

Baseline proposal for PMAs with 200G per lane signaling

Adee Ran (Cisco), Daniel Koehler (Synopsys), Gary Nicholl (Cisco), Jeff Slavick (Broadcom), Mark Gustlin (Cisco), David Ofelt (Juniper), Xiang He (Huawei), Zvi Rechtman (NVIDIA), Shawn Nicholl (AMD), Matt Brown (Huawei)

Support

- Eric Maniloff, Ciena
- Ted Sprague, Infinera
- Leon Bruckman, Huawei
- Yan Zhuang, Huawei
- Kent Lusted, Intel
- Rick Rabinovich, Keysight Technologies
- Viet Tran, Keysight Technologies
- Ali Ghiasi, Ghiasi Quantum LLC
- Shimon Muller, Enfabrica
- Henry Wong, Alphawave
- David Cassan, Alphawave
- Vasudevan Parthasarathy, Broadcom
- Eugene Opsasnick, Broadcom

Background

- PCS and PMAs for 800G at 100 Gb/s per lane have been defined by 802.3df (clauses 172 and 173)
- The 1.6TBASE-R PCS in [gustlin 3dj 01b 230206](#) has been adopted by motion #10 in the P802.3dj January 2023 electronic meeting
 - Supplement proposal for 1.6T PCS Lane Formation: [opsasnick 3dj 01 2303](#)
- [ran 3dj 01a 230206](#) presented the motivation, architecture concept, and FEC performance results for symbol-pair muxing PMA for 200G, 400G, and 800G PHYs at 200 Gb/s per lane
 - Additional earlier contributions are listed in that presentation
- **Straw poll #7 in the P802.3dj January 2023 electronic meeting indicated support for the symbol-pair multiplexing direction**
 - **Y: 43 , N: 4 , NMI: 32**

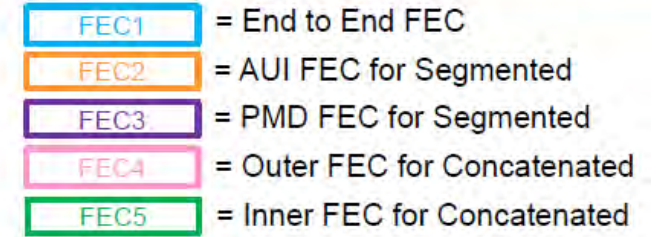
This presentation

- A full baseline proposal for multiple PMAs in 802.3dj: 200G, 400G, 800G, 1.6T
- Symbol muxing (in pairs or quartets) on all 200G per lane AUIs (and possibly on some PMD interfaces)

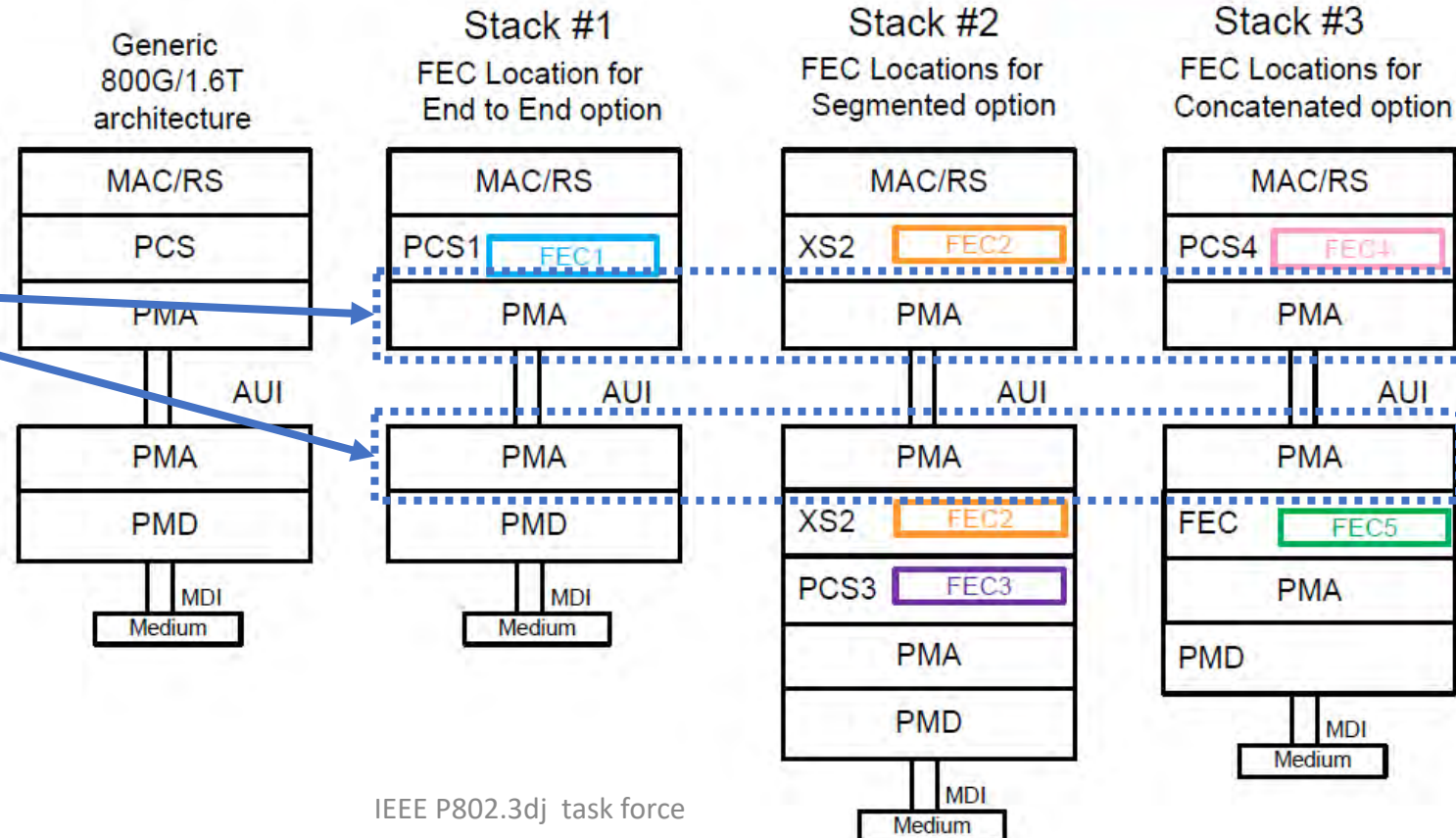
Adopted Logic Architecture for Reference

Proposed 800GbE/1.6TbE Architecture

- How various FEC schemes fit into the architecture
- FECs might or might not be reused across schemes



This proposal covers the PMA definition for PMAs surrounding the 200G/lane AUI



New proposed clause structure

1. Overview
 - Scope, sublayer positioning, general concepts, delay constraints
 2. Nomenclature and conventions
 3. Service interface and interface below
 - Including subclauses related to physical instantiations (CDR, clocking, signal drivers) – with references to AUI annexes
 - Tolerance of skew variation
 4. PMA functions (detailed in subsequent slides)
 5. State diagrams and related variables, functions, etc.
 6. MDIO mapping
 7. PICS
- Not included in this proposal:
 - Informative annex: PHY and XS partitioning examples
 - Normative annex(es): AUIs

Green denotes subclauses not included in full detail in this presentation. Editorial license is assumed

Nomenclature and conventions

- Proposed definitions
 - Use n and m as indexes of input and output lanes, with N and M as the total numbers, respectively; p as the index of PCSs on each lane, with P as the total
 - Group “forward” PMAs into families using lane muxing ratios (R8F, R2F, R1)
 - For example, R2f means that in the Tx direction, 2 PMA lanes at the input are muxed to 1 PMA lane at the output, and therefore $N=2M$
 - Denote the “backward” PMAs as R8B and R2B (the reversed versions of R8F and R2F respectively, e.g., for PMAs attached to a PHY XS)
 - “Backward PMAs” are practically the same physical devices as “Forward PMAs” but described differently in standard language!
- There are 18 named PMAs
 - 7 “Forward” PMAs
 - R8F: 32:4, 16:2, 8:1 (800G, 400G, and 200G, respectively, attached to a PCS or DTE XS)
 - R2F: 8:4, 4:2, 2:1 (800G, 400G, and 200G, respectively)
 - 16:8 (1.6T) (ratio is 2, but unlike R2F, symbol muxing on both input and output)
 - 7 “Backward” PMAs
 - R8B, R2B: Same as R8F and R2F but in reversed direction
 - 8:16 (1.6T) : Same as 16:8 but in reversed direction
 - 4 R1: 8:8, 4:4, 2:2, 1:1 (retimers for 1.6T, 800G, 400G, and 200G, respectively)

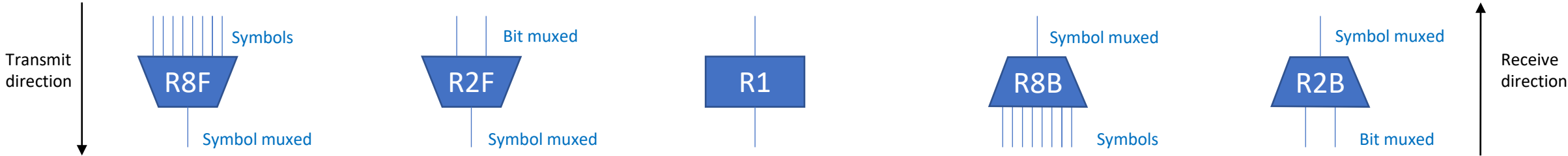
Nomenclature table

PMA family	Specific PMAs	Sublayer/interface above	Sublayer/Interface below
R8F	PMA(8:1)	200GBASE-R PCS / DTE 200GXS	200GAUI-1 / PMD
	PMA(16:2)	400GBASE-R PCS / DTE 400GXS	400GAUI-2 / PMD
	PMA(32:4)	800GBASE-R PCS / DTE 800GXS	800GAUI-4 / PMD
R2F	PMA(2:1)	200GAUI-2	200GAUI-1 / PMD
	PMA(4:2)	400GAUI-4	400GAUI-2 / PMD
	PMA(8:4)	800GAUI-8	800GAUI-4 / PMD
R1	PMA(1:1)	200GAUI-1	200GAUI-1 / PMD
	PMA(2:2)	400GAUI-2	400GAUI-2 / PMD
	PMA(4:4)	800GAUI-4	800GAUI-4 / PMD
	PMA(8:8)	1.6TAUI-8	1.6TAUI-8 / PMD
	PMA(16:8)	1.6TBASE-R PCS / DTE 1.6TXS / 1.6TAUI-16	1.6TAUI-8 / PMD

Nomenclature table – Backward PMAs

PMA family	Specific PMAs	Sublayer/interface above	Sublayer/Interface below
R8B	PMA(1:8)	200GAUI-1	PHY 200GXS
	PMA(2:16)	400GAUI-2	PHY 400GXS
	PMA(4:32)	800GAUI-4	PHY 800GXS
R2B	PMA(1:2)	200GAUI-1	200GAUI-2
	PMA(2:4)	400GAUI-2	400GAUI-4
	PMA(4:8)	800GAUI-4	800GAUI-8
	PMA(8:16)	1.6TAUI-8	1.6TAUI-16 / PHY 1.6TXS

Backward PMAs are equivalent to their forward counterparts with transmit and receive directions swapped



PMA functions

- Multiplexing methods
 - Bit-wise multiplexing and demultiplexing
 - Symbol-wise multiplexing and demultiplexing
- Symbol-pair and symbol-quartet alignment
 - Text, illustration, state diagram
- PMA input processing
 - Flow diagrams
 - Subclause per PMA category (e.g., 8:1) or per direction (Tx/Rx)
- PMA output processing
 - Flow diagrams
 - Subclause per PMA category (e.g., 8:1) or per direction (Tx/Rx)
- PAM4 encoding and decoding (based on existing PMAs)
 - Precoding capability in all physically instantiated interfaces is “Tx:required, Rx:optional” (see 135.5.7.2)
- Signal status (based on existing PMAs)
- Loopback modes (based on existing PMAs)
- Test patterns (based on existing PMAs)

Green denotes subclauses not included in full detail in this presentation. Editorial license is assumed

Multiplexing methods

- **Symbol-pair multiplexing**

- R2F and R8F PMAs in the transmit direction

- **Bitwise multiplexing**

- R2F PMAs in the receive direction

- **Symbol-quartet multiplexing**

- PMA(16:8) in the transmit direction

- **Symbol-pair de-multiplexing**

- R2F and R8F PMAs in the receive direction

- **Bitwise de-multiplexing**

- R2F PMAs in the transmit direction

- **Symbol-quartet de-multiplexing**

- PMA(16:8) in the receive direction

Backward PMAs are equivalent to their forward counterparts with transmit and receive directions swapped

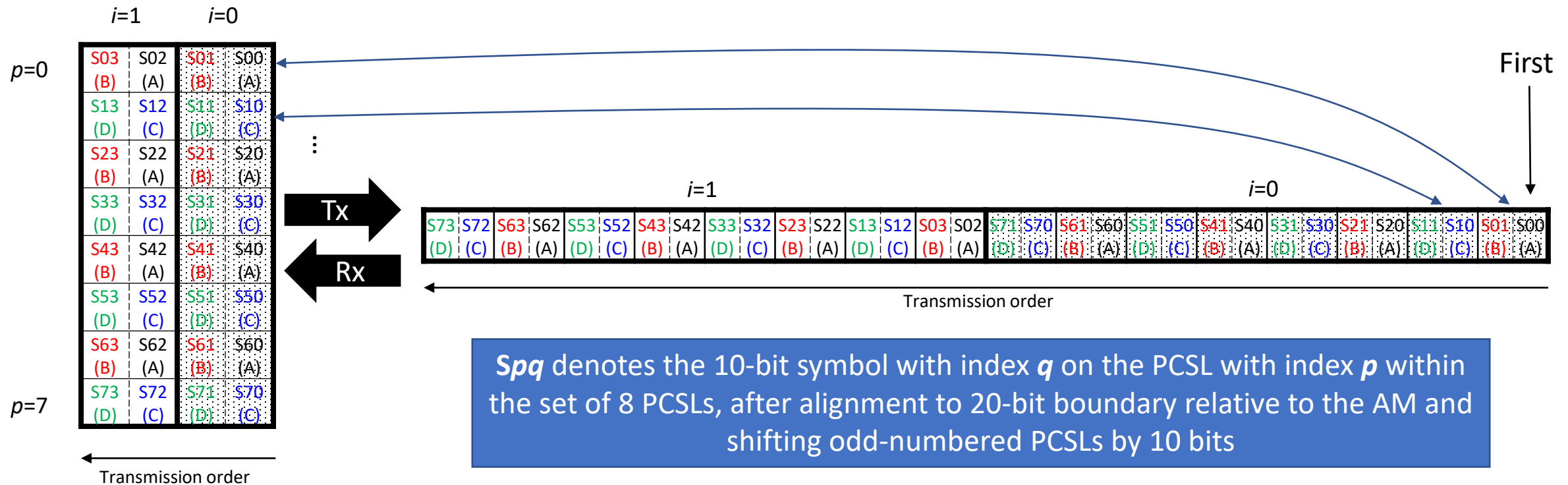
- **PAM4 symbol relay from input to output**

- R1 PMAs in both transmit and receive directions

Symbol-pair multiplexing illustration

PMA(32:4) – 800G PHYs

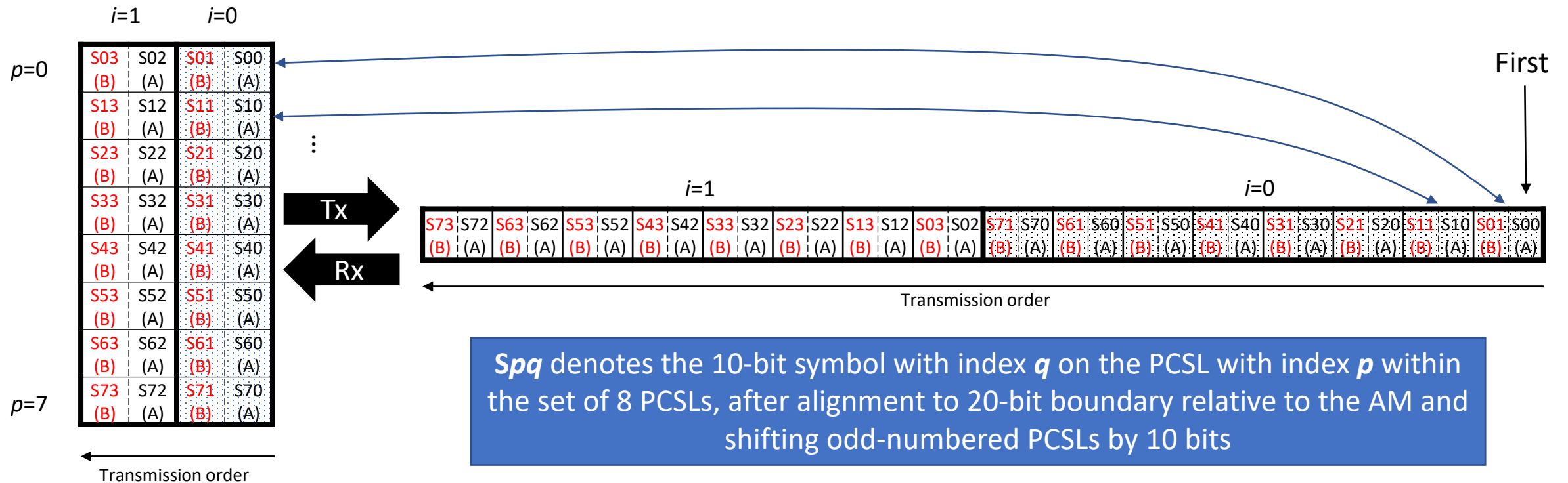
(1 out of 4 lanes illustrated)



Symbol-pair multiplexing illustration

PMA(16:2), PMA(8:1) – 400G and 200G PHYs

(For PMA(16:2), 1 out of 2 lanes illustrated)



R8F/R8B and R2F/R2B symbol-pair multiplexing rules

Putting the diagrams into standard language...

R8F and R2F Transmit direction

The following rule applies for multiplexing eight PCSLs into a PMA output lane in R8F and R2F PMAs.

In R2F PMAs, the bit streams corresponding to PCSLs are obtained by bitwise de-multiplexing of the input lanes. In R8F PMAs, the bit streams corresponding to PCSLs are obtained directly from the service interface. The bit streams are shifted as required to align to a 20-bit boundary relative to the AM on each PCSL. Odd-numbered PCSLs are shifted by additional 10 bits relative to the AM.

Denote the index of the PCSL within the eight PCSLs as p ($p=0$ to 7), where **for 800GBASE-R PMAs, even values of p are from PCS flow 0 (PCSLs 0 to 15) and odd values of p are from PCS flow 1 (PCSLs 16 to 31)**. Denote the index of the aligned input 20-bit block on PCSL p as i and the bit index within block i as j ($j=0$ to 19). The mapping between the bit sequences on the eight PCSLs and the output bit sequence is:

$$output[160i + 20p + j] = PCSL[p, 20i + j]$$

R8F and R2F Receive direction

The following rule applies for de-multiplexing a PMA input lane into eight PCSLs in R8F and R2F PMAs.

The input bit stream is shifted as required to align to a 20-bit boundary relative to the AMs on all PCSLs. Denote the index of the PCSL within the eight PCSLs as p ($p=0$ to 7), the index of the aligned output 20-bit block on PCSL p as i , and the bit index within block i as j ($j=0$ to 19). The mapping between input bit sequence and the bit sequences on the eight PCSLs is:

$$PCSL[p, 20i + j] = input[160i + 20p + j]$$

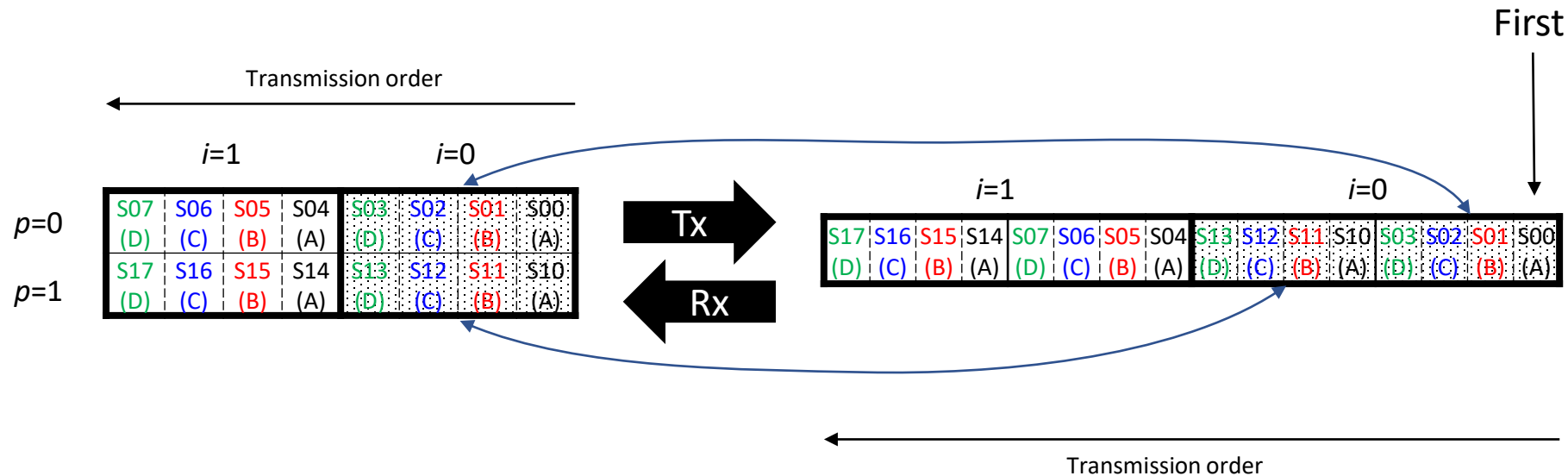
For R8B and R2B PMAs, the rules are the same as those for R8F and R2F PMAs, respectively, with the transmit and receive directions swapped.

Restricted

Symbol-quartet multiplexing illustration

PMA(16:8) – 1.6T PHYs

(1 out of 8 lanes illustrated)



S_{pq} denotes the 10-bit symbol with index q on the PCSL with index p within the set of 2 PCSLs, after alignment to 40-bit boundary relative to the AM

PMA(16:8) and PMA(8:16) symbol-quartet multiplexing rules

PMA(16:8) Transmit direction

The following rule applies for multiplexing two PCSLs into a PMA output lane in PMA(16:8).

The bit streams corresponding to PCSLs are obtained directly from the service interface. The bit streams are shifted as required to align to a 40-bit boundary relative to the AM on each PCSL.

Denote the PCSL index within a set of two PCSLs as p ($p=0$ to 1), the index of a 40-bit input block on PCSL p as i , and the bit index within block i as j ($j=0$ to 39). The mapping between the bit sequences on the two input PCSLs and the output bit sequence is:

$$output[80i + 40p + j] = PCSL[p, 40i + j]$$

PMA(16:8) Receive direction

The following rule applies for de-multiplexing a PMA input lane into two PCSLs in PMA(16:8).

The input bit stream is shifted as required to align to a 40-bit boundary relative to the AMs on all PCSLs. Denote the index of the PCSL within the set of PCSLs from the same input lane as p ($p=0$ to 1), the index of the aligned output 40-bit block on PCSL p as i , and the bit index within block i as j ($j=0$ to 39). The mapping between the input bit sequence and the bit sequences on the two output PCSLs is:

$$PCSL[p, 40i + j] = input[80i + 40p + j]$$

For PMA(8:16), the mappings are the same as for PMA(16:8) with the transmit and receive directions swapped.

R2F and R2B bitwise multiplexing rules

For R2F PMAs in the transmit direction, denote the lane index within a set of two input lanes as n ($n=0$ to 1), the PCSL index within the set of four input PCSLs on lane n as p ($p=0$ to 3), and the bit index within PCSL p as j . The mapping between the two input bit sequences and the bit sequences on the eight output PCSLs is:

$$PCSL[4n + p, j] = input[n, 4j + p]$$

For R2F PMAs in the receive direction, denote the lane index within a set of two output lanes as m ($m=0$ to 1). Denote the PCSL index within the set of four PCSLs allocated to output lane m as p ($p=0$ to 3), where **for 800GBASE-R, $p \in \{0, 1\}$ are from PCS flow 0 (PCSLs 0 to 15) and $p \in \{2, 3\}$ are from PCS flow 1 (PCSLs 16 to 31)**. Denote the bit index within PCSL p as j . The mapping between the eight input PCSLs and the two output bit sequences is:

Restricted

$$output[m, 4j + p] = PCSL[4m + p, j]$$

For R2B PMAs, the mappings are the same as for R2F PMAs with the transmit and receive directions swapped.

Symbol-pair alignment illustration

R2F PMA in the transmit direction

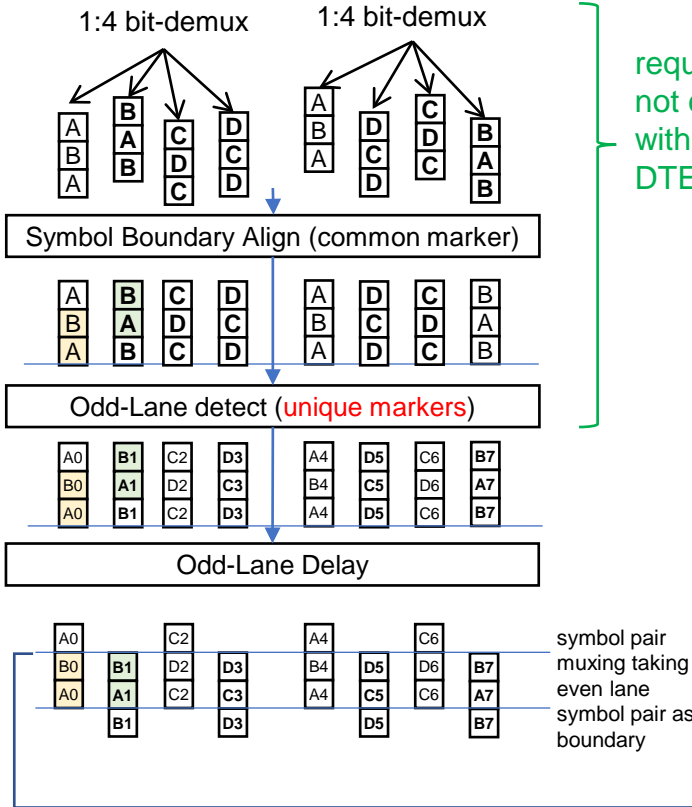
800G: 2-flows with 4 FEC symbols (A,B,C,D)

A0

A := codeword e.g. A,B,C,D

0..n := identify symbols from same PCSL, used for identifying odd/even, not a specific lane number

AB, CD
pairs
arbitrary



required only if
not co-located
with PCS or
DTE XS

Notes:

An 800G implementation would have four instances independent from each other.

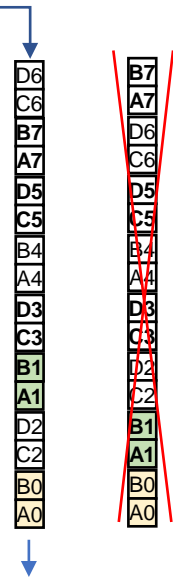
It is not possible to guarantee a mix of even and odd lanes exist in every such 1:4 de-muxed stream.

The odd lane delay must occur always, even if all lanes happen to be odd lanes only.

checkerboard
(even/odd arbitrary)

checkerboard
(even/odd identified)

striped
symbol
pattern



800G PMA must be able to identify and re-order incoming PCSLs to alternate between flows. Compliant input has both flows on every input lane.

For 200G and 400G PMAs, any muxing order can be used.

Symbol-pair alignment

- The following process is specified for R2F and R8F PMAs. It is applied to each group of physical lanes that are multiplexed together into one lane. The R2B and R8B process is equivalent, with input and output directions swapped.
- In the transmit direction, on each input lane:
 - In R2F PMAs only: De-multiplex to obtain a group of 4 PCSLs
 - Align the bit sequence on each PCSL to symbol-pair (20-bit) boundary relative to the AM
 - In odd-numbered PCSLs, delay the bit sequence by one symbol (10 bits)
 - This turns the “checkerboard” into a striped pattern
 - Perform symbol-pair multiplexing to the output lane
 - In 800G, muxing is required to **alternate symbol-pairs between flow 0 and flow 1** ← **Restricted**
 - In 200G/400G – any order.
- In the receive direction, on each input lane:
 - Find the symbol-pair alignment by locking on AM for all 8 PCSLs
 - Can be found using just the common part of the AMs.
 - There are 20 possible offsets that need to be checked. Only one will match all PCSLs.
 - Distribute symbol pairs to PCSLs based on the obtained alignment
 - In R2F PMAs only: bitwise-multiplex groups of 4 PCSLs.

Align to symbol pair/quartet (AM Lock)

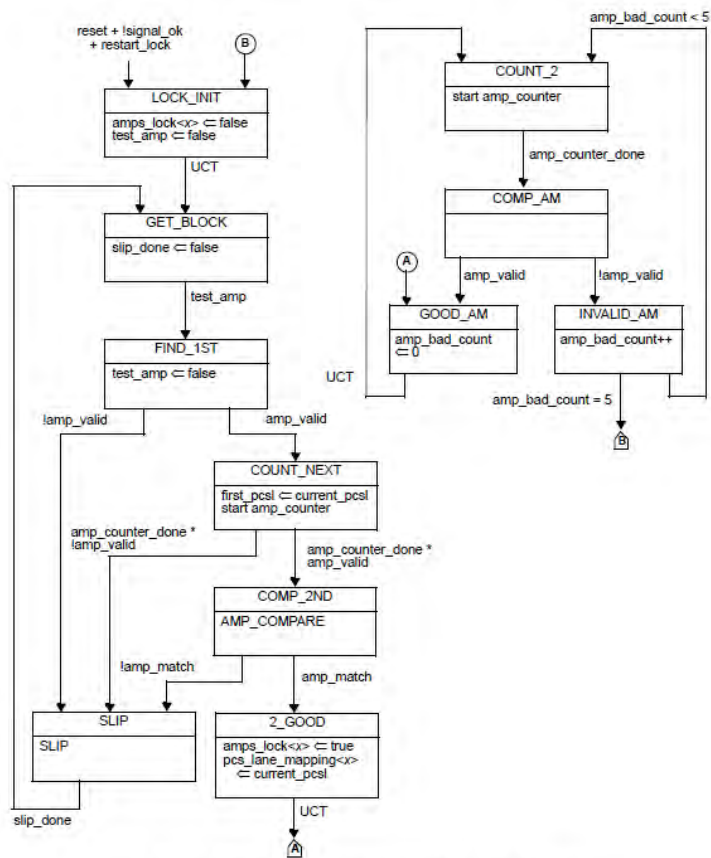
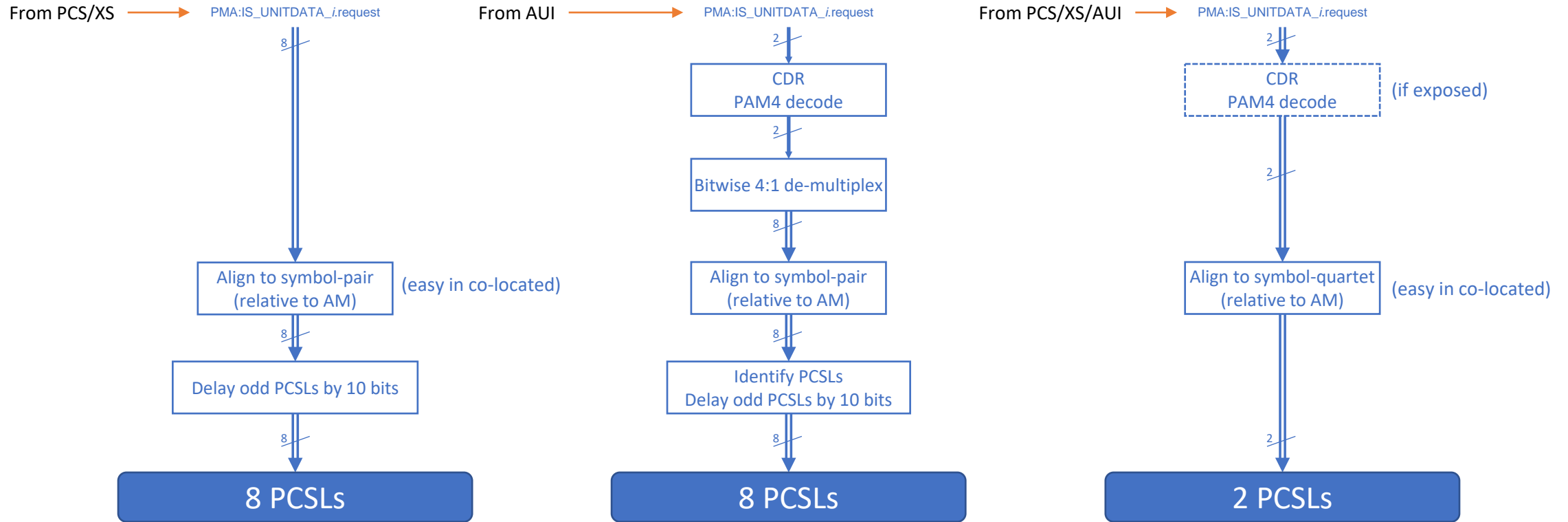


Figure 119-12—Alignment marker lock state diagram

- State diagram instance per PMA (physical) lane
- Specification is based on the Clause 119 AM lock state diagram
 - Exception is that the conditions for **amp_valid** and **amp_match** are that *AMs have been identified on all PCSLs on the PMA lane*
 - Counters should be extended as required.

PMA input processing – Transmit direction

(obtaining the aligned PCSLs)



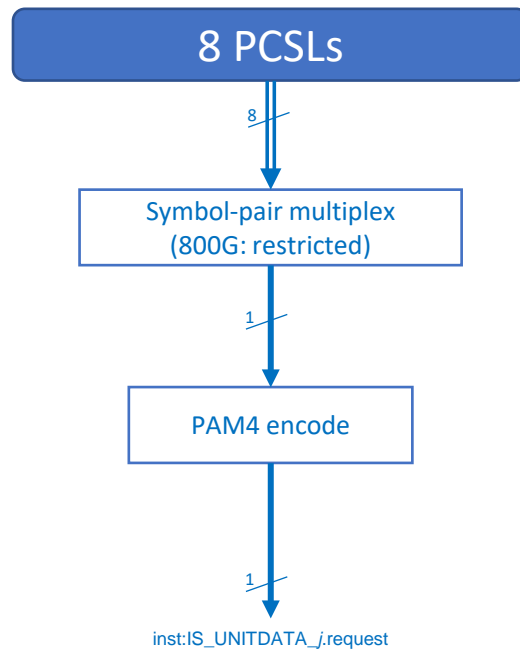
R8F PMA 1, 2, or 4 instances
(200G, 400G, or 800G)

R2F PMA 1, 2, or 4 instances
(200G, 400G, or 800G)

PMA(16:8) 8 instances
(1.6T)

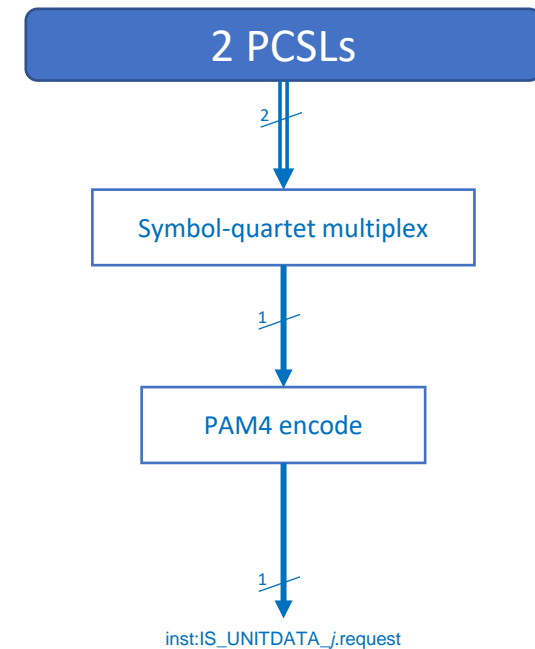
PMA output processing – Transmit direction

(creating the output lane sequences from the aligned PCSLs)



R8F and R2F PMAs

1, 2, or 4 instances
(200G, 400G, or 800G)

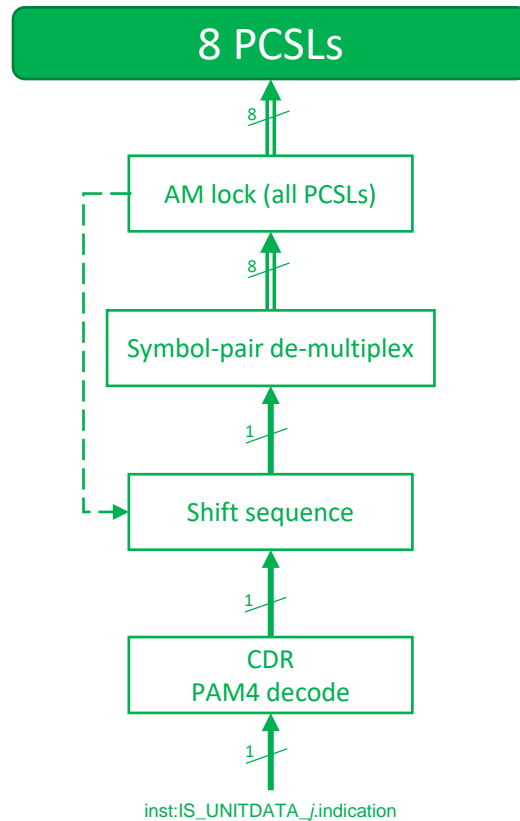


PMA(16:8)

8 instances
(1.6T)

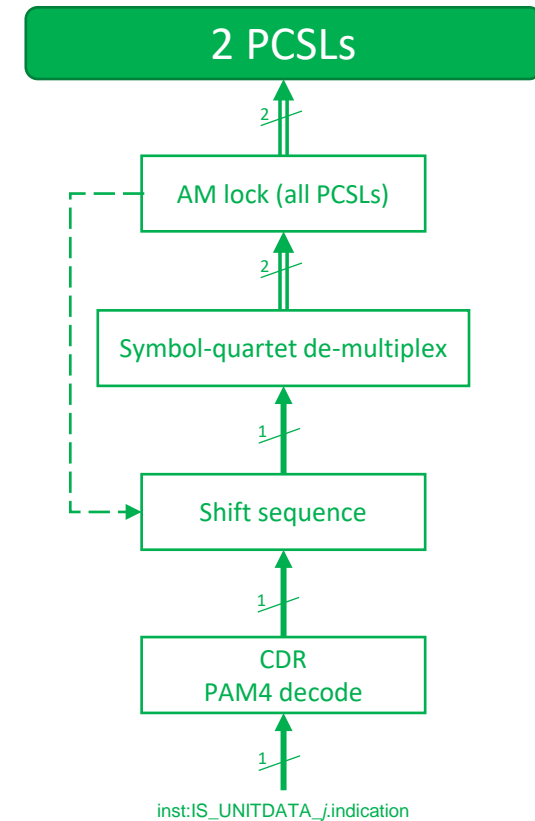
PMA input processing – Receive direction

(obtaining the aligned PCSs)



R8F and R2F PMAs

1, 2, or 4 instances
(200G, 400G, or 800G)

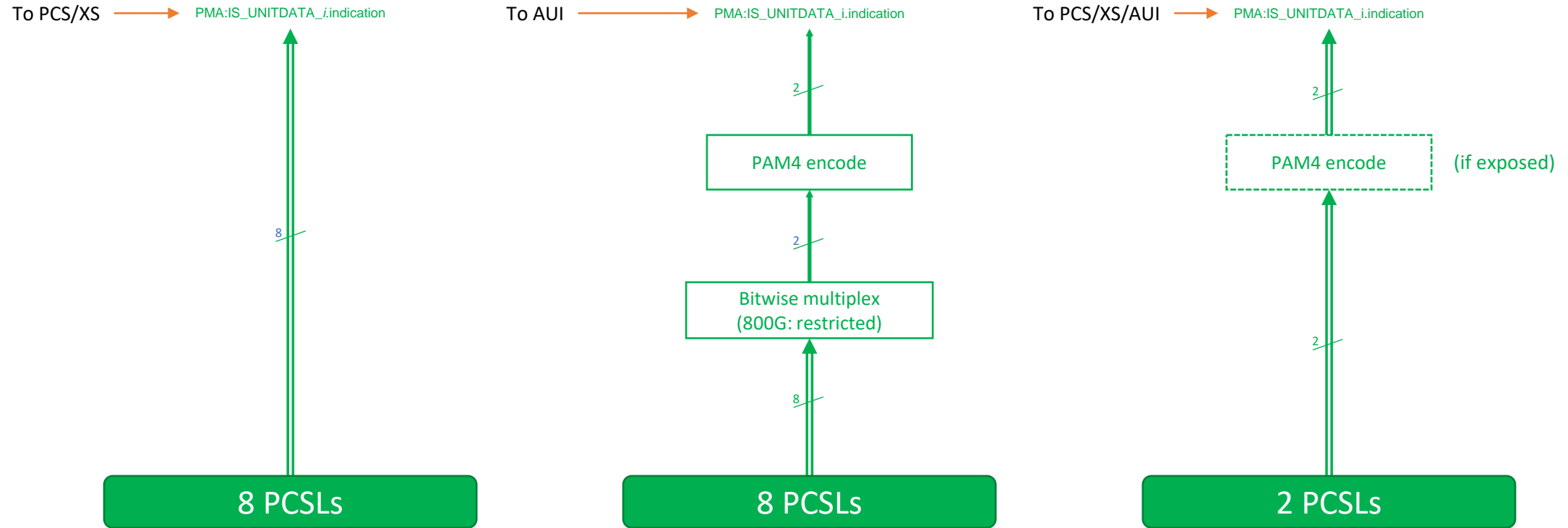


PMA(16:8)

8 instances
(1.6T)

PMA output processing – Receive direction

(creating the output lane sequences from the aligned PCSLs)



R8F PMA 1, 2, or 4 instances
(200G, 400G, or 800G)

R2F PMA 1, 2, or 4 instances
(200G, 400G, or 800G)

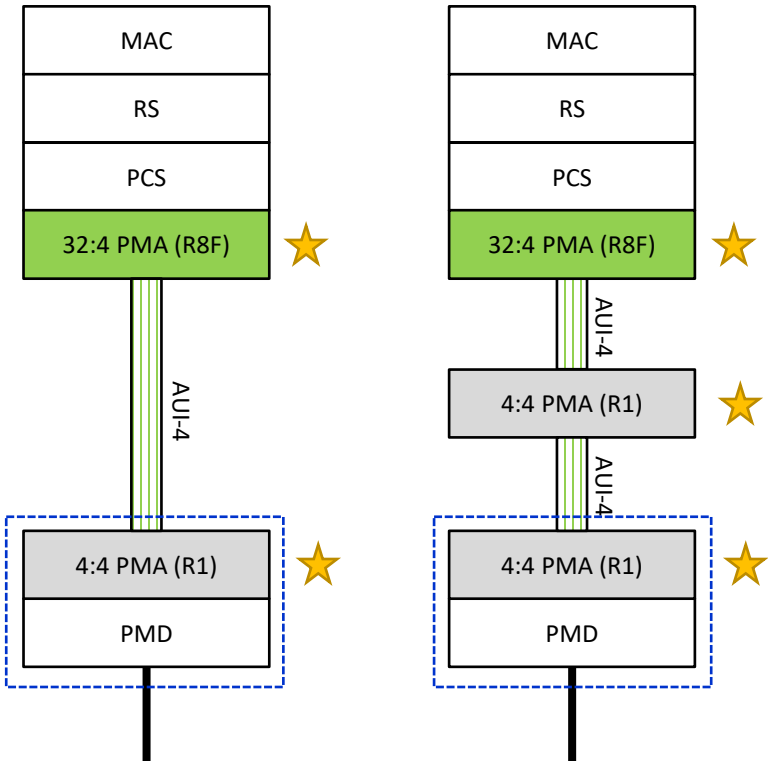
PMA(16:8) 8 instances
(1.6T)

PMA partitioning examples

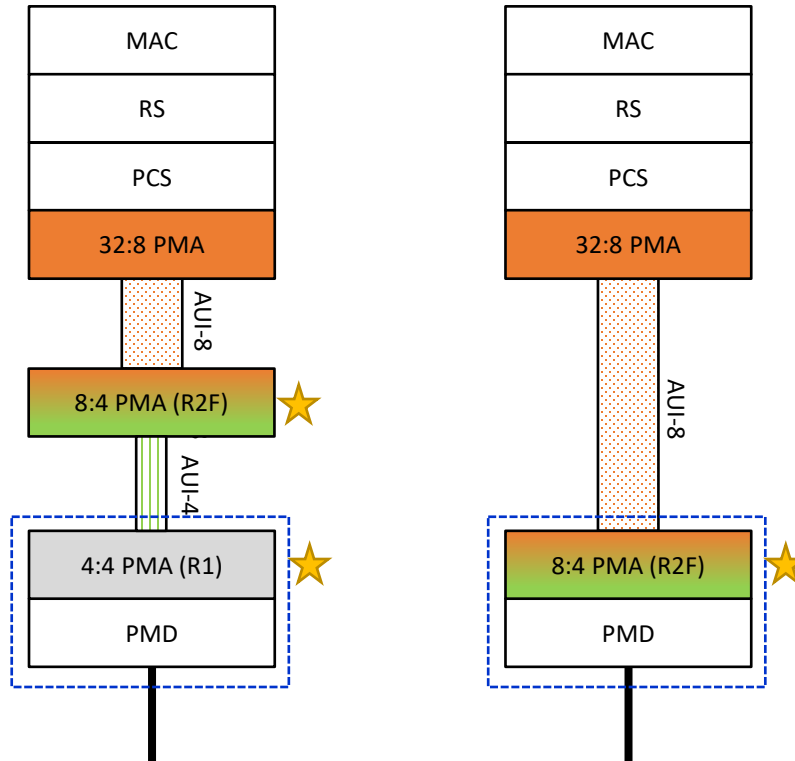
Informative annex can be based on these figures, with editorial license

Sublayer stacks for 800GBASE-*R4 "Type 1"

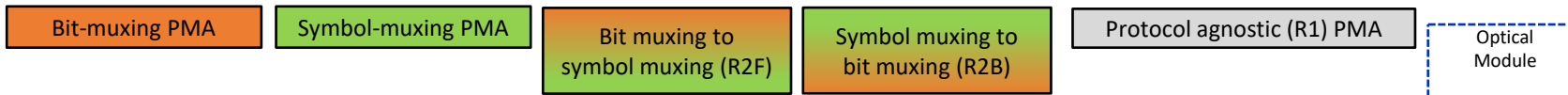
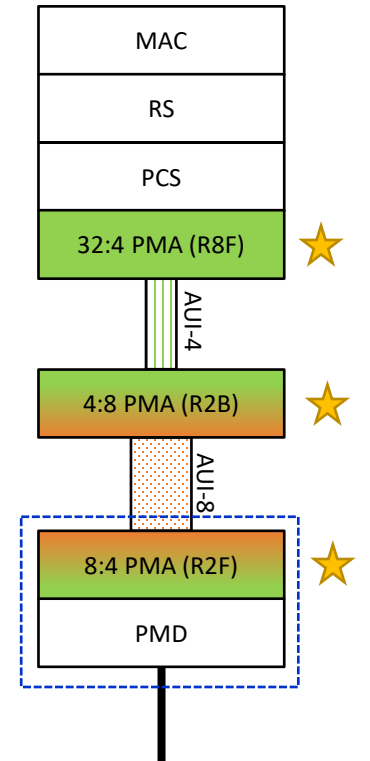
New Host ASIC



Legacy Host ASIC

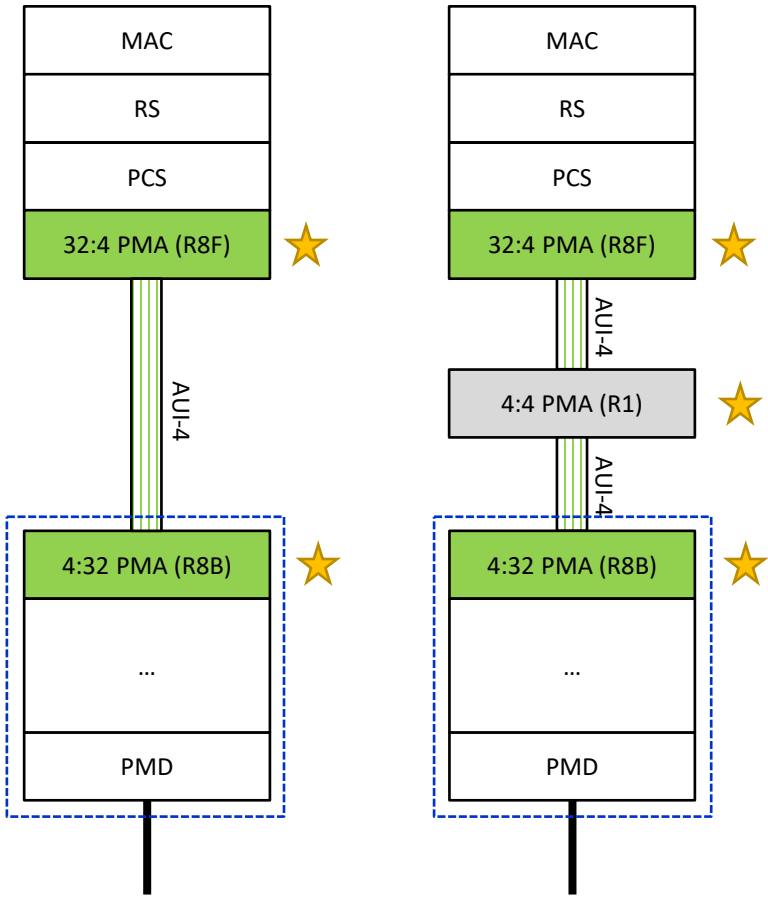


New Host ASIC + Legacy Module Interface?

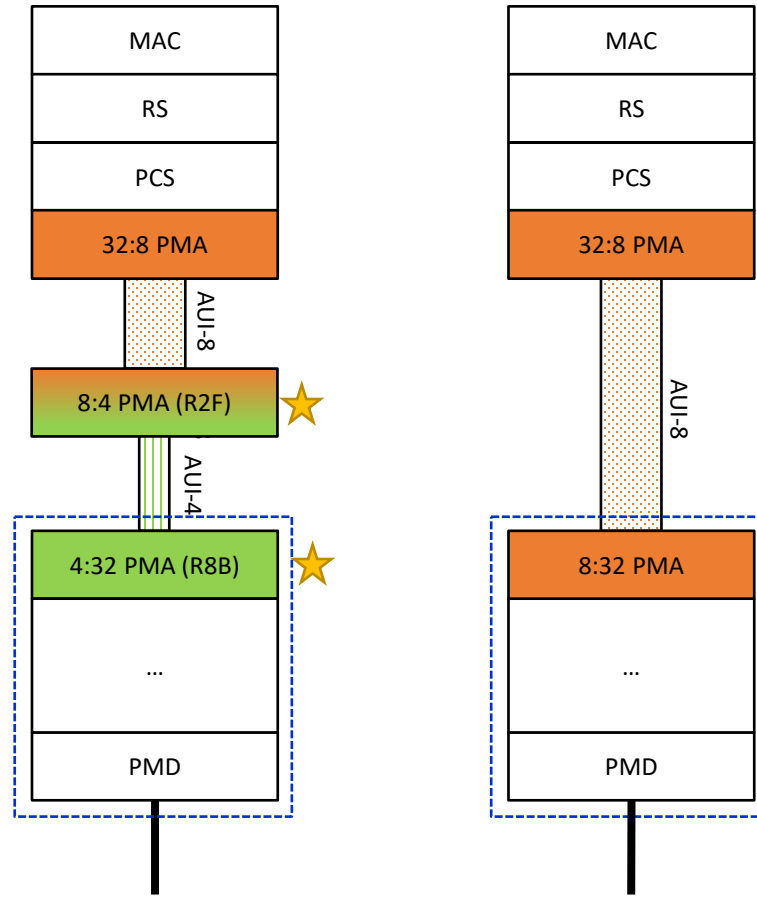


Sublayer stacks: 800GBASE-R, "Type 2"

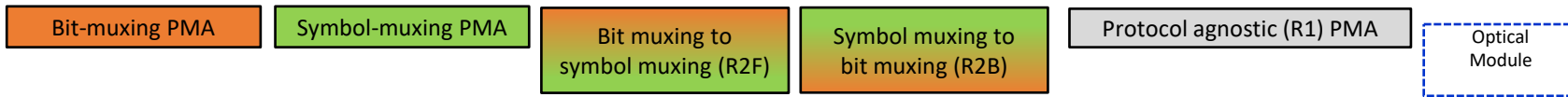
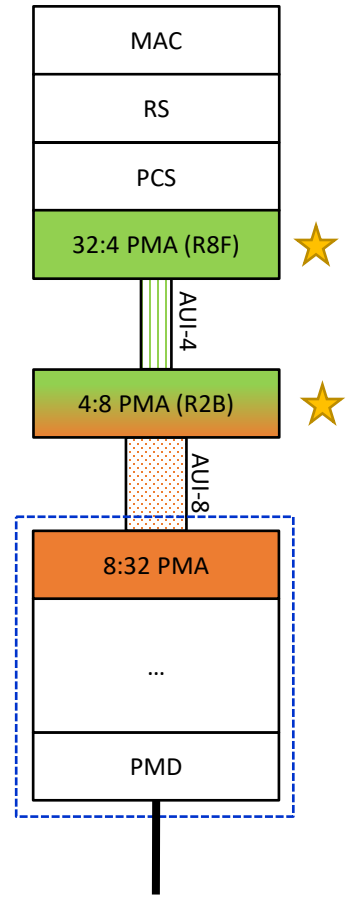
New Host ASIC



Legacy Host ASIC

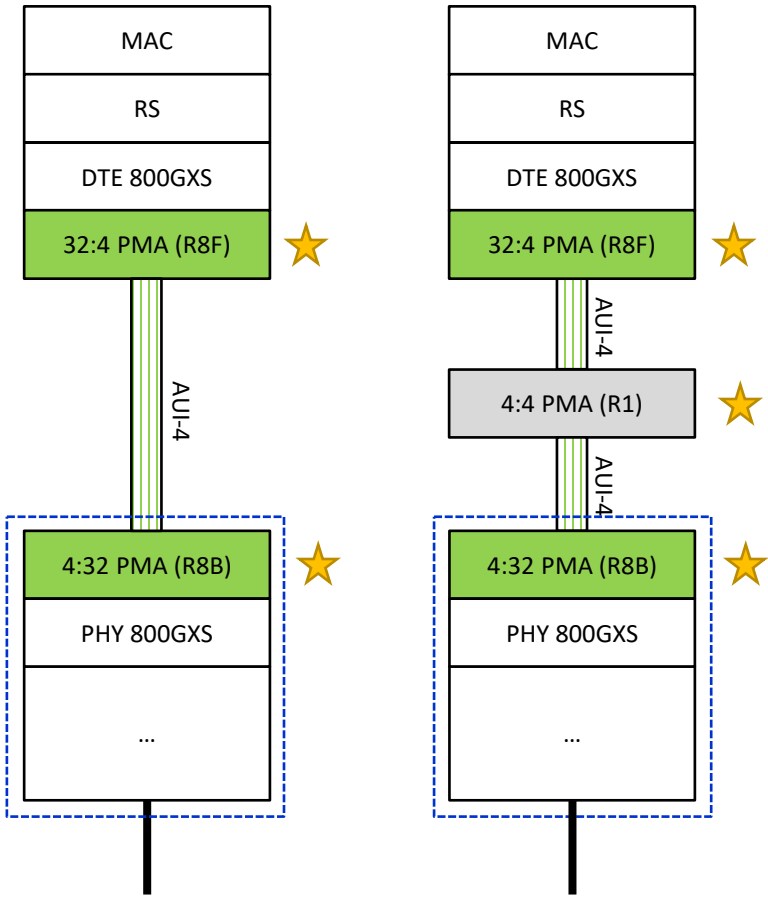


New Host ASIC + Legacy Module Interface?

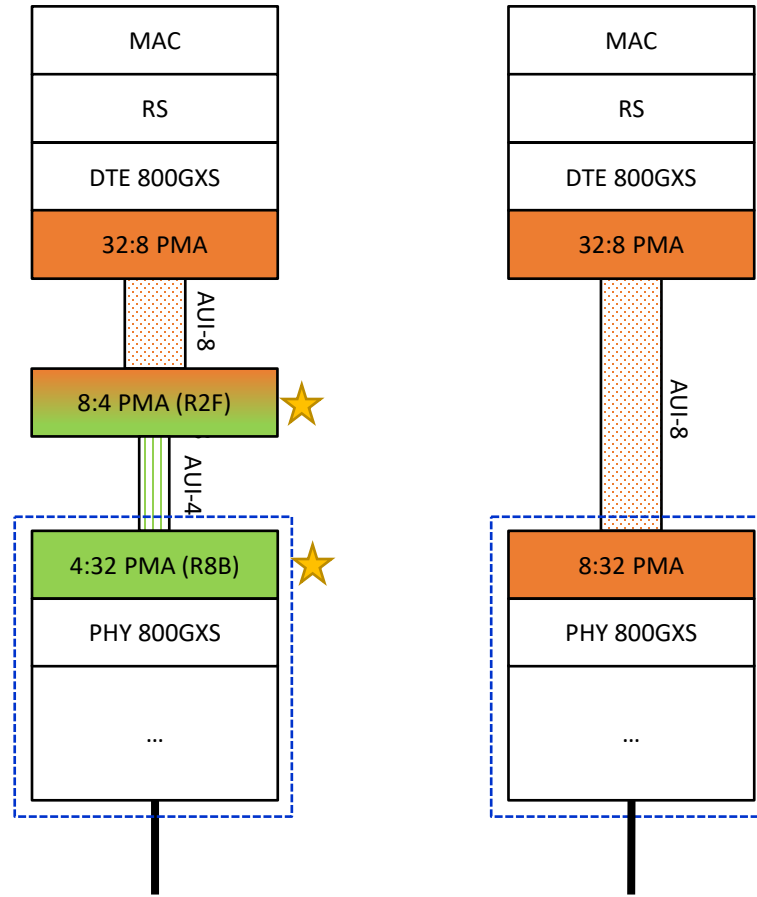


Sublayer stacks: 800GBASE-R with 800GMII Extender

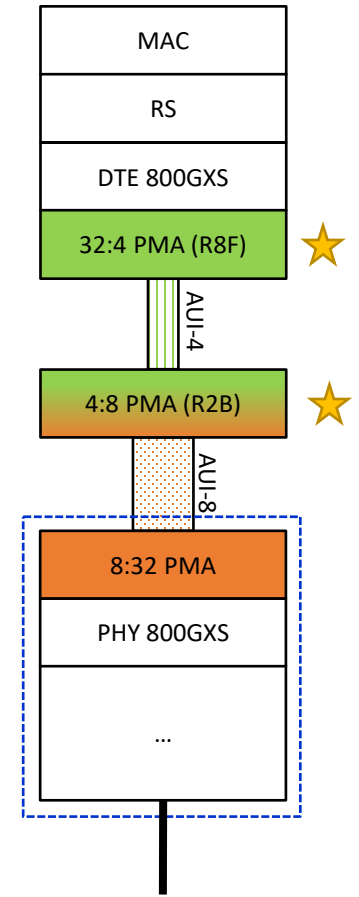
New Host ASIC



Legacy Host ASIC



New Host ASIC + Legacy Module Interface?

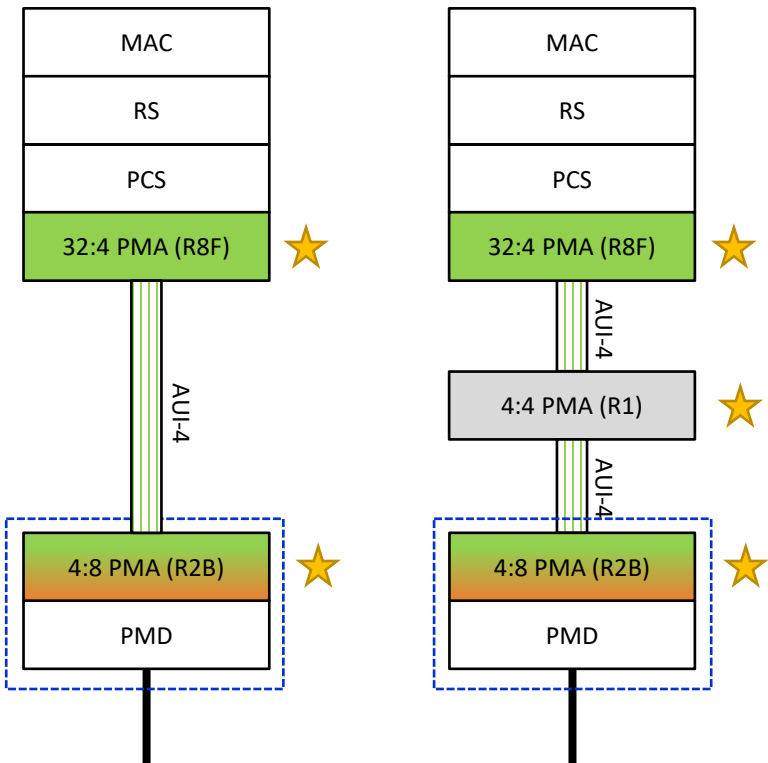


Legend for sublayer stacks:

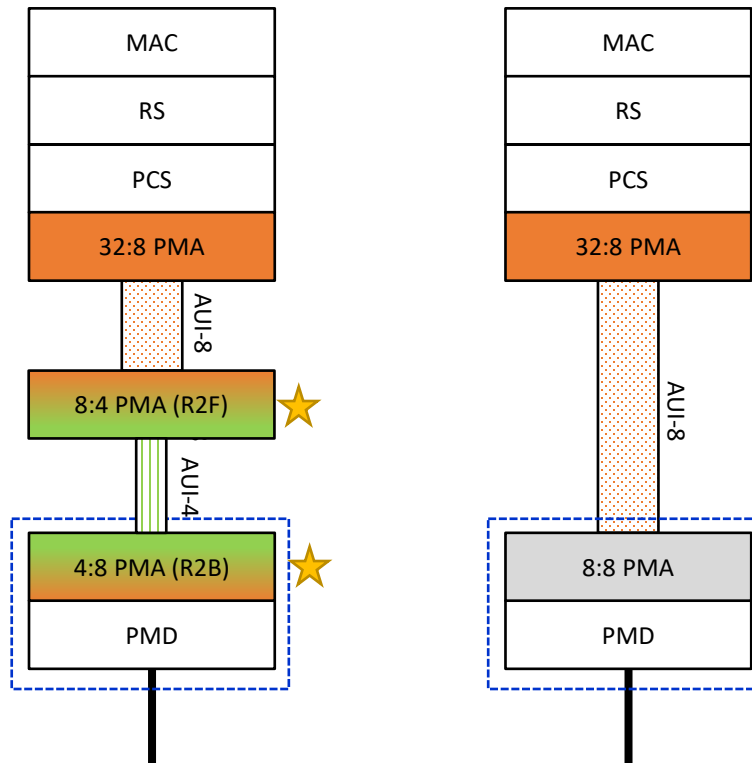
- Bit-muxing PMA (Orange box)
- Symbol-muxing PMA (Green box)
- Bit muxing to symbol muxing (R2F) (Orange box)
- Symbol muxing to bit muxing (R2B) (Green box)
- Protocol agnostic (R1) PMA (Grey box)
- Optical Module (Dashed blue box)

Sublayer stacks for 800GBASE-*8 “Type 1”

New Host ASIC



Legacy Host ASIC



New Host ASIC + Legacy Module Interface?

