

Addressing delay values

Comments 45, 89, 91, 92, 137

Matt Brown, Alphawave Semi, 802.3df Editor-in-Chief

Adee Ran, Cisco

Piers Dawe, Nvidia

IEEE P802.3df Task Force

September 2023

Contributors

- Gary Nicholl, Cisco
- Kapil Shrikhande, Marvell
- Jeff Slavick, Broadcom

Supporters

<name>

Introduction

- Several comments report that the delay allocations for PMA and/or PMD are too low relative to real implementations and are not correctly distributed.
- The presentation first looks at the sub-elements of each sublayer.
- Then a new set of delay values are proposed that better reflect the functionality and meet the proposed delay.

Comments

CI 0 SC 0 P104 L12 # I-45

Ran, Adee Cisco Systems, Inc.

Comment Type T Comment Status D delay values

The PMD delay constraint for 800G optical PMDs should be the same in ns terms to those of similar PMDs at the same signaling rate with fewer lanes (viz., 20.48 ns rather than 40.96 ns).

To allow the total delay for 800G modules as has been adopted in response to comment #82 against D2.0, an extra delay of 20.48 ns can be allocated to the PMA instead, to create the same total delay of 87.04 ns (for PMD+PMA). Note that the delay could be added only for the PMA(8:8), but currently, there is no distinction between PMA types.

This comment affects clauses 124, 167, 169, and 173.

SuggestedRemedy

in 124.3.1 and in 167.3.1 Change "32 768 bit times (64 pause_quanta or 40.96 ns)" to "16384 bit times (32 pause_quanta or 20.48 ns)".

In 173.5.4, Change the values in Table 173-1 to "53 248", "104", and "66.56".

Change the corresponding entries in Table 169-4 accordingly.

Proposed Response Response Status W

PROPOSED ACCEPT IN PRINCIPLE.

The delay number should be revisited one more time.
A related presentation will be provided for review.

CI 124 SC 124.3.1 P104 L13 # I-89

Dawe, Piers J G NVIDIA

Comment Type TR Comment Status D delay values

The delay for 800GBASE-DR8 or 800GBASE-DR8-2 PMD including 2 m of fiber in one direction should be the same 20.48 ns as 400GBASE-DR4 and all other 200GBASE-R and 400GBASE-R optical PMDs (see tables 116-6 and 7). It was changed "because modern PMDs contain DSP": but that is semantics. We should not have different specification methods for 800GBASE-DR8 and 400GBASE-DR4 PMA/PMD: they are the same modules! For a typical retimed module, the PMA-PMD interface is internal so it doesn't matter (if we say it doesn't matter), but as linear and co-packaged optics become more popular, the interface is accessible, and a spec that has given the time for the A to D to the part that doesn't contain it becomes a problem. See comment against 169.3.3. Also note that a 32:8 or 8:32 PMA is "a SerDes" but an 8:8 PMA may be implemented as two SerDes back to back, with additional delay. See dawe_3df_01a_2307 Module and PMA delay limits, and other comments on delay.

SuggestedRemedy

Revert the PMD allowance to 16,384 bit times (32 pause_quanta or 20.48 ns) for all 8x100G optical, consistent with all 1/2/4x100G optical. With another comment, this gives a module with one PMD and one PMA $20.48+92.16 = 112.64$ ns. vs. D2.1 $40.96+46.08 = 87.04$ ns and 802.3-2018 $20.48 + 92.16/2$ (maybe) = 66.56 ns which seems to be tight for some DSP.

Proposed Response Response Status W

PROPOSED ACCEPT IN PRINCIPLE.

Resolve using the response to comment #I-45.

CI 169 SC 169.4 P182 L28 # I-91

Dawe, Piers J G NVIDIA

Comment Type ER Comment Status D delay values

The delay allowance for an 8:8 PMA is too low, and the allowance for an optical PMD is out of step with other optical PMDs. (The allowance for CR or KR PMD+AN may be wrong too, but it doesn't matter much as they are always combined with PMAs.) See dawe_3df_01a_2307 Module and PMA delay limits, and other comments on delay

SuggestedRemedy

Change "800GBASE-R PMA" to "32:8 or 8:32 800GBASE-R PMA". Add a row "8:8 800GBASE-R PMA, 73,728 BT, 144 PQ, 92.16 ns (exactly twice that for the 32:8 or 8:32 PMA). Revert the VR8, SR8, DR8 and DR8-2 PMD allowances to 16,384 BT, 32 PQ, 20.48 ns.

Proposed Response Response Status W

PROPOSED ACCEPT IN PRINCIPLE.
Resolve using the response to comment #I-45.

CI 173 SC 173.6.4 P240 L46 # I-92

Dawe, Piers J G NVIDIA

Comment Type TR Comment Status D delay values

This new delay allocation per PMA-instance may be OK where a PMA is packaged with a PCS, XS or PMD, but it is tight for a standalone PMA (e.g. "on-board retimer"). It is unlikely that a PMA will be packaged with an exposed 32x25G PMA interface except in a prototype.

SuggestedRemedy

Double the allowance for the 8:8 PMA only, from 36,864 BT, 72 PQ, 46.08 ns to 73,728 BT, 144 PQ, 92.16 ns. No need to change the delay allocation for 32:8 and 8:32 PMA.

Proposed Response Response Status W

PROPOSED ACCEPT IN PRINCIPLE.
Resolve using the response to comment #I-45.

CI 169 SC 169.4 P182 L28 # I-137

Maki, Jeffery Juniper Networks, Inc.

Comment Type TR Comment Status D delay values

800GBASE-R PMA Delay + 800GBASE-DR8 PMD Delay or 800GBASE-DR8-2 PMD Delay is 87.04 ns (the optical module Delay) and is too small in relation to prevalent implementations where values are measured to be as high as 106 ns to 108 ns with the various suppliers reporting values as high as 109 ns to 129 ns.

SuggestedRemedy

Increase the allowed sum to 200 pause_quanta or 128 ns.

Proposed Response Response Status W

PROPOSED ACCEPT IN PRINCIPLE.
Resolve using the response to comment #I-45.

Related comment

Cl 169 SC 169.3.3 P182 L4 # I-90
Dawe, Piers J G NVIDIA
Comment Type TR Comment Status D PMD SI

Traditionally, the PMD limited a PAM2 signal and the PMA did timing recovery, and might include some PCB. With PAM4, the PMA does Gray mapping too. 116.3.3.2.1, Semantics of the service primitive, says that:

"each of the rx_symbol parameters can either take one of two values: zero or one; or take one of four values: zero, one, two, or three", possibly implying that the PMD makes the decisions (therefore contains any DSP equaliser and associated A to D, as well as analog equalisation). With DSP and soft decision coming to specs related to 802.3df soon, this may need to change or be clarified. We need to be careful where we assume the A to D and DSP functions are when dividing up or combining elements of the delay budget.

For EPoC, 100.2.1.2, PMD_UNITDATA.indication, says:

This primitive defines the transfer of I/Q value pair data from the Clause 100 PMD to the Clause 101 PMA. The semantics of the service primitive are PMD_UNITDATA.indication(I_value, Q_value, ChNum). The data conveyed by PMD_UNITDATA.indication is a continuous stream of I/Q value pairs and received OFDM channel. Both I_value and Q_value are encoded as 32-bit signed integers. ChNum indicates the applicable channel.

P802.3cw 156.2.1.2.1, Semantics of the primitive, says:

The PMD_UNITDATA.indication primitive conveys four "analog" signals, representing... 3cw is not binding here, but EPoC and 3cw are reasonable ways of describing the component parts, that work when more sophisticated signal processing techniques are used. But they put the A to D in different places.

SuggestedRemedy

The "PMD makes the decisions" model will put too much of the PHY in an unrecognisable "PMD sublayer". EPoC's "PMD contains the D to A" model seems un-intuitive, and it would mean that a PMA in an AUI (which obviously can contain an A to D) must have a very different delay allocation to a PMA next to the PMD. P802.3cw's "PMD may provide E/O conversion, gain, and analog EQ" model seems the most promising.

Addressing this question may be needed to set the delay limits of the sublayers.

Add an exception here, that unlike in 116.3.3.2.1, IS_UNITDATA_i.indication(rx_symbol) conveys an analog signal representing a PAM4 signal, possibly with noise and distortion. See other comments on delay.

Proposed Response Response Status W

PROPOSED REJECT.

For commonality with 100 Gb/s per lane interfaces for 100 Gb/s, 200 Gb/s, and 400 Gb/s Ethernet the PMD service interface should remain as currently defined.

The proposed changes might be worth considering for in a later project, e.g., 802.3dj, for higher signaling rate interfaces.

Observations:

Electrical PMD delay is 40.96 ns including 14 ns for medium (~3 m)
So $D_{pmd_s} = 40.96 - 14 = 26.96$ ns

The net delay for any PMA type is 46.08 ns.
Neither Table 169-4 nor the AUI annexes, specify the the interconnect delay for AUI.

Optical PMD delay is 40.96 ns, which includes 2 m of fiber (~10 ns)
So $D_{pmd_s} = 40.96 - 10 = 30.96$

Total allocation for optical module excluding fiber is:
 $46.08 + 30.96 = 77.04$ ns

Table 169–4—Sublayer delay constraints (800GBASE)

Sublayer	Maximum (bit time) ^a	Maximum (pause_quanta) ^b	Maximum (ns)	Notes ^c
800G MAC, RS, and MAC Control	196 608	384	245.76	See 170.1.4.
800GBASE-R PCS or 800GXS ^d	640 000	1250	800	See 172.5.
800GBASE-R PMA	36 864	72	46.08	See 173.5.4.
800GBASE-KR8 PMD	32 768	64	40.96	Includes allocation of 14 ns for one direction through backplane medium. See 163.5.
800GBASE-CR8 PMD	32 768	64	40.96	Includes allocation of 14 ns for one direction through cable medium. See 162.5.
800GBASE-VR8 PMD	32 768	64	40.96	Includes 2 m of fiber. See 167.3.1.
800GBASE-SR8 PMD	32 768	64	40.96	Includes 2 m of fiber. See 167.3.1.
800GBASE-DR8 PMD	32 768	64	40.96	Includes 2 m of fiber. See 124.3.1.
800GBASE-DR8-2 PMD	32 768	64	40.96	Includes 2 m of fiber. See 124.3.1.

^a For 800GBASE-R, 1 bit time (BT) is equal to 1.25 ps. (See 1.4.215 for the definition of bit time.)

^b For 800GBASE-R, 1 pause_quantum is equal to 640 ps. (See 31B.2 for the definition of pause_quanta.)

^c Should there be a discrepancy between this table and the delay requirements of the relevant sublayer clause, the sublayer clause prevails.

^d If an implementation includes the 800GMII Extender, the delay associated with the 800GMII Extender includes two 800GXS sublayers.

See 80.4 for the calculation of bit time per meter of fiber or electrical cable.

See 31B.3.7 for PAUSE reaction timing constraints for stations at operating speeds of 800 Gb/s.

80.4 Delay constraints

Predictable operation of the MAC Control PAUSE operation (Clause 31, Annex 31B) demands that there be an upper bound on the propagation delays through the network. This implies that MAC, MAC Control sublayer, and PHY implementations conform to certain delay maxima, and that network planners and administrators conform to constraints regarding the cable topology and concatenation of devices. Table 80–7 contains the values of maximum sublayer delay (sum of transmit and receive delays at one end of the link) in bit times as specified in 1.4 and pause_quanta as specified in 31B.2. If a PHY contains an Auto-Negotiation sublayer, the delay of the Auto-Negotiation sublayer is included within the delay of the PMD and medium.

Equation (80–1) specifies the calculation of cable delay in nanoseconds per meter of fiber or electrical cable, based upon the parameter n , which represents the ratio of the speed of electromagnetic propagation in the fiber or electrical cable to the speed of light in a vacuum, $c = 3 \times 10^8$ m/s.

$$\text{cable delay} = \frac{10^9}{nc} \text{ ns/m} \quad (80-1)$$

The value of n should be available from the fiber or electrical cable manufacturer; but if no value is known, then a conservative delay estimate can be calculated using a default value of $n = 0.66$, which yields a default cable delay of 5 ns/m.

Legacy delay, electrical 802.3ck, Clause 80/116

Table 80-7—Sublayer delay constraints

Sublayer	Maximum (bit time) ^a	Maximum (pause_quantum) ^b	Maximum (ns)	Notes ^c
...				
100GBASE-R RS-FEC	40 960	80	409.60	See 91.4.
100GBASE-P RS-FEC-Int	51 200	100	512.00	See 161.4.
Inverse RS-FEC	40 960	80	409.60	See 152.4.
...				
100GBASE-P PMA	9 216	18	92.16	See 135.5.4.
100GBASE-KR1 PMD	4 096	8	40.96	Includes allocation of 14 ns for one direction through backplane medium. See 163.5.
100GBASE-KR2 PMD	4 096	8	40.96	Includes allocation of 20 ns for one direction through backplane medium. See 137.5.
100GBASE-KR4 PMD	2 048	4	20.48	Includes delay of one direction through backplane medium. See 93.4.
100GBASE-KP4 PMA/PMD	8 192	16	81.92	Includes delay of one direction through backplane medium. See 94.2.5.
100GBASE-CR1 PMD	4 096	8	40.96	Includes allocation for 14 ns for one direction through cable medium. See 162.5.
100GBASE-CR2 PMD	4 096	8	40.96	Includes allocation for 20 ns for one direction through cable medium. See 136.5.
...				

^a For 40GBASE-R, 1 bit time (BT) is equal to 25 ps and for 100GBASE-R, 1 bit time (BT) is equal to 10 ps. (See 1.4.215 for the definition of bit time.)

^b For 40GBASE-R, 1 pause_quantum is equal to 12.8 ns and for 100GBASE-R, 1 pause_quantum is equal to 5.12 ns. (See 31B.2 for the definition of pause_quantum.)

^c Should there be a discrepancy between this table and the delay requirements of the relevant sublayer clause, the sublayer clause prevails.

Table 116-6—Sublayer delay constraints (200GBASE)

Sublayer	Maximum (bit time) ^a	Maximum (pause_quantum) ^b	Maximum (ns)	Notes ^c
...				
200GBASE-R PMA	18 432	36	92.16	See 120.5.4.
200GBASE-KR2 PMD	8 192	16	40.96	Includes allocation of 14 ns for one direction through backplane medium. See 163.5.
200GBASE-KR4 PMD	8 192	16	40.96	Includes allocation of 20 ns for one direction through backplane medium. See 137.5.
200GBASE-CR2 PMD	8 192	16	40.96	Includes allocation of 14 ns for one direction through cable medium. See 162.5.
200GBASE-CR4 PMD	8 192	16	40.96	Includes allocation of 20 ns for one direction through cable medium. See 137.5.
...				

^a For 200GBASE-R, 1 bit time (BT) is equal to 5 ps. (See 1.4.215 for the definition of bit time.)

^b For 200GBASE-R, 1 pause_quantum is equal to 2.56 ns. (See 31B.2 for the definition of pause_quantum.)

^c Should there be a discrepancy between this table and the delay requirements of the relevant sublayer clause, the sublayer clause prevails.

KR4/CR4 values are in error. Address in maintenance. See next slide.

Table 116-7—Sublayer delay constraints (400GBASE)

Sublayer	Maximum (bit time) ^a	Maximum (pause_quantum) ^b	Maximum (ns)	Notes ^c
...				
400GBASE-R PMA	36 864	72	92.16	See 120.5.4.
400GBASE-KR4 PMD	8 192	16	20.48	Includes allocation of 14 ns for one direction through backplane medium. See 163.5.
400GBASE-CR4 PMD	8 192	16	20.48	Includes allocation for 14 ns for one direction through cable medium. See 162.5.
400GBASE-VR4 PMD	8 192	16	20.48	Includes 2 m of fiber. See 167.3.1.
...				

^a For 400GBASE-R, 1 bit time (BT) is equal to 2.5 ps. (See 1.4.215 for the definition of bit time.)

^b For 400GBASE-R, 1 pause_quantum is equal to 1.28 ns. (See 31B.2 for the definition of pause_quantum.)

^c Should there be a discrepancy between this table and the delay requirements of the relevant sublayer clause, the sublayer clause prevails.

Legacy delay

802.3ck, Clause 162/163

From IEEE 802.3ck-2022...

162.5 Delay constraints

The sum of the transmit and the receive delays at one end of the link contributed by the PMD and AN sublayers including the medium in one direction shall be no more than the maximum delays listed in Table 162-4. It is assumed that the one-way delay through the medium is no more than 14 ns.

Table 162-4—Delay constraints

PMD	Maximum (bit times) ^a	Maximum (pause_quanta) ^b	Maximum (ns)
100GBASE-CR1	4 096	8	40.96
200GBASE-CR2	8 192	16	40.96
400GBASE-CR4	16 384	32	40.96

^a One bit time is equal to 10 ps for 100GBASE-CR1, 5 ps for 200GBASE-CR2, and 2.5 ps for 400GBASE-CR4. (See 1.4.215 for the definition of bit time.)

^b One pause_quantum is equal to 5.12 ns for 100GBASE-CR1, 2.56 ns for 200GBASE-CR2, and 1.28 ns for 400GBASE-CR4. (See 31B.2 for the definition of pause_quantum.)

Descriptions of overall system delay constraints can be found in 80.4 for 100GBASE-CR1 and in 116.4 for 200GBASE-CR2 and 400GBASE-CR4.

From IEEE 802.3ck-2022...

163.5 Delay constraints

The sum of the transmit and receive delays at one end of the link contributed by the PMD and AN including the medium in one direction shall be no more than the maximum delays listed in Table 163-4. It is assumed that the one-way delay through the medium is no more than 14 ns.

Table 163-4—Delay constraints

PMD	Maximum (bit times) ^a	Maximum (pause_quanta) ^b	Maximum (ns)
100GBASE-KR1	4 096	8	40.96
200GBASE-KR2	8 192	16	40.96
400GBASE-KR4	16 384	32	40.96

^a One bit time is equal to 10 ps for 100GBASE-KR1, 5 ps for 200GBASE-KR2, and 2.5 ps for 400GBASE-KR4. (See 1.4.215 for the definition of bit time.)

^b One pause_quantum is equal to 5.12 ns for 100GBASE-KR1, 2.56 ns for 200GBASE-KR2, and 1.28 ns for 400GBASE-KR4. (See 31B.2 for the definition of pause_quantum.)

Descriptions of overall system delay constraints can be found in 80.4 for 100GBASE-KR1 and in 116.4 for 200GBASE-KR2 and 400GBASE-KR4.

Same as current specifications for 800GBASE-CR8/KR8

Optical PMD delay, new and old

Clause 124/167

It is rather odd that the same electro-optics function is different for each Ethernet rate, even though the per lane function is identical.

From IEEE 802.3df D3.0...

124.3 Delay and Skew

124.3.1 Delay constraints

Change 124.3.1 as follows:

The sum of the transmit and receive delays at one end of the link contributed by the 400GBASE-DR4 or 400GBASE-DR4-2 PMD including 2 m of fiber in one direction shall be no more than 8192 bit times (16 pause_quanta or 20.48 ns). A description of overall system delay constraints and the definitions for bit times and pause_quanta can be found in 116.4 and its references.

The sum of the transmit and receive delays at one end of the link contributed by the 800GBASE-DR8 or 800GBASE-DR8-2 PMD including 2 m of fiber in one direction shall be no more than 32 768 bit times (64 pause_quanta or 40.96 ns).

Descriptions of overall system delay constraints and the definitions for bit times and pause_quanta can be found in 116.4 for 400GBASE-DR4 and 400GBASE-DR4-2, and in 169.4 for 800GBASE-DR8 and 800GBASE-DR8-2.

From IEEE 802.3db-2022...

167.3.1 Delay constraints

An upper bound to the delay through the PMA and PMD is required for predictable operation of the MAC Control PAUSE operation.

The sum of the transmit and receive delays at one end of the link contributed by the 100GBASE-VR1 or 100GBASE-SR1 PMD including 2 m of fiber in one direction shall be no more than 2048 bit times (4 pause_quanta or 20.48 ns).

The sum of the transmit and receive delays at one end of the link contributed by the 200GBASE-VR2 or 200GBASE-SR2 PMD including 2 m of fiber in one direction shall be no more than 4096 bit times (8 pause_quanta or 20.48 ns).

The sum of the transmit and receive delays at one end of the link contributed by the 400GBASE-VR4 or 400GBASE-SR4 PMD including 2 m of fiber in one direction shall be no more than 8192 bit times (16 pause_quanta or 20.48 ns).

Descriptions of overall system delay constraints and the definitions for bit times and pause_quanta can be found in 80.4 for 100GBASE-VR1 and 100GBASE-SR1, and in 116.4 and its references for 200GBASE-VR2, 200GBASE-SR2, 400GBASE-VR4, and 400GBASE-SR4.

From IEEE 802.3df D3.0...

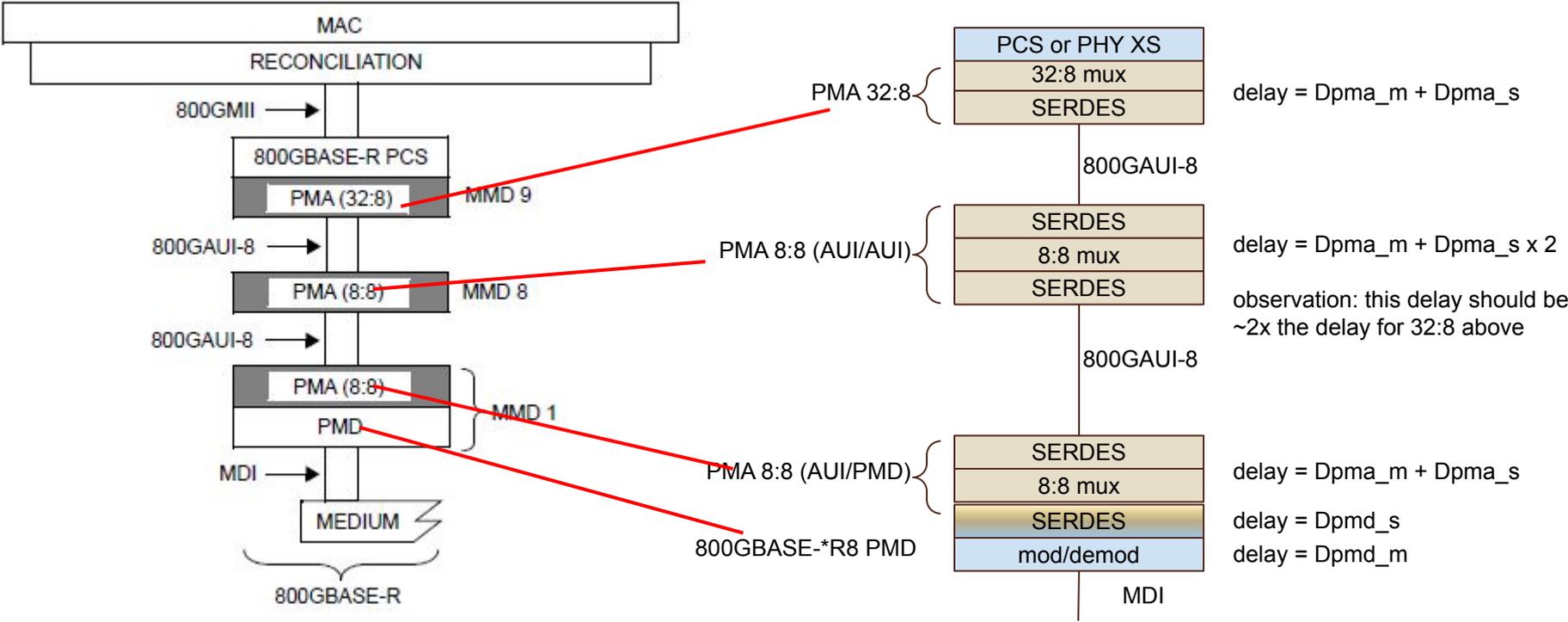
167.3 Delay and Skew

167.3.1 Delay constraints

Insert a new paragraph after the fourth paragraph in 167.3.1 as follows:

The sum of the transmit and receive delays at one end of the link contributed by the 800GBASE-VR8 or 800GBASE-SR8 PMD including 2 m of fiber in one direction shall be no more than 32 768 bit times (64 pause_quanta or 40.96 ns).

Subdivided delay contributors in PMD and PMA



$D_{\text{pma_m}}$ = delay of the PMA mux function
 $D_{\text{pms_s}}$ = delay of the AUI SERDES on one side of PMA
 $D_{\text{pmd_s}}$ = delay of the SERDES and mod/demod in PMD
 $D_{\text{pmd_m}}$ = delay of optical modulation/demodulation

Proposal (option 1, 112.64 ns module)

To address the comments:

- Update Table 169-4 as shown below.
- Update each of the associated clauses 124, 162, 163, 167, and 173 to reflect the changes below.

Sublayer	Maximum (bit time) ^[1]	Maximum (pause_quanta) ^[2]	Maximum (ns)	Notes ^[3]
800G MAC, RS, and MAC Control	196 608	384	245.76	See 170.1.4.
800GBASE-R PCS or 800GXS ^[4]	640 000	1250	800	See 172.5.
800GBASE-R PMA	36 864	72	46.08	See 173.5.4.
800GBASE-R PMA 32:8 or 8:32	36 864	72	46.08	See 173.5.4.
800GBASE-R PMA 8:8	72 728	144	92.16	See 173.5.4.
800GBASE-KR8 PMD	32 768	64	40.96	Includes allocation of 14 ns for one direction through backplane medium. See 163.5.
800GBASE-CR8 PMD	32 768	64	40.96	Includes allocation of 14 ns for one direction through cable medium. See 162.5.
800GBASE-VR8 PMD	32 768 16 384	64 32	40.96 20.48	Includes 2 m of fiber. See 167.3.1.
800GBASE-SR8 PMD	32 768 16 384	64 32	40.96 20.48	Includes 2 m of fiber. See 167.3.1.
800GBASE-DR8 PMD	32 768 16 384	64 32	40.96 20.48	Includes 2 m of fiber. See 124.3.1.

Proposal (option 2, 122.88 ns module)

To address the comments:

- Update Table 169-4 as shown below.
- Update each of the associated clauses 124, 162, 163, 167, and 173 to reflect the changes below.

Sublayer	Maximum (bit time) ^[1]	Maximum (pause_quanta) ^[2]	Maximum (ns)	Notes ^[3]
800G MAC, RS, and MAC Control	196 608	384	245.76	See 170.1.4.
800GBASE-R PCS or 800GXS ^[4]	640 000	1250	800	See 172.5.
800GBASE-R PMA	36 864	72	46.08	See 173.5.4.
800GBASE-R PMA 32:8 or 8:32	40 960	80	51.2	See 173.5.4.
800GBASE-R PMA 8:8	81 920	160	102.4	See 173.5.4.
800GBASE-KR8 PMD	32 768	64	40.96	Includes allocation of 14 ns for one direction through backplane medium. See 163.5.
800GBASE-CR8 PMD	32 768	64	40.96	Includes allocation of 14 ns for one direction through cable medium. See 162.5.
800GBASE-VR8 PMD	32 768 16 384	64 32	40.96 20.48	Includes 2 m of fiber. See 167.3.1.
800GBASE-SR8 PMD	32 768 16 384	64 32	40.96 20.48	Includes 2 m of fiber. See 167.3.1.
800GBASE-DR8 PMD	32 768 16 384	64 32	40.96 20.48	Includes 2 m of fiber. See 124.3.1.

Summary

- Background provided for delay allocations for PMA and PMD.
- Two options proposed.
- Recommend option 1.

Thanks!