Comparison and Proposal for Multi-PCS Lane Distribution Path Delay Variance

MICROCHIP

A Leading Provider of Smart, Connected and Secure Embedded Control Solutions

SMART | CONNECTED | SECURE

Richard Tse

IEEE 802.3cx Teleconference Nov 17, 2020

Supporters:

- Andras de Koos, Microchip Technology
- Clark Carty, Cisco Systems
- Denny Wong, Xilinx
- Dino Pozzebon, Microchip Technology
- Marek Hajduczenia, Charter Communications
- Mark Bordogna, Intel
- Richard Tse, Microchip Technology
- Sriram Natarajan, Cisco Systems
- Steve Carlson, High Speed Design Inc.
- Ulf Parkholm X, Ericsson

Proposed Text (originally from tse_3cx_02a_0920.pdf)

- **Enhance existing text in 90.7 on FEC so it also deals with multi-lane PCS.**
- **Replace “SFD” with “message timestamp point” throughout 90.7 (not all are shown below). Definition of “message timestamp point” to be added later. See tse_3cx_02_1120.pdf.**
- **Insertions are highlighted in** blue **and deletions are highlighted in** ~~red~~**.**

For a PHY that includes an FEC and/or multilane distribution functions, the transmit and receive path data delays may show significant variation depending upon the position of the ~~SFD~~message timestamp point within the FEC block and in the multilane distribution sequence. However, since the variation due to this effect in the transmit path is expected to be compensated by the inverse variation in the receive path, it is recommended that the transmit and receive path data delays be reported as if the ~~SFD~~message timestamp point is at the start of the FEC block and multilane distribution sequence. For PHYs with both FEC and multilane distribution, the start of the FEC block is guaranteed to coincide with the start of a multilane distribution sequence.

Comparison of Proposals

Pros and Cons Summary

- Comparing solutions for timestamping on multi-PCS lane PHYs recommended by tse_3cx_02a_0520.pdf and he_3x_01_0920.pdf and described by tse_3cx_02a_0420.pdf.

Characteristic	Advantage		
	"Option A + Method 1"	"Option B + Method 1"	"Option C + Method 2"
Has intrinsic timestamp granularity limit of "1 bit"	✓	✓	✓
Satisfies zero Tx skew recommendation	✗	✓	✓
Compatible with other PHY functions with variable delays	✗	✗	✓
Allows 802.3 PCS delay registers to be used for high accuracy applications	✗	✗	✓
Allows high accuracy with separated MAC and PHY	✗	✗	✓

Timestamp granularity of 1 bit

- Discussed at Oct 17, 2020 ad-hoc meeting
  - See wong_3cx_01_1020.pdf and he_3cx_01_1020.pdf
  - “Option A + Method 1”
    - Each xMII word gets a unique timestamp
    - PCS delay is constant regardless of the PCS lane
  - “Option B + Method 1”
    - Timestamps at all the Tx PCS output are identical, thus multiple successive xMII words have the same timestamp
    - However, each PTP message will still get a unique timestamp
  - “Option C + Method 2”
    - Each xMII word gets a unique timestamp
    - PCS delay is modelled as a constant regardless of the PCS lane
  - All of the above choices have a timestamp granularity of 1 bit per PTP message

Zero Tx Skew Recommendation

- There was general agreement to recommend targeting zero Tx skew to get maximum timestamp accuracy
  - See dekoos_3cx_01_1020.pdf from the last ad-hoc meeting
  - "Option A + Method 1" inherently adds Tx skew of one 66B block for each successive lane and, thus, does not satisfy this recommendation
  - "Option B + Method 1" and "Option C + Method 2" satisfy this recommendation

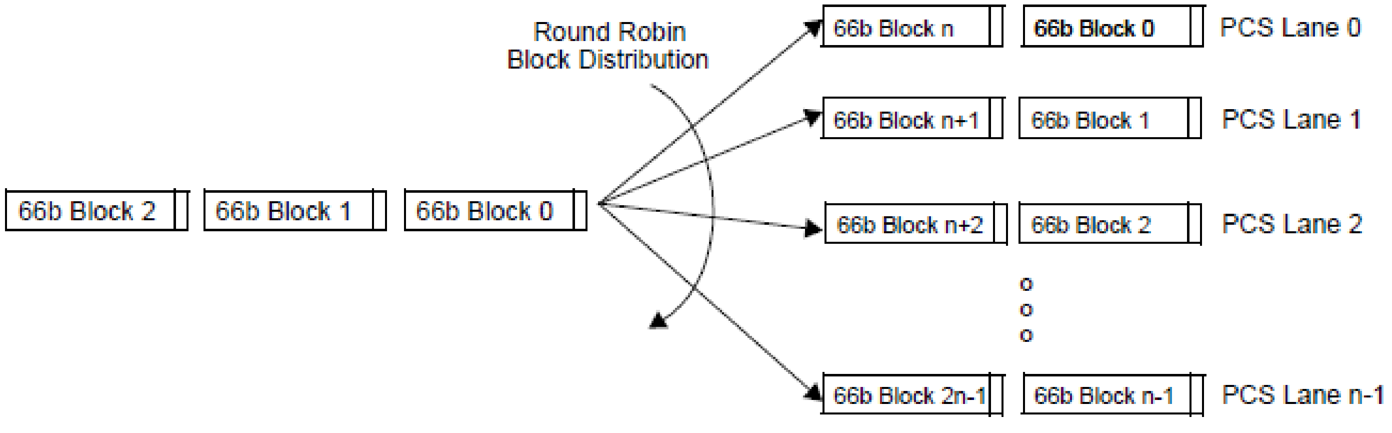


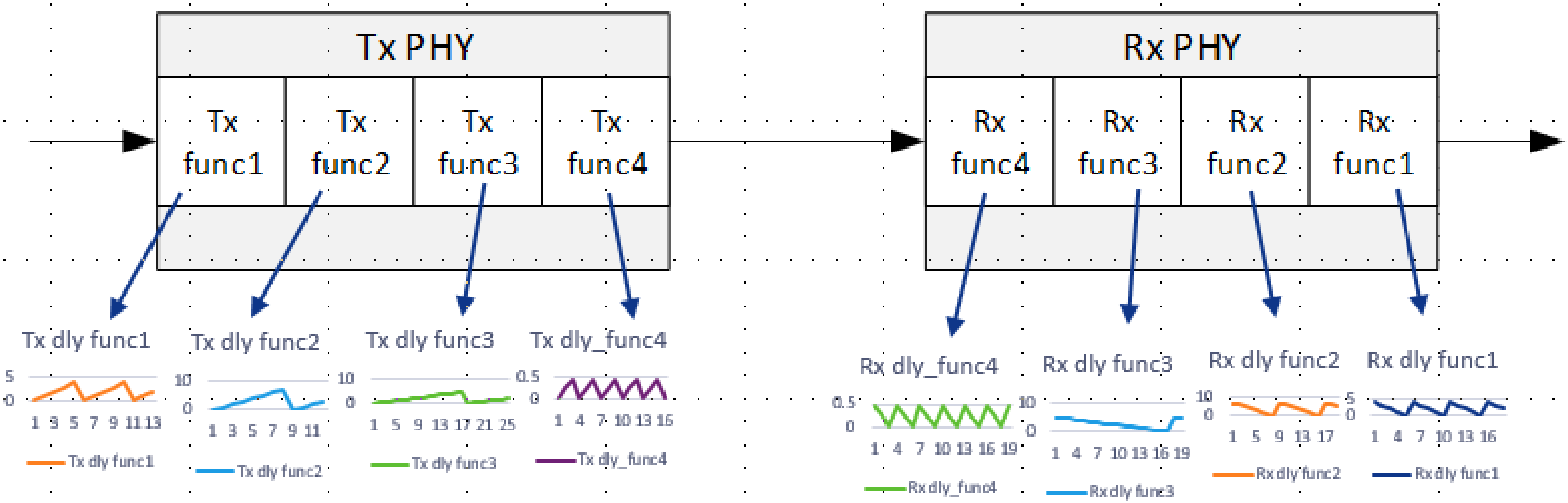
Figure 82–6—PCS Block distribution

Compatibility with Other PHY Functions with Variable Delays (1/3)

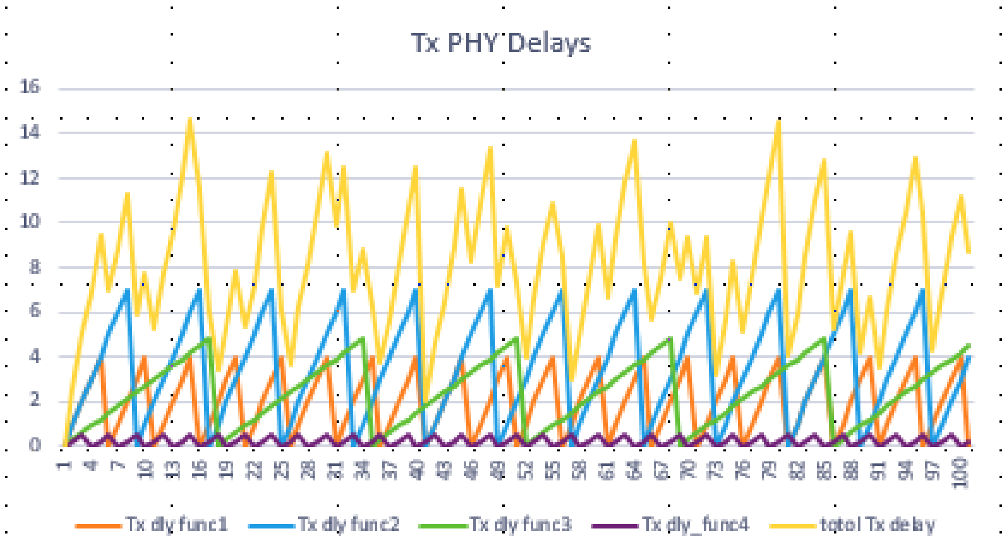
- Proposed solution (“Option C + Method 2) complements the specified solution for dealing with FEC delays
  - See tse_3cx_01_1020.pdf for details on FEC delays (transcoding and FEC lane distribution)
  - Variable FEC delay is modelled as a constant value as this specified solution takes advantage of the fact that the Tx FEC delay and the Rx FEC delay sum to a constant value
- Alternate solution (“Option B + Method 1) is contradictory to the specified solution for dealing with FEC delays
  - Multi-PCS lane distribution delays are different for each lane instead of being modelled as a constant value as per FEC lane distribution

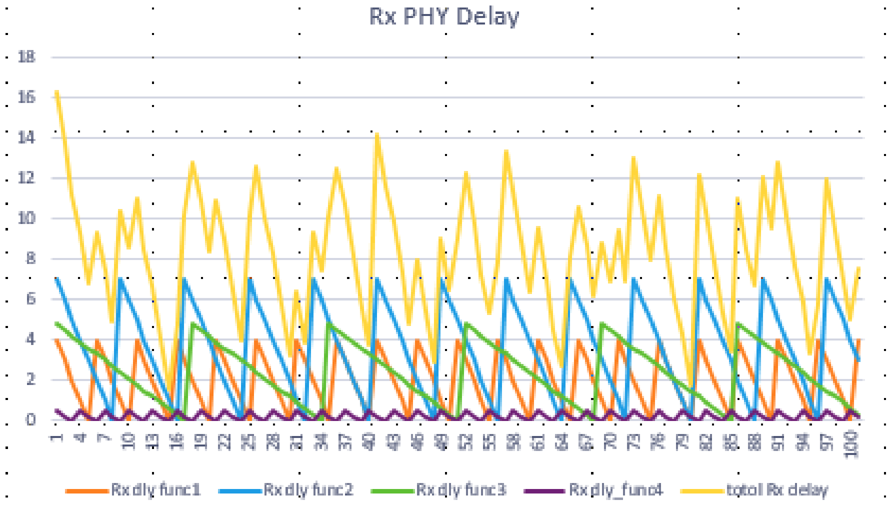
Compatibility with Other PHY Functions with Variable Delays (2/3)

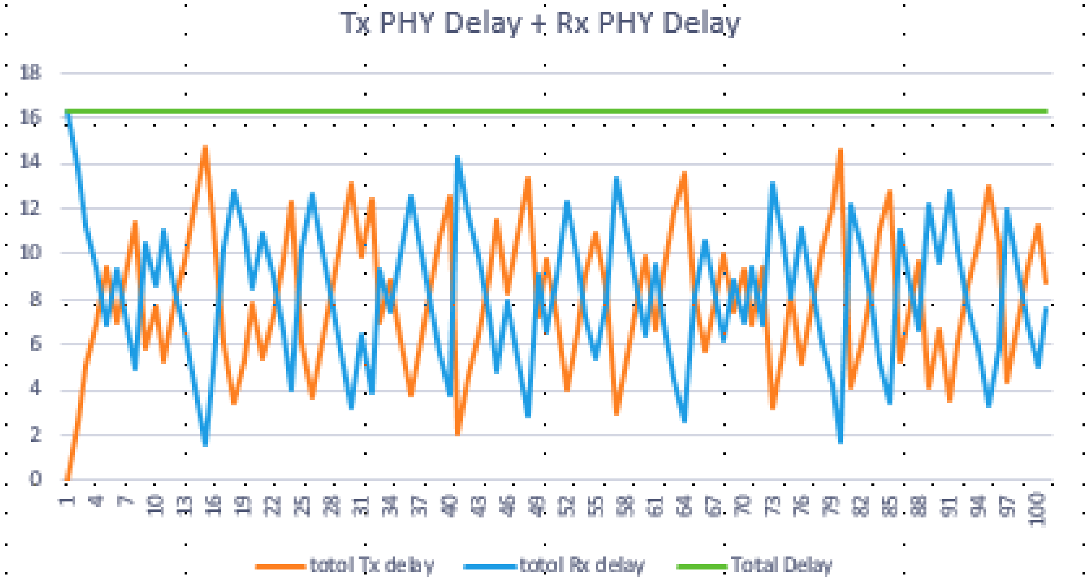
- tse_3cx_03_0520.pdf shows that the delays of cascaded PHY functions, which each have varying delays, can be modelled as an aggregated constant value



Compatibility with Other PHY Functions with Variable Delays (3/3)

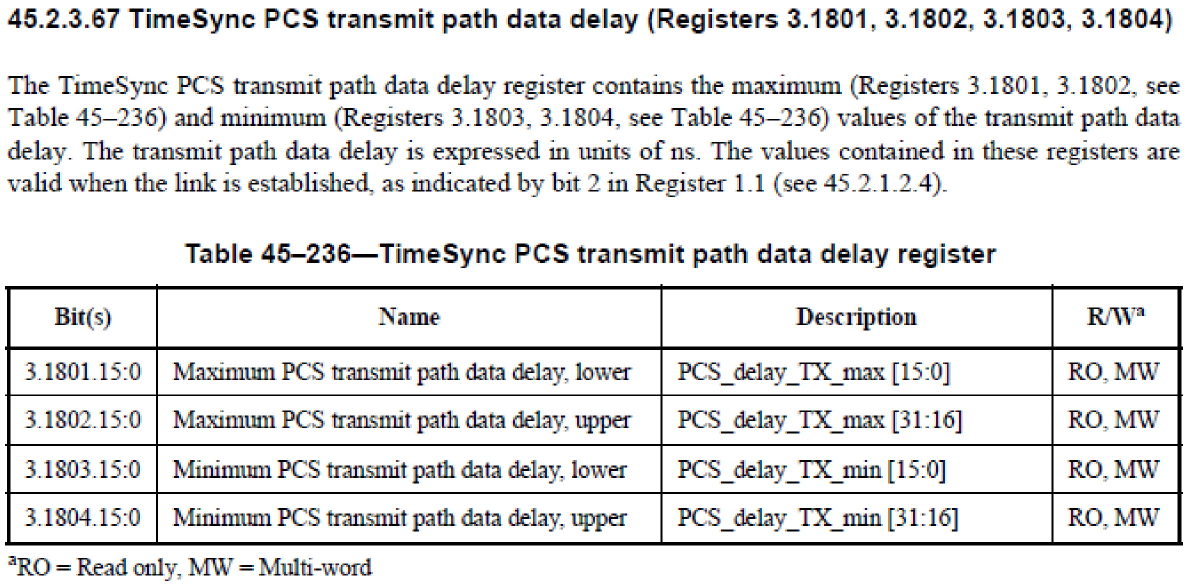






Allows 802.3 PCS Delay Registers to be Used (1/2)

- Existing PCS path data delay registers specify static values for the minimum and maximum PCS delays



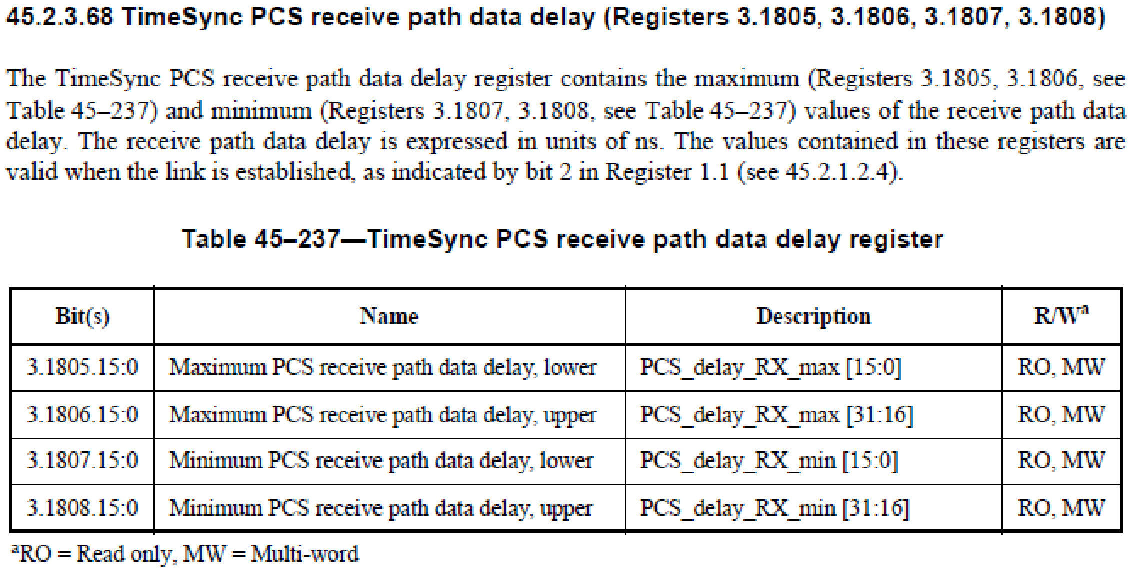
45.2.3.67 TimeSync PCS transmit path data delay (Registers 3.1801, 3.1802, 3.1803, 3.1804)

The TimeSync PCS transmit path data delay register contains the maximum (Registers 3.1801, 3.1802, see Table 45–236) and minimum (Registers 3.1803, 3.1804, see Table 45–236) values of the transmit path data delay. The transmit path data delay is expressed in units of ns. The values contained in these registers are valid when the link is established, as indicated by bit 2 in Register 1.1 (see 45.2.1.2.4).

Table 45–236—TimeSync PCS transmit path data delay register

Bit(s)	Name	Description	R/W[a]
3.1801.15:0	Maximum PCS transmit path data delay, lower	PCS_delay_TX_max [15:0]	RO, MW
3.1802.15:0	Maximum PCS transmit path data delay, upper	PCS_delay_TX_max [31:16]	RO, MW
3.1803.15:0	Minimum PCS transmit path data delay, lower	PCS_delay_TX_min [15:0]	RO, MW
3.1804.15:0	Minimum PCS transmit path data delay, upper	PCS_delay_TX_min [31:16]	RO, MW

[a] RO = Read only, MW = Multi-word



45.2.3.68 TimeSync PCS receive path data delay (Registers 3.1805, 3.1806, 3.1807, 3.1808)

The TimeSync PCS receive path data delay register contains the maximum (Registers 3.1805, 3.1806, see Table 45–237) and minimum (Registers 3.1807, 3.1808, see Table 45–237) values of the receive path data delay. The receive path data delay is expressed in units of ns. The values contained in these registers are valid when the link is established, as indicated by bit 2 in Register 1.1 (see 45.2.1.2.4).

Table 45–237—TimeSync PCS receive path data delay register

Bit(s)	Name	Description	R/W[a]
3.1805.15:0	Maximum PCS receive path data delay, lower	PCS_delay_RX_max [15:0]	RO, MW
3.1806.15:0	Maximum PCS receive path data delay, upper	PCS_delay_RX_max [31:16]	RO, MW
3.1807.15:0	Minimum PCS receive path data delay, lower	PCS_delay_RX_min [15:0]	RO, MW
3.1808.15:0	Minimum PCS receive path data delay, upper	PCS_delay_RX_min [31:16]	RO, MW

[a] RO = Read only, MW = Multi-word

Allows 802.3 PCS Delay Registers to be Used (2/2)

- Proposed solution (“Option C + Method 2”) models the variable PCS delay as a constant value
  - Difference between minimum and maximum PCS delay register values can be small (difference could, conceptually, be 0ns)
  - Register can be used for high accuracy timestamping applications

- Alternate solution (“Option B + Method 1”) has dynamically varying PCS delay
  - Difference between the minimum and maximum PCS delay register values will be larger (e.g., 100GE has intrinsic PCS delay variance of 12.16ns)
  - Register might not be compatible with high accuracy timestamping applications

Allows High Accuracy with Separated MAC and PHY

- Proposed solution (Option C + Method 2)
  - PCS delay is modelled as a constant value, which can be used by an external MAC to timestamp accurately at its xMII
- Alternate solution (Option B + Method 1)
  - A MAC that is separated from the PCS is not inherently able to know which lane the message timestamp point arrived on (for Rx) or will appear on (for Tx), thus it cannot determine the dynamic delay of the PCS function

SMART | CONNECTED | SECURE

Thank You

Alternate implementation option (1/2)

- **Add the following text to Clause 90.7**

  Block distribution in a multi-lane PCS causes variance in the path data delay. Because the data stream crossing the transmit xMII is the same as the data stream crossing the receive xMII, the sum of the transmit block distribution functional delay and the receive block distribution functional delay is the same for every PCS lane.

  For a transmit PHY that performs block distribution from the xMII to multiple PCS lanes (e.g., the 100GBASE-R PCS in clause 82), the path data delay variance experienced by blocks transiting from the xMII to different PCS lanes is treated as a constant value. The constant value that represents the block distribution function's delay is equal to half of the difference between the shortest distribution time from the xMII to a PCS lane (e.g., for lane N of an N-lane PCS) and the largest distribution time from the xMII to a PCS lane (e.g., for lane 0).

Alternate implementation option (2/2)

- **Add the following text to Clause 90.7, continued…**

  For a receive PHY that performs block distribution from multiple PCS lanes to the xMII (e.g., the 100GBASE-R PCS in clause 82), the path data delay variance experienced by blocks transiting from the per-lane outputs of the deskew buffer to the xMII is treated as a constant value. The constant value that represents the block distribution function's delay is equal to half of the difference between the shortest distribution time from the output of a deskew buffer lane to the xMII (e.g., for lane 0) and the largest distribution time from the output of a deskew buffer lane to the xMII (e.g., for lane N of an N-lane PCS).

  The constant value for the receive PHY is equal to the constant value for the transmit PHY. This constant value can be used to represent the multi-lane block distribution function's portion of the PCS delay when using the TimeSync PCS transmit path data delay and the TimeSync receive path data delay.