800GBASE-ER1 Architecture (comments 36 and 218)

Tom Huber (Nokia), Gary Nicholl (Cisco), Jeff Slavick (Broadcom), Eugene Opsasnick (Broadcom), Andras deKoos (Microchip), Mike Sluyski (Cisco), Dave Ofelt (Juniper)

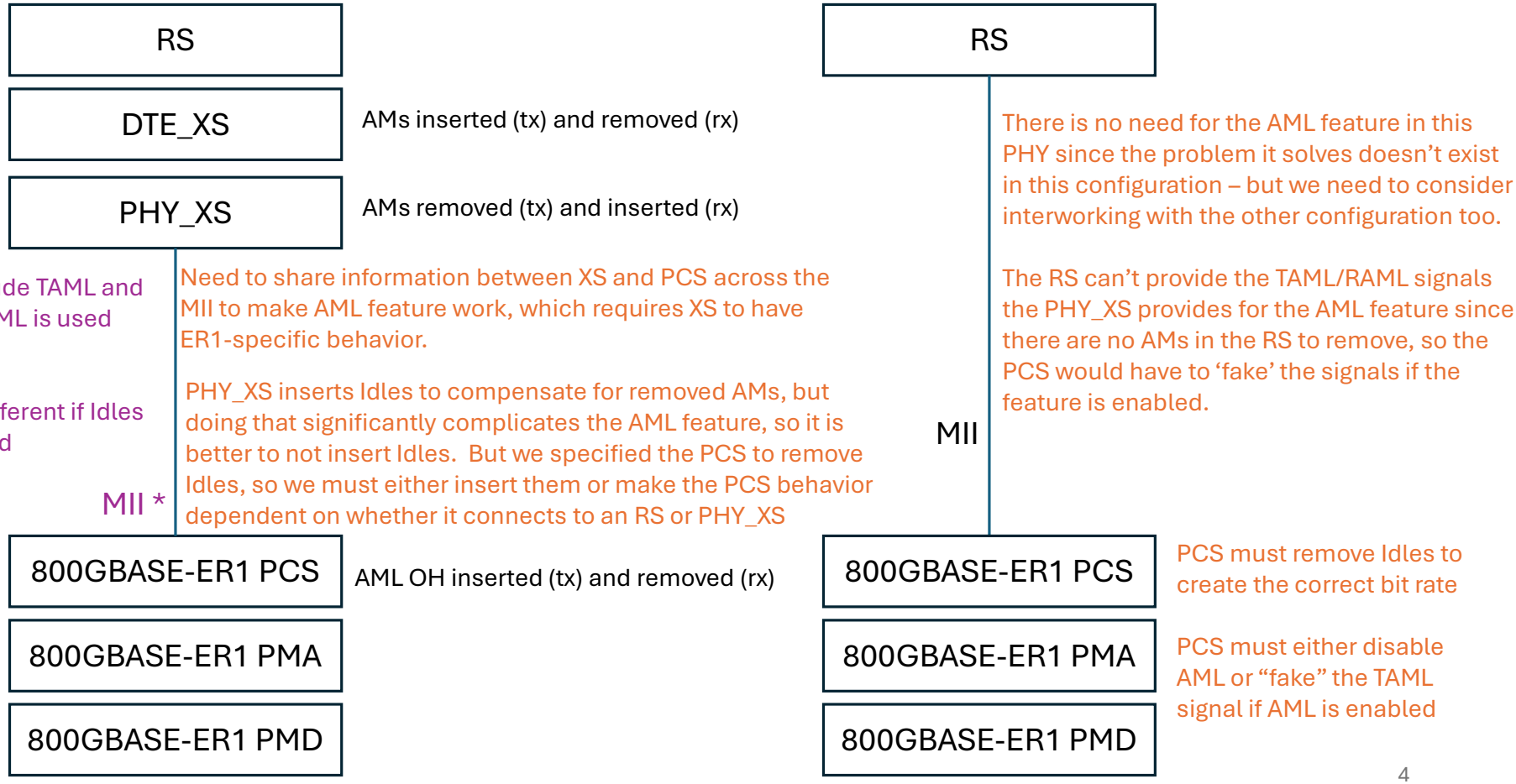
Supporters

- John D’Ambrosia – Futurewei, US subsidiary of Huawei
- Matt Brown – Alphawave

How did we get here?

- As we continued to fill in the details for 800GBASE-ER1 and 800GBASE-ER1-20, using a separate ER1 PCS, the following challenges were discovered:
  - 800GBASE-ER1 and –ER1-20 are a separate PHY family from the other 800G PHYs
  - The feature for conveying alignment marker location requires introducing new modes of operation to the 800G XS
  - The description of the functions in the XS and ER1 PCS is not well-aligned with anticipated implementations because the interface between the XS and the PCS is the MII, whereas practical implementations are mapping 257b blocks directly into the ER1 frame structure
  - The use of the XS cannot be mandated, which increases complexity in the PCS
- These challenges introduce potential schedule risk for WG ballot in May 2025
- There is an alternative architecture that avoids the complexity and mitigates the schedule risk

Inconsistent behavior in current 800GBASE-ER1 architecture

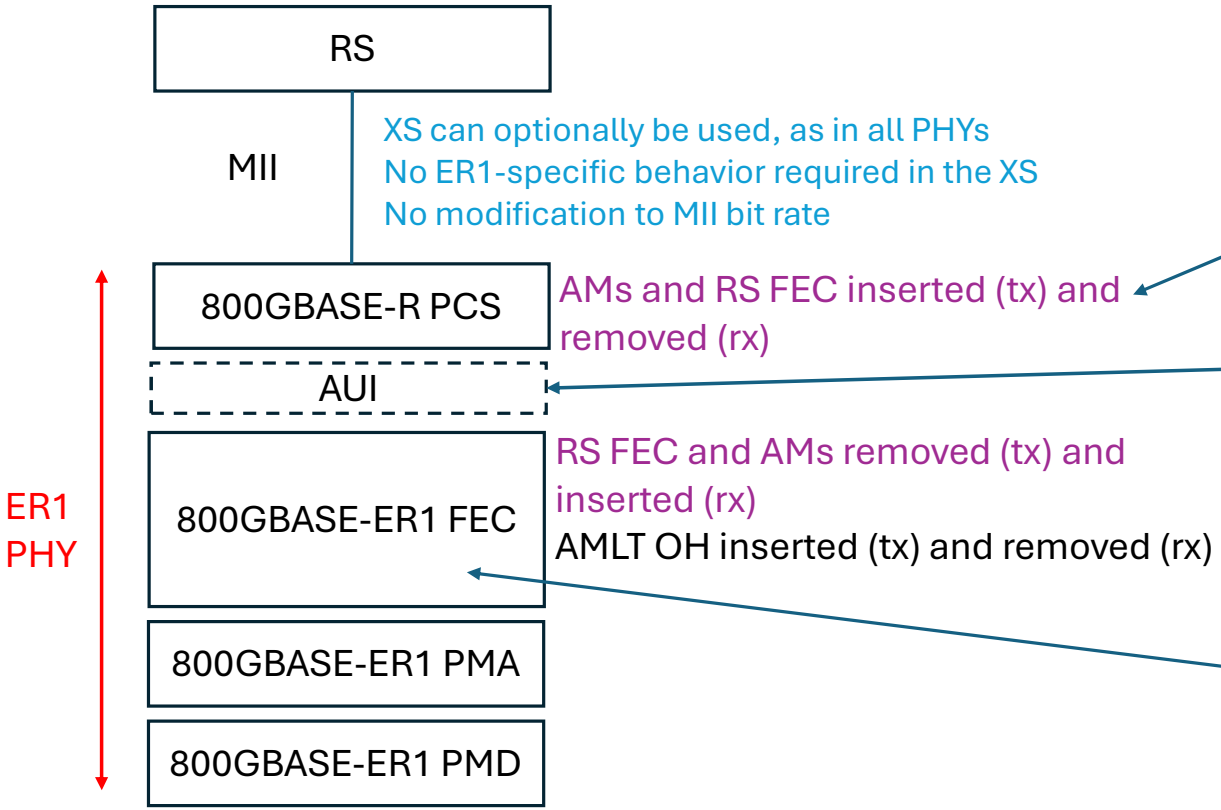


Alternative architecture for 800GBASE-ER1

- The original rationale for treating ER1 as a separate PCS was based on ER1 using a different FEC from 800GBASE-R on the ER1 link
  - Implementations using pluggable modules would utilize a segmented FEC architecture since the AUI will use RS(544,514) FEC
  - Use of the XS and a separate PCS was seen as a natural way to support this segmented FEC architecture
- An alternative is to use the 800GBASE-R PCS and define an 800GBASE-ER1 FEC sublayer that:
  - Terminates the RS FEC
  - Maps 257b blocks into the ER1 frame and adds the ER1 FEC
- The alternative doesn't require changing the technical baselines; it is purely a change in how we document the PHY to better align with how people will implement it

Alternative model for 800GBASE-ER1 based on FEC sublayer

This architecture aligns with 800GBASE-FR4, -LR4, and -LR1

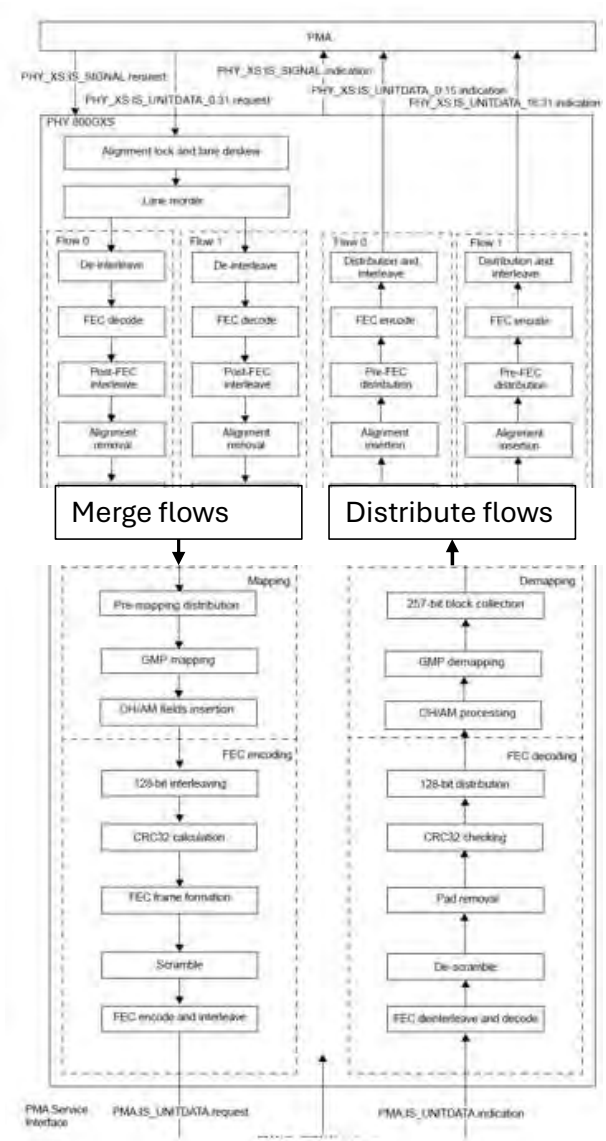


If PCS and FEC sublayers are in the same device (no AUI), purple functions can be omitted in implementation

Most implementations will have at least one AUI, which requires the RS FEC.

Description of ER1-specific functions (including AML feature) is contained in one sublayer and can be done based on 257b blocks, as it will be implemented, rather than based on MII transactions

New ER1 FEC sublayer

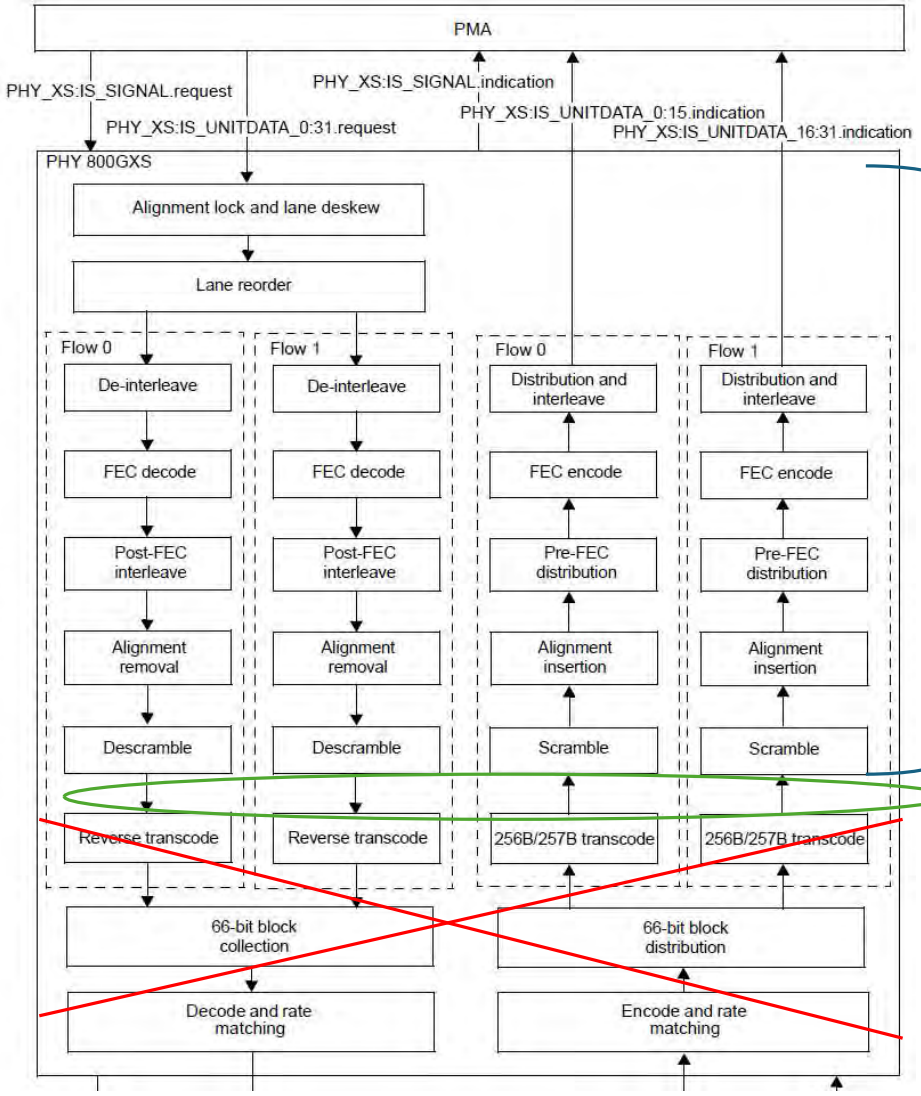


Functions defined by reference to the PHY_XS

New functions to 802.3dj, but defined already in OIF and ITU-T

Functions currently in 800GBASE-ER1 PCS

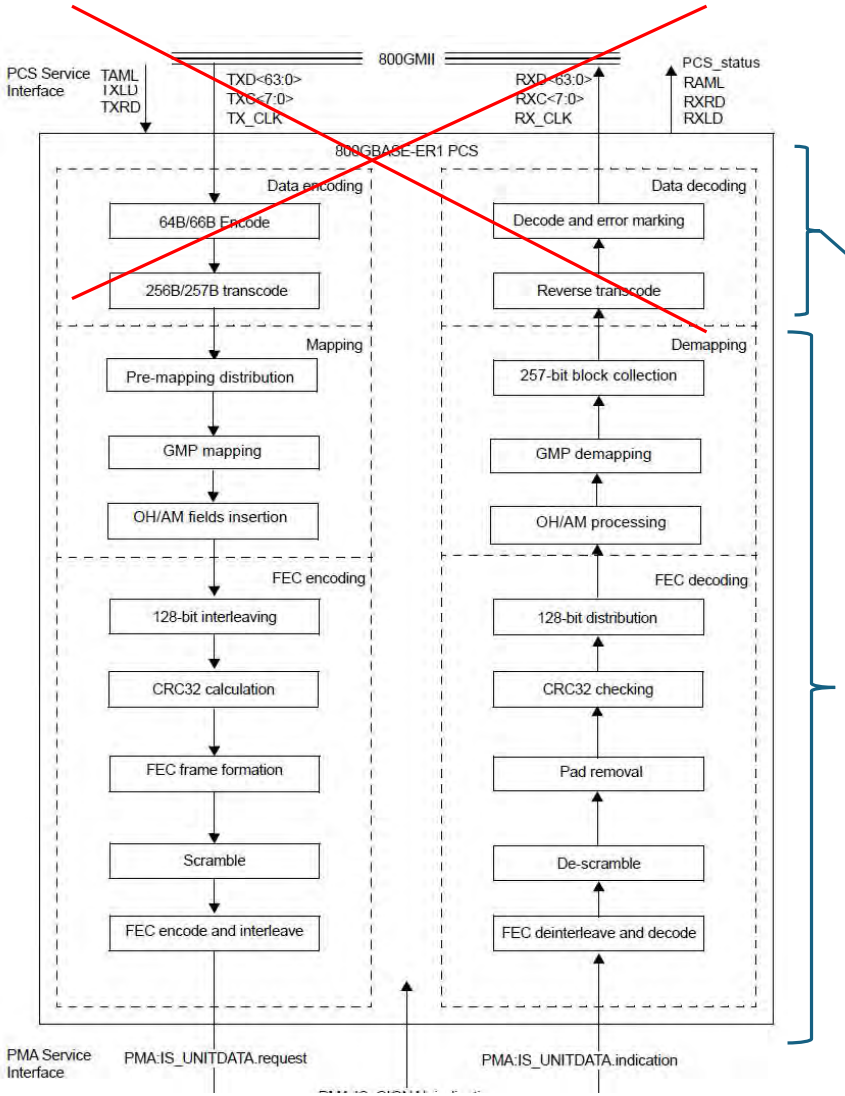
PHY_XS functions used in ER1 FEC



These functions are providing an “inverse RS FEC” function within the the ER1 FEC sublayer.

They can all be specified by reference to the PHY_XS.

Add “flow merge”/”flow distribute” functions here to merge the two flows to a single stream of 257b blocks in the Tx path and distribute the single path to two flows in the Rx path



ER1 PCS functions used in ER1 FEC

Replace this Data encoding/decoding block with a “Inverse RS FEC” block that uses functions from the PHY_XS (see previous slide) plus a “flow merge” function to combine the two flows into a single 257b block stream

No changes needed as a result of changing the architecture to a FEC sublayer, other than in description of AML field in the OH/AM fields insertion function

AML feature details

- The AML feature is part of the “OH/AM fields insertion” and “OH/AM processing” functions
- With the architecture change, the feature can be described based on 257b blocks rather than 66b blocks
- We still need to do many of the things in https://www.ieee802.org/3/dj/public/adhoc/optics/1224_OPTX/slavick_3dj_optx_01_241219.pdf to complete specification of the AML feature, but they will be done in the context of the single FEC sublayer rather than split across the XS and PCS with communication via signals at the MII

Proposal

- Change the model of ER1 from a PCS to a FEC sublayer
  - All 800G PHYs use the same PCS
  - All 800G PHYs use the same optional XS
  - FEC-specific aspects are captured in FEC sublayers for all PHYs that use FEC beyond/other than RS(544,514)
  - Better alignment with real-world implementation enables simpler description
- This is purely a change in documentation structure, not a change in functionality
- Implement based on slides 13-29, with editorial license

Details of changes to be made

Clause 169 changes (1)

- Update Table 169-1 to indicate that the 800GBASE-ER1 and ER1-20 PHYs use 800GBASE-R encoding
  - Change “using 800GBASE-ER1 PCS and PMA encoding” to “using 800GBASE-R and 800GBASE-ER1 PMA encoding”
- Update Table 169-3a:
  - Change the title of clause 186
  - In the ER1-20 and ER1 rows, include clauses 172, 173, 120F, 120G, 176, 176C, 176D with the same status as for LR1
- Remove all changes to 169.2.3 (i.e., remove 169.2.3 from the document completely since it doesn’t need to change)

Clause 169 changes (2)

- Add a new 169.2.4c to discuss the ER1 FEC sublayer

  169.2.4c 800GBASE-ER1 FEC sublayer

  The 800GBASE-ER1 FEC sublayer terminates the FEC added by the 800GBASE-R PCS and adds a different FEC for use over the 800GBASE-ER1 or 800GBASE-ER1-20 PMD. The 800GBASE-ER1 FEC is specified in Clause 186.

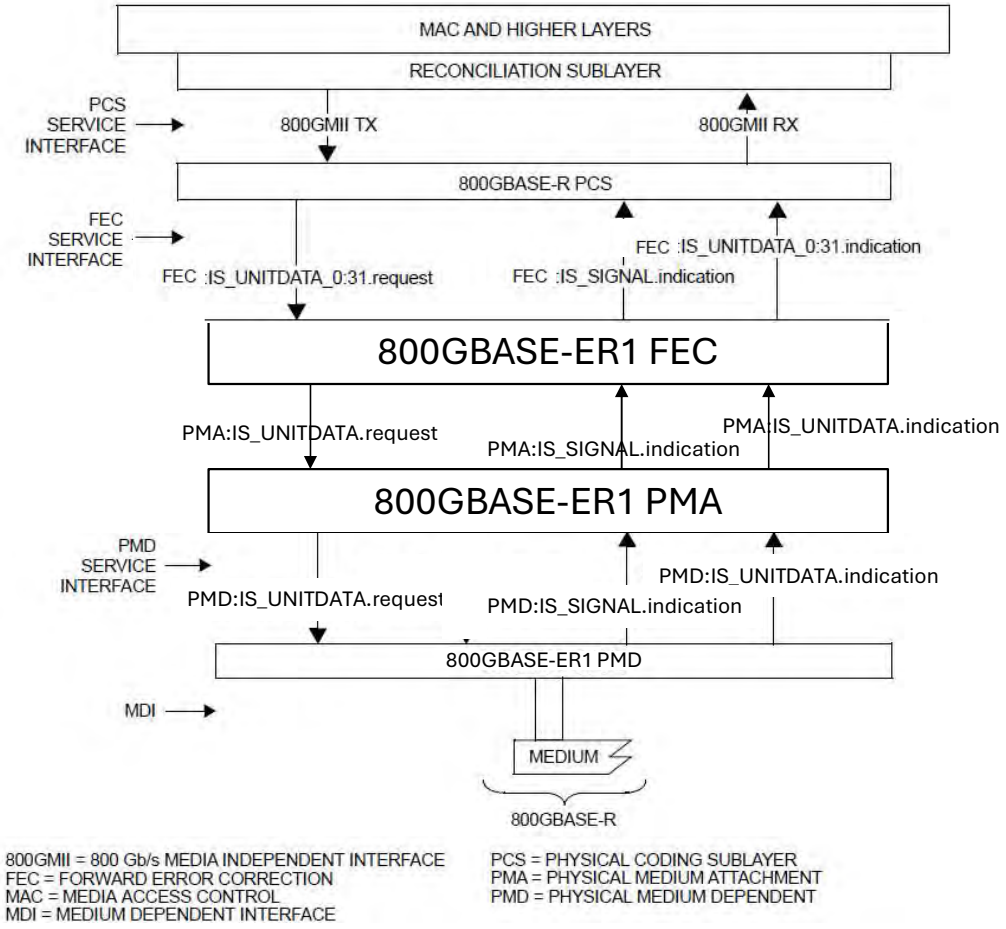
- Remove the sentence inserted in 169.3 about PHYs that are not part of the 800GBASE-R family
- In the new FEC bullet in 169.3.2, add “or 800GBASE-ER1 FEC” after “Inner FEC”

Clause 169 changes (3)

- Add a new Figure 169-2b illustrating the layer stack for ER1 and ER1-20 (see slide 16)
- Add a row to Table 169-4 for ER1 FEC and ER1 PMA:

Sublayer	Maximum (bit time)	Maximum (pause_quanta)	Maximum (ns)	Notes
800GBASE-ER1 FEC and 800GBASE-ER1 PMA	TBD	TBD	TBD	See 186.5

Figure 169-2b



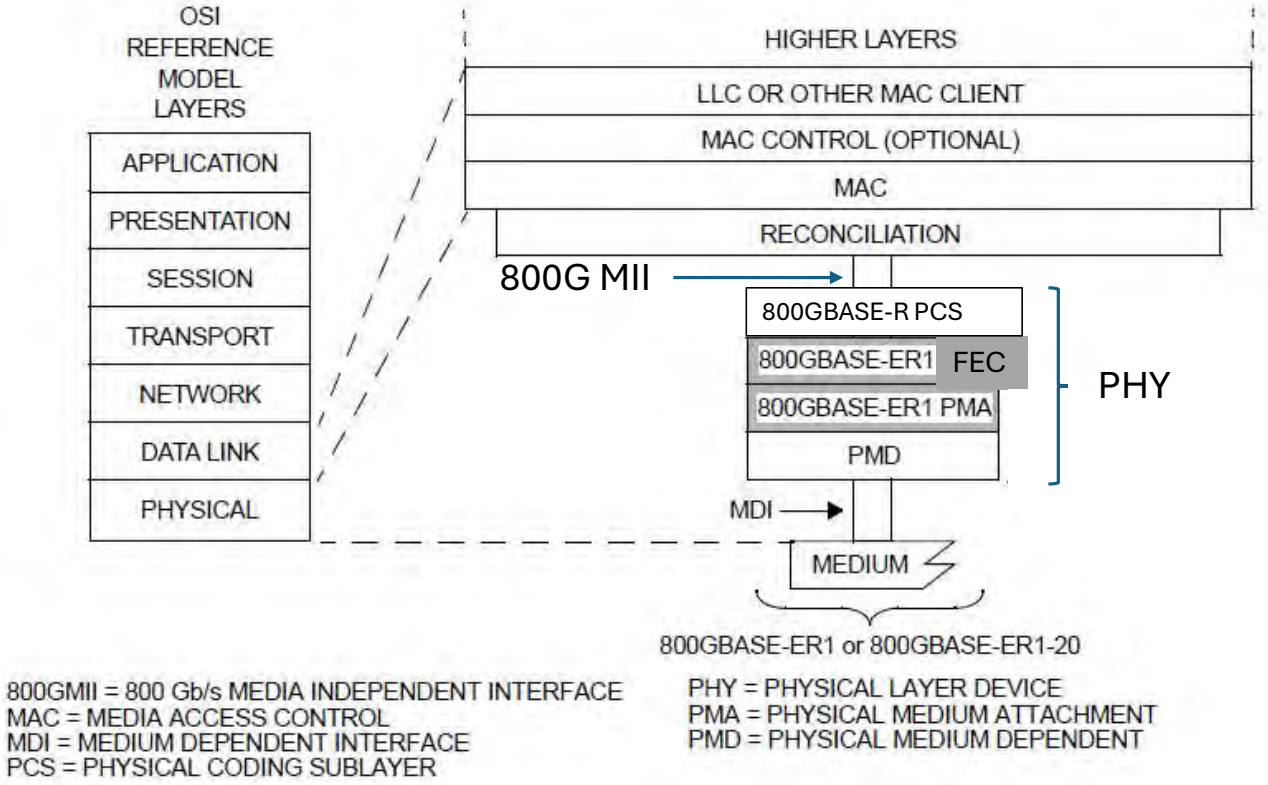
Clause 171 changes

- Remove everything related to the ER1 PCS alignment marker location transparency feature in these subclauses:
  - 171.1.1: remove the bullet in the first list about optional support for enhanced PTP accuracy
  - 171.3:
    - Remove the bullet in the first list about PHY 800GXS
    - Remove the instruction to replace Figure 171-2 and the replacement figure
    - Remove 171.3.3 (there are no longer any changes being made)
  - Remove new subclause 171.6a, renumber new subclause 171.6b to 171.6a
  - Remove Table 171-2 (there are no longer any changes being made)
  - Remove “Enhanced PTP accuracy ability” row from Table 171-3
  - Remove new subclause 171.9.4.6a, renumber new subclause 171.9.4.6b to 191.9.4.6a

Clause 186, 186.1.1, 186.1.2

- Change title to “FEC sublayer (FEC) and Physical Medium Attachment (PMA), type 800GBASE-ER1”
- 186.1.1: Change to “This clause specifies the Forward Error Correction (FEC) sublayer and Physical Medium Attachment (PMA) sublayer for the Physical Layer implementations known as 800GBASE-ER1-20 and 800GBASE-ER1. The 800GBASE-ER1-20 and 800GBASE-ER1 PHYs (see 187.1) use the 800GBASE-R PCS defined in clause 172 and the 800GBASE-ER1 FEC and 800GBASE-ER1 PMA defined in this clause.”
- 186.1.2:
  - Change first sentence: “Figure 186-1 shows the relationship of the 800GBASE-ER1 FEC and 800GBASE-ER1 PMA sublayers (shown shaded), along with other Ethernet sublayers, …”
  - Revise Figure 186-1 to replace the 800GBASE-ER1 PCS with the 800GBASE-R PCS (not shaded) and 800GBASE-ER1 FEC (shaded) sublayers

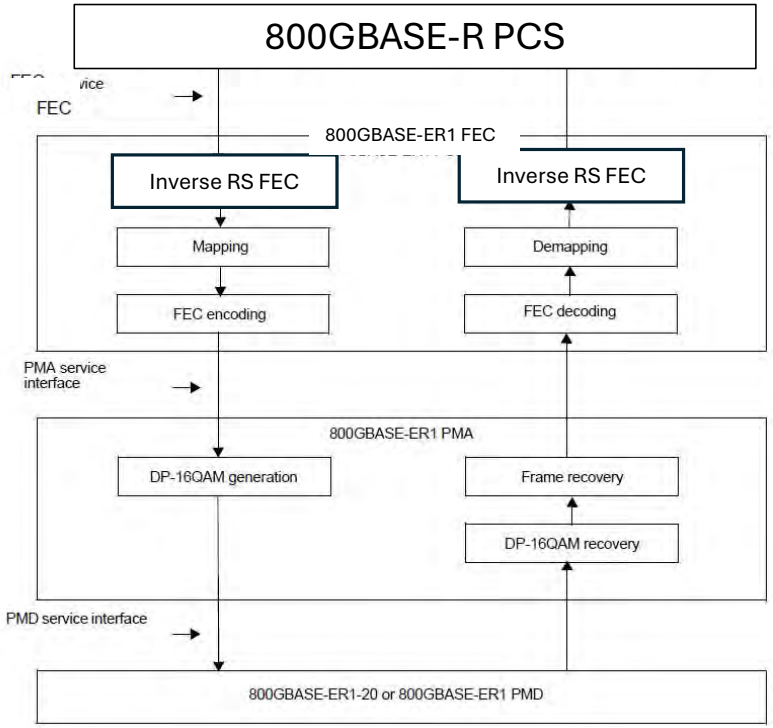
Revised figure 186-1



Clause 186.1.3 (1)

- Rewrite the text: “In the transmit direction, the 800GBASE-ER1 FEC and 800GBASE-ER1 PMA together provide termination of the RS(544,514) FEC, mapping into the ER1 FEC frame, FEC encoding, and generation of dual polarization, 16-state quadrature amplitude modulation (DP-16QAM) symbols at the 800GBASE-ER1-20 or 800GBASE-ER1 PMD service interface. In the receive direction the 800GBASE-ER1 PMA and 800GBASE-ER1 FEC together decode DP-16QAM symbols from the 800GBASE-ER1-20 or 800GBASE-ER1 PMD service interface, perform FEC error detection and correction, and de-mapping from the ER1 FEC frame. A high-level block diagram of the 800GBASE-ER1 FEC and the 800GBASE-ER1 PMA is shown in Figure 186-2.
- Revise Figure 186-2 to reflect the new architecture
  - The 800GBASE-R PCS should be at the top, with an optional AUI.
  - Change PCS to FEC, and change the first block from “Data encoding” to “Inverse RS FEC”

Revised figure 186-2…



Clauses 186.2, 186.2.1, 186.2.2 (FEC sublayer overview and service interface)

- Change title to “800GBASE-ER1 FEC sublayer”
- 186.2.1: change the service interface to align with how the inner FEC service interface is specified in clauses 177 and 184
- 186.2.2:
  - Change Figure 186-3 to align with what is on slide 7
  - Revise the text to align with the new figure:
    - The transmit receives 32 PCS lanes (which can come from the PCS, PHY_XS, or n:32 PMA), aligns and reorders as necessary, terminates the RS FEC, removes alignment markers, merges the two flows, distributes to 8 streams and maps into the ER1 frame, adds the ER1 FEC, and interleaves to a single stream.
    - The receive direction deinterleaves to 8 streams, terminates the ER1 FEC, de-maps from the ER1 frame, interleaves to one stream, distributes to two flows, adds alignment markers, and adds RS FEC.

Clause 186.2.3 (FEC sublayer TX direction)

- Replace subclause 186.2.3.1 with an “inverse RS FEC” subclause that describes alignment lock, deskew, lane reordering, and per-flow deinterleaving, FEC decoding, and re-interleaving, alignment marker removal, and descrambling by reference to the appropriate clauses in the PHY_800GXS
- Replace 186.2.3.2 with a “Flow merge” subclause that merges the two flows into a single stream of 257b blocks
- Change “PCS” to “FEC” in all other subclauses 186.2.3.x
- In 186.2.3.4, clarify that the first block in every GMP frame comes from flow0

Clause 186.2.3.6.10 (alignment marker location OH)

- Change feature name from ‘enhanced PTP accuracy’ to ‘alignment marker location’
- Remove references to the XS
- If the AML feature is implemented (ability = 1) and enabled (enabled = 1):
  - The FEC sublayer maintains a counter of 257b blocks inserted into GMP frames
  - This counter is reset whenever AMs are removed by the FEC sublayer
  - The value of the counter corresponding to the first payload block in the GMP multi-frame is inserted into the AML overhead.
  - The AML overhead provides, for each GMP multi-frame, the number of 257b blocks since AMs were removed prior to the start of that GMP multi-frame
- If the feature is not implemented or not enabled, the value 0 is inserted in the AML overhead of each GMP multi-frame
- Update figure 186-7 to remove 66b blocks at the top, and change the label in the top right

Clause 186.2.4 (FEC sublayer RX direction)

- Change PCS to FEC in all 186.2.4.x subclauses
- Replace 186.2.4.9 with a “Flow distribution” function to distribute the de-mapped 257b blocks to the two flows (first to flow0, then to flow1)
- Replace 186.2.4.10 with an “inverse RS FEC” subclause that specifies per-flow scrambling, AM insertion, pre-FEC distribution, FEC encoding, and post-FEC distribution and interleaving via reference to the PHY_800GXS

Clause 186.2.4.6.5 (alignment marker location OH)

- A counter of 257b blocks is maintained to control insertion of AMs (this is necessary whether or not AML is implemented, since the RX has to insert AMs)
  - If the AML feature is not implemented or not enabled, the AM insertion counter simply free-runs
- If the AML feature is implemented (ability = 1) and enabled (enabled = 1):
  - AML overhead is extracted
  - Since the number of 257b blocks between AMs is fixed, the AML value can be predicted once the first value has been received; to mitigate against bit errors, multiple “good” values should be used to confirm a new AM position
  - A state machine is used to lock the counter to the received AML overhead values (based on slide 20 of https://www.ieee802.org/3/dj/public/adhoc/optics/1224_OPTX/slavick_3dj_optx_01_241219.pdf)
- Remove references to the XS

Clauses 186.5 – 186.7 (Delay constraints, Path data delay, Management)

- 186.5: Change PCS to FEC
- 186.6: Change PCS to FEC
- 186.6.1:
  - Change PCS to FEC in title
  - Update all the register names from PCS_delay_… to FEC_delay_…
- 186.7.1:
  - Change PCS to FEC in text and register names
  - Change “enhanced_ptp_accuracy_enable” to “am_location_feature_enable”

Clause 186.8.3 (PICS)

- 186.8.3: change title from PCS to FEC
- 186.8.3.2 (transmit function)
  - Remove *SE and TF1 – TF4
  - Insert new TF1 for inverse RS FEC
  - Insert new TF2 for flow merge
  - Renumber TF5 and later to TF3 and later
- 186.8.3.3 (receive function)
  - Remove *SD and RF9 – RF12
  - Add new RF9 for flow distribution
  - Add new RF10 for inverse RS FEC
- 186.3.4: Change PCS to FEC

Clause 187.1

- Update Figure 187-1 based on new Figure 186-1, replacing 800GBASE-ER1 PCS with 800GBASE-R PCS and 800GBASE-ER1 FEC sublayers

Thank you!