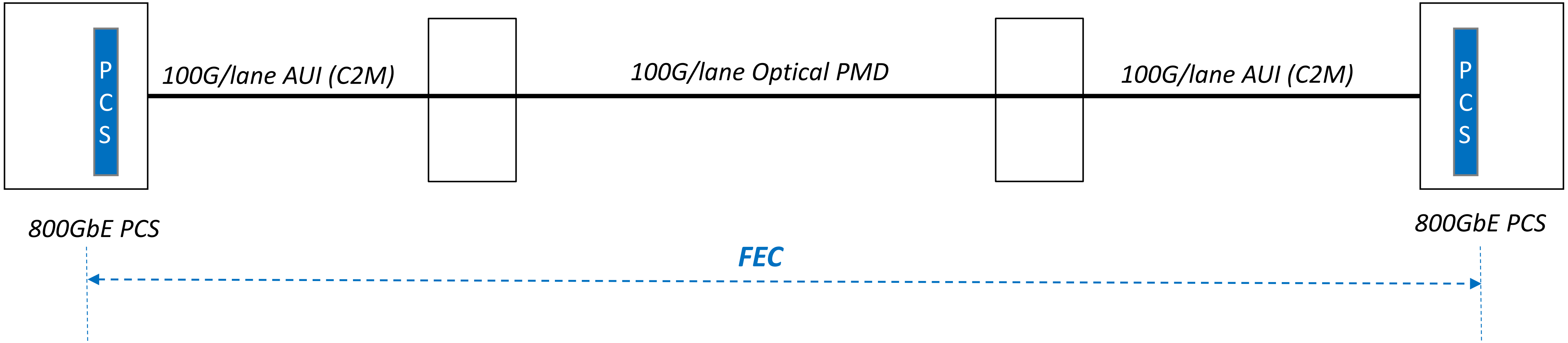


***** *PCS and FEC are in the PCS sub-layer (same as CL119)*

Note : Not showing layering diagram for Cu PMD (will be same as other Cu PMD layering diagrams in 802.3)

End-End PCS/FEC scheme for 800GbE (8 x 100G) PMDs

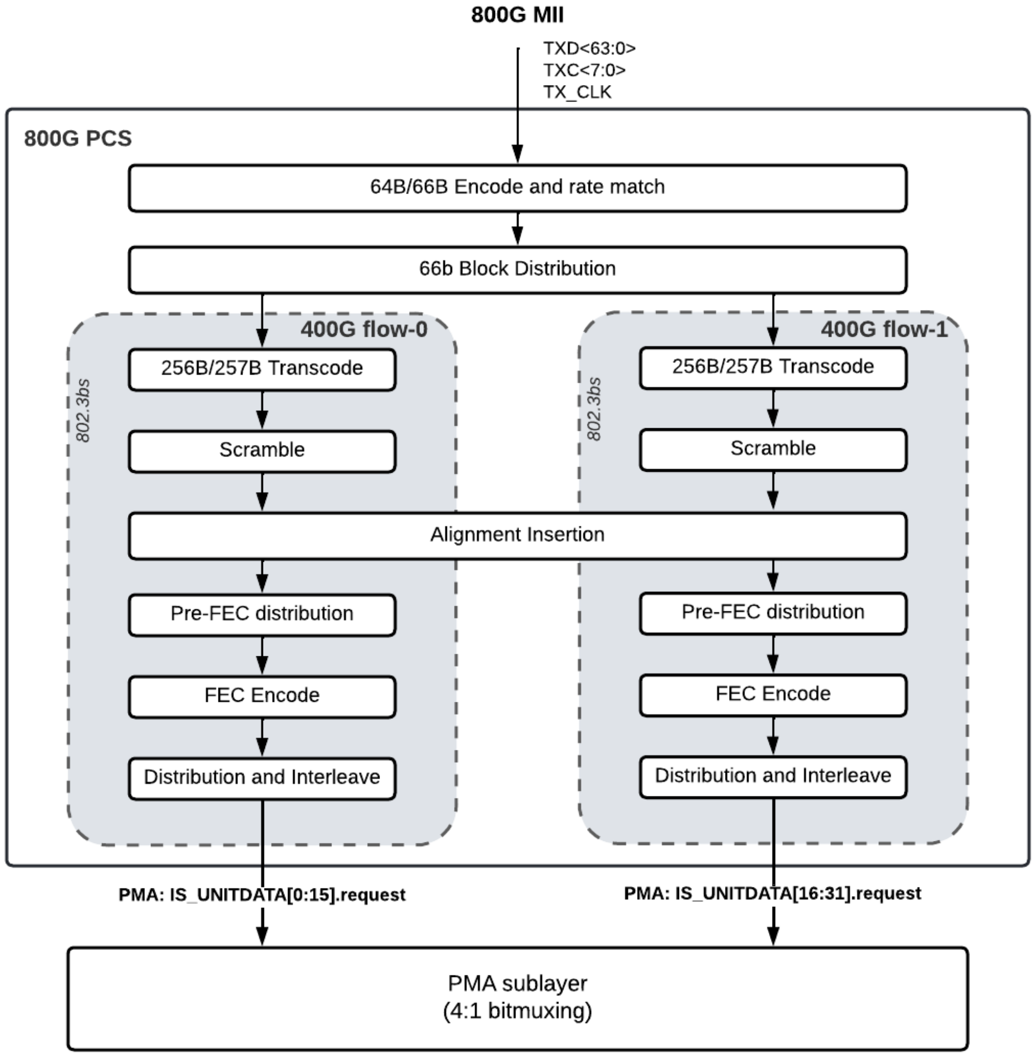




Note : This End-End PCS/FEC works with optional Chip to Chip AUIs and a combination of Chip to chip and Chip to module *(same as 400GAUI-4 in 802.3ck)*

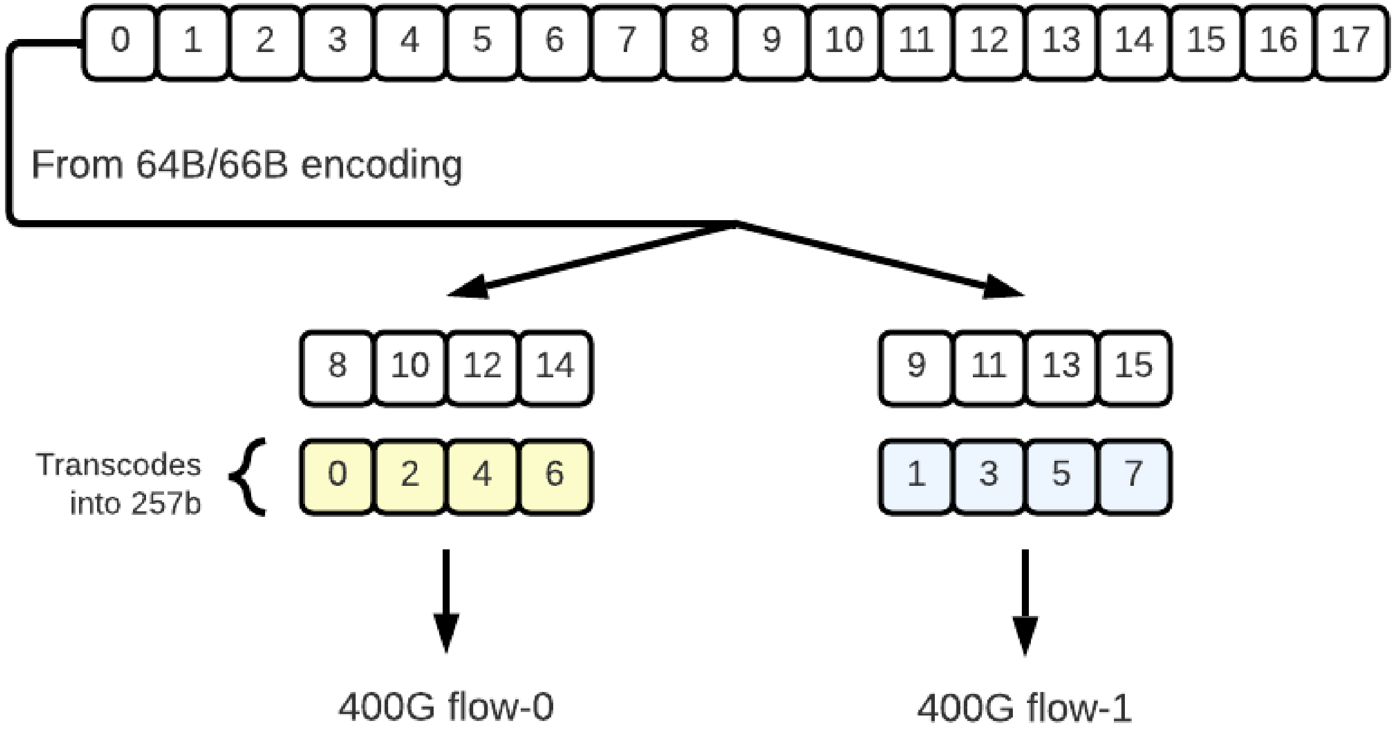
Tx PCS/FEC Data Flow

- Based on two 802.3bs, CL119 sub-layers in parallel
  - Two 400G FEC flows (flow-0 and flow-1)
- 66b round robin distribution into two 400G flows after 64B/66B encode
- Sub-blocks shown within each flow are identical to CL119, except :
  - AM values are made unique across the two flows
  - AM insertion is aligned across the two flows
- 32 PCS lanes per 800GbE PCS
  - 16 PCS lanes per 400G flow
- Any 4 PCS lanes to any PMA output lane
  - 4:1 bitmuxing



Tx 66b Block Distribution

- Round Robin among two ‘400G Flows’



Alignment Marker Insertion

- 802.3bs 400G AM structure
  - AM size = 8 x 257b
  - Spacing = 160k x 257b = 8192 CWs
- AM total sizing for 800G = 2x400G
  - AM size = 16 x 257b
  - Spacing = 320k x 257b = 16384 CWs
- Markers inserted at consecutive 257b blocks across both 400G flows
  - Flow-0 is first in time carrying the even encoded 4x66b blocks
  - Flow-1 carries odd encoded 4x66b blocks

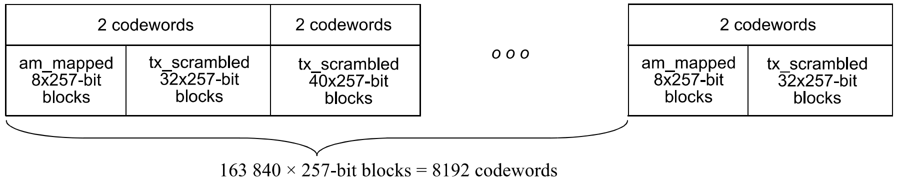
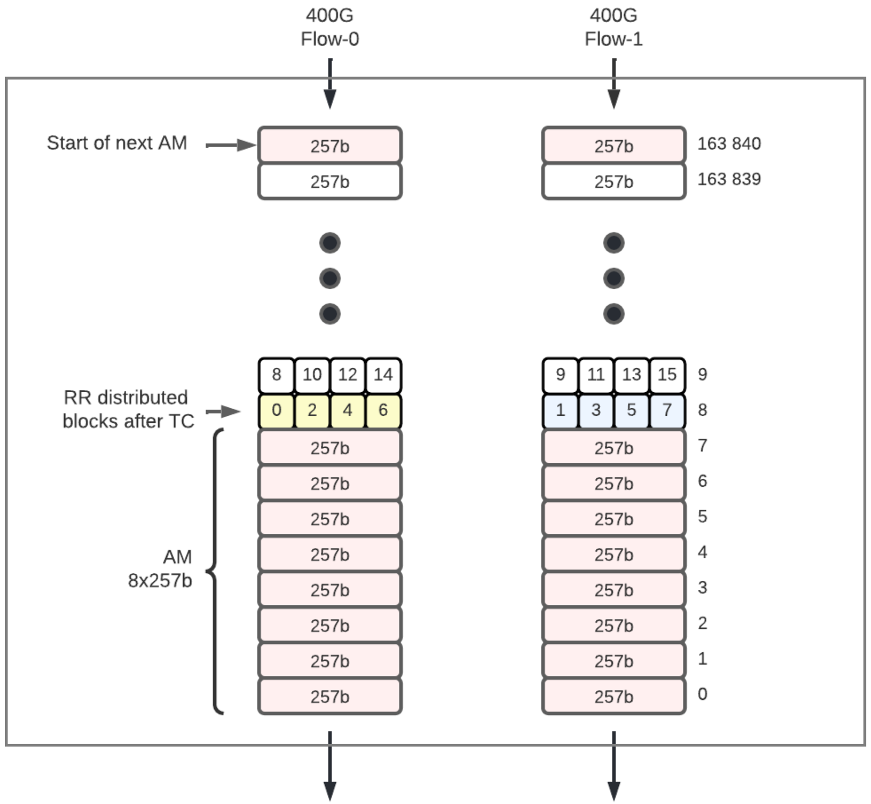

2 codewords
2 codewords
2 codewords
am_mapped 8x257-bit blocks
tx_scrambled 32x257-bit blocks
tx_scrambled 40x257-bit blocks
o o o
am_mapped 8x257-bit blocks
tx_scrambled 32x257-bit blocks
163 840 × 257-bit blocks = 8192 codewords

Figure 119–8—400GBASE-R alignment marker insertion period

Source: IEEE Std 802.3-2018



AM Marker Encoding

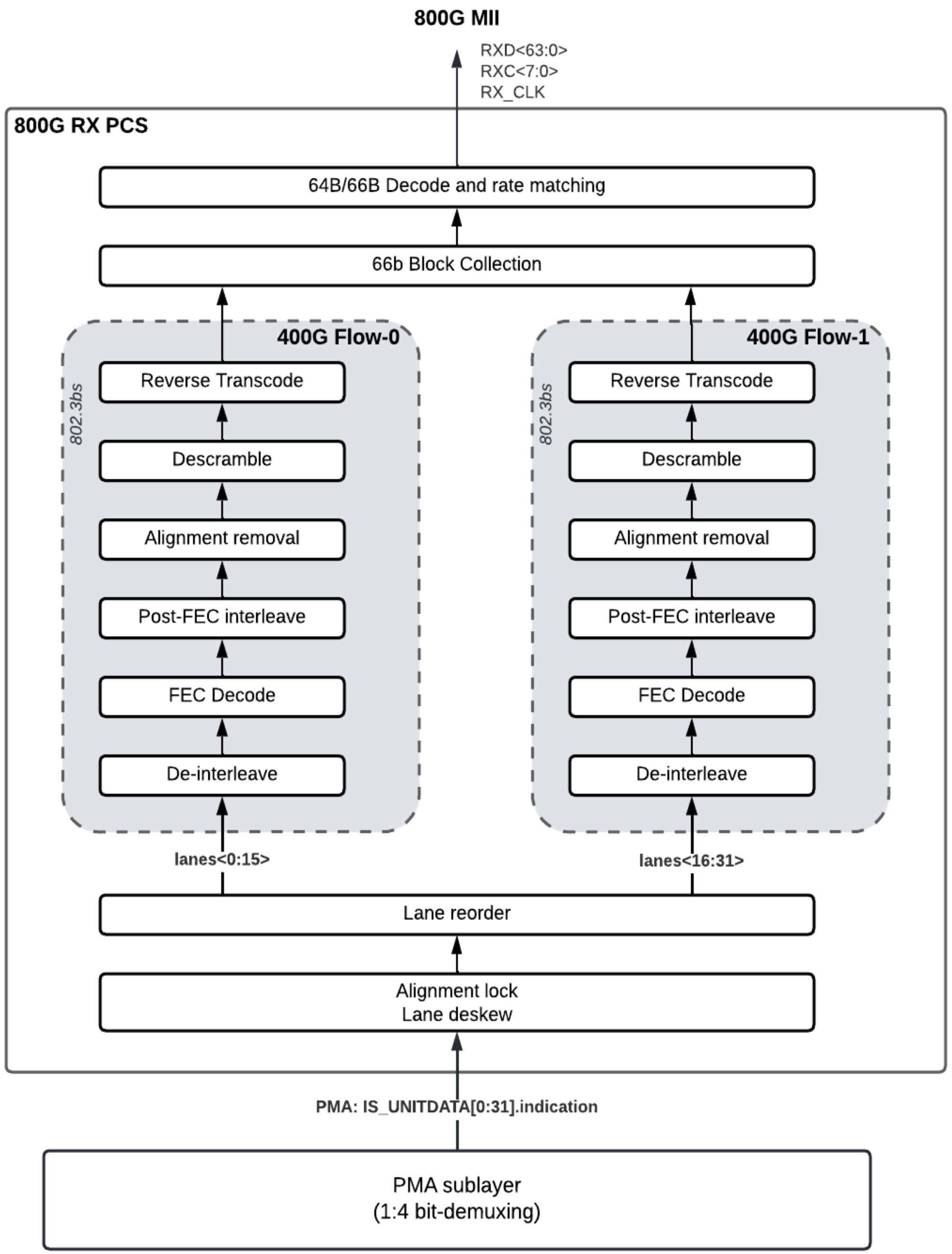
- CM0-CM5 and UP0-UP2 are unchanged from 400GbE CL119
- UM0/UM3 for PCS lanes 0-15 are inverted from 400GbE
- UM1/UM2/UM4/UM5 for PCS lanes 16-31 are inverted from 400GbE
- Prevents lock with 400GbE ports
- Maintains DC balance

PCS Lane #	Encoding														
	CM0	CM1	CM2	UP0	CM3	CM4	CM5	UP1	UM0	UM1	UM2	UP2	UM3	UM4	UM5
0	0x9A	0x4A	0x26	0xB6	0x65	0xB5	0xD9	0xD9	**0xFE**	0x71	0xF3	0x26	**0x01**	0x8E	0x0C
1	0x9A	0x4A	0x26	0x04	0x65	0xB5	0xD9	0x67	**0xA5**	0xDE	0x7E	0x98	**0x5A**	0x21	0x81
2	0x9A	0x4A	0x26	0x46	0x65	0xB5	0xD9	0xFE	**0xC1**	0xF3	0x56	0x01	**0x3E**	0x0C	0xA9
3	0x9A	0x4A	0x26	0x5A	0x65	0xB5	0xD9	0x84	**0x79**	0x80	0xD0	0x7B	**0x86**	0x7F	0x2F
4	0x9A	0x4A	0x26	0xE1	0x65	0xB5	0xD9	0x19	**0xD5**	0x51	0xF2	0xE6	**0x2A**	0xAE	0x0D
5	0x9A	0x4A	0x26	0xF2	0x65	0xB5	0xD9	0x4E	**0xED**	0x4F	0xD1	0xB1	**0x12**	0xB0	0x2E
6	0x9A	0x4A	0x26	0x3D	0x65	0xB5	0xD9	0xEE	**0xBD**	0x9C	0xA1	0x11	**0x42**	0x63	0x5E
7	0x9A	0x4A	0x26	0x22	0x65	0xB5	0xD9	0x32	**0x29**	0x76	0x5B	0xCD	**0xD6**	0x89	0xA4
8	0x9A	0x4A	0x26	0x60	0x65	0xB5	0xD9	0x9F	**0x1E**	0x73	0x75	0x60	**0xE1**	0x8C	0x8A
9	0x9A	0x4A	0x26	0x6B	0x65	0xB5	0xD9	0xA2	**0x8E**	0xC4	0x3C	0x5D	**0x71**	0x3B	0xC3
10	0x9A	0x4A	0x26	0xFA	0x65	0xB5	0xD9	0x04	**0x6A**	0xEB	0xD8	0xFB	**0x95**	0x14	0x27
11	0x9A	0x4A	0x26	0x6C	0x65	0xB5	0xD9	0x71	**0xDD**	0x66	0x38	0x8E	**0x22**	0x99	0xC7
12	0x9A	0x4A	0x26	0x18	0x65	0xB5	0xD9	0x5B	**0x5D**	0xF6	0x95	0xA4	**0xA2**	0x09	0x6A
13	0x9A	0x4A	0x26	0x14	0x65	0xB5	0xD9	0xCC	**0xCE**	0x97	0xC3	0x33	**0x31**	0x68	0x3C
14	0x9A	0x4A	0x26	0xD0	0x65	0xB5	0xD9	0xB1	**0x35**	0xFB	0xA6	0x4E	**0xCA**	0x04	0x59
15	0x9A	0x4A	0x26	0xB4	0x65	0xB5	0xD9	0x56	**0x59**	0xBA	0x79	0xA9	**0xA6**	0x45	0x86
16	0x9A	0x4A	0x26	0xB6	0x65	0xB5	0xD9	0xD9	0x01	**0x8E**	**0x0C**	0x26	0xFE	**0x71**	**0xF3**
17	0x9A	0x4A	0x26	0x04	0x65	0xB5	0xD9	0x67	0x5A	**0x21**	**0x81**	0x98	0xA5	**0xDE**	**0x7E**
18	0x9A	0x4A	0x26	0x46	0x65	0xB5	0xD9	0xFE	0x3E	**0x0C**	**0xA9**	0x01	0xC1	**0xF3**	**0x56**
19	0x9A	0x4A	0x26	0x5A	0x65	0xB5	0xD9	0x84	0x86	**0x7F**	**0x2F**	0x7B	0x79	**0x80**	**0xD0**
20	0x9A	0x4A	0x26	0xE1	0x65	0xB5	0xD9	0x19	0x2A	**0xAE**	**0x0D**	0xE6	0xD5	**0x51**	**0xF2**
21	0x9A	0x4A	0x26	0xF2	0x65	0xB5	0xD9	0x4E	0x12	**0xB0**	**0x2E**	0xB1	0xED	**0x4F**	**0xD1**
22	0x9A	0x4A	0x26	0x3D	0x65	0xB5	0xD9	0xEE	0x42	**0x63**	**0x5E**	0x11	0xBD	**0x9C**	**0xA1**
23	0x9A	0x4A	0x26	0x22	0x65	0xB5	0xD9	0x32	0xD6	**0x89**	**0xA4**	0xCD	0x29	**0x76**	**0x5B**
24	0x9A	0x4A	0x26	0x60	0x65	0xB5	0xD9	0x9F	0xE1	**0x8C**	**0x8A**	0x60	0x1E	**0x73**	**0x75**
25	0x9A	0x4A	0x26	0x6B	0x65	0xB5	0xD9	0xA2	0x71	**0x3B**	**0xC3**	0x5D	0x8E	**0xC4**	**0x3C**
26	0x9A	0x4A	0x26	0xFA	0x65	0xB5	0xD9	0x04	0x95	**0x14**	**0x27**	0xFB	0x6A	**0xEB**	**0xD8**
27	0x9A	0x4A	0x26	0x6C	0x65	0xB5	0xD9	0x71	0x22	**0x99**	**0xC7**	0x8E	0xDD	**0x66**	**0x38**
28	0x9A	0x4A	0x26	0x18	0x65	0xB5	0xD9	0x5B	0xA2	**0x09**	**0x6A**	0xA4	0x5D	**0xF6**	**0x95**
29	0x9A	0x4A	0x26	0x14	0x65	0xB5	0xD9	0xCC	0x31	**0x68**	**0x3C**	0x33	0xCE	**0x97**	**0xC3**
30	0x9A	0x4A	0x26	0xD0	0x65	0xB5	0xD9	0xB1	0xCA	**0x04**	**0x59**	0x4E	0x35	**0xFB**	**0xA6**
31	0x9A	0x4A	0x26	0xB4	0x65	0xB5	0xD9	0x56	0xA6	**0x45**	**0x86**	0xA9	0x59	**0xBA**	**0x79**

Note: in table above, bolded text indicates changes from CL 119 AM values

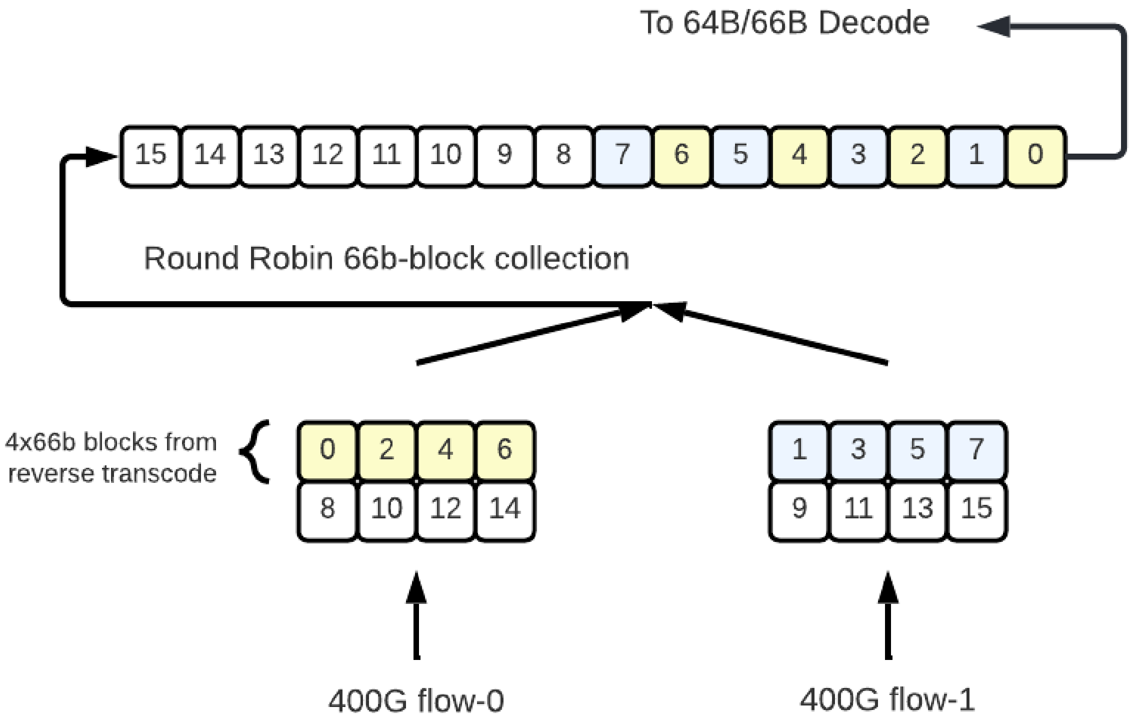
Rx PCS/FEC Data Flow

- Alignment Lock and Deskew
  - AM lock : per lane, same as CL119
  - De-skew : across 32 PCS lanes
- Lane reorder (and split)
  - Reorder and split 32 PCS lanes into 2 groups of 16
    - Lanes 0-15 : Flow-0
    - Lanes 16-31 : Flow-1
- FEC decode, de-scramble, transcode decode – same as CL119
- Round robin block collection must be aligned across Flow-0/1 based on Alignment Marker location



Rx 66b Block Collection

- Round Robin 66b Block Collection is opposite of Tx Block Distribution



Re-use CL119 State Diagrams

- Re-use all of the following
  - Figure 119–12—Alignment marker lock state diagram
  - Figure 119–13—PCS synchronization state diagram
  - Figure 119–14—Transmit state diagram
  - Figure 119–15—Receive state diagram

- Minor modification to the following
  - Add restart_lock<y> variable per 400G flow
    - restart_lock = restart_lock<0> OR restart_lock<1>
  - Add hi_ser<y> variable per 400G flow
    - hi_ser = hi_ser<0> OR hi_ser<1>

PMA

- PMA functions as defined in CL120, with latest 802.3ck updates for 100G/lane
  - Bit-multiplexing (4:1)
  - Modulation (PAM4)
  - AUI Physical lane instantiation (8 lane)
  - Signaling lane rate (106.25Gb/s)
  - Coding (Gray, precoding)
  - Clock and data recovery
  - Loopbacks
  - Test patterns

- Per lane AUI specifications from 802.3ck

PMA Muxing

- Any 32 PCS Lanes to Any 8 PMA Lanes
    - 4:1 Bit-multiplexing of data from any 4 PCS lanes to any 1 PMA lane
    - The receiver can receive PMA lanes in any order and has a full 32 lane reorder block
    - Clock content is same as a 400GE CL119 stream
        - Analysis completed and presented in wong_3df_logic_220630

Outline

- Introduction
- PCS/FEC/PMA Baseline proposal
- **Implementation considerations**
- Architecture considerations
- Summary of work since May’22 interim
- Conclusions

Latency considerations

- Two 400GbE FEC encode/decode engines in parallel
- FEC latency for this baseline proposal same as 400GbE FEC latency

PCS lanes for 8 x 100G

- Many 800G implementations will support 100/200/400/800GbE Ethernet ports
  - 32 PCS lanes already exist to support 2 x 400GbE / 4 x 200GbE / 8 x 100GbE !
  - Reuse of per lane PCS alignment logic

800Gb/s Block config	PCS/FEC lanes per Ethernet port	Total PCS/FEC lanes per 800Gb/s block
1 x 800GbE port	32 lanes @ 25G	32
2 x 400GbE ports	16 lanes @ 25G	32
4 x 200GbE ports	8 lanes @ 25G	32
8 x 100GbE ports	4 lanes @ 25G	32

- Choice of 32 PCS lanes can enable implementations over 16 x 50G AUI lanes
  - If needed (e.g. test equipment)

Other Implementation Considerations

- This baseline benefits from the use of two 400GbE PCSs in parallel
  - Reuse of logic blocks from 400GbE PCS possible
  - FEC engines, transcoder, scramblers running at same bandwidth as 400GbE
  - Per lane alignment lock running at same speed as 400GbE
  - Minimizes new development and verification

- This baseline follows the approach taken by the adopted 800GbE 8-lane AUIs and PMD baselines
  - 800GbE 8-lane AUIs and PMDs are doubling number of lanes from 400GbE
    - Example 1: 800GAUI-8 is 2 x 400GAUI-4 in parallel
    - Example 2: 800GBASE-DR8 is 2 x 400GBASE-DR4 in parallel
  - Allows re-use of specifications, maximize use of technology and investment from 400GbE

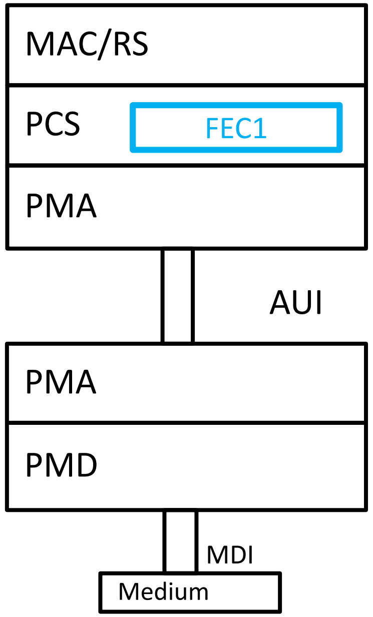
Outline

- Introduction
- PCS/FEC/PMA Baseline proposal
- Implementation considerations
- **Architecture considerations**
- Summary of work since May’22 interim
- Conclusions

IEEE P802.3df Architecture : FEC schemes

End-to-End FEC scheme

(FEC1 used for AUIs and PMD)



MAC/RS

PCS FEC1

PMA

AUI

PMA

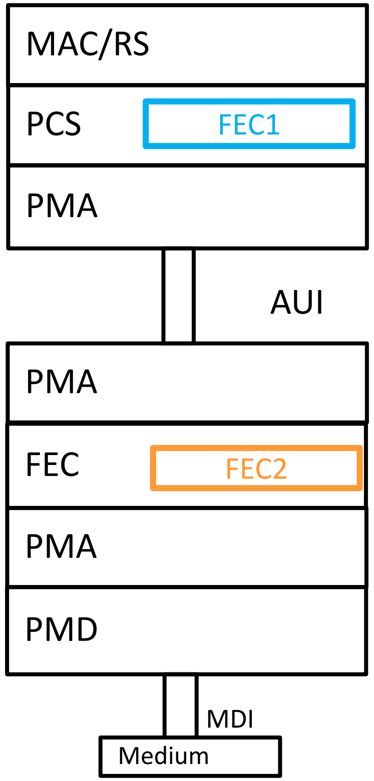
PMD

MDI

Medium

Concatenated FEC scheme

(FEC2 is added on top of FEC1.
FEC 1 for AUIs, FEC1+FEC2 for PMD)



MAC/RS

PCS FEC1

PMA

AUI

PMA

FEC FEC2

PMA

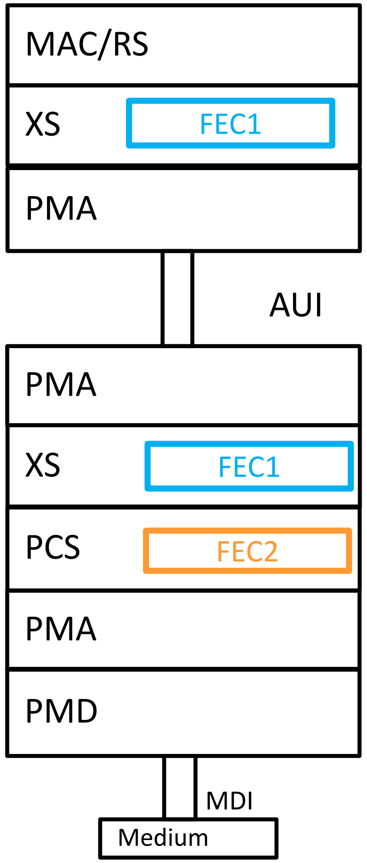
PMD

MDI

Medium

Segmented FEC scheme

(FEC2 replaces FEC1. FEC1 used for
local AUI only. FEC2 for PMD only)



MAC/RS

XS FEC1

PMA

AUI

PMA

XS FEC1

PCS FEC2

PMA

PMD

MDI

Medium

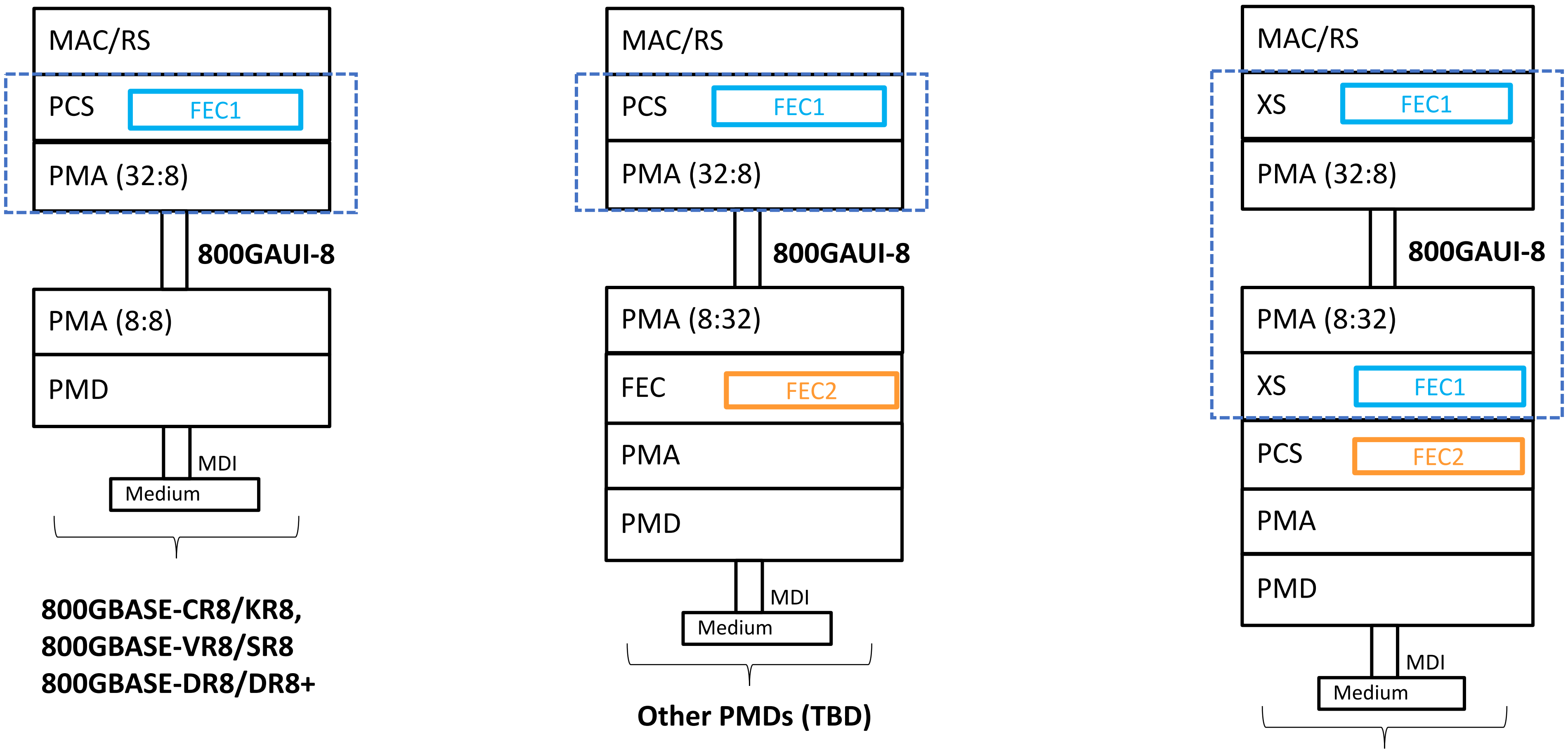
Included in this baseline

800GbE Architecture : FEC schemes over AUI-8

End-to-End FEC scheme | **Concatenated FEC scheme** | **Segmented FEC scheme**

Targeted by this Baseline

Other FEC schemes / evolution remains open



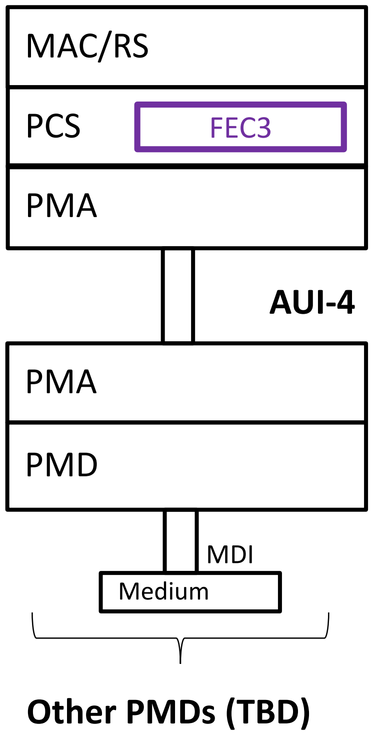
800GBASE-CR8/KR8,
800GBASE-VR8/SR8
800GBASE-DR8/DR8+

Other PMDs (TBD)

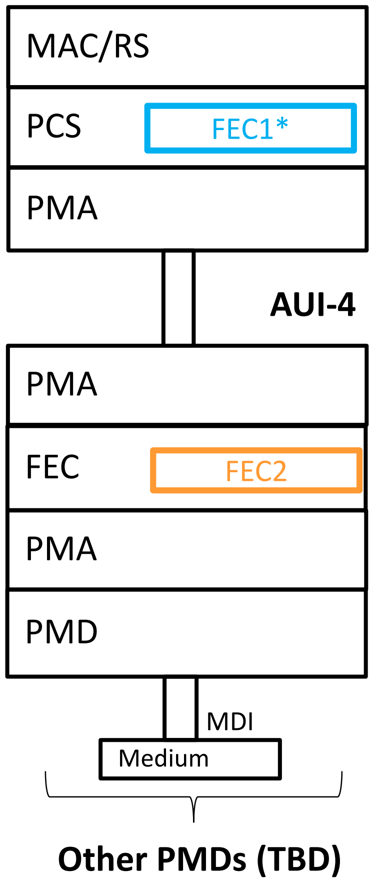
Other PMDs (TBD)

800GbE Architecture : FEC schemes over AUI-4

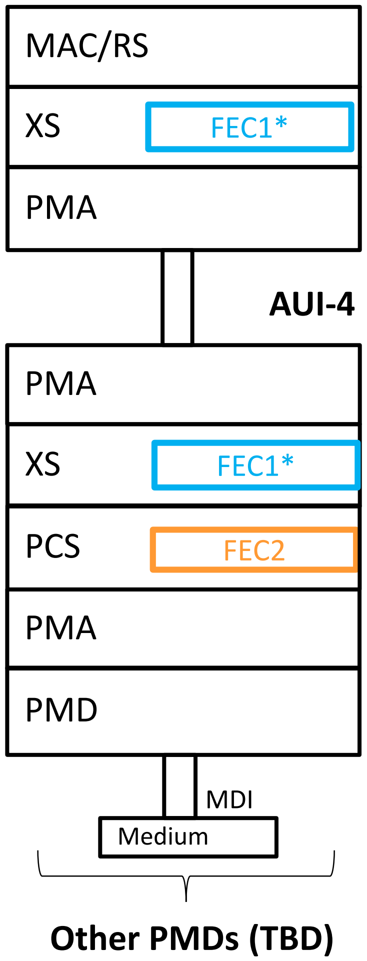
End-to-End FEC scheme



Concatenated FEC scheme



Segmented FEC scheme



** FEC1 could be the FEC proposed in this Baseline, or it could be a different FEC. Evolution options remain open.*

Outline

- Introduction
- PCS/FEC/PMA Baseline proposal
- Implementation considerations
- Architecture considerations
- **Summary of work since May’22 interim**
- Conclusions

Summary of work since May’22

- FLR analysis completed and presented in Logic Ad hoc (06/30/22)
  - See opsasnick_3df_logic_220630a
  - Baseline meets the 6.2E-11 FLR requirement corresponding to the 1E-13 BER objective
  - Addressed questions raised by X. Wang
- FLR analysis using burst error model completed and presented in Logic Ad hoc (06/30/22)
  - See opsasnick_3df_logic_220630a
  - Burst error performance looks good, no FLR floor observed
  - Some FEC gain is possible using a cross-flow bit-muxing to interleave bits from 4 codewords
- Clock content analysis completed and presented in Logic Ad hoc (06/30/22)
  - See wong_3df_logic_220630
  - Clock content is same as a 400GE CL119 stream
  - Analysis was pending from May’22 baseline presentation

Outline

- Introduction
- PCS/FEC/PMA Baseline proposal
- Implementation considerations
- Architecture considerations
- Summary of work since May’22 interim
- **Conclusions**

Conclusions

- This Baseline: 800GbE PCS, FEC and PMA for 8 x 100G PMDs and 8 x 100G AUIs
- Supports all adopted 802.3df copper and optical PMDs baselines using 100G/lane
- Highly leverages existing 400GbE specifications
  - 2 x 400GbE (Clause 119) with minor modifications to the specifications
- Highly leverages existing 400GbE implementations
  - Enable re-use of per-lane AM lock, FEC interleaving, FEC encode/decode, scrambler, transcoder
- Meets the FLR requirement corresponding to 1E-13 BER objective
- Clock Content is same as a 400GbE CL119 stream
- Enables faster time-market for 800GbE (8 x 100G/lane) implementations
  - Maximizing technology reuse and existing industry investments
- Fits into an overall 800GbE Logic Architecture, and does not constrain future FEC schemes using 200G/lane AUIs and PMDs and/or Coherent PMDs
- 1.6TbE PCS/FEC can be chosen independently of 800GbE
  - Decisions made in this baseline will not restrict options / choices for 1.6TbE

Thanks !

Backup – FLR Analysis Data for Random and Burst Errors

- FLR data from opsasnick_3df_logic_220630a
- Additional data added for 400GbE for comparison

BER$_{in}$ and SNR Requirements with Random Errors

RS(544,514) FEC	FLR Target	FSF	CER Required	BER$_{in}$ Required	PAM4 DER Required	SNR (dB) Required
No Interleave	6.2E-11	1.125	5.49E-11	3.20E-4	6.40e-4	17.45
2 CW Interleave	6.2E-11	2.125	2.92E-11	3.06E-4	6.13E-4	17.48
4 CW Interleave	6.2E-11	4.125	1.50E-11	2.93E-4	5.85E-4	17.52

- 100G/lane PMDs assume BER$_{in}$ = 2.4E-4 or better.
  - See “Bit Error Ratio” in Clauses 124, 140, 151, etc.
  - Expand requirement to include two AUI on each end of the link, adds 4 * 1E-5 = 2.8E-4

- Even if BER$_{in}$ is worse than 2.8E-4, all Interleaves meet the 6.2E-11 FLR Target
- SNR increase to meet the same FLR from 2-way to 4-way FEC interleave is ≈ 0.04dB (negligible)

Burst Error Results for 8x100 PCS Options

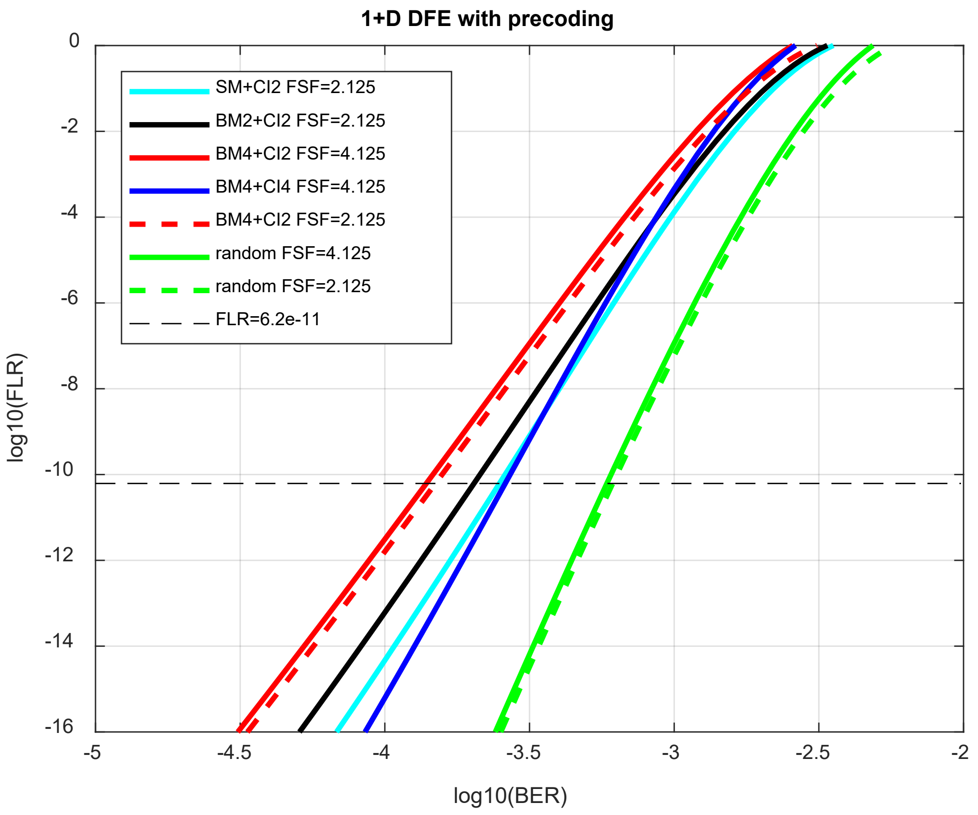
Option	Required FLR	1+0.1D no precoding (a=0.01)		1+0.5D no precoding (a=0.375)		1+D with precoding (a=0.75)	
		Required SNR	Required DER	Required SNR	Required DER	Required SNR	Required DER
1.a	6.20E-11	17.49	6.09E-04	17.57	5.40E-04	18.09	2.49E-04
1.b	6.20E-11	17.49	6.07E-04	17.79	3.96E-04	18.22	2.03E-04
2.a	6.20E-11	17.52	5.79E-04	18.41	1.48E-04	18.44	1.39E-04
2.b	6.20E-11	17.52	5.80E-04	17.83	3.69E-04	18.06	2.61E-04
400GbE	6.20E-11	17.49	6.07E-4	18.36	1.61E-04	18.40	1.50E-04

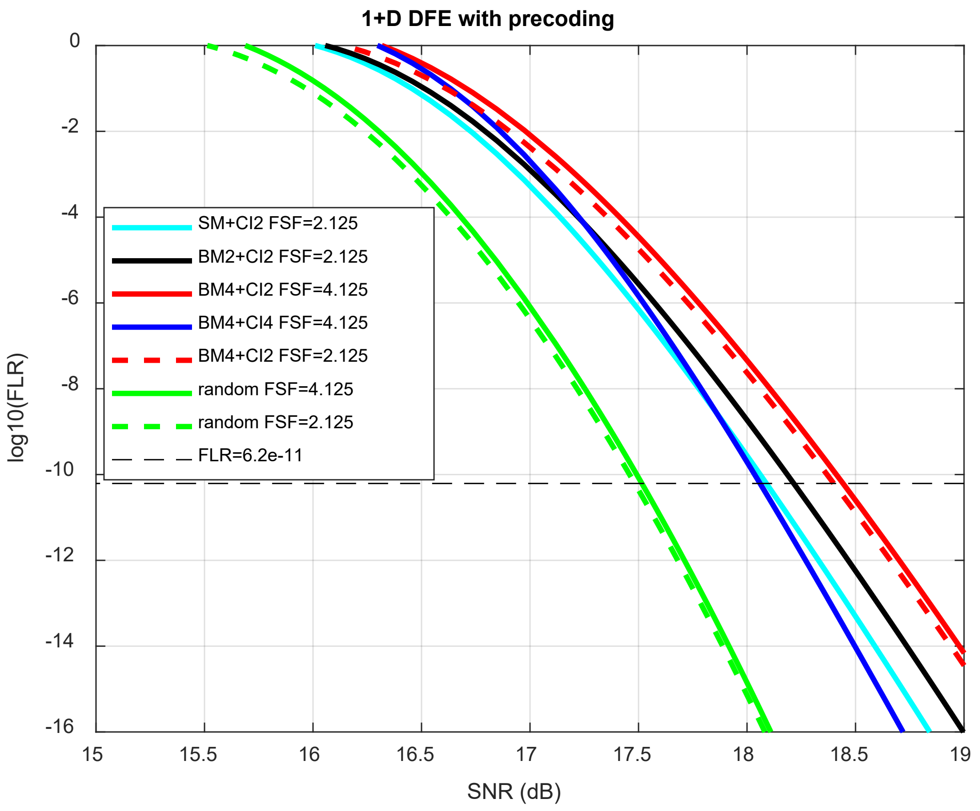
- 1+0.1D : Nearly random (a=0.01)
  - Option 1.a, 2 CW interleave, is 0.03dB better than 4 CW Interleave

- 1+D : High burst correlation (a=0.75)
  - Option 2.b, 4:1 bitmux across 4 CW, is best option by 0.03dB

Option 1.a (SM + CI2, FSF=2.125) is for proposal from wang_3df_logic_220623a.pdf
Option 2.a (BM4 + CI2, FSF=4.125) and Option 2.b (BM4 + CI4, FSF=4.125) is for this baseline.
400GbE uses (BM4 + CI2, FSF=2.125)

FLR for Random and Burst Errors with High Correlation





Option 1.a (random, FSF=2.125) & (SM + CI2, FSF=2.125) (light blue) is for proposal from wang_3df_logic_220623a.pdf
Option 2 (random, FSF=4.125) & 2.a:(BM4 + CI2, FSF=4.125) (solid red) and 2.b:(BM4 + CI4, FSF=4.125) (dark blue) is for this baseline.
400Gbe uses (BM4 + CI2, FSF 2.125) (dashed red)