Skew limits for 800 Gb/s Ethernet

(supporting comments 15, 16, 118)

Adee Ran

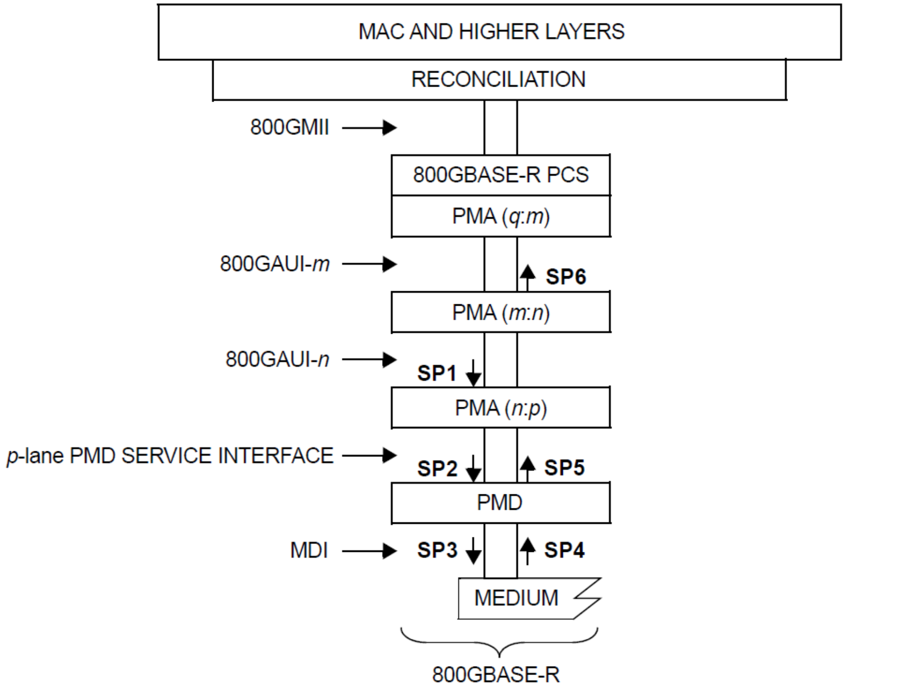
Support

- Piers Dawe, NVIDIA

Clarification of scope

- This presentation and comments 15, 16, and 118 pertain to **skew between multiple physical lanes** that carry differential signals.
- Skew between two single-ended signals is a different (and important) topic, but it is beyond the scope of this discussion.

Skew points and budget in D1.1

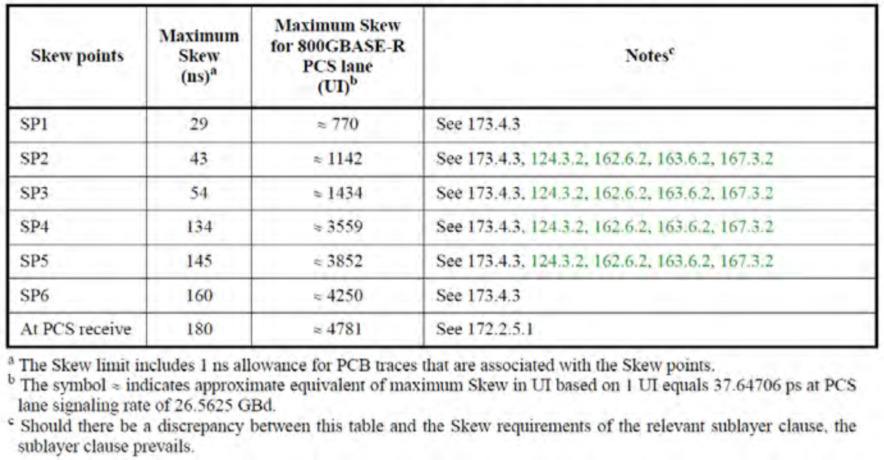


800GAUI-n = 800 Gb/s ATTACHMENT UNIT INTERFACE
800GMII = 800 Gb/s MEDIA INDEPENDENT INTERFACE
MAC = MEDIA ACCESS CONTROL
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PMA = PHYSICAL MEDIUM ATTACHMENT

PMD = PHYSICAL MEDIUM DEPENDENT
q=32
m=8
n=8
p=8

Figure 169–5—800GBASE-R Skew points for a PHY with multiple 800GAUI-n

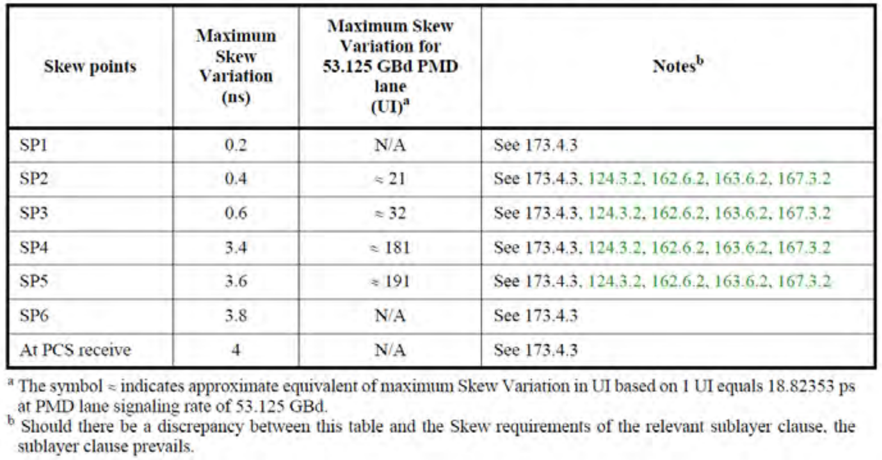
Table 169–5—Summary of Skew constraints



Skew points	Maximum Skew (ns)[a]	Maximum Skew for 800GBASE-R PCS lane (UI)[b]	Notes[c]
SP1	29	≈ 770	See 173.4.3
SP2	43	≈ 1142	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP3	54	≈ 1434	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP4	134	≈ 3559	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP5	145	≈ 3852	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP6	160	≈ 4250	See 173.4.3
At PCS receive	180	≈ 4781	See 172.2.5.1

[a] The Skew limit includes 1 ns allowance for PCB traces that are associated with the Skew points.
[b] The symbol ≈ indicates approximate equivalent of maximum Skew in UI based on 1 UI equals 37.64706 ps at PCS lane signaling rate of 26.5625 GBd.
[c] Should there be a discrepancy between this table and the Skew requirements of the relevant sublayer clause, the sublayer clause prevails.

Table 169–6—Summary of Skew Variation constraints



Skew points	Maximum Skew Variation (ns)	Maximum Skew Variation for 53.125 GBd PMD lane (UI)[a]	Notes[b]
SP1	0.2	N/A	See 173.4.3
SP2	0.4	≈ 21	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP3	0.6	≈ 32	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP4	3.4	≈ 181	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP5	3.6	≈ 191	See 173.4.3, 124.3.2, 162.6.2, 163.6.2, 167.3.2
SP6	3.8	N/A	See 173.4.3
At PCS receive	4	N/A	See 173.4.3

[a] The symbol ≈ indicates approximate equivalent of maximum Skew Variation in UI based on 1 UI equals 18.82353 ps at PMD lane signaling rate of 53.125 GBd.
[b] Should there be a discrepancy between this table and the Skew requirements of the relevant sublayer clause, the sublayer clause prevails.

Expanding on comment #15

- ***The skew constraints for 800 Gb/s in ns are the same as those for earlier generations, as early as 40 Gb/s, Table 80-8.***
  - The origin of the skew limits will be explored in this presentation.
- ***The size of PCS buffers required for deskewing grows linearly with the data rate; the size is quite large even at 400G, and would be doubled at 800G, due to the doubling of the number of PCS lanes. The current skew limit of 160 ns at the PCS receive requires about 150 kilobits per 800G port just for deskewing. This affects both latency and power consumption across the industry.***
  - In 800GBASE-R the *PCS UI* is ~37.65 ps, so 160 ns is ~4250 *PCS UI*, and each PCS lane needs a separate deskew buffer; for 32 PCSLs, the total buffer size is at least **136 kilobits**
  - When defined in 802.3ba, the total buffer size was 18.5 kilobits for 100GBASE-R and 7.5 kilobits for 40GBASE-R
  - Note that actual delay is caused by actual skew; but the current limits allow a high delay of 160 ns.
- ***The original skew limits were probably exaggerated even for 40G, and there is no need to carry them on for new technologies and new PCS designs.***
  - We will show that the numbers are vastly exaggerated.
- ***The numbers we set in 802.3df will also affect hosts and modules (with XS) in 802.3dj, so are worth considering carefully now.***
  - P802.3df defines the new 800G PCS; **now is the time to define the skew limits – they can't be changed in dj**

The origin of skew and skew variation limits

- A lot of discussion in 802.3ba
  - anslow_01_0508 – discussed dynamic skew.
  - giannakopoulos_01_0508 – discussed how skew can be introduced by endpoints and media
  - kolesar_01_0508 suggested max skew of 4.5 ns per 100m of OM3 parallel fiber, and notes that “The actual skew observed in real cables is far lower”
  - giannakopoulos_01_0708
  - giannakopoulos_01_1108 suggested maximum values at each skew point.
  - isono_01_0109 suggested increasing skew allocations for PMDs to account for Thin Film Filter optical mux/demux.
- Additional discussion in 802.3cd
  - wertheim_010417_3cd_adhoc – suggested reducing skew variation for PHYs with 25G PCS/FEC lanes
  - brown_112316_3cd_adhoc – suggested reducing skew and skew variation for single-lane PHYs

Digging into the skew budget

- The initial skew budget is described in giannakopoulos_01_1108
- The major parts of that budget are
  - PMA skew – stated as **25.5 ns** in the Tx (including PCS output) and **14.3 ns** in the Rx (up to the PCS input)
  - Medium skew – stated as 13.6 ns for parallel fiber
    - We have a much larger value today – see below
- PMD skew allowance of additional **11 ns in each direction** was added later.
- Other contributions stated are negligible in comparison.

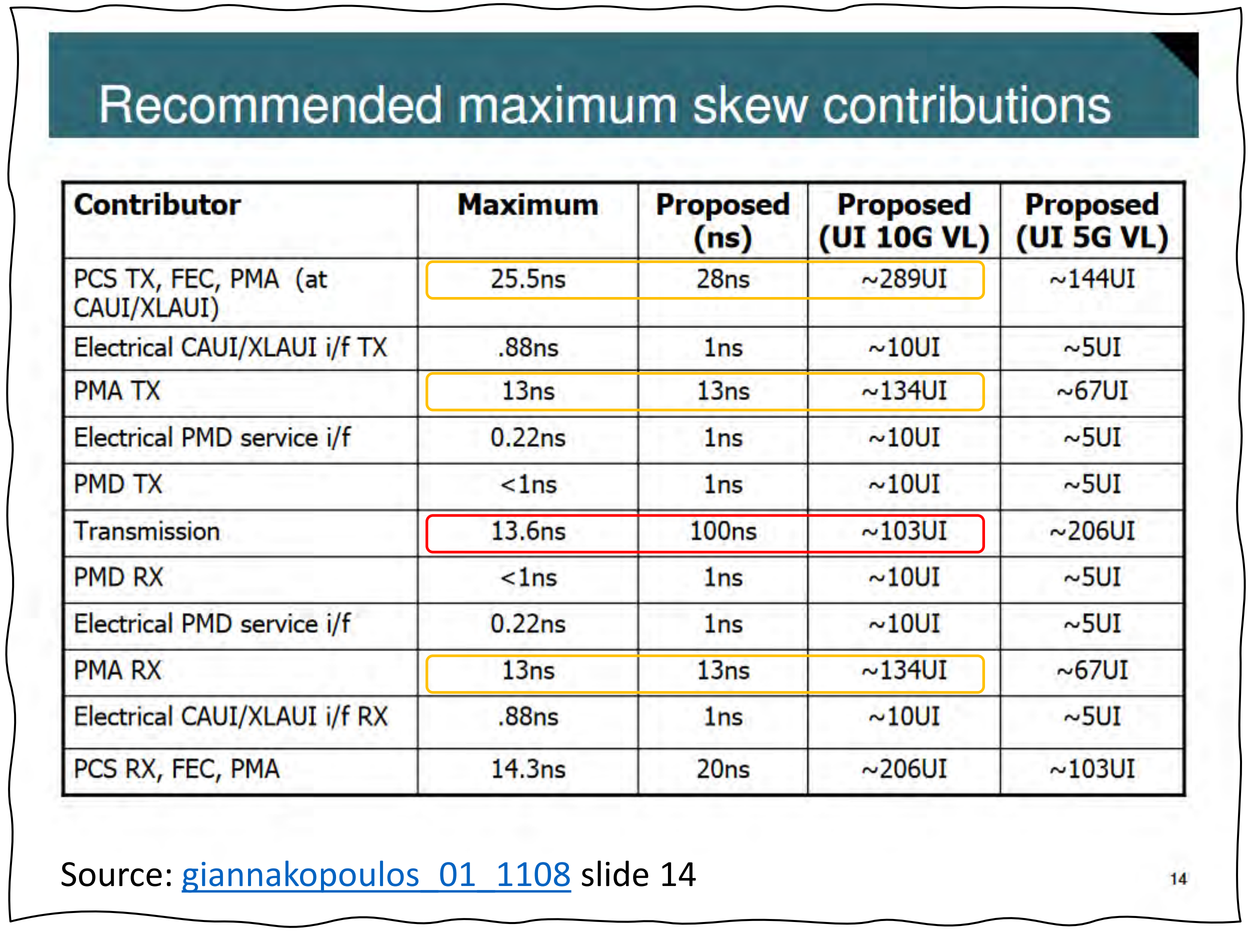
PMA skew allowance is exaggerated

- giannakopoulos_01_0508 states (in slide 7) that in "ASIC" implementation the Tx skew is up to 2 ns and Rx skew is similar
  - Based on 644 MHz SerDes interface, this is still realistic today
- It then describes "FPGA solution with external 10G SerDes devices" (slides 8-10)
  - Interface to external SerDes is assumed to be a 16-bit bus per lane (as in XSBI from 10G Ethernet)
  - Skew for stages feeding the "Internal SerDes" for the 16-bit busses is allocated 12.8 ns in each direction
  - Tx is allocated an additional 11.2 ns due to FIFOs for the 16-bit wide interfaces
  - This scenario is completely obsolete nowadays; we should not carry it forward and tax the whole market
- Also assumes 4" difference in PCB routing between lanes of the same port (CAUI-10?), adding 1.76 ps in both Rx and Tx
  - 800G ports will likely have even smaller routing differences – but this component is relatively small
- The maximum skew contributed by modern PMAs is estimated as ***64 PCS UI*** **(128 UI for 100 Gb/s per lane SerDes)** or **~2.4 ns per Tx or Rx**
  - A retimer or module has both Tx and Rx, therefore **128 *PCS UI* or ~4.8 ns**
- The proposed limits are based on this value per PMA; They can be recalculated for other values

Medium skew allowance is exaggerated

It is unclear how “transmission” (medium) was allocated 100 ns

- No data supporting this large skew
- Earlier slides have 13.6 ns for SR (considered worst-case), and ~1 ns for LR
- From comparison to other rows (highlighted), it might be a calculation error (~103 UI corresponds to **10 ns**)
- At some later point this number was reduced to **80 ns** (20 ns was moved to optical skew in the PMDs)
- This skew allocation for the medium (SP3-SP4) has been carried over up to 400GbE



Recommended maximum skew contributions

Contributor	Maximum	Proposed (ns)	Proposed (UI 10G VL)	Proposed (UI 5G VL)
PCS TX, FEC, PMA (at CAUI/XLAUI)	25.5ns	28ns	~289UI	~144UI
Electrical CAUI/XLAUI i/f TX	.88ns	1ns	~10UI	~5UI
PMA TX	13ns	13ns	~134UI	~67UI
Electrical PMD service i/f	0.22ns	1ns	~10UI	~5UI
PMD TX	<1ns	1ns	~10UI	~5UI
Transmission	13.6ns	100ns	~103UI	~206UI
PMD RX	<1ns	1ns	~10UI	~5UI
Electrical PMD service i/f	0.22ns	1ns	~10UI	~5UI
PMA RX	13ns	13ns	~134UI	~67UI
Electrical CAUI/XLAUI i/f RX	.88ns	1ns	~10UI	~5UI
PCS RX, FEC, PMA	14.3ns	20ns	~206UI	~103UI

Source: giannakopoulos_01_1108 slide 14

14

Is this 80 ns allocation really needed?

- WDM skew is far less than 80 ns
- Parallel MMF – maximum estimated as 13.6 ns
- We also have an objective for 2 km over parallel SMF
  - The adopted baseline proposal, welch_3df_01a_220222, mentions the 80 ns as maximum skew (slide 9)
  - However, on the same slide it is stated that "Skew for unbent fiber usual low ~3ps/m ~6ns for 2km" and "Bent fiber <…> would only be expected to occur over a short net effective length out of a 2km span."
  - No data was provided to justify a need for 80 ns.
- Maybe a consideration for very large reach (80 km?)
  - PHYs for very high reaches (e.g., ZR) use different PCS/FEC and do not expose skew to the BASE-R PCS.
- Allocation of **20 ns** for medium skew is proposed (with guard band over the quoted values 13.6 ns and 6 ns).

PMD skew allowance is exaggerated

- Skew contributions of the PMD (SP3-SP2 and SP5-SP4) are 11 ns each
- The origin of that skew could be the Thin Film Filter optical mux/demux mentioned in isono_01_0109
  - The presentation (and related comment #280 against 802.3ba D1.1) asked for 10 ns for each Tx and Rx (2 m difference between fibers in the PMD?), which was accepted
  - Unclear when and why 10 ns was changed to 11 ns
  - These are significant allowances
  - Whether this large skew occurs anywhere in practice is questionable
- PMD skew allocation of **4.8 ns** in each direction is proposed
  - Same as the PMA
  - This would even allow an additional two-SerDes retimer inside the PMD (in addition to the module's PMA)

Is the proposed reduction too aggressive?

- Every component listed has its own guard band.
- Not all systems have all skew contributors (for example, two AUIs on each side and internal PMD skew and media skew).
- The “digital” skew components are stochastic, and may change across resets for every component.
    - These are more than 50% of the total maximum skew budget.
- The maximum skews of each of the components are unlikely to happen together and constructively.
- **Reducing the maximum total skew at the PCS input should be safe.**

Proposed skew limits (to update Table 169-8)

Skew point	Contributor	Maximum (PCS UI)	Cumulative (PCS UI)	Cumulative (ns)	Reason
SP1	Tx PCS/PMA and possible external PMA	192	~~≈770~~ 192	~~29~~ ≈7.2	PCS/PMA Tx + PMA Rx + PMA Tx
SP2	Module PMA	128	~~≈1142~~ 320	~~43~~ ≈12	PMA Rx + PMA Tx
SP3	Module PMD Tx	128	~~≈1434~~ 448	~~54~~ ≈16.9	In case the module has an additional internal 2-SerDes retimer
SP4	Medium	512	~~≈3559~~ 960	~~134~~ ≈36.1	~20 ns maximum parallel fiber skew
SP5	Module PMD Rx	128	~~≈3852~~ 1088	~~145~~ ≈41	As in PMD Tx
SP6	Module PMA and external PMA	256	~~≈4250~~ 1344	~~160~~ ≈50.6	Two PMAs, each with Rx and Tx
PCS input	Rx PCS/PMA	64	~~≈4781~~ 1408	~~180~~ ≈53	PMA Rx

Note: PCS UI is ~37.6 ps; for 800GBASE-R, the maximum UI on any physical interface is ~18.8 ps

Skew variation

- Has a lower impact on PCS buffer size than maximum skew, but increases delay, and can affect gearbox design
- The current limit is 4 ns at the PCS Rx input, dominated by the media contribution (SP4-SP3 = 2.8 ns)
- Comment #16 suggests reducing this component to 0.7 ns (25%)
- After examining previous work (as described in the next slide), a modified suggested remedy is a more modest reduction to 1.4 ns (50%)

Skew variation – previous work

- Slide 13 of anslow_01_0508 shows optical dynamic skew of less than 1 ns for all WDM fibers
  - The case of 80 km with 1550 nm (estimated as ~9.5 ns) is another exception, which didn't make it into the current skew variation specification. Even if Ethernet defines PHYs for this range, they will likely use a different PCS, which will handle skew variation. All other sublayers (which may be used within a PHY extender) should not be burdened.
- The parallel fiber PMDs are listed with higher total dynamic skew, up to 1.07 ns at the Rx PCS input
  - Total (optical and electrical contributions) are all much less than 4 ns
  - The highest value is for 300 m, a total of 2.42 ns at the Rx PCS input
  - The presentation refers to kolesar_01_0508; where slide 13 suggests 6.8 ps/m maximum dynamic skew for the parallel MMF
  - 802.3df objectives include MMF for reaches only up to 100 m
  - Taking 200 m as a guard band, we can cap the SP4-SP3 skew variation at **1.4 ns or ~106 UI.**

Proposed skew variation limits (to update Table 169-6)

Skew point	Contributor	Maximum (PMD UI)	Cumulative (PMD UI)	Cumulative (ns)
SP1	Tx PCS/PMA and possible external PMA	10.5	**N/A**	**0.2**
SP2	Module PMA	10.5	**≈21**	**0.4**
SP3	Module PMD Tx	10.5	**≈32**	**0.6**
SP4	Medium	74	~~**≈181**~~ **≈106**	~~3.4~~ **2**
SP5	Module PMD Rx	10.5	~~**≈191**~~ **≈116**	~~3.6~~ **2.2**
SP6	Module PMA and external PMA	10.5	**N/A**	~~3.8~~ **2.4**
PCS input	Rx PCS/PMA	10.5	**N/A**	4 **2.6**

The only change is the medium contribution – but that affects all subsequent rows in the table

Note: PMD UI for all PMDs defined in 802.3df is ~18.8 ps