800GbE PCS and PMA Baseline Proposals for 100 Gb/s per lane PHYs

Xinyuan Wang, Huawei Technologies
Matt Brown, Huawei Technologies

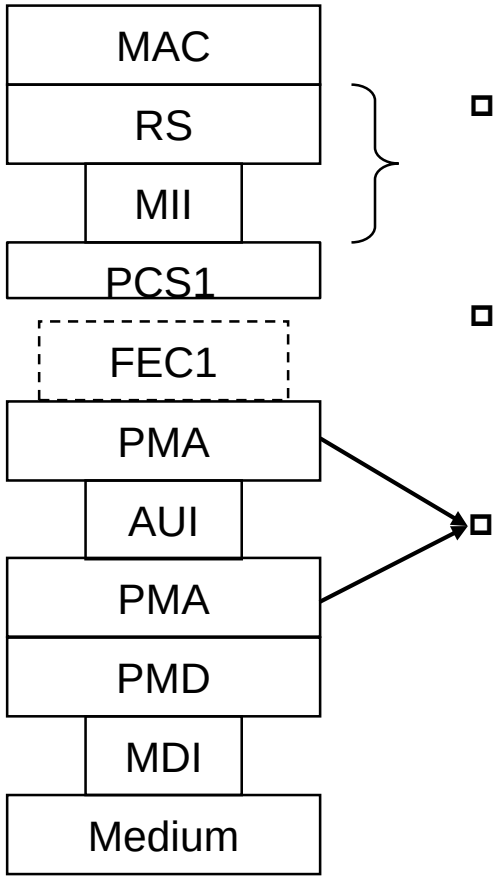
Supporters:

- Weiqiang Cheng, China Mobile
- Haojie Wang, China Mobile
- Ruibo Han, China Mobile
- Xiang He, Huawei Technologies
- Yu Xu, Huawei Technologies

Introduction

- For 800 Gb/s Ethernet using 100 Gb/s per lane technology there is general agreement on the logic sublayers except for the PCS/PMA.
- Two proposals are being considered by the task force for the PCS/PMA.
- Previous presentations have provided details on these two approaches as well as the advantages and disadvantages.
- This presentation continues the discussion comparing the advantages and disadvantages of the two approaches.
- Option 2 (**Speed-up Clause 119**) is the way to go.

Related 800GbE Logic Baseline



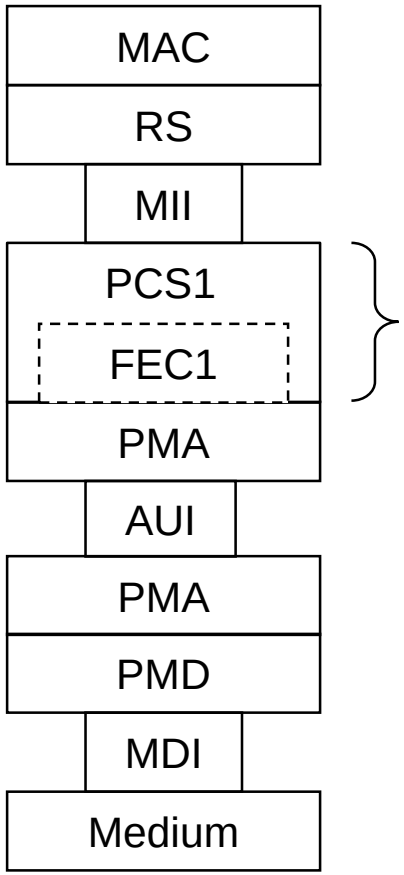
- No objections for **speed-up Clause 117** for 200/400GbE in nicholl_3df_logic_220623
- **PCS with FEC baseline** under discussion
- No objections for PMA using **100 Gb/s per lane signaling** based on **Clause 120** for 200/400GbE in nicholl_3df_logic_220623

Refer to: Minutes of Architecture and logic Ad Hoc at June 23th meeting:
https://www.ieee802.org/3/df/public/adhoc/logic/22_0623/minutes.pdf

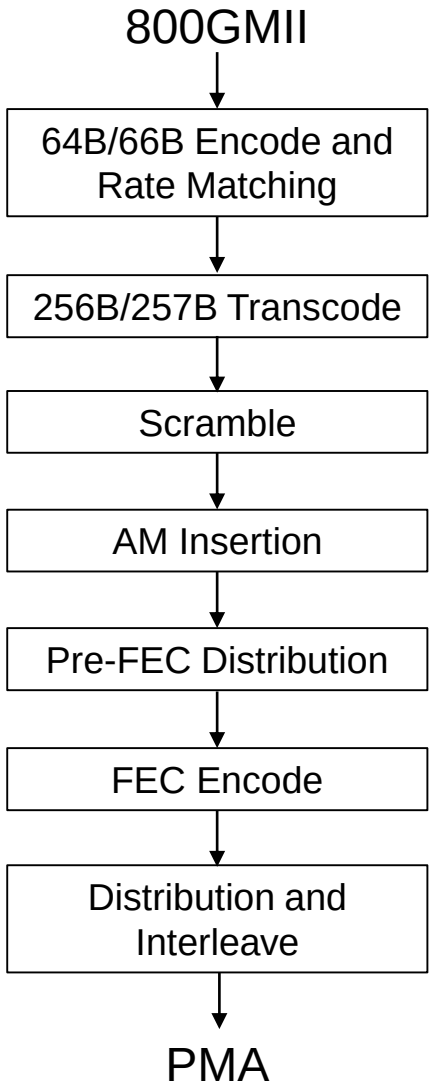
Proposal Overview: Speed-up CL119

- This presentation proposes to adopt PCS/PMA sublayer equivalent to the 200GBASE-R PCS/PMA sublayers defined in Clause 119/120 for 100Gb/s per lane PHYs.
  - Proposed in wang_3df_logic_220623a and wang_3df_logic_220630.
  - PCS with FEC equivalent to the 200GBASE-R PCS defined in Clause 119.
  - Total data rate increased from 200 Gb/s to 800 Gb/s (increase 4X).
  - Eight PCS lanes at 106.25 Gb/s each (increase 4X).
    - RS(544,514) FEC symbol interleave from 2 codewords in each PCS lane.
  - Data 64B/66B encoded, 256B/257B encoded, scrambled.
  - Distributed to two RS(544,514) codewords and encoded.
  - Codewords interleaved then distributed to 8 PCS lanes.
  - No changes to alignment markers since rate will differentiate form 200GBASE-R lanes.
    - No analysis required.

800GbE PCS Baseline Advantage (Speed-up CL119)



- 4X speed-up Clause 119 based on 200GbE with 8 PCS lanes.
- Straight forward evolution.
  - **No new technical work needed.**
- Meets all adopted objectives.
  - With all adopted 8X100 Gb/s AUIs/PMD baselines.
- With **low latency advantage**.



Is Latency Important?

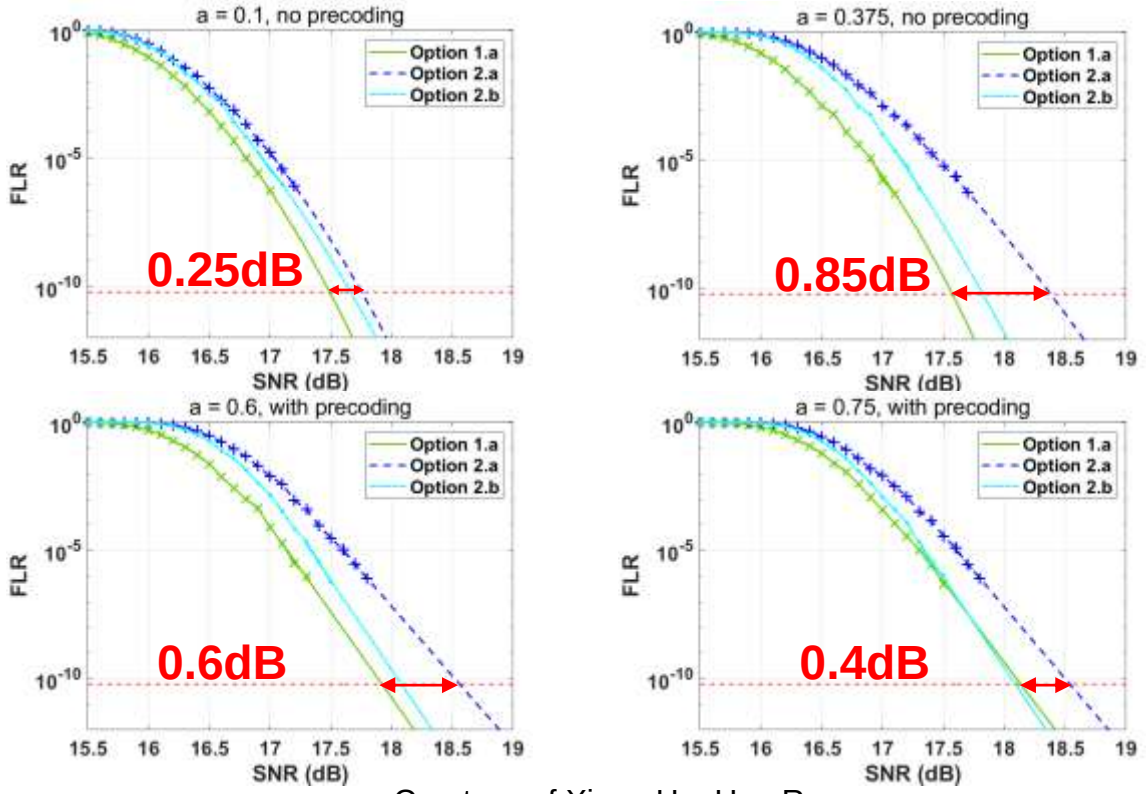
- ETC seems to think so…
  - Low Latency FEC from Ethernet Technology Consortium:
  - https://ethernettechnologyconsortium.org/wp-content/uploads/2020/03/LL-FEC-Specification-1.0-25G-Consortium.pdf
  - https://www.hpcwire.com/off-the-wire/25-gigabit-ethernet-consortium-offers-low-latency-specification-for-50gbe-100gbe-and-200gbe-hpc-networks/
  - https://www.fibre-systems.com/news/low-latency-fec-specification-offered-25-gigabit-ethernet-consortium
  - https://www.hpcwire.com/2020/04/07/ethernet-technology-consortium-launches-800-gigabit-ethernet-specification/
- PCISIG seems to think so…
  - PCIE 7.0 has a low latency target, from PCISIG.com
  - https://pcisig.com/blog/announcing-pcie%C2%AE-70-specification-doubling-data-rate-128-gts-next-generation-computing

Clear Advantage of Speed Up CL119 for Single Part Link with Burst Errors

- In opsasnick_3df_logic_220630, the term “option 1, 2” is reversed as general descriptions in TF.
- The option **1a**, corresponding to our proposal has **better** FEC performance than option **2a** as proposed by shrikhande_3df_01b_220517, across the whole range of burst errors, especially for a=0.375.
- Simulation was done on a single part link, and shows that option **1a** (proposed in this presentation) has clear advantage over option **2a** and 2b, except at a = 0.75 when option 2b has slight advantage.
- In Slide #10 of anslow_3ck_01_1118, 1.1E-4 BER is required for multi-tap DFE with precoding.
  - shrikhande_3df_01b_220517 proposal will lead to even worse FLR due to 4 codewords rather than 2.

Option 1 and 2 are flipped in this page only to match opsasnick_3df_logic_220630.

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



Courtesy of Xiang He, Hao Ren.

Reference from 802.3bs and FLR Risk Analysis for 800GbE

Multi-part link results

The BER of the electrical sub-links for a penalty of ~0.1 dB optical in the optical sub-link are shown in the table below.

	At slicer output for FLR = 6.2E-11			
	Total electrical		Optical	
Same cwd (1), a = 0.75	Burst	2.9E-7*	Random	2.4E-4
Same cwd, symbol interleave (2), a = 0.75	Burst	7.5E-7*	Random	2.4E-4
Same cwd (1), a = 0.5	Burst	1.6E-5*	Random	2.4E-4
1:4 Pre-interleaved (4), a=0.75	Burst	2.2E-5*	Random	2.4E-4
1:2 Pre-interleaved (8), a=0.75	Burst	3.5E-5*	Random	2.4E-4
Diff cwd (FOM) (7), a = 0.75	Burst	4E-5*	Random	2.4E-4
Same cwd elec only precoded, a=0.75	Burst	5.1E-5*	Random	2.4E-4
Same cwd end-to-end precoded, a=0.75	Burst	6.9E-5*	Random	4.9E-5
1:4 Pre-interleaved (6), a=0.75	Burst	5.7E-5*	Random	2.4E-4
1:2 Pre-int, sym mux (10), a=0.75	Burst	7.6E-5*	Random	2.4E-4
1:4 Pre-int, sym mux (9), a=0.75	Burst	1E-4*	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 and 2 when a = 0.5.

Solution adopted by 802.3bs with **3.5E-5 BER** for up to 4X AUIs to support 6.2E-11 FLR.

Option 1 (**2a** in opsasnick_3df_logic_220630) **can't meet FLR objective due to 4 codewords.**

Option 1 (**2b** in opsasnick_3df_logic_220630) of 2X parallel CL119 proposal with 4X codeword and additional 4:1 bit multiplexing will **require lower than 5.7E-5 BER** for up to 4X AUIs to support 6.2E-11 FLR.

More analysis is needed as PCS lanes are formed differently.

Option 2 (**1a** in opsasnick_3df_logic_220630) of 4X speed-up CL119 proposal **allows 7.6E-5 BER** for up to 4X AUIs to support 6.2E-11 FLR.

More margin can be provided.

https://www.ieee802.org/3/bs/public/15_09/anslow_3bs_03_0915.pdf

Clock Content Issue of Option 1

- In wong_3df_logic_220630 discussion, it is pointed out:
  - This contribution **did not cover all combinations** of PCS lanes multiplexing, such as 1 lane from one 400Gbps flow and 3 lanes from the other 400Gbps flow.
  - It did not consider the initial state of the two parallel scramblers.

Only partial (5%) any-to-any 4:1 combinations of Option 1 (2a in opsasnick_3df_logic_220630)

Only partial (1.3%) restricted 4:1 combinations of Option 1 (2b in opsasnick_3df_logic_220630)

800GbE search

- Combinations of 800GbE PCS lanes for 4:1 bit interleaving to form 100Gb/s lanes were searched to find any unusual clock content with delays between -10 to +10 bits. Gray coding and PAM4 were assumed.
  - For all possible combinations from lanes 0-15 (flow #1), 2356 out of 404,520,480 possibilities were identified as having unusual clock content similar to the plots shown in the previous slides.
  - For all possible combinations from lanes 16-31 (flow #2), the search gave the same results as lanes 0-15 as they are two 400G PCS flows.
  - For all natural pairs from lanes 0-15 and lanes 16-31 (e.g. [0 1 16 17],[2, 3, 16, 17],.....,[14 15 30 31]), the search did not find any with unusual clock content out of 14,224,896 possibilities
  - For all pairs from lanes 0-15 and the "equivalent" pairs from 16-31 (e.g. [0 1 16 17],[0 2 16 18],.....,[14 15 30 31]), the search did not find any with unusual clock content out of 26,671,680 possibilities

DC (baseline) Wander Challenge for Option 1

Baseline wander

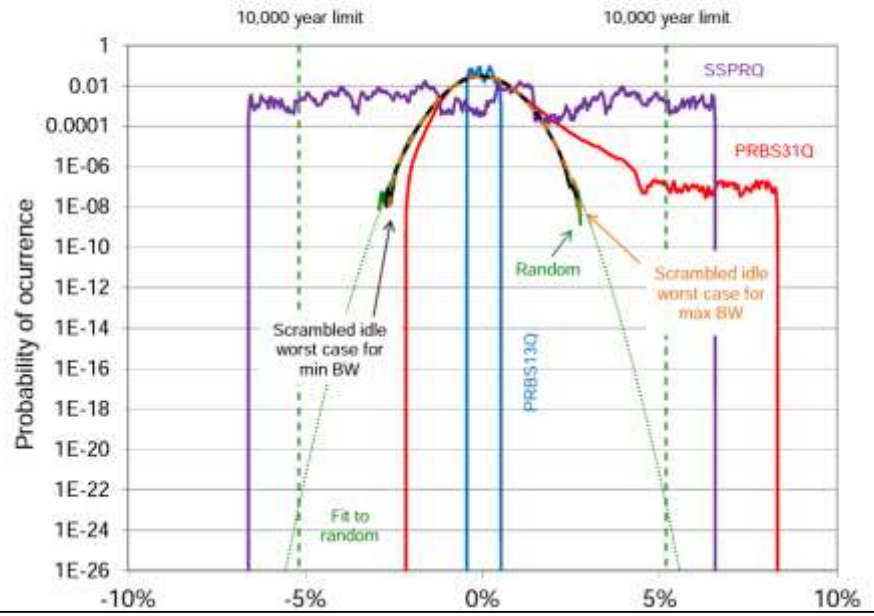
Previous NRZ contributions have used a "baseline wander" parameter

This was defined as:

Baseline wander is the instantaneous offset (in %) in the signal generated by AC coupling at the Baud rate / 10,000.

This analysis re-uses this definition unmodified, but it should be noted that for PAM4, the eye height is 1/3 that of NRZ so the effects of a given amount of baseline wander will be greater.

Baseline wander, 100G lanes, new markers



- DC wander analysis for Option 2 is not needed.
  - Completely reuse of 200 GbE AM.
  - Proven field success without issues.
- DC wander analysis for Option 1 is not performed yet.
  - **32 new AM patterns** could introduce DC wander when multiplexing, especially for 8:1.

https://www.ieee802.org/3/bs/public/adhoc/logic/apr28_16/anslow_03_0416_logic.pdf

Comparison of Candidate Options: Parameter View

	Option 1: 2X CL119	Option 2: Speed-up CL119
Number of FEC codewords	4X	2X
Interleave	2X	2X
FEC decode latency	⩾a+12.8ns	a
BER(Bit Error Ratio)	1.00E-13	1.00E-13
FLR(Frame Loss Ratio) at equivalent BER	2X	1X
Number of PCS lanes	32 PCS lanes @ 25 Gb/s	8 PCS lanes @ 100 Gb/s
Support 25/50 Gb/s per lane	Yes	No
Number of Alignment Markers	32	8
PMA	4:1 for 100 Gb/s per lane	1:1 for 100 Gb/s per lane
Clock content challenge	Need more work for all PCS lanes multiplexing combinations	No
DC wander challenge	Need more work	No
Reuse of implementation	Largely (replication/muxing)	Largely (speed-up)
Reuse of standard specification	Largely	Completely
Fast time to market	Yes	Yes

Observations:

- 800 Gb/s Ethernet is only now being defined and has a long life ahead.
  - 32 PCS lanes with 4:1/8:1 multiplexing is the critical penalty for PCS/PMA baseline Option 1 due to its impact on FEC performance, clock content and DC wander challenges, especially for future 200 Gb/s per lane evolution.
  - Need to evaluate broad market potential for supporting 25 and 50 Gb/s lane rates.
    - Could be useful for early stage tests for logic functions, but no real product will use them.
  - Latency is an important criteria. Lower latency is always desirable.
  - A common architecture across all 4 Ethernet rates is agreed on when adopting the logic architecture baseline.
    - Option 2 can lead to similar speed-up CL119 solution for 1.6TbE, but what is the solution from Option 1?

Conclusion:

- The technical advantages of option 2 (4X speed-up 200GBASE-R PCS/PMA) are numerous.
- This baseline proposes 800GbE PCS and PMA based on sped up of Clause 119/120 to support 8X100 Gb/s per lane AUIs, electrical and optical PMDs.

Thanks!

Backup

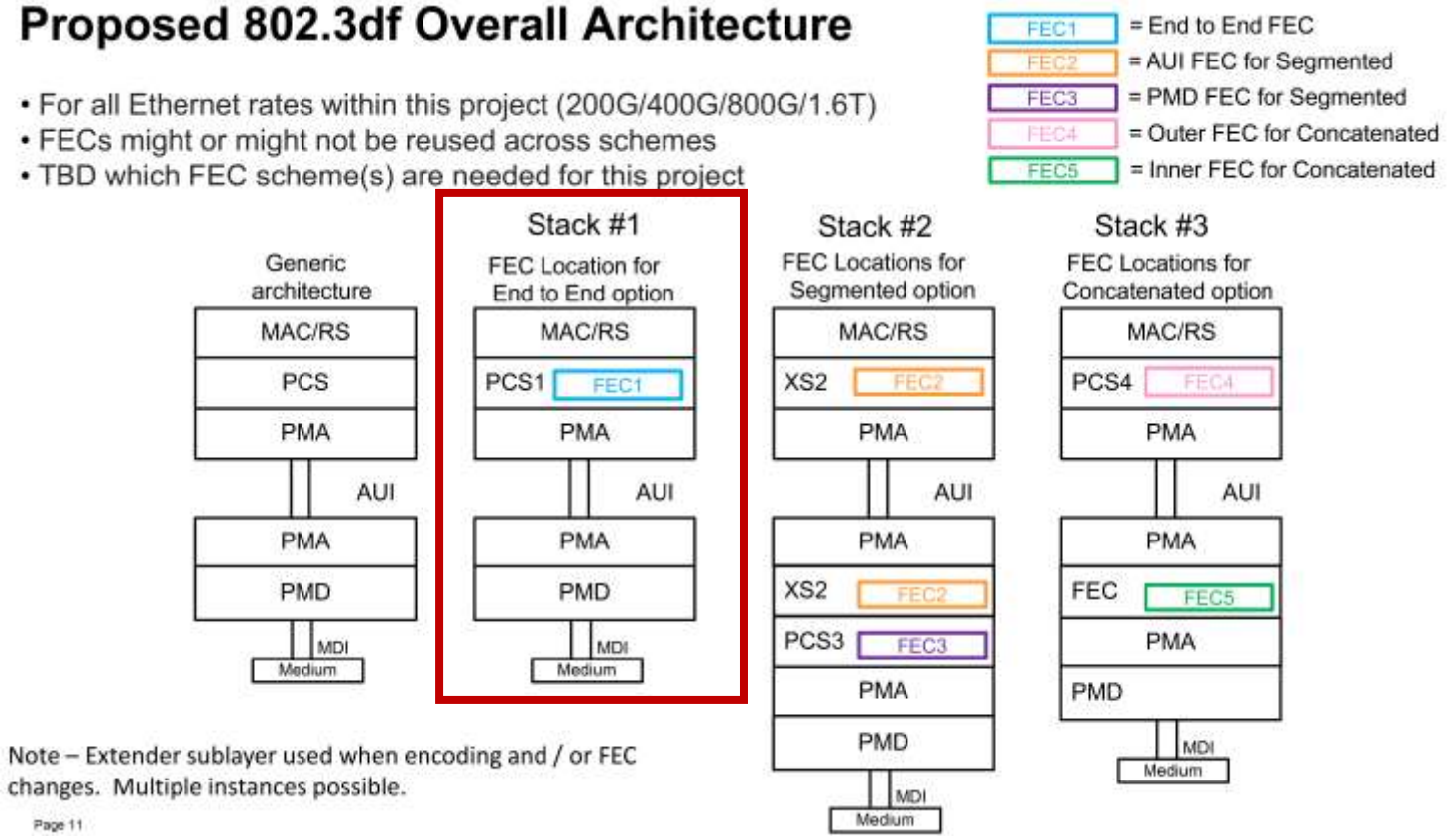
The following slides is detailed baseline for 4X speed-up CL119.

References

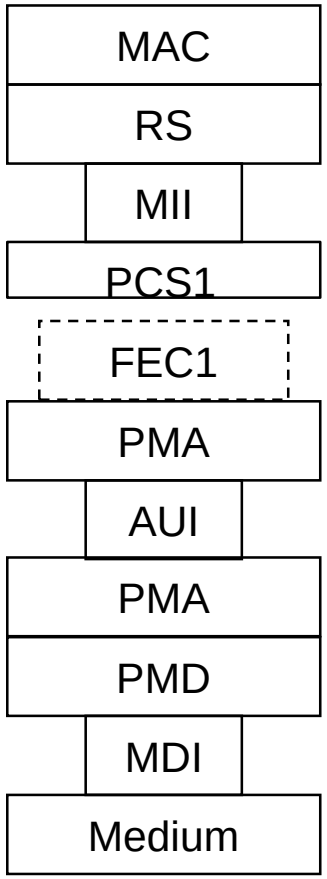
- Previous contributions in Task Force relating to the 800 Gb/s PCS/PMA baseline are listed here:
  - Option 1: 2X parallel Clause 119 (400GBASE-R PCS/PMA)
    - https://www.ieee802.org/3/df/public/22_03/shrikhande_3df_01_220329.pdf
    - https://www.ieee802.org/3/df/public/22_05/22_0517/shrikhande_3df_01a_220517.pdf
  - Option 2: Sped up Clause 119 (200GBASE-R PCS/PMA)
    - https://www.ieee802.org/3/df/public/22_03/wang_3df_01a_220308.pdf
    - https://www.ieee802.org/3/df/public/22_05/22_0517/he_3df_01_220517.pdf

References: (Cont'd)

- Adopted logic architecture baseline at May meeting
  - https://www.ieee802.org/3/df/public/22_05/motions_3df_2205_0524.pdf

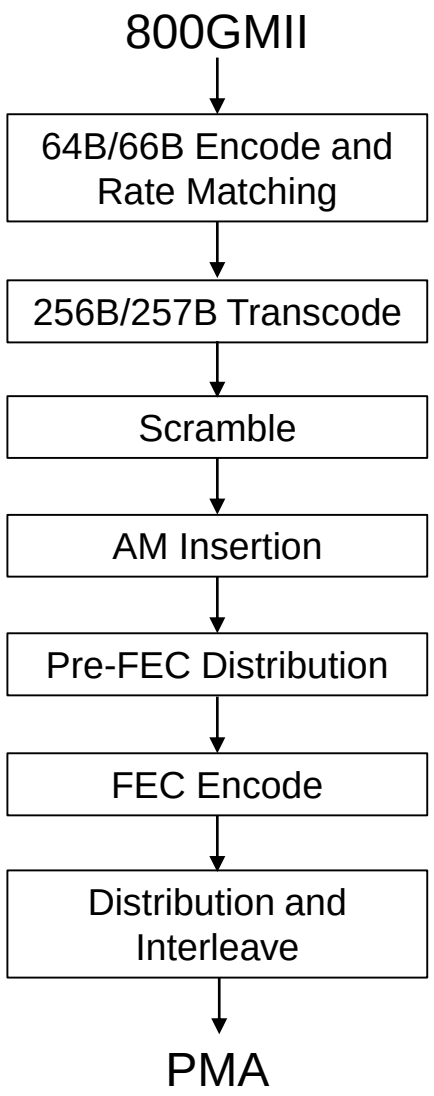


PCS and PMA in 800GbE Logic Architecture



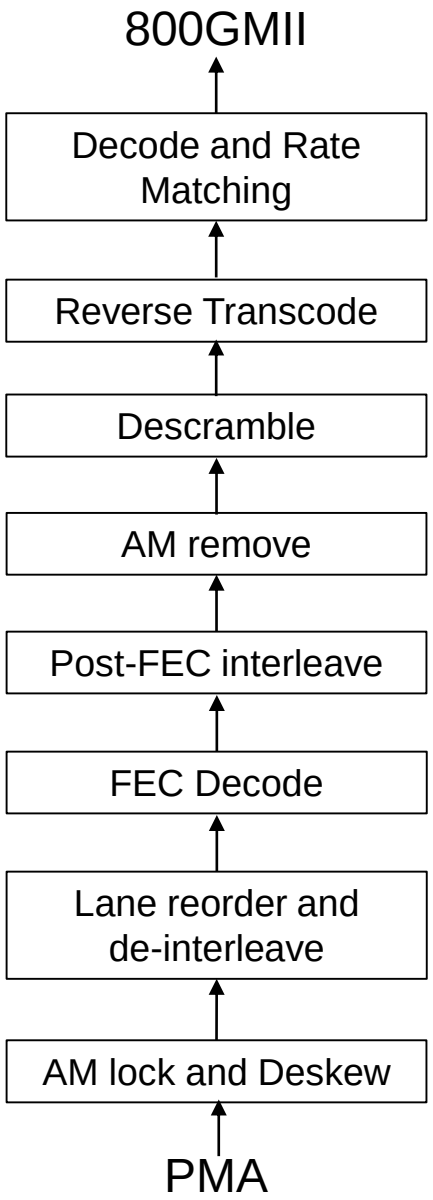
- The 800GBASE-R PCS are composed of Transmit and Receive processes, which shields the Reconciliation Sublayer (and MAC) from the specific nature of the underlying channel.
  - Communicating with the 800GMII: the PCS uses an eight octet-wide, synchronous data path, with frame delineation being provided by transmit control signals (TXC<n> = 1) and receive control signals (RXC<n> = 1).
- The PMA sublayer operates independently of block and frame boundaries.
  - The PCS provides the functions necessary to map frames between the 800GMII format and the PMA service interface format.

TX PCS Data Flow



- 64B/66B encode based on Clause 119.2.3/82.2.3.
- Transcode to 256B/257B based on Clause 119.2.4.2/91.5.2.5.
  - Allow direct encode from 64B/66B.
- Scramble based on Clause 119.2.4.3/82.2.5.
- FEC Encoder is RS(544,514,15,10) with 2-way interleave based on Clause 119.
  - All FEC processing is same as in Clause 119.2.4, including error correction and detection modes at RX.
- 8 PCS lanes @ 100 Gb/s.
- Support for any PCS lane on any physical lane.
- Compensation for any rate differences caused by the insertion or deletion of alignment markers or due to any rate difference between the 800GMII and PMA through the insertion or deletion of idle control characters.

RX PCS Data Flow



- Reverse of TX PCS data flow.
- Arbitrary PCS lanes order arrival from PMA.

Alignment Marker

- In order to support deskew and reordering of the individual PCS lanes at the RX PCS, alignment markers corresponding to PCS lanes are periodically inserted after being processed by the alignment marker mapping function.
- Refer to Clause 119.2.4.4.1, identical format as Figure 119-4 for 200GbE with 8 PCS lanes.

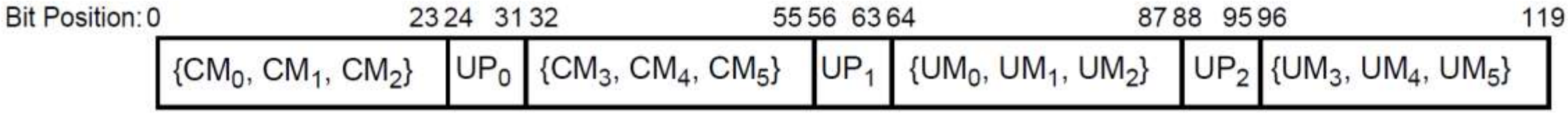


Figure 119–4—Alignment marker format

Alignment Marker Encoding

- Refer to Clause 119.2.4.4.1, identical encodings as Table 119-1 for 200GbE with 8 PCS lanes.

Table 119–1—200GBASE-R alignment marker encodings

PCS lane number	Encoding[a] $\{CM_0, CM_1, CM_2, UP_0, CM_3, CM_4, CM_5, UP_1, UM_0, UM_1, UM_2, UP_2, UM_3, UM_4, UM_5\}$
0	0x9A,0x4A,0x26,0x05,0x65,0xB5,0xD9,0xD6,0xB3,0xC0,0x8C,0x29,0x4C,0x3F,0x73
1	0x9A,0x4A,0x26,0x04,0x65,0xB5,0xD9,0x67,0x5A,0xDE,0x7E,0x98,0xA5,0x21,0x81
2	0x9A,0x4A,0x26,0x46,0x65,0xB5,0xD9,0xFE,0x3E,0xF3,0x56,0x01,0xC1,0x0C,0xA9
3	0x9A,0x4A,0x26,0x5A,0x65,0xB5,0xD9,0x84,0x86,0x80,0xD0,0x7B,0x79,0x7F,0x2F
4	0x9A,0x4A,0x26,0xE1,0x65,0xB5,0xD9,0x19,0x2A,0x51,0xF2,0xE6,0xD5,0xAE,0x0D
5	0x9A,0x4A,0x26,0xF2,0x65,0xB5,0xD9,0x4E,0x12,0x4F,0xD1,0xB1,0xED,0xB0,0x2E
6	0x9A,0x4A,0x26,0x3D,0x65,0xB5,0xD9,0xEE,0x42,0x9C,0xA1,0x11,0xBD,0x63,0x5E
7	0x9A,0x4A,0x26,0x22,0x65,0xB5,0xD9,0x32,0xD6,0x76,0x5B,0xCD,0x29,0x89,0xA4

[a] Each octet is transmitted LSB to MSB.

Alignment Marker Mapping into FEC Codewords and PCS Lanes

- Refer to Clause 119.2.4.4.1, identical mapping as Figure 119-5 for 200GbE with 8 PCS lanes.

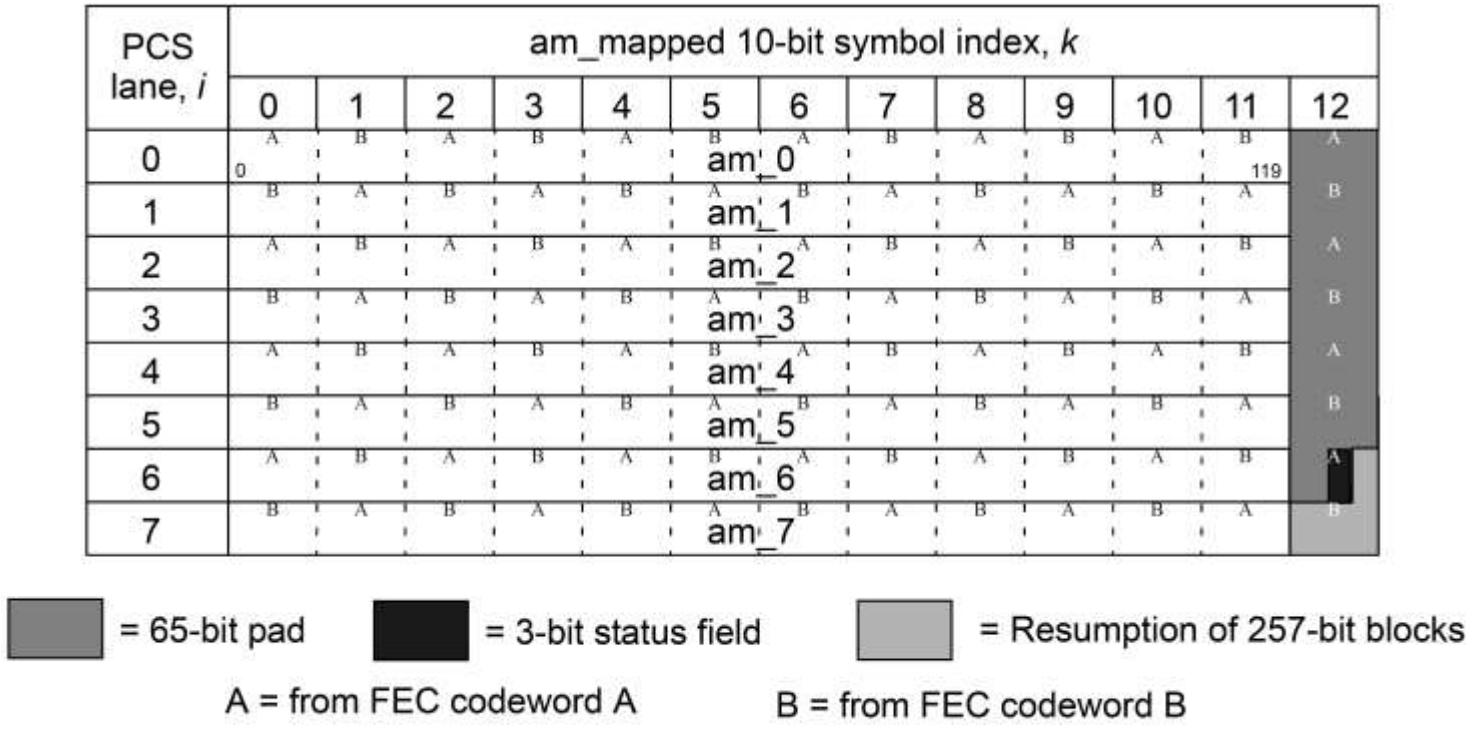
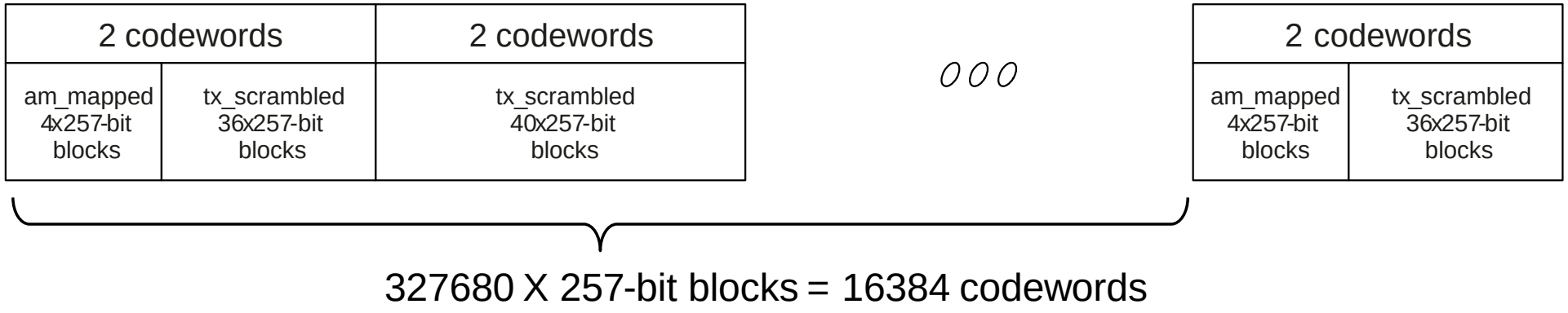


Figure 119–5—200GBASE-R alignment marker mapping to PCS lanes

Alignment Marker Insertion Period

- Refer to Clause 119.2.4.4.1/2, AMs are always aligned to the beginning of a RS FEC codeword, repetition distance is 16384 FEC codewords comparing to 8192 for 400GbE and 4096 for 200GbE in IEEE 802.3bs.



RX Process Function

- Refer to Clause 119.2.5
- Alignment lock and deskew:
  - The RX PCS forms 8 separate bit streams from PMA and obtains lock to the alignment markers as specified by the alignment marker lock state diagram shown in Figure 119–12.
  - After alignment marker lock is achieved on each of the 8 PCS lanes (bit streams), all inter-lane skew is removed as specified by the PCS synchronization state diagram shown in Figure 119–13.
- Reorder and De-interleave:
  - The RX PCS shall order the PCS lanes according to the PCS lane number. The PCS lane number is defined by the unique portion (UM0 to UM5) of the alignment marker that is mapped to each PCS lane.
  - After all PCS lanes are aligned, deskewed, and reordered, the two FEC codewords are de-interleaved to reconstruct the original stream of two FEC codewords.

RS FEC Decode Process

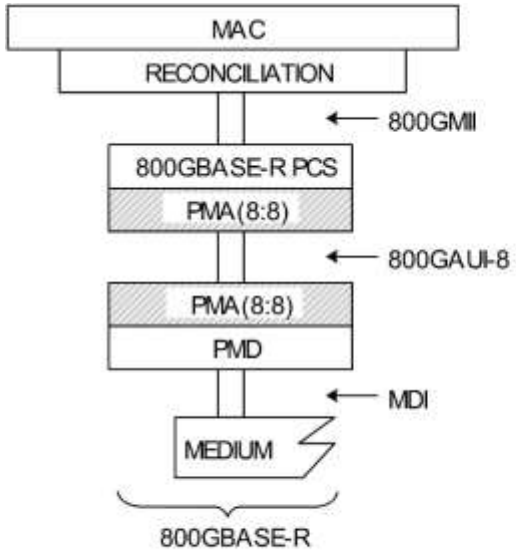
- RS FEC Decode:
  - Extracts the message symbols from the codeword, corrects them as necessary, and discards the parity symbols.
  - Capable of indicating when an errored codeword was not corrected.
  - When decoder determines that a codeword contains errors that were not corrected, it shall cause the RX PCS to set every 66-bit block within the two associated codewords to an error block (set to EBLOCK_R).
- Post-FEC Distribution:
  - After decode, data is interleaved on a 10-bit basis into rx_scrambled_am from two codewords corresponding to 40 transcoded blocks for RS(544,514) in order to recreate the transmitted data stream.
  - The first 1028 bits of rx_scrambled_am blocks is the vector am_rx<1027:0> where bit 0 is the first bit received.
  - The vector am_rx shall be removed from rx_scrambled_am to create rx_scrambled prior to descrambling.

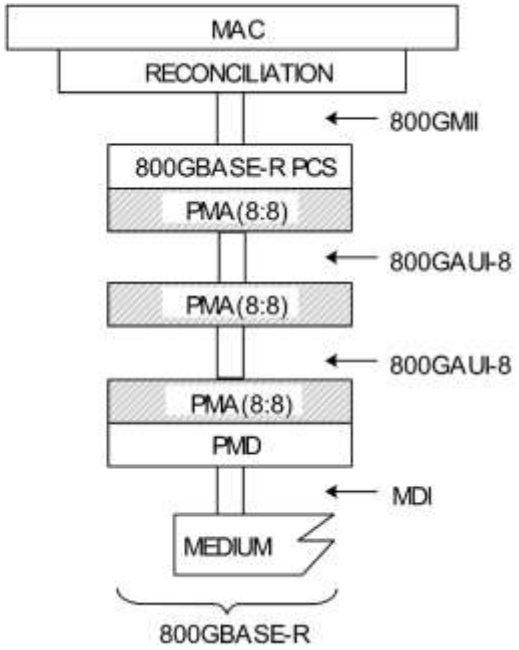
PMA Function

- Refer to Clause 120, both the transmit and receive directions:
  - Adapt the PCSL formatted signal to the appropriate number of abstract or physical lanes.
  - Provide per input-lane clock and data recovery.
  - Provide bit-level multiplexing.
  - Provide clock generation.
  - Provide signal drivers.
  - Optionally provide local loopback to/from the PMA service interface.
  - Optionally provide remote loopback to/from the PMD service interface.
  - Optionally provide test-pattern generation and detection.
  - Tolerate skew variation.
  - Perform PAM4 encoding and decoding for 800GBASE-R PMAs where the number of physical lanes is 8.

PMA Sublayer Positioning

- An implementation may use one or more PMA sublayers to adapt the number and rate of the PCS lanes to the number and rate of the PMD lanes.
- Potential forward looking compatible to Segmented, Concatenated FEC schemes.





PMA Multiplexing

- The PMA will support bit multiplexing, without regard to skew or PMA lane identity.
  - All skew is only handled in the RX PCS process as in Slide #14.
  - Refer to Clause 120.5.3 for skew variant
- The maximum cumulative delay contributed by up to four PMA stages in a PHY (sum of transmit and receive delays at one end of the link) shall meet the values specified in Table 120-1 for 92.16ns.

Standard Specification:

- New subclauses can explicitly refer to the existing subclauses in Clause 119/120.