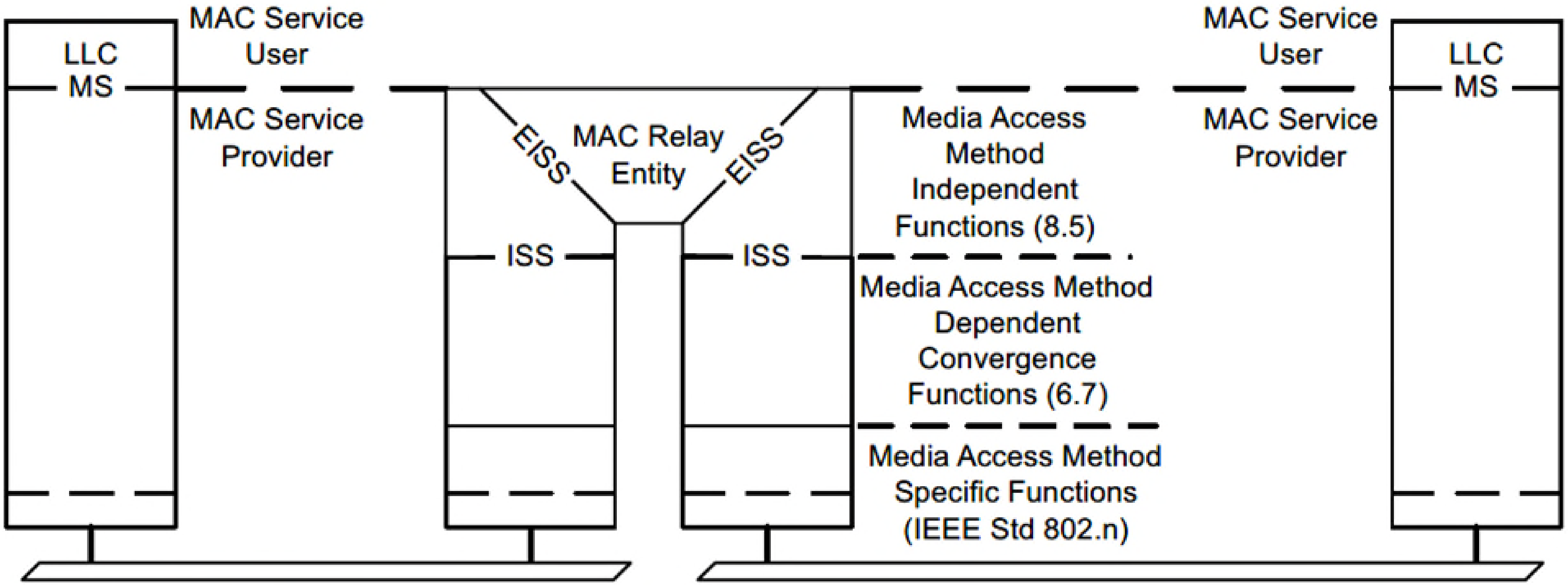
Effects of delays between the transmission selection point and the on-the-wire timing point

Alon Regev, Keysight Technologies

Mar 11, 2025

802.1Q view of the world


LLC
MS
MAC Service User
MAC Service Provider
EISS
MAC Relay Entity
EISS
ISS
ISS
Media Access Method Independent Functions (8.5)
Media Access Method Dependent Convergence Functions (6.7)
Media Access Method Specific Functions (IEEE Std 802.n)
MAC Service User
MAC Service Provider
LLC
MS

NOTE-The notation IEEE Std 802.n in this figure indicates that the specifications for these functions can be found in the relevant standard for the media access method concerned; for example, n would be 3 (IEEE Std 802.3) in the case of Ethernet.

Figure 6-1—Internal organization of the MAC sublayer

802.1Q view continued

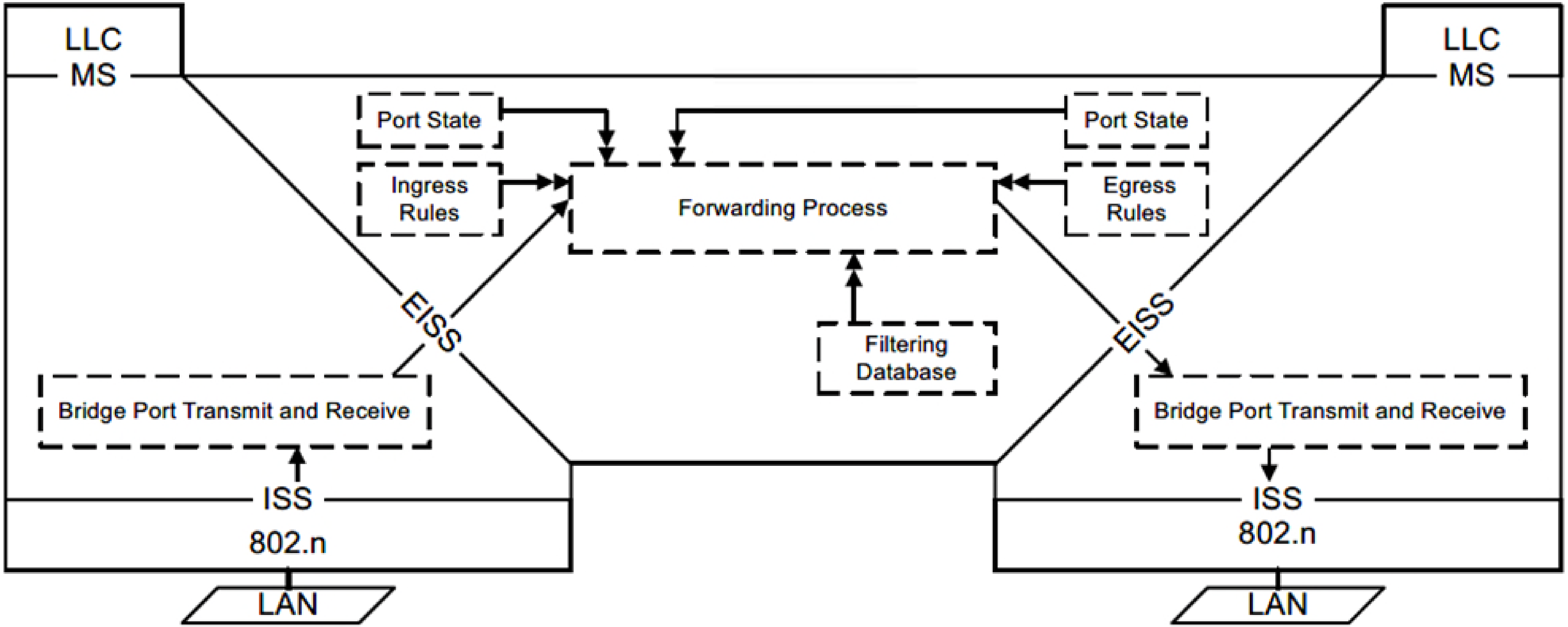


Figure 8-4—Relaying MAC frames

802.1Q view continued

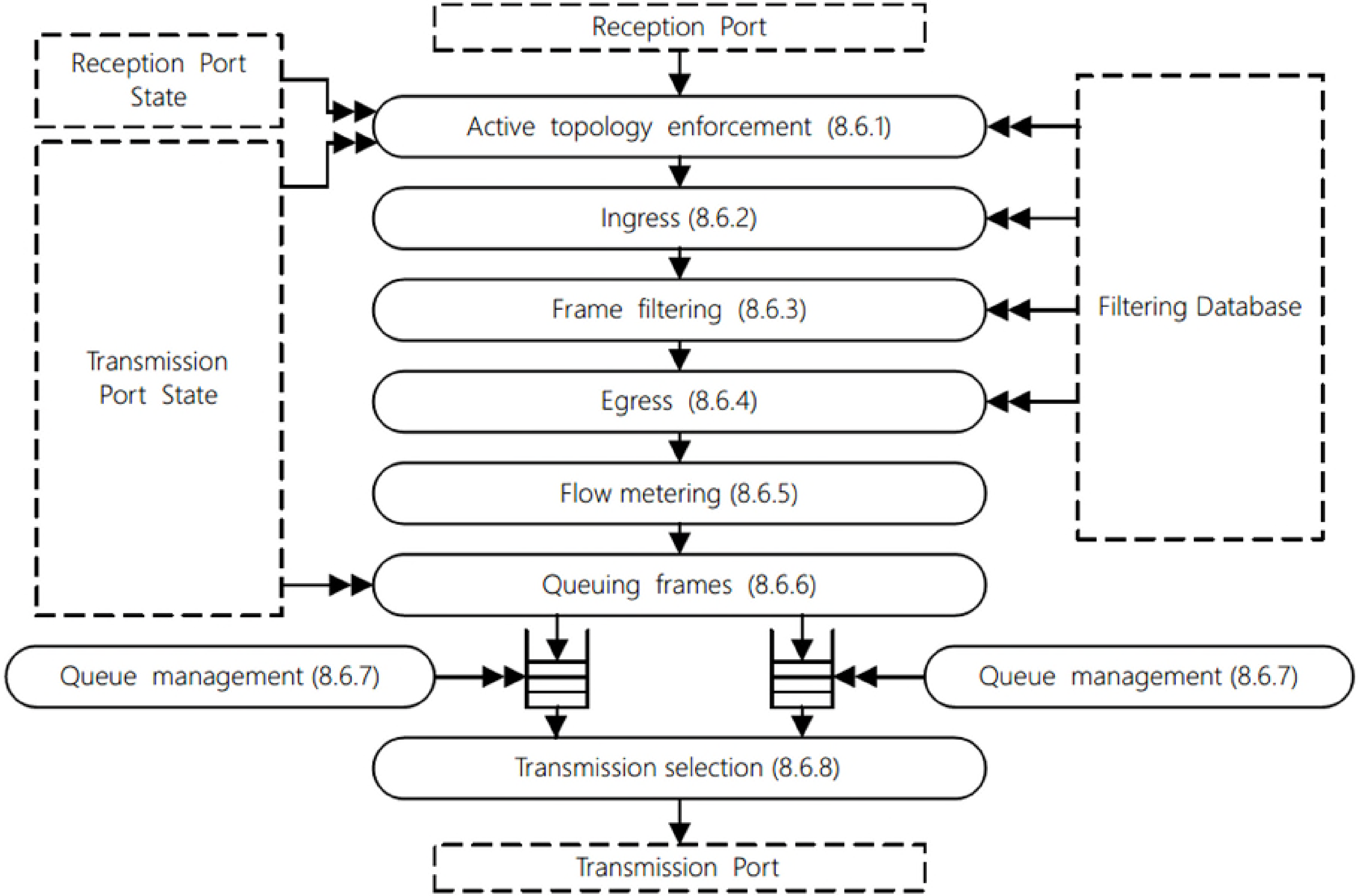
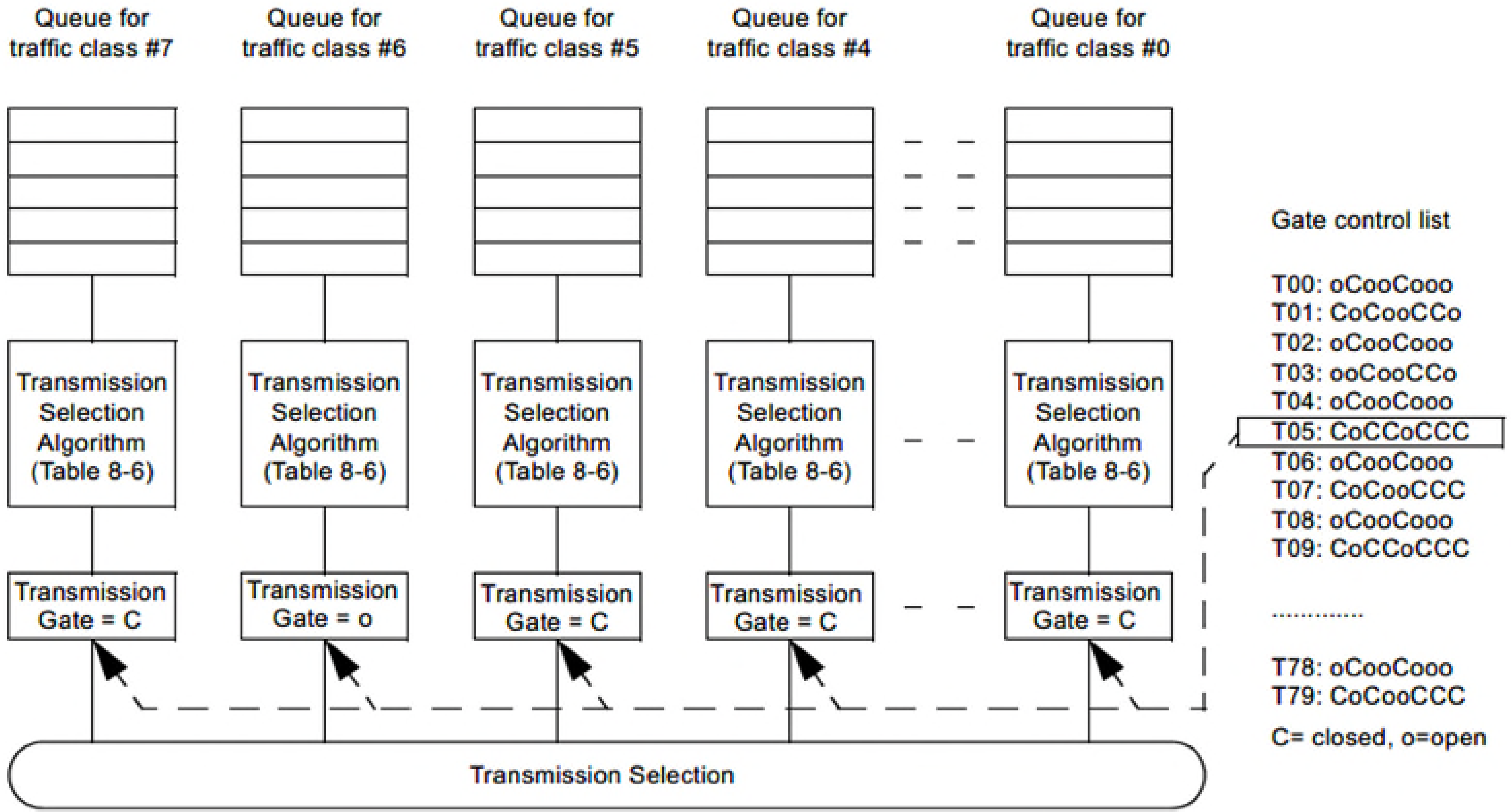


Figure 8-12—Forwarding process functions

802.1Q view continued


Queue for traffic class #7
Queue for traffic class #6
Queue for traffic class #5
Queue for traffic class #4
Queue for traffic class #0
Transmission Selection Algorithm (Table 8-6)
Transmission Selection Algorithm (Table 8-6)
Transmission Selection Algorithm (Table 8-6)
Transmission Selection Algorithm (Table 8-6)
Transmission Selection Algorithm (Table 8-6)
Transmission Gate = C
Transmission Gate = o
Transmission Gate = C
Transmission Gate = C
Transmission Gate = C
Transmission Selection

Gate control list

T00: oCooCooo
T01: CoCooCCo
T02: oCooCooo
T03: ooCooCCo
T04: oCooCooo
T05: CoCCoCCC
T06: oCooCooo
T07: CoCooCCC
T08: oCooCooo
T09: CoCCoCCC

....

T78: oCooCooo
T79: CoCooCCC

C= closed, o=open

Figure 8-17—Transmission selection with gates

1588-2019 latency view

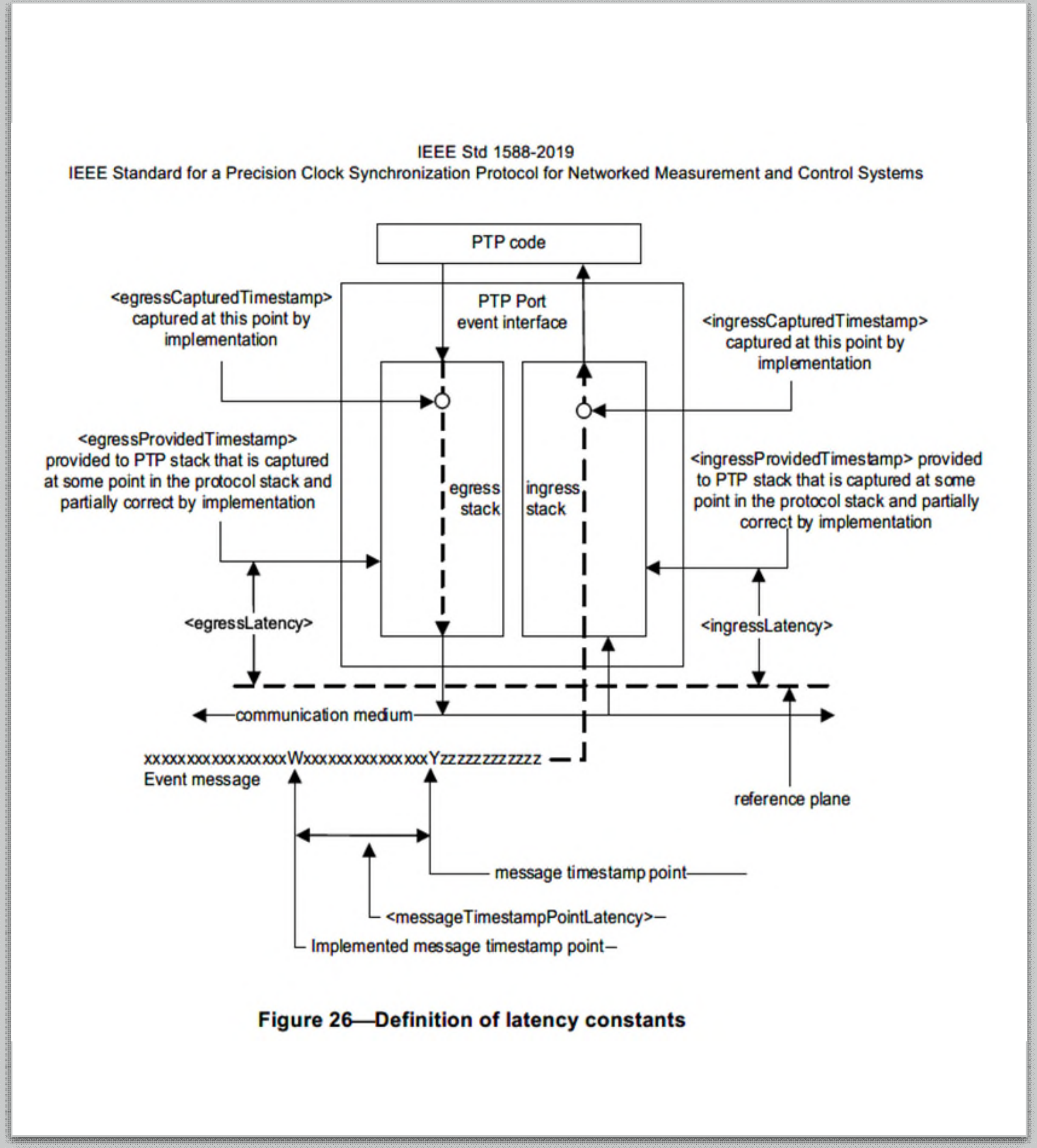
- IEEE Std 1588-2019 timestamps are supposed to be aligned to the reference plane (where the reference plane is "*the boundary between PTP Instance hardware and the PTP Network medium*")
  - I believe that the 1588 "reference plane" maps to the 802.3 "MDI" (Media Dependent Interface, effectively the connector where the Ethernet fiber/cable/wire in plugged-in)
- IEEE Std 1588-2019 allows the timestamp to be captured at a different point than the reference plane but recommends (using "should") that the time be corrected, specifically:

  The implementation-specific corrections of the captured timestamps are specified as follows:

  <egressProvidedTimestamp> = <egressCapturedTimestamp> + <implementation-specific correction of egressLatency and messageTimestampPointLatency>

  <ingressProvidedTimestamp> = <ingressCapturedTimestamp> – < implementation-specific correction of ingressLatency and messageTimestampPointLatency>

  (see IEEE Std 1588-2019 subclause 7.3.4.2)
- IEEE Std 1588-2019 Figure 26 (shown to the right) shows the relationship between these different points



IEEE Std 1588-2019
IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems

PTP code
PTP Port event interface
<egressCapturedTimestamp> captured at this point by implementation
<ingressCapturedTimestamp> captured at this point by implementation
<egressProvidedTimestamp> provided to PTP stack that is captured at some point in the protocol stack and partially correct by implementation
<ingressProvidedTimestamp> provided to PTP stack that is captured at some point in the protocol stack and partially correct by implementation
egress stack
ingress stack
<egressLatency>
<ingressLatency>
communication medium
xxxxxxxxxxxxxxxxxWxxxxxxxxxxxxxxYzzzzzzzzzzz
Event message
reference plane
message timestamp point
<messageTimestampPointLatency>
Implemented message timestamp point

Figure 26—Definition of latency constants

IEEE 802.3-2018 / p802.3cx latency view

- IEEE Std 802.3-2018 clause 90 defines how this is done in IEEE Standard 802.3
  - subclause 90.5 defines a generic Reconciliation Sublayer (gRS) with functions to detect the SFD and timestamp packets

- Figure 90-3 in IEEE Std 802.3 shows the relationship between this gRS (the timestamping capture point) and the MDI (the reference plane)
  - IEEE p802.3cx corrects "data delay" to "path data delay"

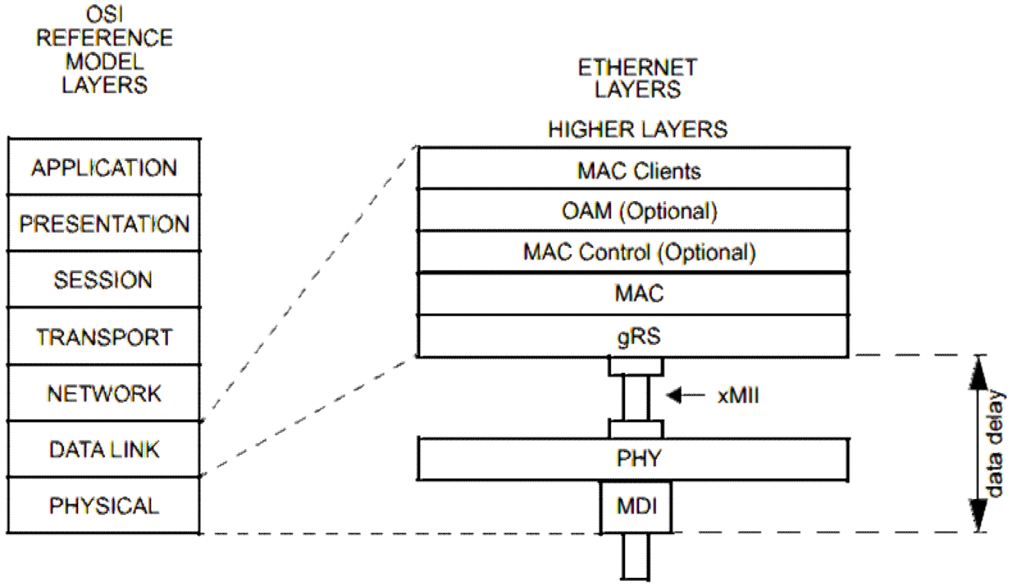
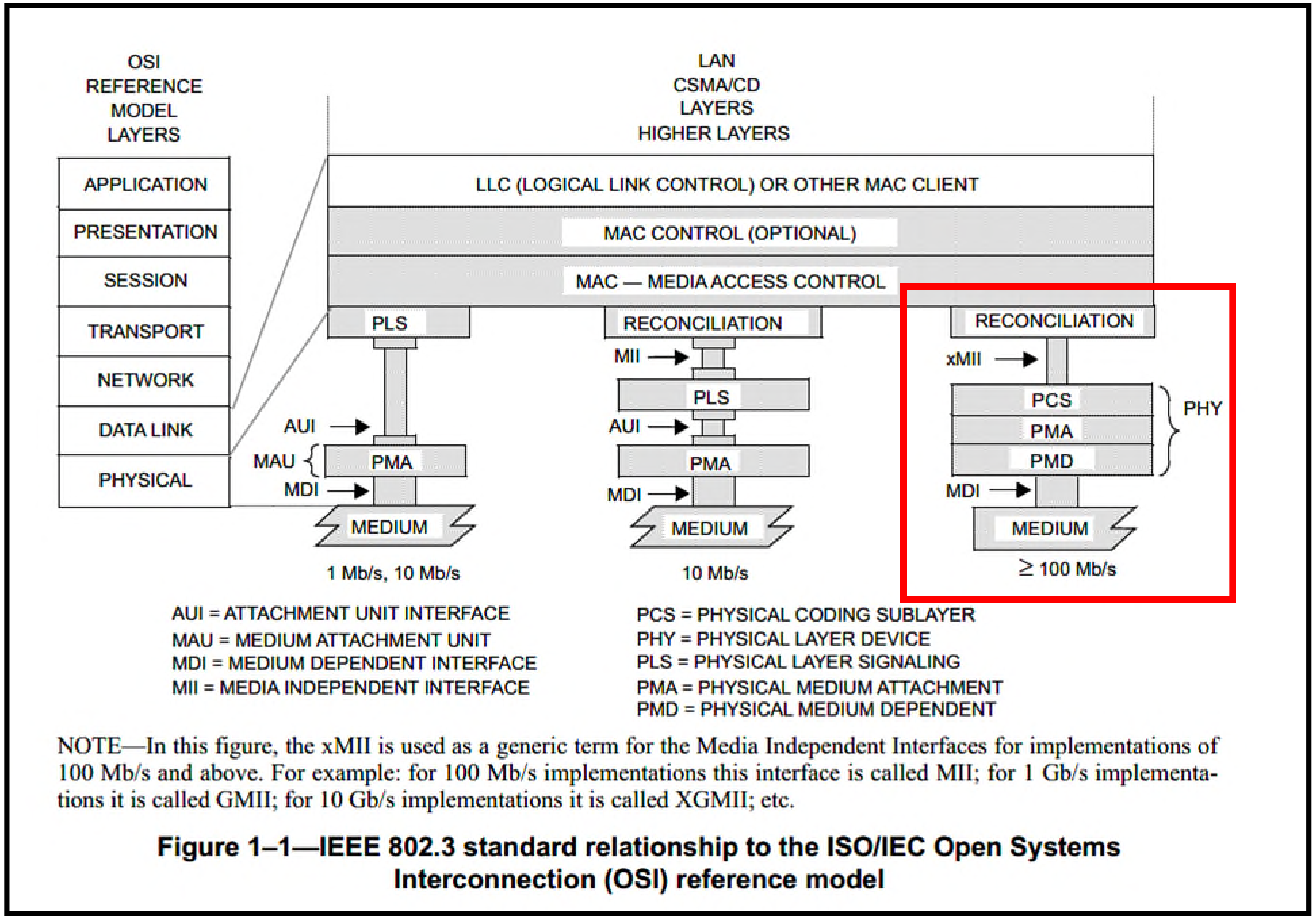


Figure 90–3—Data delay measurement

IEEE Std 802.3 shows additional details of the layers between the Reconciliation Sublayer and the MDI



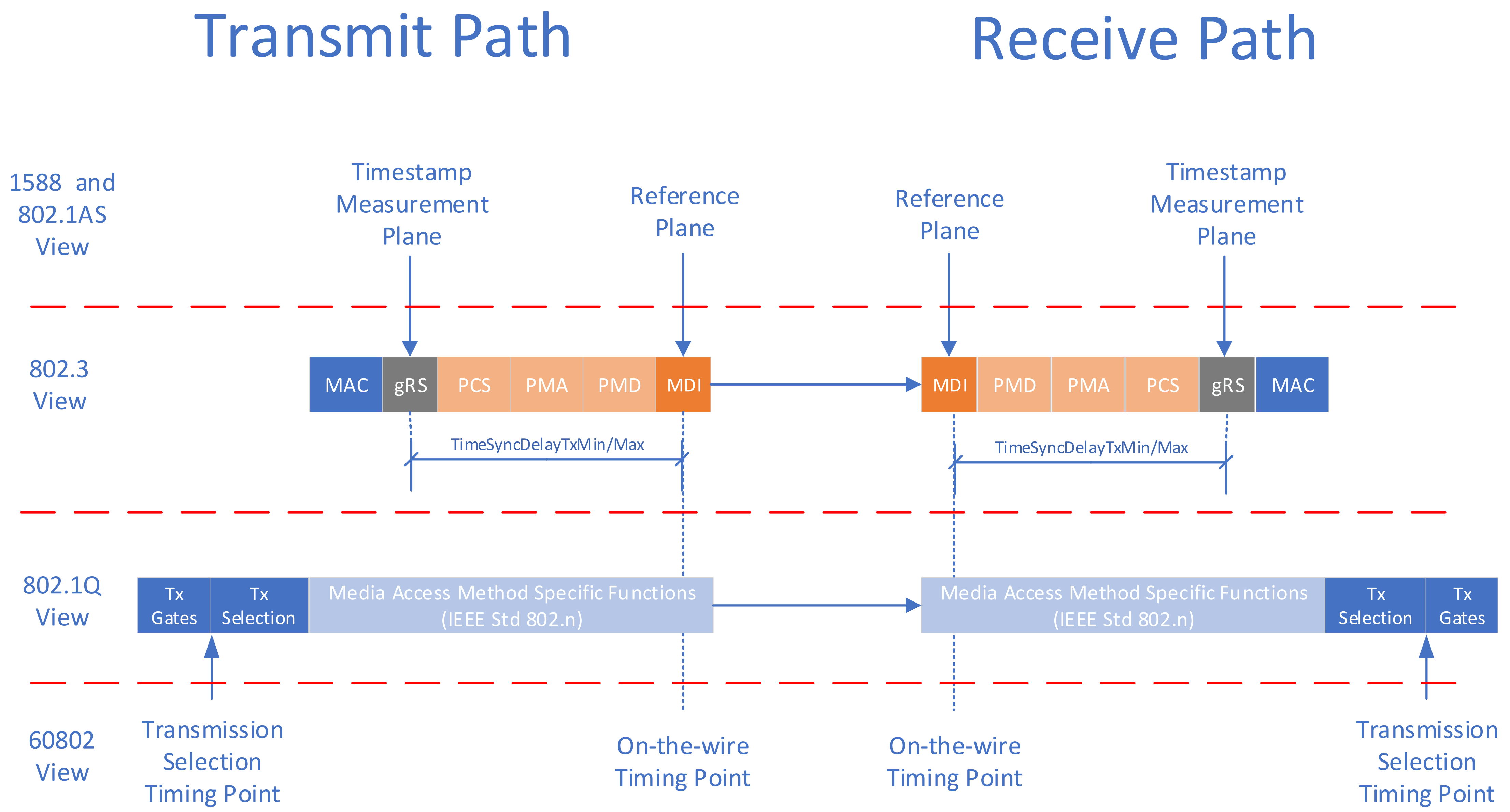
AUI = ATTACHMENT UNIT INTERFACE
MAU = MEDIUM ATTACHMENT UNIT
MDI = MEDIUM DEPENDENT INTERFACE
MII = MEDIA INDEPENDENT INTERFACE

PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE
PLS = PHYSICAL LAYER SIGNALING
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT

NOTE—In this figure, the xMII is used as a generic term for the Media Independent Interfaces for implementations of 100 Mb/s and above. For example: for 100 Mb/s implementations this interface is called MII; for 1 Gb/s implementations it is called GMII; for 10 Gb/s implementations it is called XGMII; etc.

Figure 1–1—IEEE 802.3 standard relationship to the ISO/IEC Open Systems Interconnection (OSI) reference model

Combining views from different standards

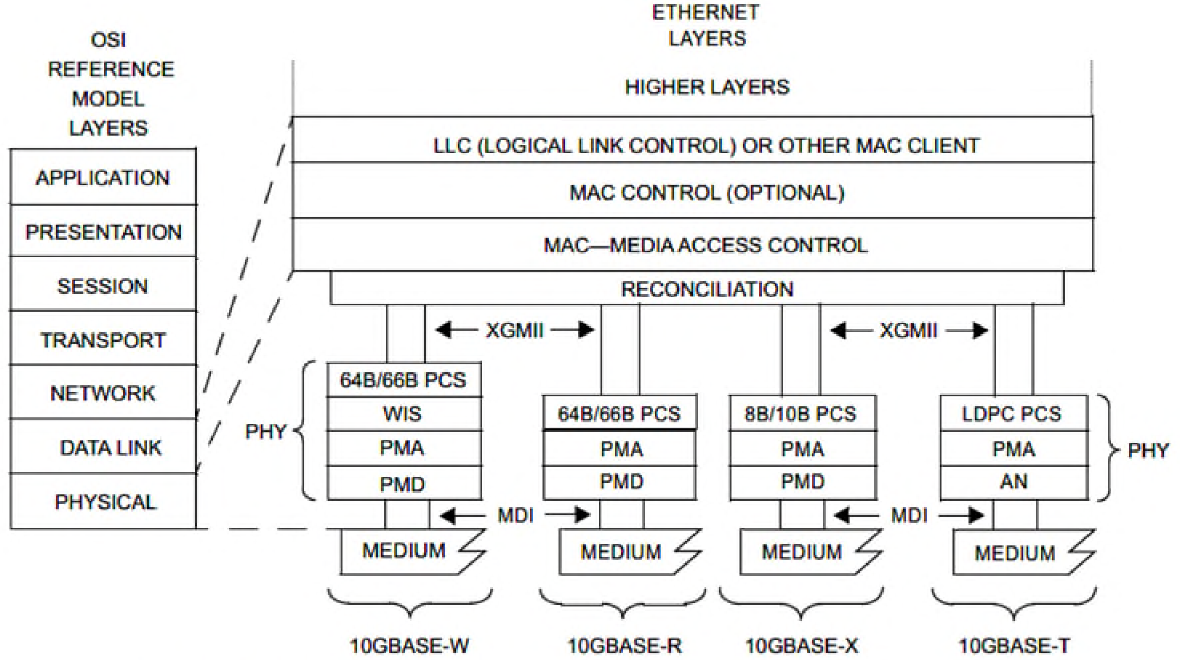


Comment R1-49: Observations regarding delays between transmission selection timing point and on-the-wire timing point

- There are multiple layers between transmission selection point and PHY that add variable delays
  - Tx Selection has output / multiplexing delays
  - MAC has variable delays to add preamble / FCS and enforce gaps
  - gRS may change IPG (and preamble in some cases), align to 32/64-bit boundaries, and add/remove alignment markers
- Compensation is only defined for the PHY
- Compensation can only be done for static (fixed) delays.
- Limiting delays through these layers (excluding PHY) to 10ns is not achievable using today’s technology
- Limiting delay variations through these layers (excluding PHY) to 10ns is achievable, but I am not aware of commercial implementations that meet this

802.3: Different layers in different PHYs

- Some PHY types do not have a PMD layer (this is the case for most twisted-pair “copper” PHYs)
- An example of this is shown in IEEE Std 802.3 Figure 44-1 that shows the different layering for different 10GE PHY types
- Note that 10GBASE-T which is a copper PHY has “AN” (Auto-Negotiation) instead of a PMD
    - In my view, AN is not a layer in the PHY but rather a separate function within the PHY that is parallel to the PMA and PCS…

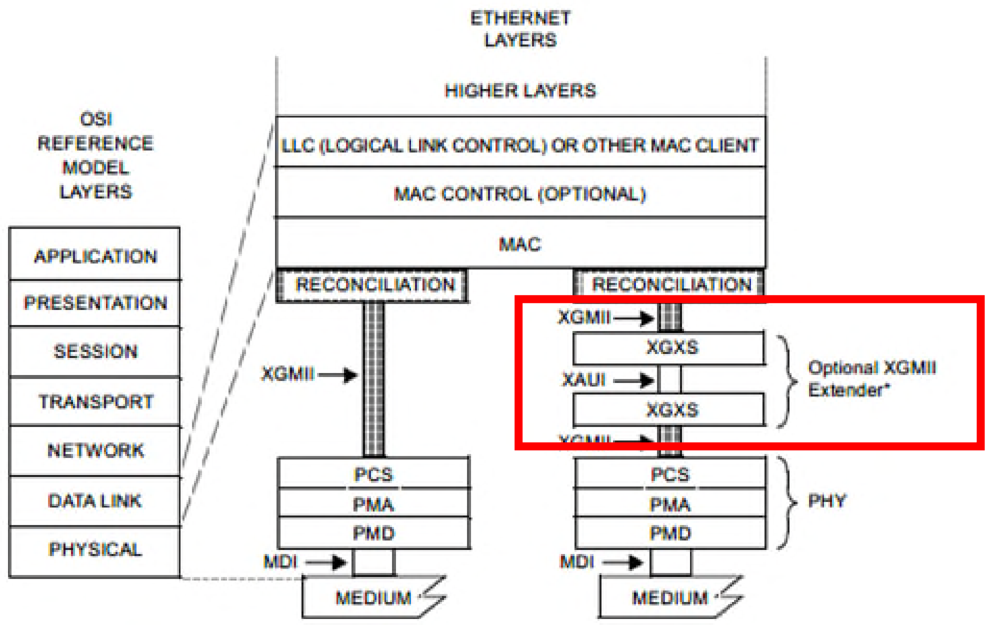


AN = AUTO-NEGOTIATION SUBLAYER
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE

PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
WIS = WAN INTERFACE SUBLAYER
XGMII = 10 GIGABIT MEDIA INDEPENDENT INTERFACE

Figure 44–1—Architectural positioning of 10 Gigabit Ethernet

802.3: More Sublayers…

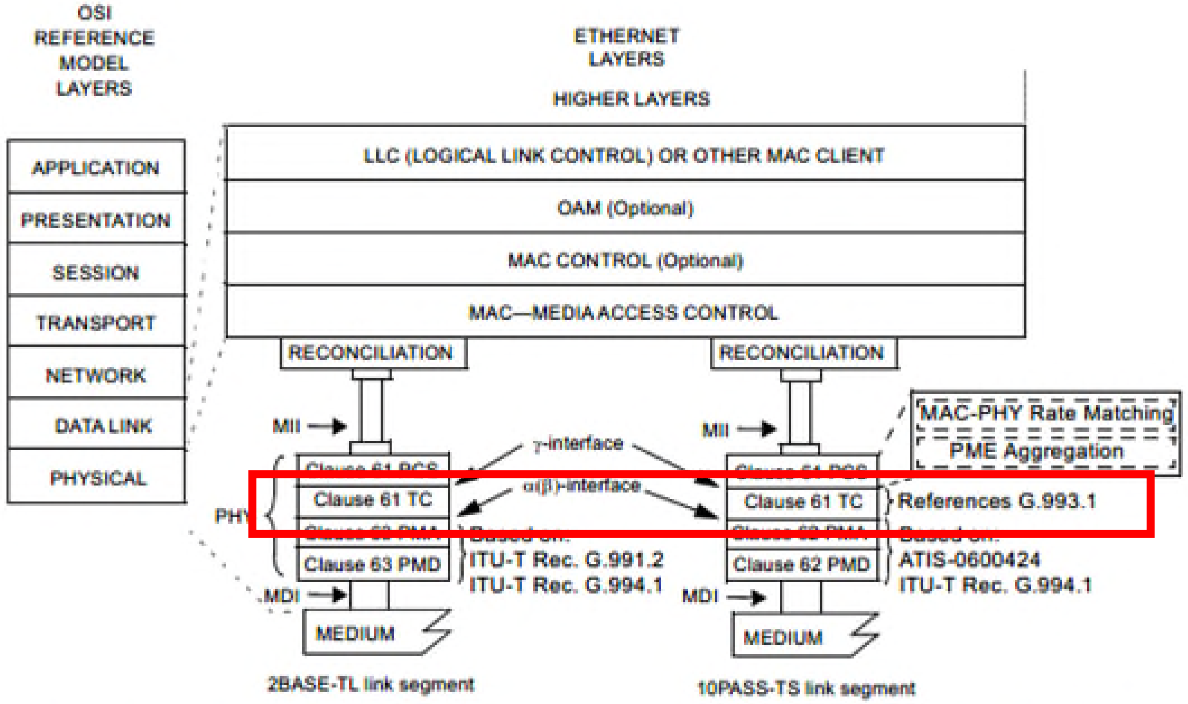


MAC = MEDIA ACCESS CONTROL
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE

PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
XAUI = 10 GIGABIT ATTACHMENT UNIT INTERFACE
XGMII = 10 GIGABIT MEDIA INDEPENDENT INTERFACE
XGXS = XGMII EXTENDER SUBLAYER

*specified in Clause 47

Figure 46–1—XGMII relationship to the ISO/IEC Open Systems Interconnection (OSI) reference model and the IEEE 802.3 Ethernet model



MDI = MEDIUM DEPENDENT INTERFACE
MII = MEDIA INDEPENDENT INTERFACE
OAM = OPERATIONS, ADMINISTRATION & MAINTENANCE
TC = TRANSMISSION CONVERGENCE

PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
PME = PHYSICAL MEDIUM ENTITY

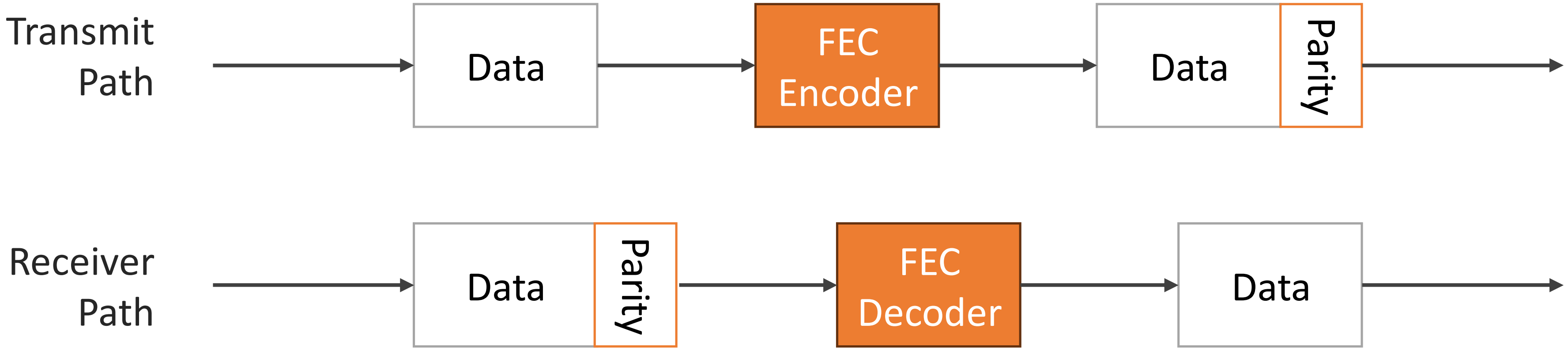
Figure 61–1—Relation of this clause to other standards

What causes delay variations in PHY

- FEC (see following slides)
- Changes in alignment (i.e. 8 bit alignment in MII, but 16 or 32 bit alignment on the wire)
- IDLE insertion/removal
- Lane marker and alignment marker insertion/removal
- Different delays in different lanes
- Clock domain crossings
- Other (non-802.3 PHY) functions being done in the PHY (and affects on buffering)

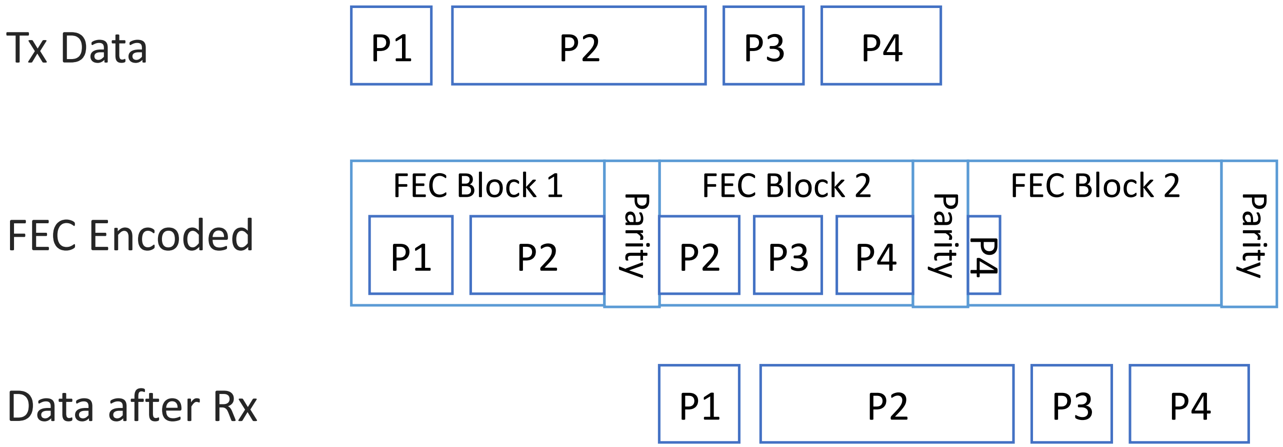
What is FEC

- FEC (Forward Error Correction) refers to additional bits added to a data transmission to correct for communication errors (typically caused by noisy environments)
  - Additional bits can be used to both detect and correct errors
  - Enables getting similar error rates using high-speed transmission as were possible using lower speed transmissions
- In Ethernet, the most common type of FEC is RS-FEC
  - NRZ transmissions typically use RS-FEC(528,14) which adds 140 code bits to each 5140 data bits
  - PAM4 typically uses RS-FEC(544,14) which adds 300 code bits to each 5140 data bits
  - PAM4 for 2.5G/5G/10G BASE-T1 uses RS-FEC (360, 10), which adds 340 code bits to 3260 data bits
  - 1000BASE-T1 (PAM3) uses RS-FEC (450,9) which adds 396 code bits to each 3654 data bits
  - PAM16 typically uses LDPC(1723,2048) which adds 325 LDPC bits to each 1625 data bits



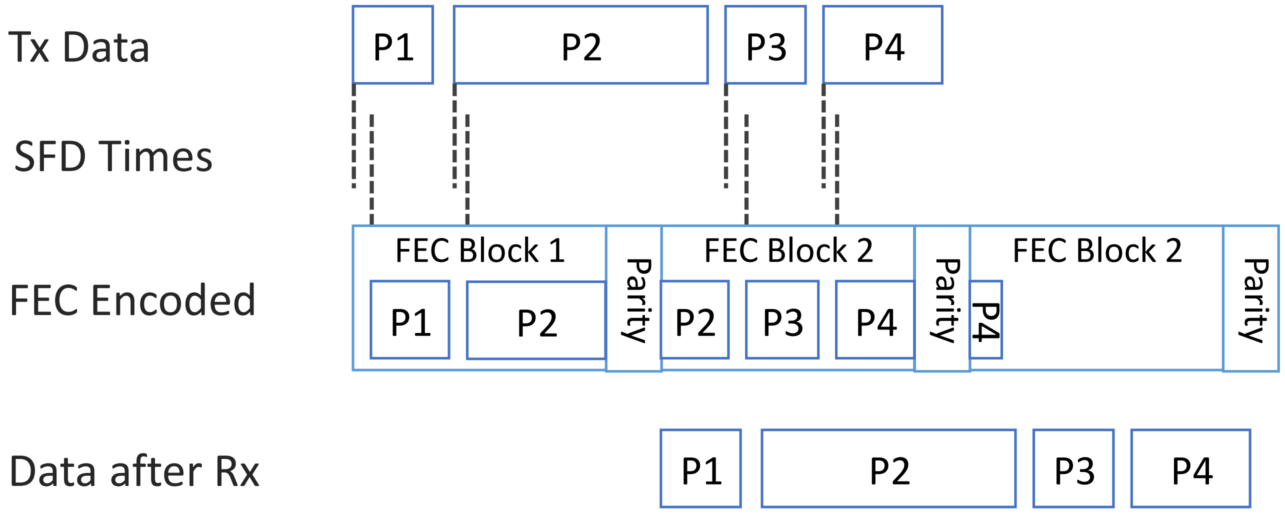
FEC and Ethernet Packets

- FEC blocks recur at a constant rate
- Ethernet packets may have different alignments vs. the FEC block
- On transmit, a packet may be delayed waiting for FEC to overhead to be encoded before the packet itself can be transmitted
- On receive, the entire FEC block needs to be received and processed before any packets included in the FEC block can be processed
- There may be clock-domain-crossings in the FEC block adding dynamic latency (typically related to the distance from the FEC block start)



Why this matters – effect of PHY delay variation on 1588 timestamping

- The delay from the MAC Tx Data (at the gRS) to transmitting the FEC encoded data (at the MII) is variable
  - Packets starting close to the start of the FEC block have greater delays (at the SFP) than packets starting close to the end of the FEC block
- As timestamping happens before (and typically without internal knowledge about) the FEC encoding, the delay between the timestamp capture point and the reference plane is not constant
- Different implementations have different ways of compensating, including using the “average” delay, assuming alignment with start or end of FEC block, or dynamically updating the delay compensation

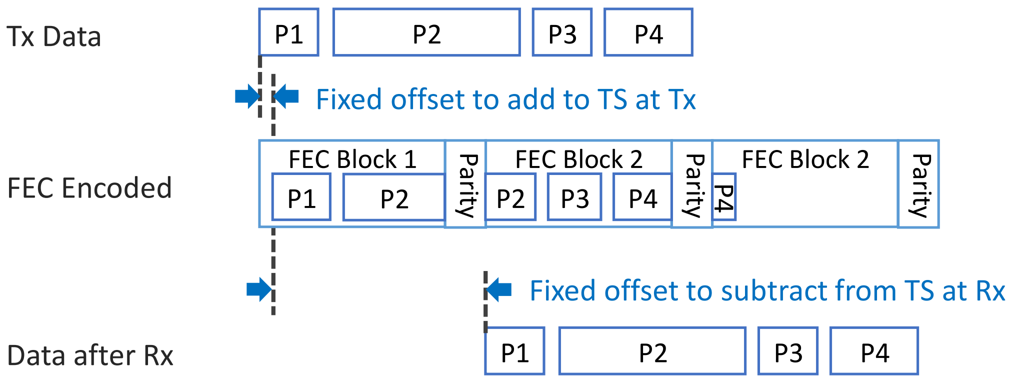


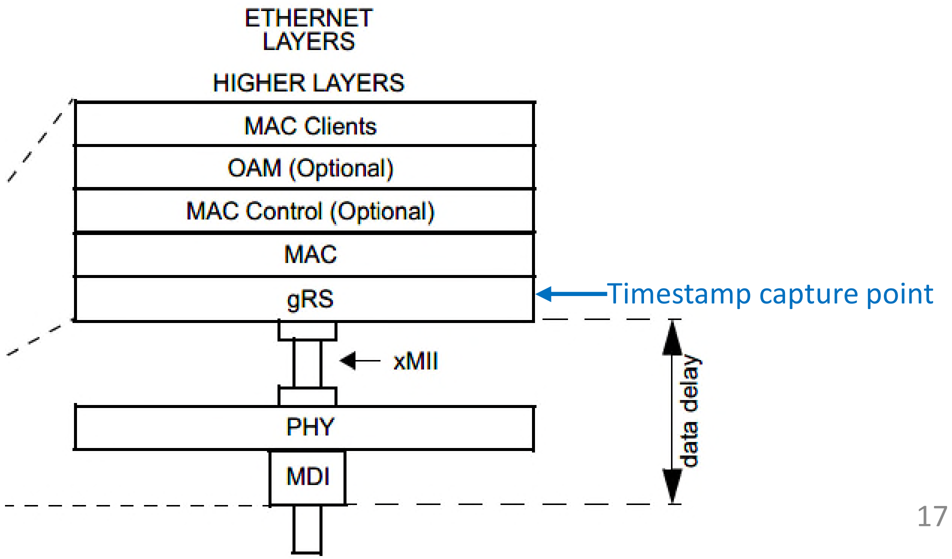
802.3 clause 90: Ethernet timestamping improvements

- 802.3 Clause 90 covers “Ethernet support for the time synchronization protocols”. It states:

> 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

  - This means that a fixed delay can be assumed on both Tx and Rx
  - This approach works if the timestamps are inserted in the gRS and if both peers use the same approach





Summary of PHY characteristics affecting timing (for PHY types listed in 60802 subclause 5.5.1 d))

PHY type	Alignment	MII type	FEC	Lanes	Encoding
10BASE-T1L	4-bit	MII	No	1 bidirectional pair	PAM3, B43T
100BASE-TX	4-bit	MII	No	1 pair in each direction	NRZ, 4B5B
100BASE-FX	4-bit	MII	No	1 fiber in each direction	NRZ, 4B5B
1000BASE-T	16-bit	GMII	No	4 bidirectional pairs	PAM5, 4D-PAM5
1000BASE-SX	16-bit	GMII	No	1 fiber in each direction	NRZ, 8B10B
2.5GBASE-T	32-bit	XGMII	LDPC(1723,2048)	4 bidirectional pairs	PAM16, 128x4D
5GBASE-T	32-bit	XGMII	LDPC(1723,2048)	4 bidirectional pairs	PAM16, 128x4D
2.5GBASE-T1	32-bit	XGMII	RS-FEC(360,10)	1 bidirectional pair	PAM4, 64B/65B
5GBASE-T1	32-bit	XGMII	RS-FEC(360,10)	1 bidirectional pair	PAM4, 64B/65B
10GBASE-T	64-bit	XGMII	LDPC(1723,2048)	4 bidirectional pairs	PAM16, DQS128
10GBASE-SR	32-bit	XGMII	No	1 fiber in each direction	NRZ, 64B66B
100BASE-T1	8-bit	MII	No	1 bidirectional pair	PAM3, 3B2T
1000BASE-T1	16-bit	GMII	RS-FEC(450,9)	1 bidirectional pair	PAM3, 3B2T

Comment R1-56: The problem with timestamping in the “PHY”

- PHY vendors like to state that “timestamping in the PHY” reduces timestamp error
- The good about timestamping in the PHY
  - Avoids delay variations due to some clock domain crossings and buffering issues
- The bad about timestamping in the PHY
  - Not following the 802 standard recommendations for timestamping point
  - There is no standard at to where in the PHY to timestamp and different PHY vendors have different interpretations
    - Often will get good results if both sides use the same PHY implementation, but can get significantly worse results if different PHY implementations are used
    - The worst case if for PHY standards that use FEC where implementations that effectively timestamp before or after FEC (or compensate for the FEC variation) get drastically different timestamps (off by 100s of bit times)
  - These variations can be orders of magnitude greater than the delay variations fixed by timestamping in the PHY
- Timestamping at the INPUT to the PHY can often generate equivalent results to timestamping at the gRS
  - But very few PHY implementations do this
- The best choice is to recommend following the 802 standards rather than to recommend non-standard behaviors