800G Specification

Web: http://ethernettechnologyconsortium.org
or http://25gethernet.org/

Revision History

Revision	Who	Date	Change Description
0.1.3	MG/Cisco and DK/MorethanIP		Initial Draft
0.1.5	MG/Cisco		Added in proposed changes from Zvi Rechtman
0.1.6	MG / Cisco		Corrected a typo.
1.0	MG / Cisco	3/10/2020	Final Release

Definitions

800GBASE-R: An Ethernet Physical Coding Sublayer based on Clause 119 of IEEE Std 802.3, operating at a data rate of 800 Gb/s.

Table of Contents

1 Overview 2
2 Standards Reference 2
3 800 Gb/s Ethernet Specification 3
3.1 Architectural Overview 3
3.1.1 Leveraging Existing Standards 4
3.2 Detailed Specification 5
3.2.1 Media Access Layer (MAC) 5
3.2.2 Miscellaneous Requirements 5
3.2.3 Reconciliation and Media Independent Sublayer (RS/MII) 5
3.2.4 Physical Coding Sublayer (PCS/FEC) 5
3.2.5 PMA Sublayer 14
3.3 Electrical Specification 14

List of Tables

TABLE 1: 800G MARKER ENCODING 9

List of Figures

FIGURE 1: 800G MAC HIGH LEVEL BLOCK DIAGRAM 3
FIGURE 2: 800G PCS TX FLOW 6
FIGURE 3: 800G PC RX FLOW 7
FIGURE 4: 66B BLOCK ROUND ROBIN DISTRIBUTION 8
FIGURE 5: MARKER INSERTION 9

1 Overview

The 25G & 50G Ethernet Consortium standard provides specifications for an 800G implementation based on 8 lanex100 Gb/s technology, enabling adopters to deploy advanced high bandwidth interoperable Ethernet technologies.

2 Standards Reference

References are made throughout this document to IEEE 802.3-2018 Ethernet Access Method and Physical Layer [base standards], as well as 802.3ck which is still in the early draft stage.

PCS/FEC/PMA

- Clause 119 PCS
- Clause 120 PMA

Electrical

- IEEE 802.3ck Clauses TBD.

Note: IEEE 802 (e.g. IEEE 802.3-2012, etc) standards documents are available free through IEEE's Get program including IEEE 802 from http://standards.ieee.org/about/get/, six month after publication of each. 802.3ck is still in draft form and is only available to those working on the standard.

3 800 Gb/s Ethernet Specification

3.1 Architectural Overview

800Gb/s Ethernet technology is designed as an interface that uses eight 106 Gb/s lanes using a 2xClause 119 PCSs (400G) to connect a single MAC operating at 800 Gb/s (though the 400G PCSs are modified, this is just a very high level conceptual view). The following figure shows the high level architecture. Note that at least for now, we are not defining an 800G PMD in this specification. 2x400G PMDs could be used to form an 800G interface, for instance 2x400GBASE-DR4 modules, though skew needs to be managed to be within the specification. This architecture could support 8x106.25G, 16x53.125G or even slower interfaces, but the 8x106.25G is the main focus.

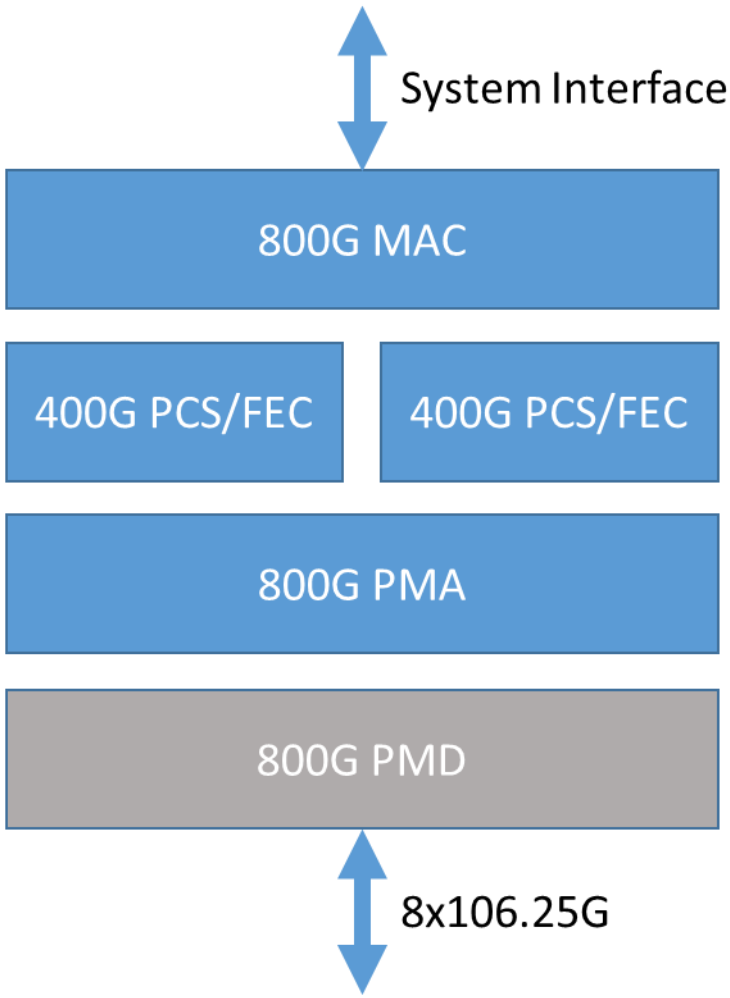

Figure 1: 800G MAC High Level Block Diagram

3.1.1 Leveraging Existing Standards

800 Gb/s capability can be supported by utilizing two 400 Gb/s PCSs (with the included FEC) and supporting 8 lanes of a 106.25G each.

The IEEE 802.3 standard for 400 Gb/s employs multi-lane distribution (MLD) to distribute data from a single Media Access Control (MAC) channel across 16 PCS lanes. This 800G standard will use a MAC scaled up to 800 Gb/s along with two 400Gb/s PCSs (with a few modifications) in order to drive 8x100G lanes. There will be a total of 32 PCS lanes (2x16 from the 400G standard), all with RS(544,514) FEC that is supported in the 400 Gb/s standard.

An important aspect of the MLD striping technique is the use of a unique alignment marker (AM) for each virtual lane. For 400Gb/s the AMs are inserted into the striped data stream every 163,840 x 257b blocks. This will continue with 800 Gb/s (and keeping the same spacing per 400G stream), but there will be twice as many AMs inserted, and AMs will have to be modified to ensure both a coherent 800 Gb/s stream and to prevent a misconfigured 400 Gb/s port from syncing up to the 800 Gb/s stream.

802.3ck will be leveraged for the C2M and C2C interfaces (operating at 106.25G per lane).

3.2 Detailed Specification

800G supports a single mode of operation:

1. Operation with FEC always enabled, RS(544,514)
2. 8x106.25G lanes

3.2.1 Media Access Layer (MAC)

The 800 Gb/s MAC inherits all attributes of the 400 Gb/s MAC, including full duplex operation only, and minimum interpacket gap of 8-bit times. See IEEE 802.3-2018 Section 4.

3.2.2 Miscellaneous Requirements

The 800 Gb/s MAC inherits all skew attributes of the 400 Gb/s MAC, see IEEE 802.3-2018 section 116.5.

3.2.3 Reconciliation and Media Independent Sublayer (RS/MII)

The 800 Gb/s MAC inherits all attributes of the 400 Gb/s MAC, including delay constraints. This includes the Deficit Idle Counter operation.

3.2.4 Physical Coding Sublayer (PCS/FEC)

800 Gb/s capability is supported by utilizing two 400 Gb/s PCSs (with the included FEC) and supporting 32 PCS lanes, each at 25Gb/s. Figure 2 shows at a high level the TX PCS data flow and functions. The 64b/66b encoder must run at 800G to create a coherent stream of data, but the rest of the processing is on a 400G slice of the data. The distribution, as shown in the diagram, is based on 1x66b blocks. The alignment marker insertion must be coordinated between the two stacks to ensure that coherent data is received and able to be processed on the RX side. Also, the alignment markers that are inserted will vary when compared to 400G to enable reception and deskew of a coherent data stream. 2x16 PCS lanes are generated from the two PCS stacks and then they are 4:1 bit multiplexed by the PMA towards the PMD in order to create 8x106G PMD lanes.

The RX flow is shown in Figure 3: 800G PC RX Flow

, it is just the reverse processing when compared to the TX, and still focused mostly on 2x400G protocol stacks. This 800G definition does allow any PCS lane to be received on any PMD lane, and so 32 PCS lanes must be locked to, and reordered appropriately to recover the data.

.

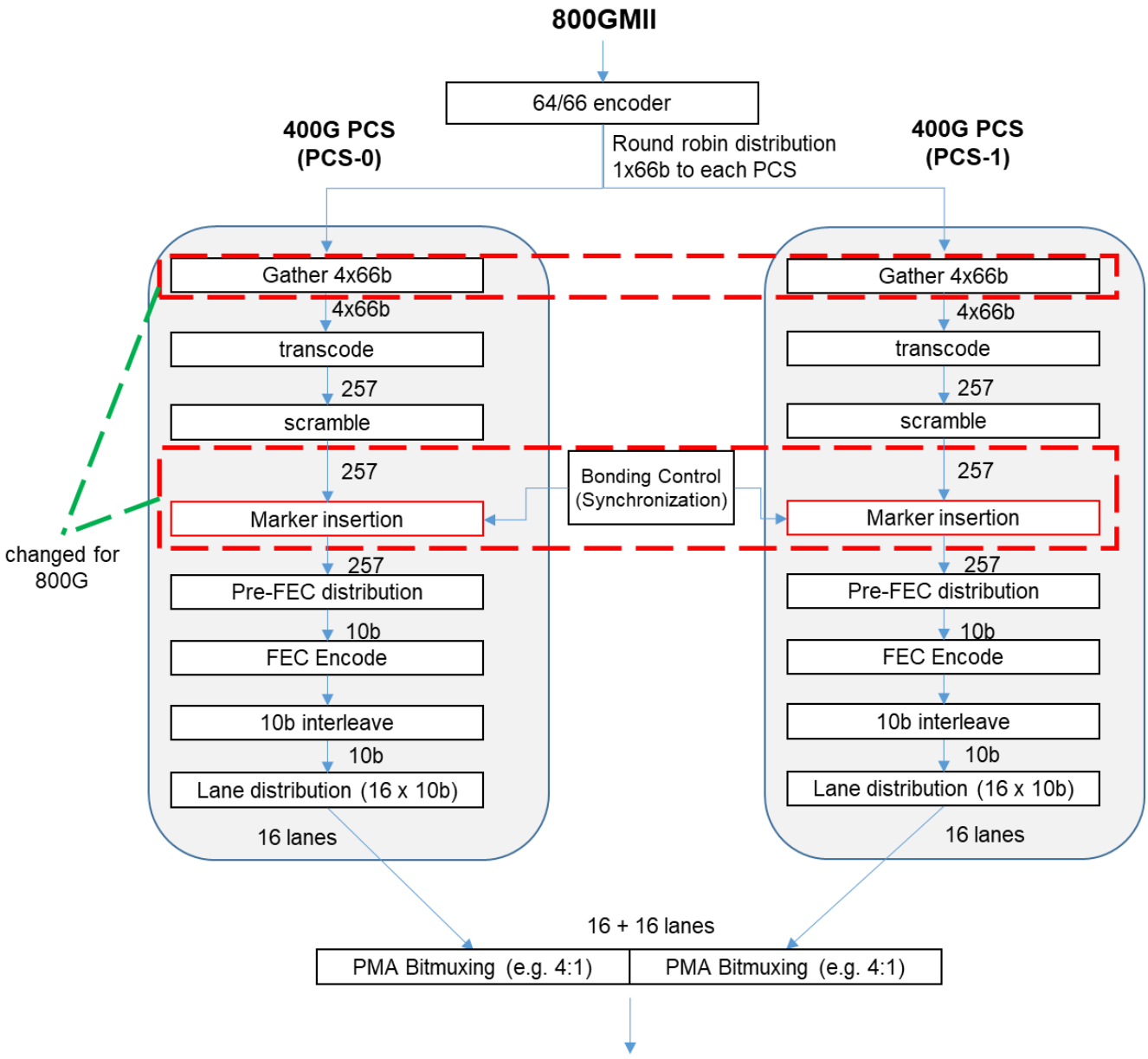

Figure 2: 800G PCS TX Flow

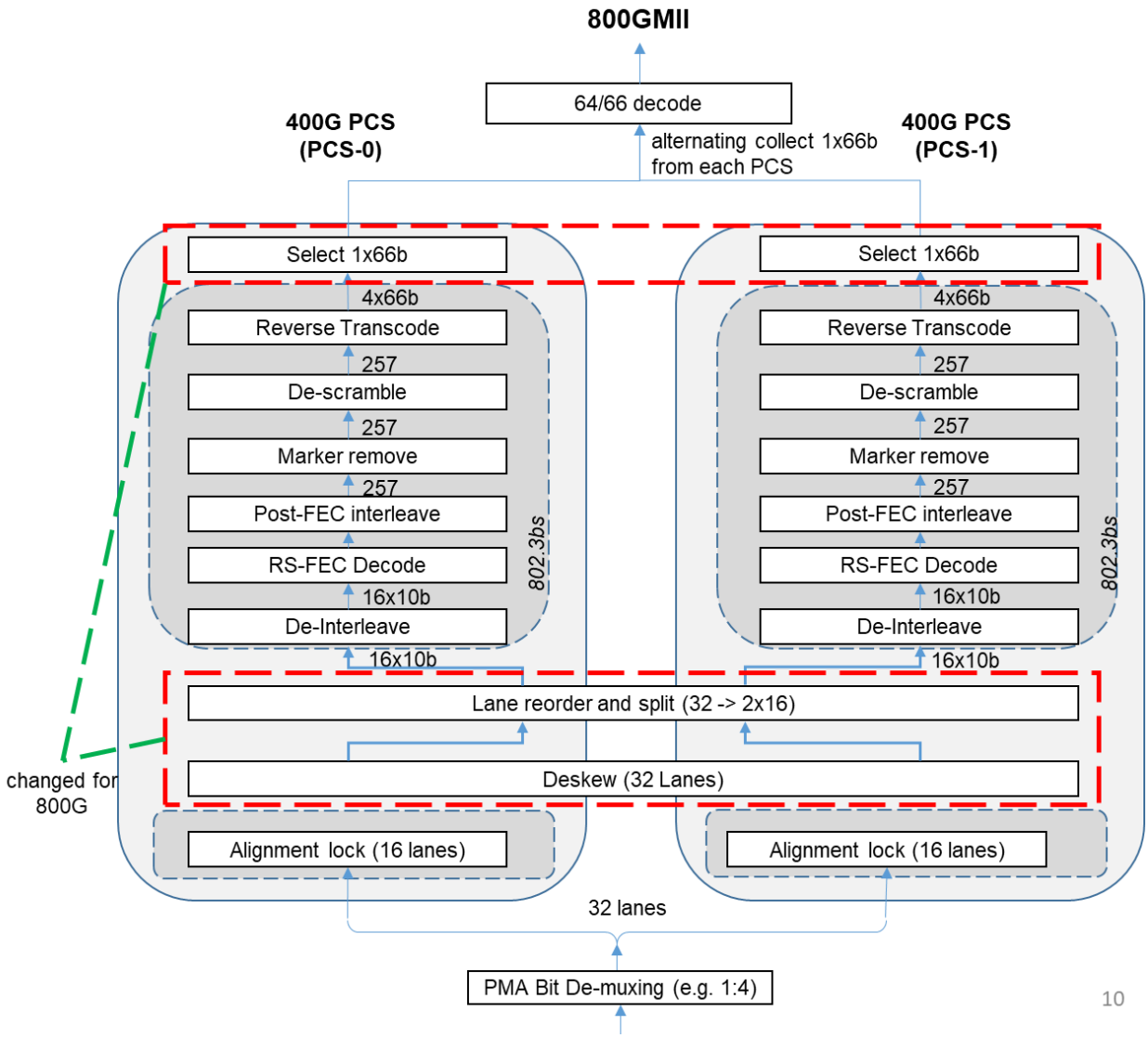

Figure 3: 800G PC RX Flow

3.2.4.1 *TX PCS Processing*

3.2.4.1.1 64B/66B Encoding

This is performed on the full 800G data stream as described in IEEE 802.3-2018 section 119.2.4.1, including figure 119-14.

3.2.4.1.2 66b block Distribution

66b blocks will be distributed to the two PCS instances 1x66b block at a time, in a round robin fashion, starting with PCS-0 and then to PCS-1 and back to PCS-0 again.

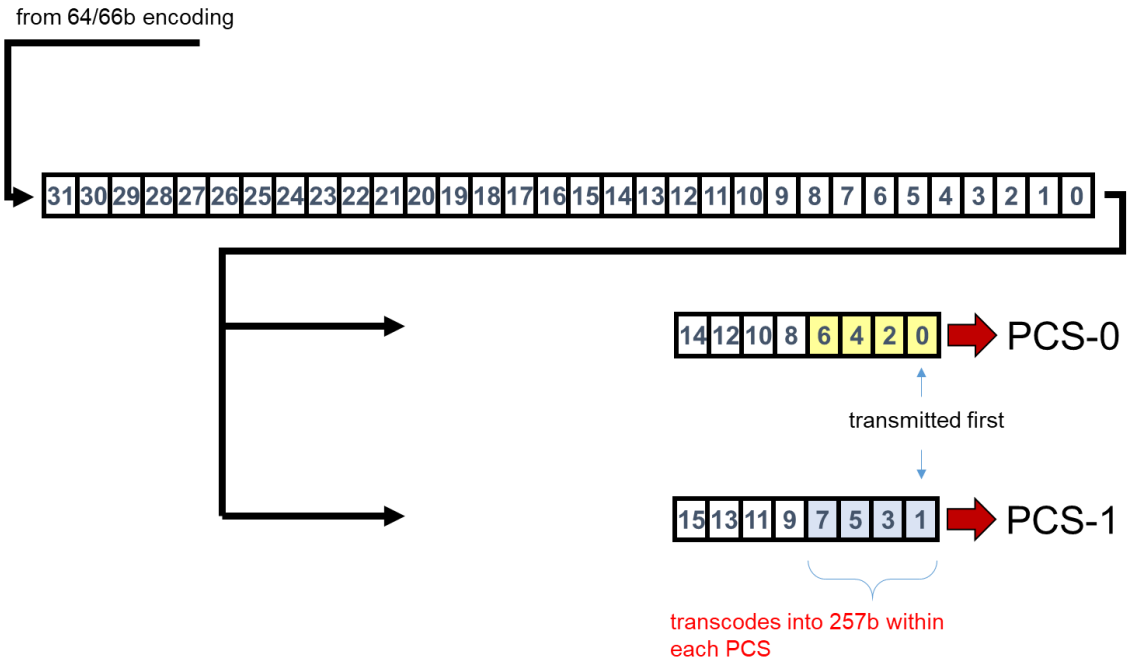

Figure 4: 66b Block Round Robin Distribution

3.2.4.1.3 Transcoding

Transcoding is performed separately on each 400G data stream, according to the rules in IEEE 802.3-2018 section 119.2.4.2.

3.2.4.1.4 Scrambling

Scrambling is performed separately on each 400G data stream, according to the rules in IEEE 802.3-2018 section 119.2.4.3.

3.2.4.1.5 Alignment Marker Insertion

In general, the rules of 802.3-2018 section 119.2.4.4 for the 400GBASE-R PCS are followed; for the spacing, AM format etc. There are a couple areas where there are differences.
To facilitate an aggregate 800G data stream over two 400G transmit channels, the marker insertion functions must be synchronized by having both transmit channels insert their markers at the exact same time (block unit), i.e. no skew above AM insertion in the protocol stack.
This allows the receiver to deskew not only between the 16 virtual lanes within each channel but also to align eventually all 32 virtual lanes to recover the aggregate data stream in correct symbol order again. Figure 5 shows this.

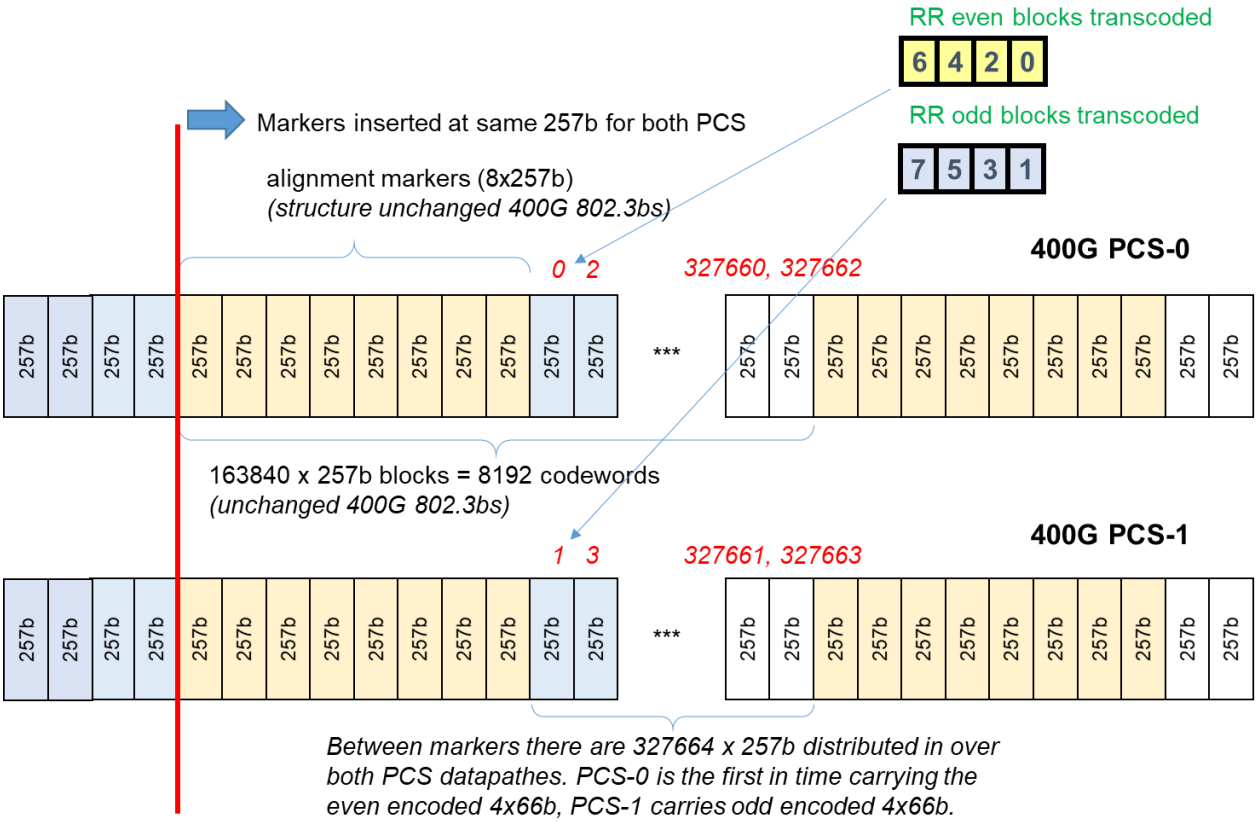

Figure 5: Marker Insertion

The actual content of the AMs will vary somewhat in order to allow reassembly of the data stream on the receive side, since we now have 32 PCS lanes vs. 16 PCS lane for 400G. One goal is to not allow a misconfigured RX (to 2x400G) to achieve alignment with an 800G TX.

The alignment marker format, the common marker (CM0-CM5) values, and the unique pad (UP0-UP2) definition is unchanged from IEEE 802.3-2018 definition.
The unique markers UM0/UM3 for PCS lanes 0-15 are inverted when compared to the IEEE 802.3-2018 definition for a 400G PHY, as indicated in the table below (marked in bold).
The unique markers UM1/UM2/UM4/UM5 for PCS lanes 16-31 are inverted when compared to the IEEE 802.3-2018 definition for a 400G PHY, as indicated in the table below (marked in bold).

Table 1: 800G Marker Encoding

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**

The transmit alignment marker status field allows the local PCS to communicate the status of the FEC degraded feature to the remote PCS. If there is no extender sublayer between the PCS and the MAC, it is set as follows:

tx_am_sf0<2:0> = {FEC_degraded_SER0,0,0}.
tx_am_sf1<2:0> = {FEC_degraded_SER1,0,0}.

The 3-bit transmit alignment marker status field is then appended to the variable am_mapped as follows:

am_mapped0<2055:2053> = tx_am_sf0<2:0>
am_mapped1<2055: 2053> = tx_am_sf1<2:0>

Alignment marker mapping is shown in Figure 2

3.2.4.1.6 Pre-FEC Distribution

Pre-FEC distribution is performed separately on each 400G data stream, according to the rules in IEEE 802.3-2018 section 119.2.4.5.

3.2.4.1.7 FEC Encode

FEC Encoding is performed separately on each 400G data stream, according to the rules in EEE 802.3-2018 section 119.2.4.6. This means that an 800G stream will have 4 FEC codewords on and interface, compared to 2 FEC codewords for 400G/200G.

3.2.4.1.8 10b Interleave and Distribution

Interleaving and distribution is performed separately on each 400G data stream, according to the rules in EEE 802.3-2018 section 119.2.4.7. This process will create 16 PCS lanes per 400G data stream. The 2x16 PCS lanes are then presented to the PMA for bit multiplexing. PCS instance 0 has PCS lanes 0-15, and PCS instance 1 has PCS lanes 16-31.

3.2.4.1.9 Test Pattern Generators

The scrambled idle test pattern as described in IEEE 802.3-2018 section 119.2.4.9 must be supported. Each 400G PCS implements the test pattern independently, therefore when operating in 800G it must be enabled in both channels and monitored in both channels respectively.

3.2.4.2 *RX PCS Processing*

3.2.4.2.1 Alignment Lock

This is performed individually on each 400G data stream as described in IEEE 802.3-2018 section 119.2.5.1, including figure 119-12.

3.2.4.2.2 Alignment Deskew, Reorder and de-interleave

This is performed across the complete 800G data stream. Processing is as described in IEEE 802.3-2018 section 119.2.5.1 and 119.2.5.2, including figure 119-13, with the difference that it is across 32 PCS lanes for this 800G interface. PCS lanes 0-15 are directed to PCS instance 0, and PCS lanes 16-31 are directed to PCS instance 1. Figure 119-13 covers the PCS synchronization state machine and an 800G implementation must behave as if there is a single state machine that acts on all 32 PCS lanes, including having 4xCWx_bad_count counters, one per FEC codeword. In addition 3 consecutive uncorrectable codewords for cwA_bad_count = 3 OR cwB_bad_count = 3 for either PCS instance will cause loss of lock, and there should be a single pcs_alignment_valid, align_status and deskew_done variable. An implementation could choose to keep two parallel state machines and then combine the results to behave as if there is a single state machine.

3.2.4.2.3 FEC Decode

This is performed individually on each 400G data stream as described in IEEE 802.3-2018 section 119.2.5.3.

3.2.4.2.4 Post FEC Interleave

This is performed individually on each 400G data stream as described in IEEE 802.3-2018 section 119.2.5.4.

3.2.4.2.5 Alignment Marker Removal

This is performed individually on each 400G data stream as described in IEEE 802.3-2018 section 119.2.5.5.

3.2.4.2.6 Descrambler

This is performed individually on each 400G data stream as described in IEEE 802.3-2018 section 119.2.5.6.

3.2.4.2.7 Transcoder

This is performed individually on each 400G data stream as described in IEEE 802.3-2018 section 119.2.5.7.

3.2.4.2.8 66b block Recombination

The 66b blocks are recombined from both 400G streams into a single coherent 800G 66b block stream, undoing the distribution that occurred in 3.2.4.1.2.

3.2.4.2.9 64b/66b Decode

The 64b/66b blocks are decoded from a single 800G based on IEEE 802.3-2018 section 119.2.5.8.

3.2.4.3 *Detailed functions and state diagrams*

3.2.4.4 *State variables*

align_status<y>

Same definition as in 119.2.6.2.2, per synchronization FSM y

restart_lock<y>

Same definition as in 119.2.6.2.2, per synchronization FSM y

hi_ser<y>

Same definition as in 119.2.6.2.2, per 400GE FEC y

rx_local_degraded<y>

Same definition as in 119.2.6.2.2, per 400GE FEC y

pcs_align_status

A Boolean variable that is set to true when align_status<y> is true for both y

and is set to false when align_status<y> is false for any y.

(=align_status0 AND align_status1)

pcs_restart_lock

A Boolean variable that is set to true when restart_lock<y> is true for any y

and is set to false when restart_lock<y> is false for both y.

(= restart_lock0 OR restart_lock1)

pcs_hi_ser

A Boolean variable that is set to true when hi_ser<y> is true for any y

and is set to false when hi_ser<y> is false for both y.

(= hi_ser0 OR hi_ser1)

3.2.4.4.1 State diagrams

The state diagrams for 800GE are based on 400GBASE-R PCS defined in 119.2.6 with the exception that there are 32 alignment marker lock processes as depicted in Figure 119–12 and two synchronization process as depicted in Figure 119–13.

3.2.4.4.1.1 Alignment marker lock state diagram

Identical to Figure 119–12 with the following expectation:

pcs_restart_lock variable is used instead of *restart_lock*.

3.2.4.4.1.2 PCS synchronization state diagram

Identical to Figure 119–13 with the following expectations:

pcs_hi_ser variable is used instead of *hi_ser*

align_status<y> variable is used instead of *align_status*

restart_lock<y> variable is used instead of *restart_lock*

3.2.4.4.1.3 Transmit state diagram

Identical to Figure 119–14.

3.2.4.4.1.4 Receive state diagram

Identical to Figure 119–15 with the following expectations:

pcs_align_status variable is used instead of *align_status*

3.2.5 PMA Sublayer

The PMA operates as defined in IEEE 802.3-2018 section 120, with the exception that there are 32 PCS lanes and only 4:1 bit muxing is performed. The PMA has complete freedom on multiplexing any PCS lanes together, PCS lanes are in no way restricted on their location on any PMD or AUI lane.

3.3 Electrical Specification

800G will typically be used with either a C2M or C2C interface in order to connect to an 800G Module. This will be defined in the emerging IEEE 802.3ck standard.