More Discussion on Multi-PCS Lane Distribution Path Delay Variance

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

Richard Tse

IEEE 802.3cx ad-hoc Teleconference Oct 20, 2020

Recall…

- At the 802.3cx meeting in Sept 2020, it was noted that the suggestion from he_3x_01_0920.pdf for dealing with multi-PCS lane path data delay variation is inconsistent with what IEEE 802.3 has already specified for multi-lane FEC.
- This contribution gives details on this topic.

MICROCHIP

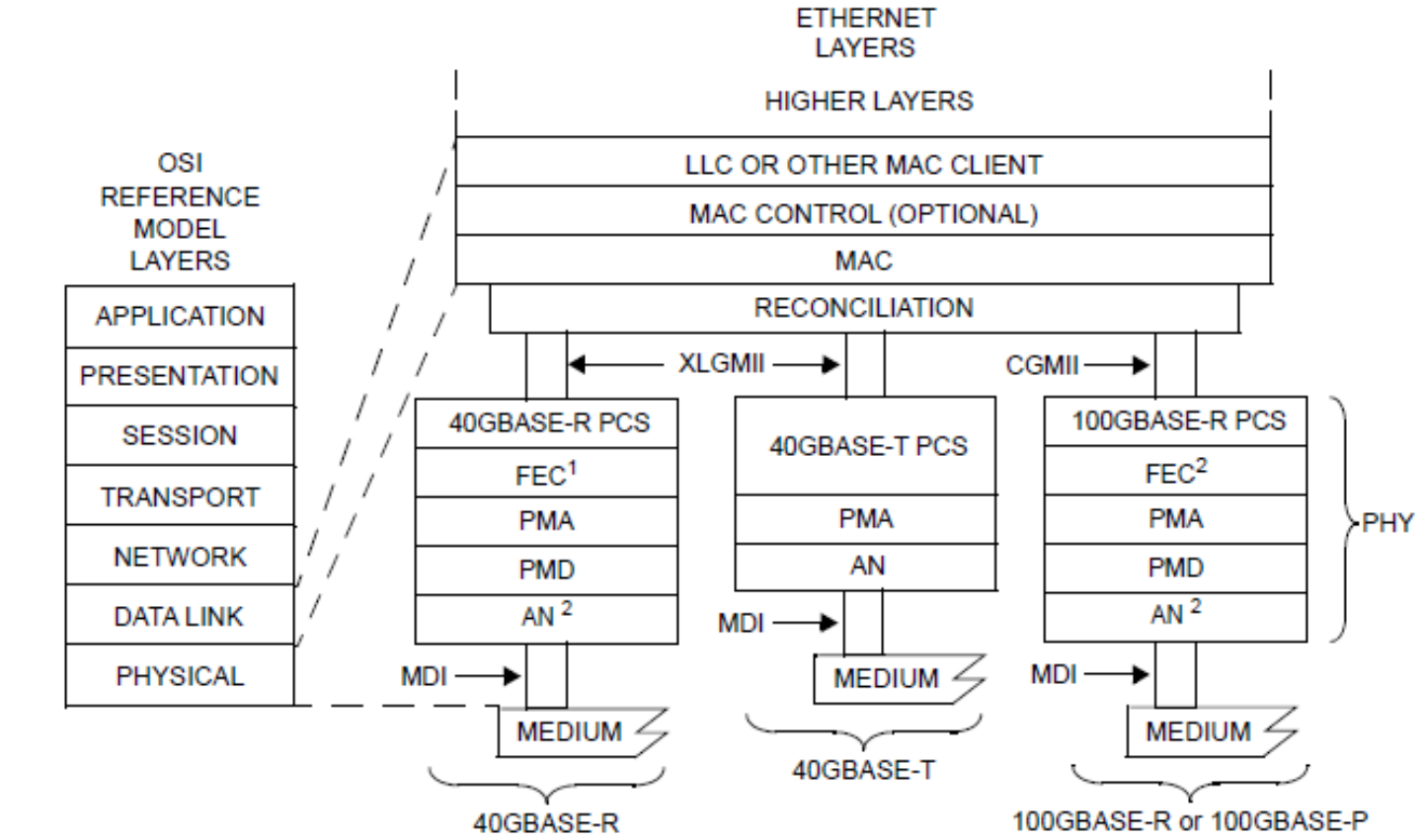
From Subclause 90.7 of IEEE 802.3-2018

For a PHY that includes an FEC function, the transmit and receive path data delays may show significant variation depending upon the position of the SFD within the FEC block. 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 is at the start of the FEC block.

MICROCHIP

40GE/100GE Architecture

- From clause 80



AN = AUTO-NEGOTIATION
CGMII = 100 Gb/s MEDIA INDEPENDENT INTERFACE
FEC = FORWARD ERROR CORRECTION
LLC = LOGICAL LINK CONTROL
MAC = MEDIA ACCESS CONTROL
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
XLGMII = 40 Gb/s MEDIA INDEPENDENT INTERFACE

NOTE 1—OPTIONAL OR OMITTED DEPENDING ON PHY TYPE
NOTE 2—CONDITIONAL BASED ON PHY TYPE

Figure 80–1—Architectural positioning of 40 Gigabit and 100 Gigabit Ethernet

MICROCHIP

40GE/100GE PCS

- From clause 82

Figure 82–2 provides a functional block diagram of the 40GBASE-R PCS and 100GBASE-R PCS.

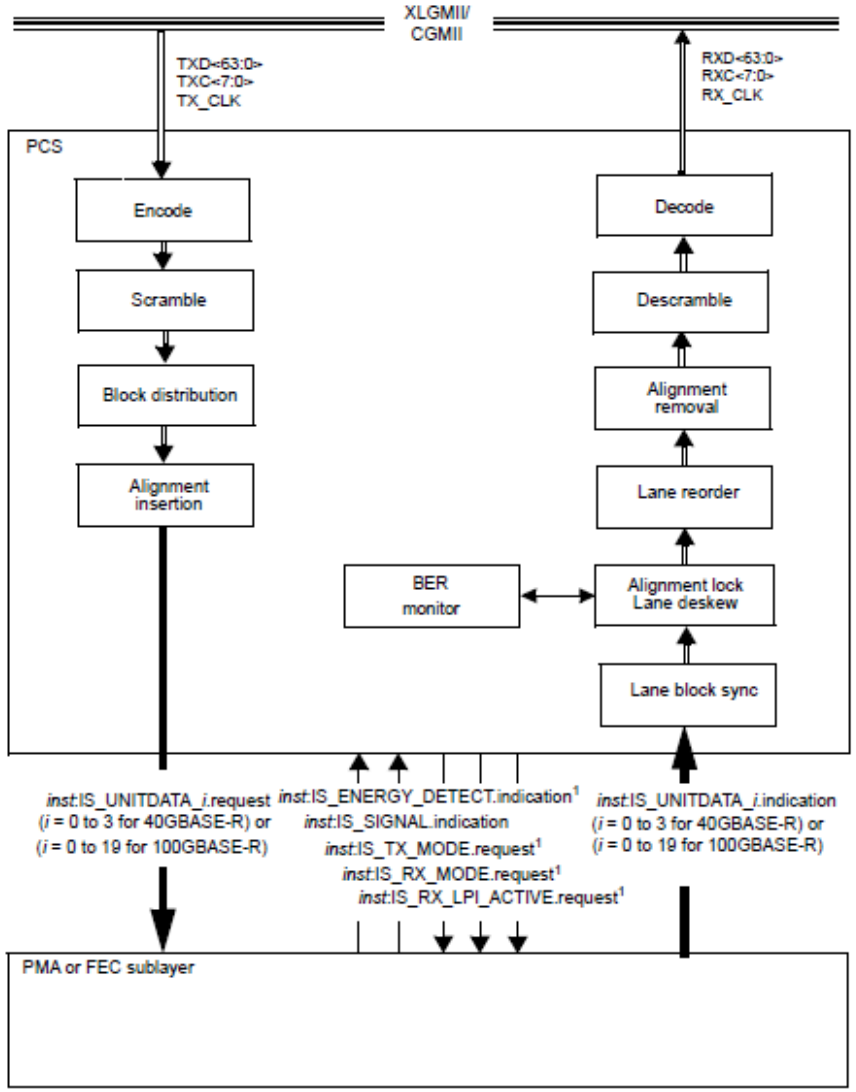


Figure 82–2—Functional block diagram

constant delay	constant delay
constant delay	constant delay
variable delay (mirrors Rx PCS peer)	AM variable delay (resolved[1])
AM variable delay (resolved[1])	variable delay (mirrors Tx PCS peer)
	semi-constant delay[2]
	constant delay

1. *See resolution for AM variable delay in draft P802.3cx/D0.2*
2. *Delay is constant but depends on start-up or system conditions*

MICROCHIP

40GE/100GE PCS block distribution

- From clause 82

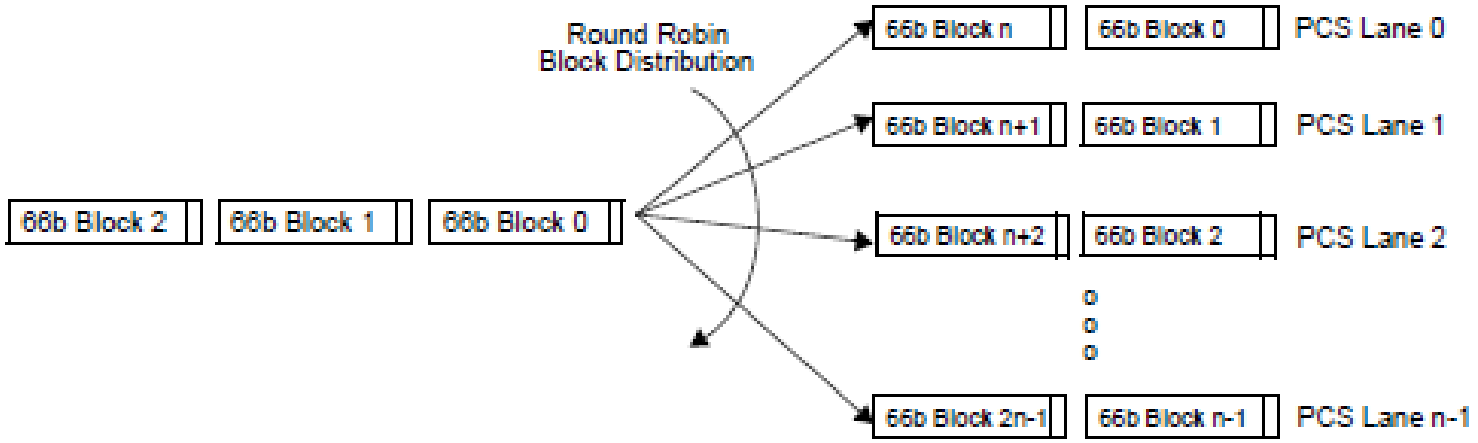


Figure 82–6—PCS Block distribution

40GE/100GE Alignment Marker Insertion

- From clause 82

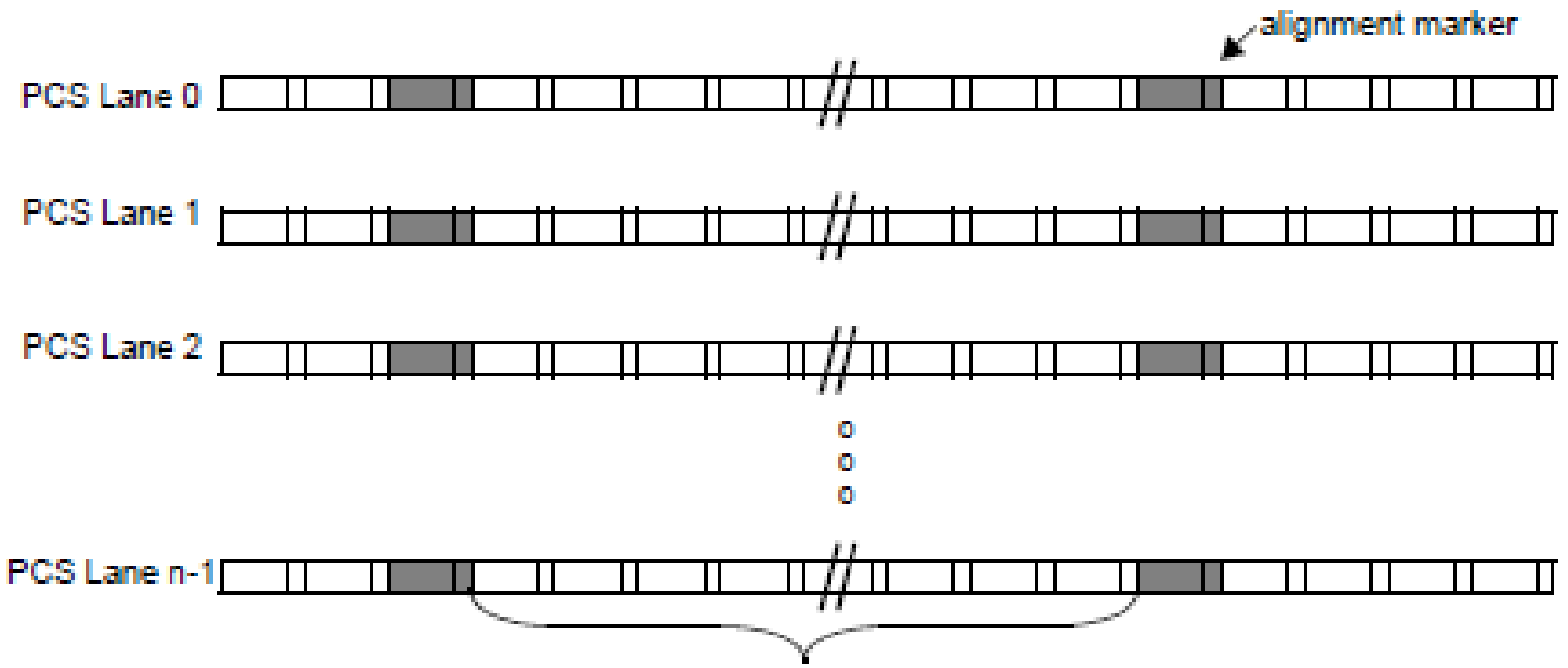


Figure 82–8—Alignment marker insertion period

40GE/100GE FEC

- From clause 91

constant delay

semi-constant delay[2]

variable delay (mirrors Tx PCS peer)

AM variable delay (resolved[1])

variable delay (mirrors Rx FEC peer)

AM variable delay (resolved)

constant delay

variable delay (mirrors Rx FEC peer)

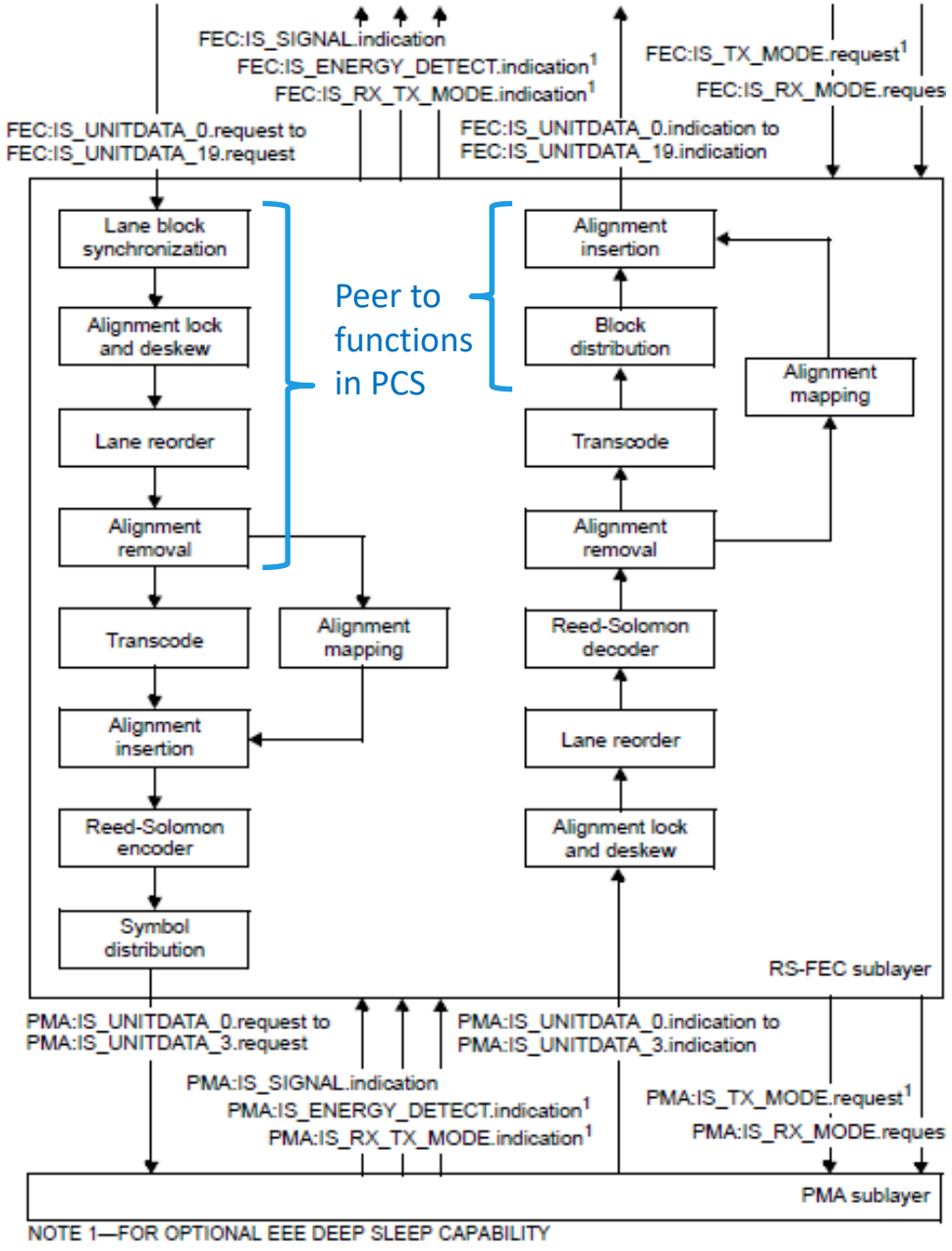


Figure 91–2—Functional block diagram

AM variable delay (resolved)

variable delay (mirrors Rx PCS peer)

variable delay (mirrors Tx FEC peer)

AM variable delay (resolved[1])

constant delay

variable delay (mirrors Tx FEC peer)

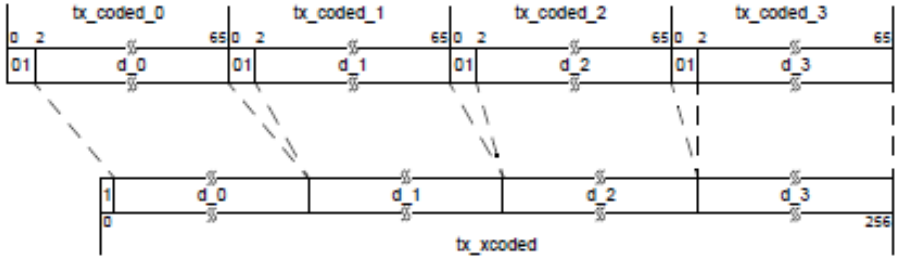
semi-constant delay[2]

1. *See resolution for AM variable delay in draft P802.3cx/D0.2*
2. *Delay is constant but depends on start-up or system conditions*

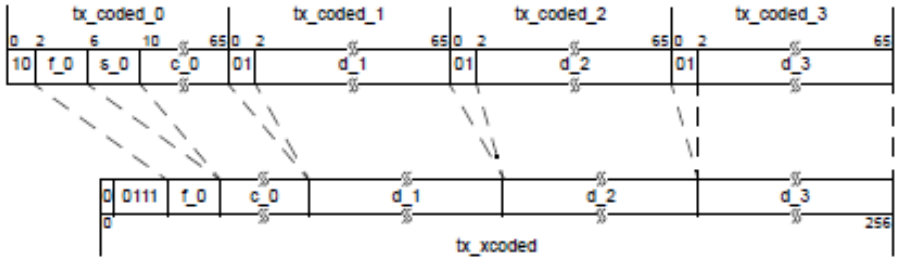
MICROCHIP

256B/257B Transcoding

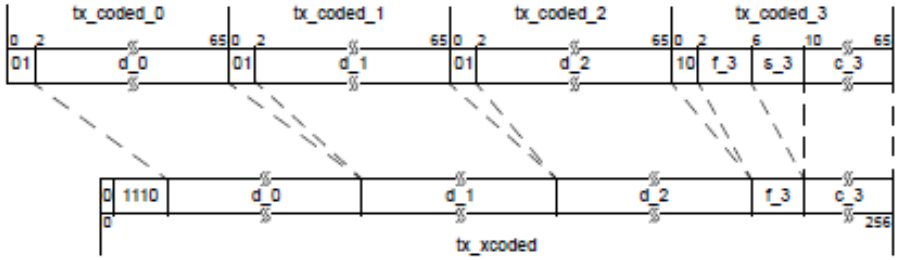
- From clause 91
- The message timestamp point could be affected by the presence of control blocks
- Any delay shift at transcoder will be mirrored by an opposite delay shift at the de-transcoder



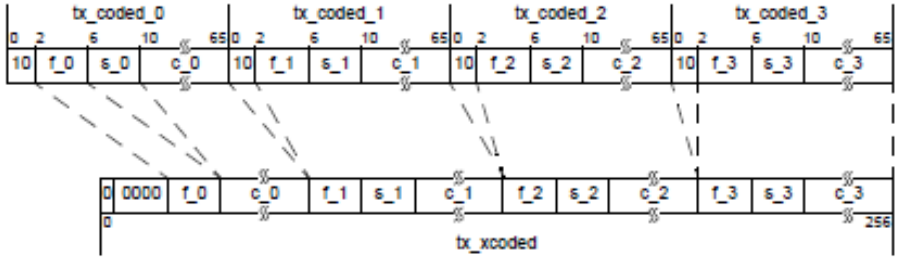
Example 1: All data blocks



Example 2: Control block followed by three data blocks



Example 3: Three data blocks followed by a control block



Example 4: All control blocks

Figure 91–3—Examples of the construction of tx_xcoded

MICROCHIP

40GE/100GE FEC

- FEC has its own lane distribution function
  - Based on 10-bit codewords instead of 66-bit blocks
  - A FEC block always starts at the lowest FEC lane and codewords are distributed in order from the lowest lane to the highest lane
- As per the statement in 90.7, the message timestamp point is specified to be moved to the start of the FEC block to which it belongs, which is on lane 0
  - Lane 0 has constant delay through combined FEC multi-lane Tx distribution + Rx multiplexing
  - Lane 0 has constant max delay for Tx FEC codeword distribution (100% of combined delay)
  - Lane 0 has constant min delay for Rx FEC codeword multiplexing (0% of combined delay)

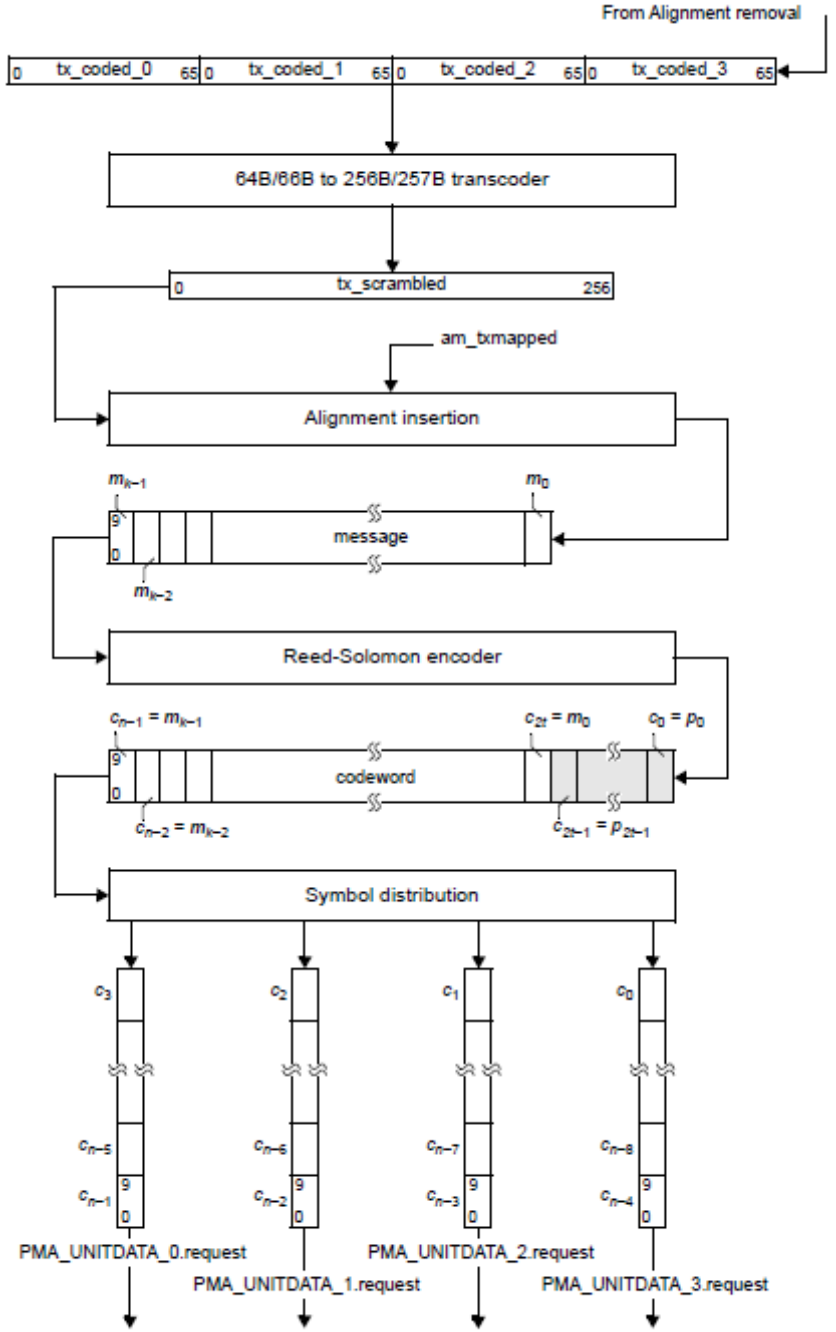


Figure 91–6—Transmit bit ordering

Conclusions

- In IEEE 802.3-2018, the method to deal with FEC's multi-lane path data delay variation is:
  - not consistent with he_3x_01_0920.pdf
  - consistent with Method 2 and Option C from tse_3cx_02a_0420.pdf (a.k.a. soln #3 in tse_3cx_01_0720.pdf and tse_3cx_01a_0920.pdf)
- It seems prudent to use the same method to deal with non-FEC multi-lane path data delay variations
- It seems practical to include register bit(s) that identify compliance to the new P802.3cx "high accuracy timestamping" methods
  - Existing implementations that do not comply to P802.3cx, would not become "non-compliant" to IEEE 802.3

MICROCHIP

Thank You

Proposed Text – implementation option 1 (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).

MICROCHIP

Proposed Text – implementation option 1 (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.

MICROCHIP

Proposed Text – implementation option 2

- **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)**
- **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.

MICROCHIP