Deskew in 200/400GE SM-PMAs (Cl 176)
(in support of comment #368)

Matt Brown - Alphawave
Shawn Nicholl - AMD
Eugene Opsasnick, Jeff Slavick - Broadcom
Gary Nicholl, Adee Ran - Cisco
Xiang He - Huawei
Mike Dudek, Kapil Shrikhande - Marvell
Dave Ofelt - Juniper Networks
Tom Huber - Nokia
Zvi Rechtman - Nvidia

Supporters

- Andras DeKoos - Microchip
- Ali Ghiasi - Ghiasi Quantum LLC
- Jerry Pepper - Keysight
- Kent Lusted - Intel
- Jeffery Maki - Juniper Networks
- Ted Sprague – Infinera
- Leon Bruckman – Huawei
- Eric Maniloff - Ciena

200G/400G SM-PMA 4 CW interleave motivation

800GbE & 1.6TbE will always have 4 CW interleave for all configurations.

Having equivalent performance on 200Gbps lanes for all Ethernet rates is desired

Therefore we need to specify the 200GbE & 400GbE SM-PMA to provide 4 CW interleave, regardless of its location within the PHY, so that the system performance of 200GbE and 400GbE will be equivalent to that of 800GbE and 1.6TbE.

4 CW interleave in 200G/400G PMAs (Cl176, D1.0)

The goal of the SM-PMA is to have a series of 4 RS-FEC symbols from 4 different RS-FEC codewords transmitted one after the other in the data stream.

However, for 200GbE/400GbE, there are only 2 RS-FEC encoders available in the stream of data.

To enable interleaving of data between 4 CWs, the 200GbE/400GbE SM-PMAs m:n (when m!=n) delay the odd numbered PCS lanes by 2 CWs

4 RS-symbols from 4 unique CWs are multiplexed to an output lane in the symbol-pair multiplexing.

- RS-FEC symbol-pair muxing alternates between odd and even PCS lanes to achieve the 4 CW interleave

These functions are present in the 200GBASE-R 8:1, 1:8 PMAs as well as 400GBASE-R 16:2, 2:16 PMAs

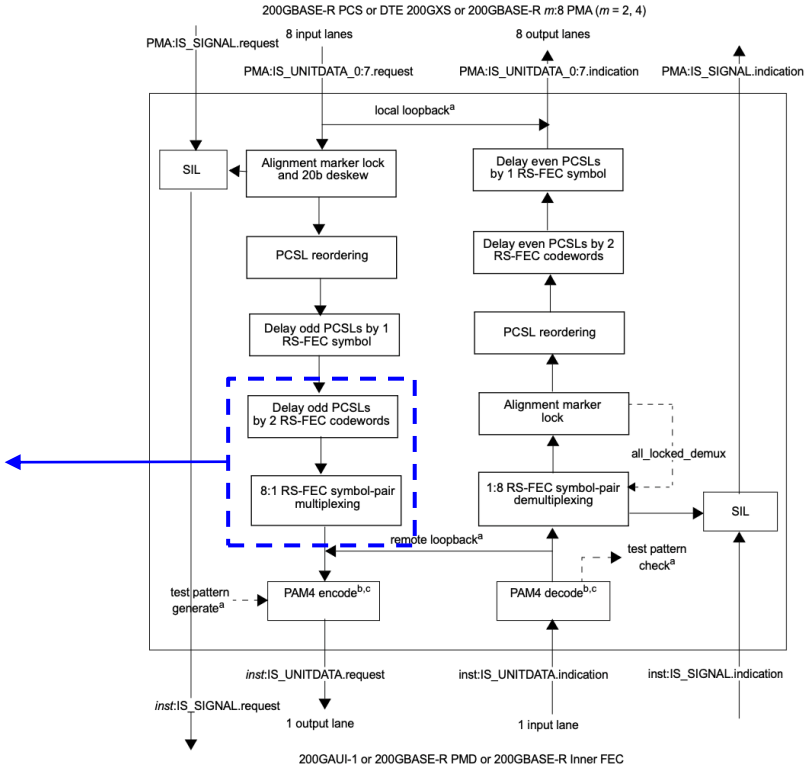

Fig 176-2 (200GBASE-R 8:1 PMA), Cl176 D1.0

Deskew function in 200G/400G PMAs (Cl176 D1.0)

20-bit deskew in the 200GbE/400GbE SM-PMAs (Cl176, D1.0) aligns RS-FEC symbol-pairs (20 bits) relative to AMs across all input lanes

- Present in 200GBASE-R 8:1, 1:8 PMAs
- Present in 400GBASE-R 16:2, 2:16 PMAs

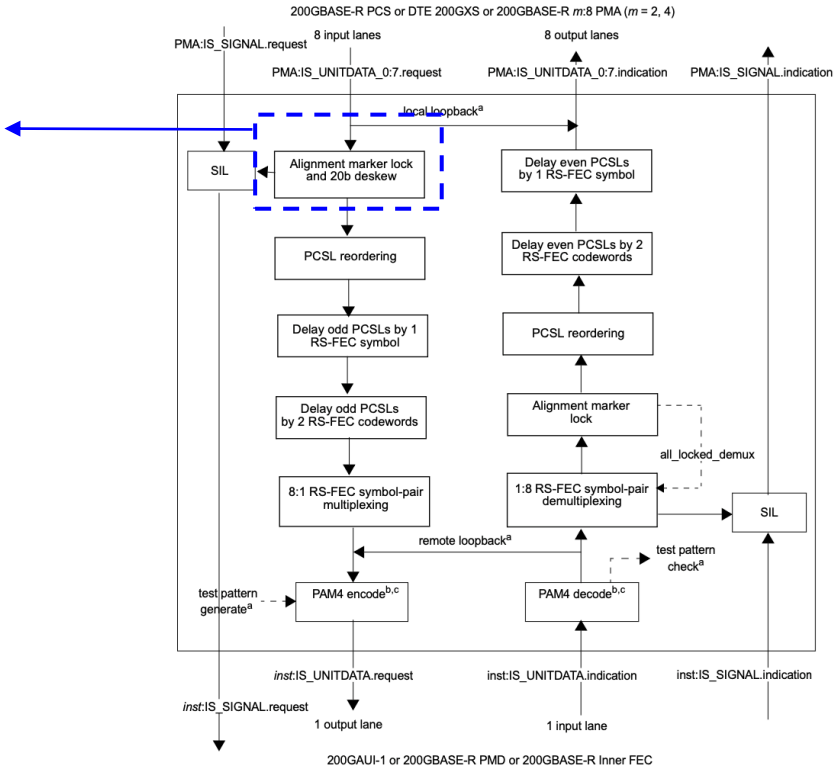

Fig 176-2 (200GBASE-R 8:1 PMA), Cl176 D1.0

Impact of skew on input lanes of the SM-PMAs

The 2 CW delay guarantees 4 CW interleaving at the output of the 8:1 symbol-pair mux when the input lanes of the SM-PMA do not have inter-lane skew (i.e. the AMs are aligned across all lanes).

- E.g. this occurs when the PMA is adjacent to the PCS and there is no inter-lane skew on PCS lanes

However, when there is inter-lane skew on the input lanes, the 2 CW delay cannot guarantee 4 CW interleaving unless the deskew function aligns codeword boundaries across the PCS lanes

- E.g. skew on input lanes of the PMA can occur when converting from 100Gbps to 200Gbps lanes

The 20b deskew in D1.0 is not sufficient to guarantee 4 CW interleaving for 200GbE/400GbE SM-PMAs

Note : 20b deskew is sufficient for 800GbE PMAs since there are 4 CWs to multiplex from (2 CWs per flow of the 800GbE PCS), so the datapath does not include a 2 CW delay (see Fig. 176-15 in D1.0)

Comment # 368 and referenced motion

Cl **176** SC **176.5.1.3.1** P **201** L **32** # 368

He, Xiang Huawei

Comment Type **TR** *Comment Status* **X**

20b deskew is incorrect. According to Motion #10 in https://www.ieee802.org/3/dj/public/23_07/motions_3cwdfdj_2307.pdf, it is required to deskew to codeword boundaries.

SuggestedRemedy

Remove the second and third paragraph in 176.5.1.3.1 and reuse 119.2.5.1.

Proposed Response *Response Status* **O**

Motion #10

Move to adopt the 4x RS codewords interleaving for 200GbE and 400 GbE using 200G/lane AUIs or PMDs, as shown in slides 4-6 and 10 of he_3dj_02a_2307 along with deskew (alignment) to codeword boundaries for 100G/lane input lanes.

M: Xiang He

S: Adee Ran

Technical (>=75%)

802.3 voters only

Result: passed by unanimous consent. 9:38 a.m.

- Motion # 10 adjustment to baseline presentation (he_3dj_02a_2307) was not implemented in D1.0
- The adjustment in the referenced motion requires deskew to codeword boundaries for 200/400GE
  - note : *100G/lane input lanes* is assumed to mean for the cases that are not n:n PMA
- Comment # 368 was submitted to correct this

Ref https://www.ieee802.org/3/dj/public/23_07/motions_3cwdfdj_2307.pdf

Alignment to codeword boundaries options

Options on how to implement the alignment to codeword boundaries per motion #10.

Note: Finest granularity of codeword alignment possible is 2 CW due to FEC-A/B interleave per PCS lane. Therefore, aligning to “codeword boundaries” is actually aligning to 2 codeword boundaries

1. Alignment to 2 CW boundaries
   - As stated in the motion #10, however this does not guarantee a 4 CW interleave in all skew conditions
2. Alignment to AM boundaries (full deskew)
   - Guarantees the 4 CW interleave under any inter-lane skew condition to comply with the intended goal of the SM-PMA symbol muxing
3. Alignment to 4 CW boundaries
   - Guarantees the 4 CW interleave under any inter-lane skew condition to comply with the intended goal of the SM-PMA symbol muxing

Option 1 : Alignment to 2 CW boundaries

Align the 8 PCS lanes that get 8:1 symbol multiplexed such that the interval between the AMs on any two PCS lanes is n * 2 CW, where n is an integer.

- The interval of 2 CW per PCS lane is 51.2ns at 200G and 25.6ns at 400G.

Problem with Option 1:

- After alignment is complete, if an odd lane precedes an even lane by exactly 2*CW, then adding the 2*CW constant delay to the odd lane brings the odd lanes exactly in phase with the even lane
- This will not result in 4 codeword interleaving as intended and instead results in a 2 CW (back to where we started)

Option 1 is not a viable option assuming 4 CW interleaving must be guaranteed under all inter-lane skew conditions and thus does not meet the intent stated on slide 3.

Option 2 : Alignment to AM boundaries (full deskew)

Align the 8 PCS lanes that get 8:1 symbol multiplexed such that all inter-lane Skew is removed.

This alignment process is similar to the deskew process used in the PCS Rx of CL119 (119.2.5.1).

After alignment is complete, the additional 2 CW delay on odd PCS lanes enables 4 CW interleave for all inter-lane skew conditions.

The data provided out of the SM-PMA will be deterministic (AM will appear with a known relationship between them)

Option 3 : Alignment to 4 CW boundaries

Align the 8 PCS lanes that get 8:1 symbol multiplexed such that the interval between the AMs on any two PCS lanes is n * 4 CW, where n is an integer.

- The interval of 4 CW per PCS lane is 102.4ns at 200GbE and 51.2ns at 400GbE.

After alignment is complete, the additional 2 CW delay on odd PCS lanes enables 4 CW interleave for all inter-lane skew conditions.

Comparison of Option 2 and 3

	Option 2 (Full deskew)[1]		Option 3 (4 CW alignment)	
	200GbE	400GbE	200GbE	400GbE
SP1 Deskew buffer Depth per PCS lane	29ns ~770 bits (~6k PMA)	29ns ~770 bits (~12k PMA)	29ns ~770 bits (~6k PMA)	29ns ~770 bits (~12k PMA)
SP6 Deskew Buffer Depth per PCS lane	160ns ~4250 bits (~33k PMA)	160ns ~4250 bits (~66k PMA)	~103ns ~2720 bits (~22k PMA)[2]	~51ns ~1360 bits (~22k PMA)
Design Complexity	Simple. Same as PCS deskew function.		Similar in complexity to full deskew.	
Verification Complexity	Lower due to single possible AM alignment		Higher due to more possible AM alignments	

[1] Full deskew is one particular instance of 4 CW alignment.
[2] Design supporting 2x200G and 1x400G operating modes would need ~44k bits

Summary

Option 1 does not guarantee 4 CW interleave for all inter-lane skew conditions.

Option 2 and Option 3 both guarantee 4 CW interleave for all inter-lane skew conditions.

Option 2 uses the same deskew process as the PCS, which is well understood in the industry

Option 3 is similar in complexity to option 2, but requires a smaller deskew buffer

Recommendation: Specify the codeword alignment process with either Option 2 or Option 3

Note: If we choose Option 3, an implementation could meet that by designing to Option 2

Backup

Skew allowed on SM-PMA input lanes for 3 different positions of the SM-PMA

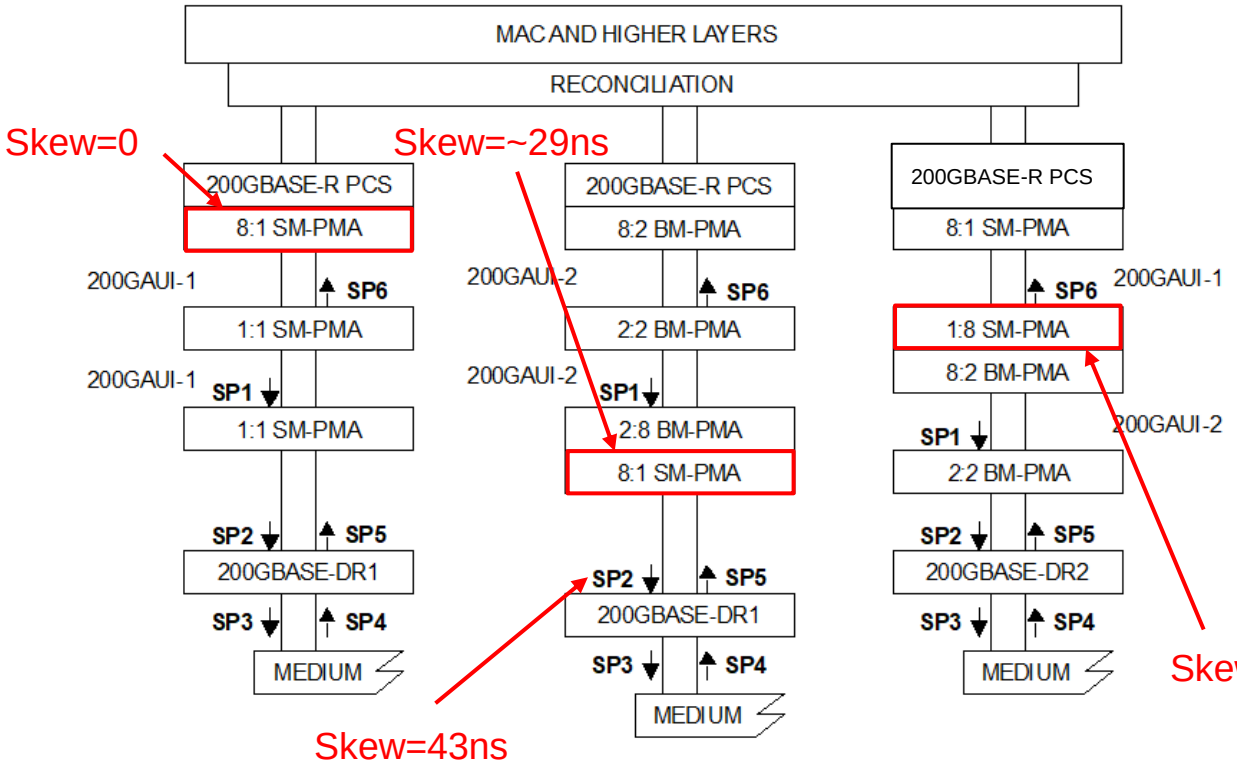

Table 116-8 (excerpt)
Same budget for 200GE and 400GE

Skew points	Maximum Skew (ns)[a]	Maximum Skew for 200GBASE-R or 400GBASE-R PCS lane (UI)[b]
SP1	29	≈ 770
SP2	43	≈ 1142
SP3	54	≈ 1434
SP4	134	≈ 3559
SP5	145	≈ 3852
SP6	160	≈ 4250
At PCS receive	180	≈ 4781

200G/400G full deskew @SP6

PCS Lane	AM arrival time	Deskew point	Delay added
0	102	154	52
1	0	154	154
2	26	154	128
3	25	154	129
4	154	154	0
5	51	154	103
6	52	154	102
7	1	154	**153**
Skew	154	0	

Assuming interleaving PCS lanes 0-7 in order 0,1,2,3,...7

Values are in ns and relative to first arriving AM prior to the deskew buffer

SP6 skew limit is 160ns at the output of the PMA

200G 4CW alignment @SP6

PCS Lane	AM arrival time	Alignment point	Delay added
0	102	154	52
1	0	51.6	51.6
2	26	51.6	25.6
3	25	51.6	26.6
4	154	154	0
5	51	51.6	0.6
6	52	154	**102**
7	1	51.6	50.6
Skew	154	102.4	

Assuming interleaving PCS lanes 0-7 in order 0,1,2,3,...7

Values are in ns and relative to first arriving AM prior to the deskew buffer

SP6 skew limit is 160ns at the output of the PMA

2 possible alignment points:
154ns [last arriving]
51.6 [4 CW before last]

400G 4CW alignment @SP6

PCS Lane	AM arrival time	Alignment point	Delay added
0	102	102.8	0.8
1	0	0.4	0.4
2	26	51.6	25.6
3	25	51.6	26.6
4	154	154	0
5	51	51.6	0.6
6	52	102.8	**50.8**
7	1	51.6	50.6
Skew	154	153.6	

Assuming interleaving PCS lanes 0-7 in order 0,1,2,3,...7

Values are in ns and relative to first arriving AM prior to the deskew buffer

SP6 skew limit is 160ns at the output of the PMA

4 possible alignment points:
154ns [last arriving]
102.8 [4 CW before last]
51.6 [8 CW before last
0.4 [12 CW before last]