Adaptive PFC Headroom and PTP

Lily Lv (Huawei)

Agenda

- Background
- Response to feedback
  - What is the measurement resolution requirement?
  - Can we leverage the existing protocol in 802.1AS or IEEE1588?
- Next steps

Background

- Adaptive PFC headroom contribution proposes a new mechanism to automatically determine the amount of memory needed for PFC headroom.
  - https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-0121-v02.pdf Adaptive PFC Headroom
  - https://www.ieee802.org/1/files/public/docs2021/new-congdon-a-pfc-h-Q-changes-0521-v01.pdf Consideration of Adaptive PFC Headroom in 802.1Q
- Motivation of adaptive PFC headroom: solve current PFC headroom configuration issue
  - Manual configuration is complex to customers
  - Vendor provided default value wastes buffer resource
    - It leads to limitation of number of queues which can enable PFC. Most commercial switches only support 2 PFC queues.
    - It has trouble in DCI scenario, as the link distance can be tens of kilometers.

Example of The Issue

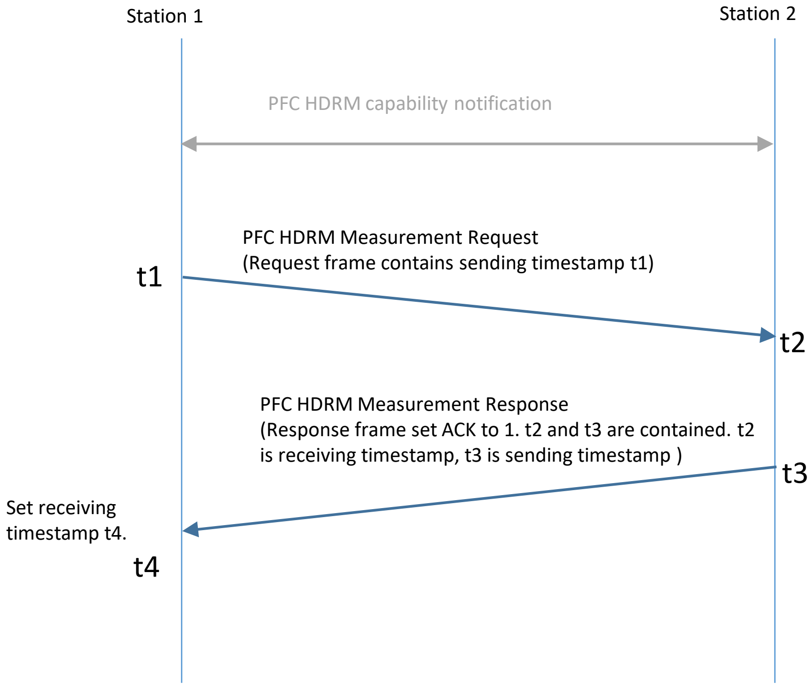
- Assumption:
  - 100Gbps
  - Default value is based on 500m and standard defined value (max value) as internal processing delay
  - Real link distance is 20m, actual internal processing delay is ¾ of the max value

$$\text{Delay Value} = \underbrace{2*(\text{Cable Delay})}_{\text{Medium delay}} + \underbrace{\text{TXds1} + \text{RXds2} + \text{HDs2} + \text{TXds2} + \text{RXds1}}_{\text{Internal Processing delay}} + \underbrace{2*(\text{Max Frame}) + (\text{PFC Frame})}_{\text{Fixed delay}}$$

		Fixed Delay	Internal Processing Delay	Medium Delay	Buffer size/queue	Queue number
Default Value	100G,500m	32992	203776	500000	92KB	2 queues
Real link distance	100G,20m	32992	203776	20000	32KB	5 queues
Real internal process delay	100G,20m	32992	203776*(3/4)=152832	20000	26KB	7 queues

Background

- Proposed mechanism of adaptive PFC headroom
  - The delay measurement procedure is similar to PTP, to measure roundtrip delay
  - The timestamp points are above MAC according to PFC delay model.
    - Internal processing delay(HD+ID) cannot be ignore, as it could be larger than link delay, hundreds of ns level or even higher depending on implementation



Station 1

Station 2

PFC HDRM capability notification

PFC HDRM Measurement Request
(Request frame contains sending timestamp t1)

t1

t2

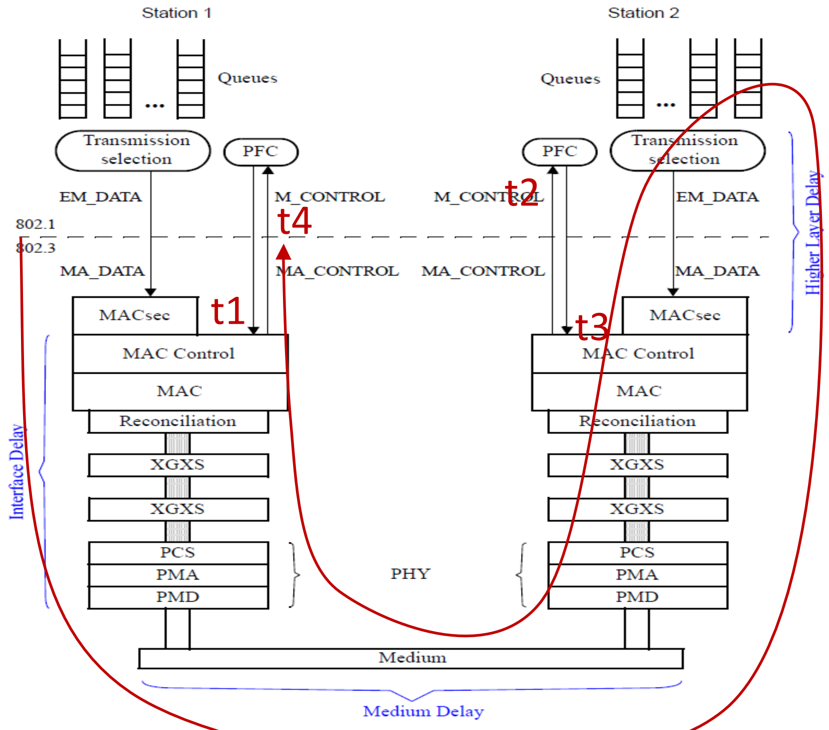
PFC HDRM Measurement Response
(Response frame set ACK to 1. t2 and t3 are contained. t2 is receiving timestamp, t3 is sending timestamp)

t3

Set receiving timestamp t4.

t4

Interface Delay(ID): the sum of MAC Control, MAC/RS, PCS, PMA, and PMD delays



High Layer Delay(HD): the time needed for a queue to go into paused state after the reception of a PFC M_CONTROL.indication that paused its priority

$$\text{Delay Value} = \boxed{\underbrace{2^*(\text{Cable Delay})}_{\text{Medium delay}} + \underbrace{\text{TXds1} + \text{RXds2} + \text{HDs2} + \text{TXds2} + \text{RXds1}}_{\text{Internal Processing delay}}} + \underbrace{2^*(\text{Max Frame}) + (\text{PFC Frame})}_{\text{Fixed delay}}$$

Q1: What is the measurement resolution requirement?

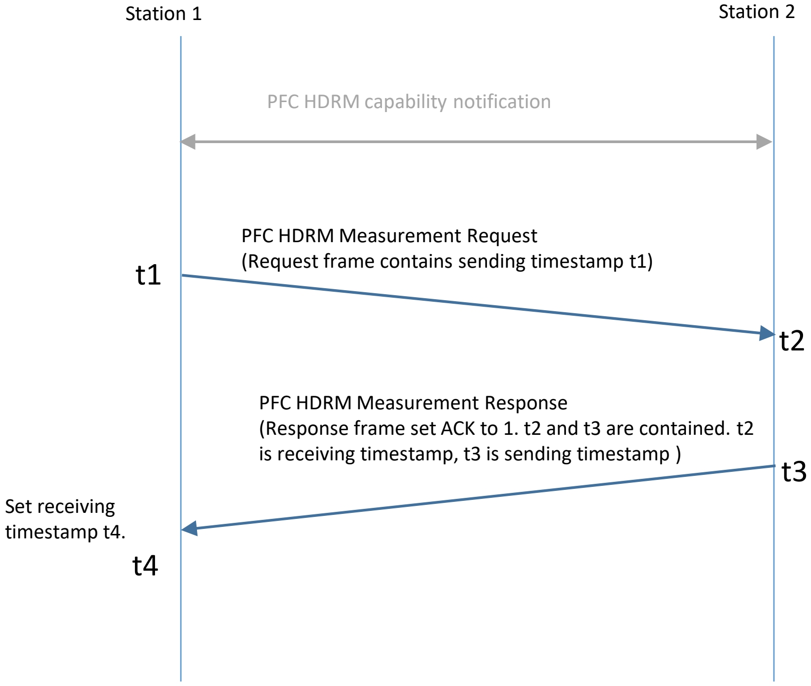
Time Accuracy Analysis of PFC Headroom Measurement

- The precision of (t4-t1) is the focus when analyzing time accuracy of PFC headroom measurement
  - What we don’t care: Peer node clock frequency offset
  - What we care: Local clock frequency drift and timestamp resolution
- Local clock frequency drift impact analysis
  - Assume 5ppm oscillator, fiber cable 100Gbps and 10km link distance
    - (t4-t1) is no more than 200us : 100us link delay plus internal processing delay)
    - 1ns time offset in 200us
    - Headroom size mismatch is about 100 bits : 1ns*100Gbps=100bit, much less than buffer chunk size.
  - So buffer chunk size (e.g. 160 bytes) could easily accommodate the inaccuracy.
- Timestamp resolution impact analysis
  - (t4-t1) is the roundtrip delay, including link delay and station internal processing delay.
    - For 100Gbps, it is above micro-seconds. Range of timestamp resolution requirement is (tens of ns ~ hundreds of ns)
    - Assume 125MHz clock, timestamp resolution is 8ns

Q2: Can we leverage the existing protocol in 802.1AS or IEEE1588?

Reuse PTP Measurement Procedure

- PTP/802.1AS supports peer-to-peer delay link measurement
- The procedure can be reused in PFC headroom delay measurement



Station 1

Station 2

PFC HDRM capability notification

PFC HDRM Measurement Request
(Request frame contains sending timestamp t1)

t1

t2

PFC HDRM Measurement Response
(Response frame set ACK to 1. t2 and t3 are contained. t2 is receiving timestamp, t3 is sending timestamp)

t3

Set receiving timestamp t4.

t4

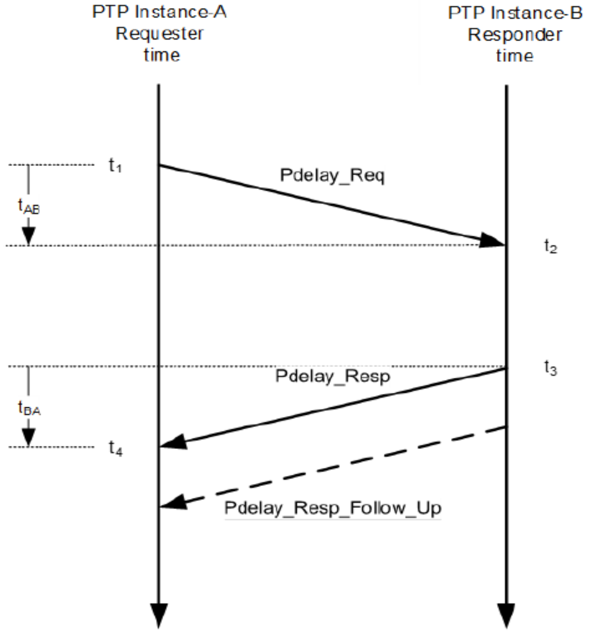


Figure 42—Peer-to-peer delay link measurement

Redefine Timestamp Points

- PTP/802.1AS focus on cable delay, it defines reference plane for message timestamp points
  - t1~t4 have same reference planes.
  - Reference plane is between PHY and medium.
  - Correction is needed if implementation captured timestamp point is not message timestamp point.

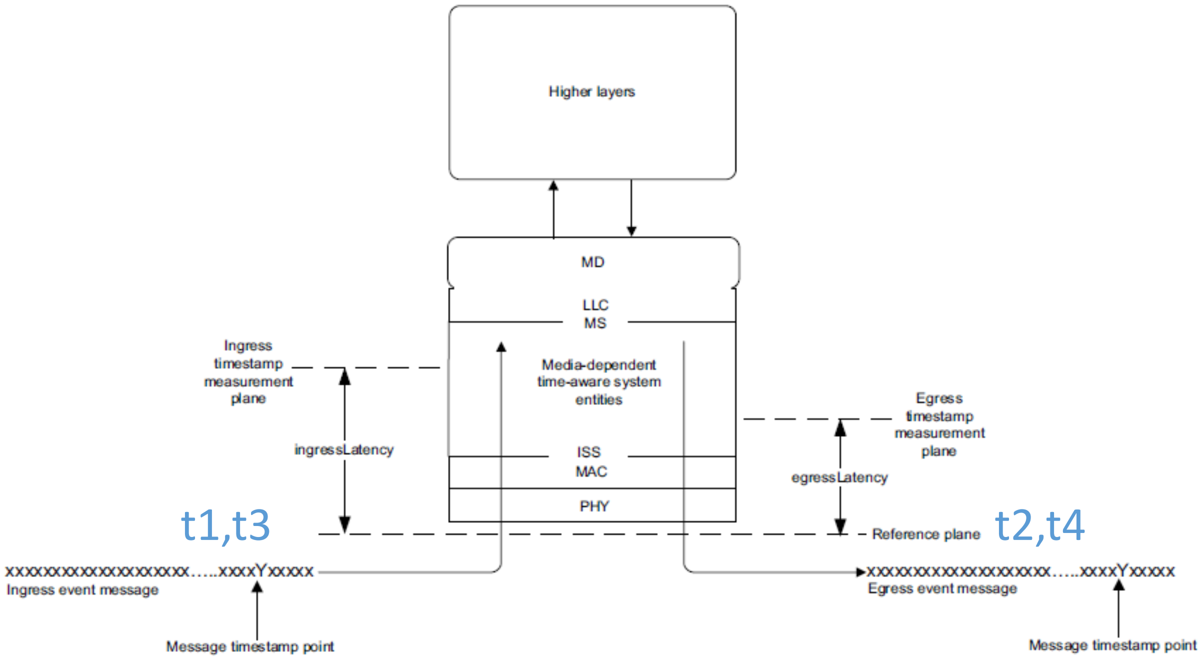
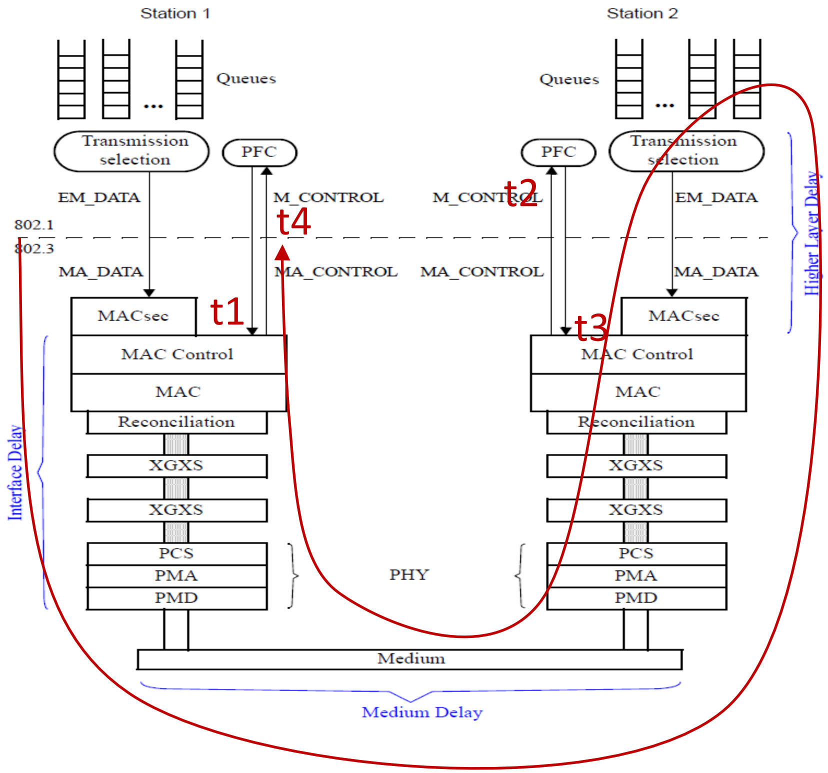


Figure 8-2—Definition of message timestamp point, reference plane, timestamp measurement plane, and latency constants

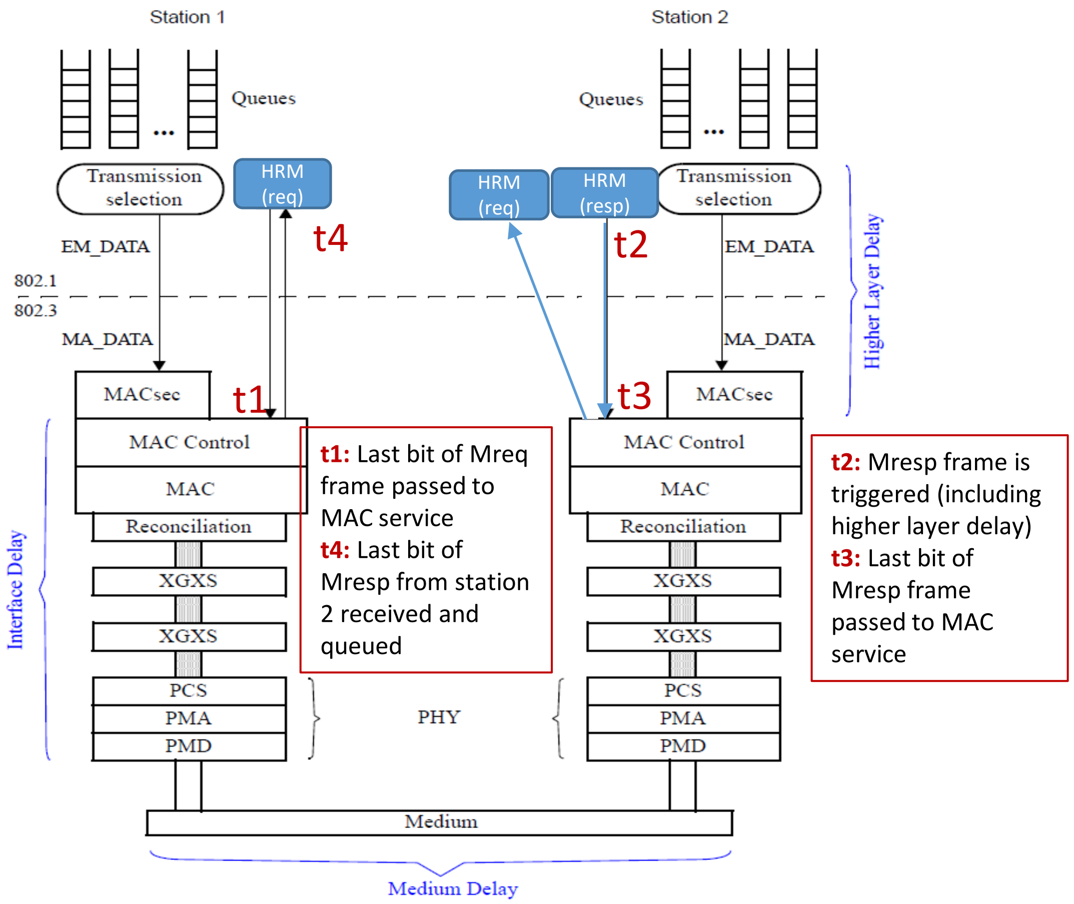
Redefine Timestamp Points

- PFC delay covers not only cable delay but also internal processing delay.
  - Message timestamp points are above MAC.
  - It is easier to capture timestamp points above MAC compared with those on PHY, less challenge on hardware.
- Refer to PTP/802.1AS, reference plane(s) for message timestamp points need to be redefined.
  - t1~t4 may have different reference planes.
  - Reference planes are above MAC
  - Correction is needed if implementation captured timestamp point is not message timestamp point.



Proposals for Implementation(1/3)

- Option 1: reuse PTP measurement procedure as well as Pdelay message, but change reference plane
  - Reference planes are above MAC
    - t1~t4 are as shown in the figure. Reference plane is not the same for all timestamps.
      - (t3-t2) is the time to generate Mresp which should be exclude from PFC headroom delay.
      - Implementation-specific correction is needed to compensate captured timestamp and message timestamp
- Pros:
  - Small changes to PTP peer-to-peer delay measurement
  - Measured delay value including internal processing delay is accurate for headroom calculation.
- Cons:
  - Need to redefine reference plane.



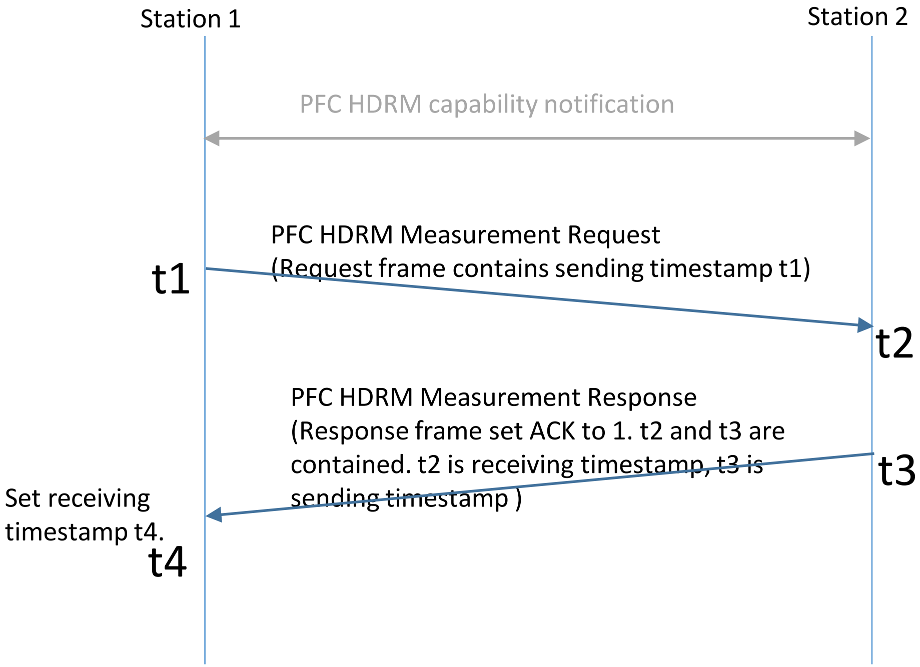
t1: Last bit of Mreq frame passed to MAC service
t4: Last bit of Mresp from station 2 received and queued

t2: Mresp frame is triggered (including higher layer delay)
t3: Last bit of Mresp frame passed to MAC service

$$\text{DV} = 2*(\text{Max Frame}) + (\text{PFC Frame}) + \underbrace{2*(\text{Cable Delay}) + \text{TXds1} + \text{RXds2} + \text{HDs2} + \text{TXds2} + \text{RXds1}}_{\text{t4-t1-(t3-t2)}}$$

Proposals for Implementation(2/3)

- Option 2: based on option 1, but design MAC control frame as measurement message
  - Internal processing delay for MAC control frame and MAC data frame may have difference.
    - PFC frame is MAC control frame, and PTP delay measurement frame is MAC data frame.
  - Design new MAC control frame for measurement
- Pros:
  - More like PFC delay procedure, can be more accurate
  - Implementation friendly, do not impact time sync module.
- Cons:
  - New design of message format
  - Need to redefine reference plane (same as option 1)



Station 1

Station 2

PFC HDRM capability notification

t1

PFC HDRM Measurement Request
(Request frame contains sending timestamp t1)

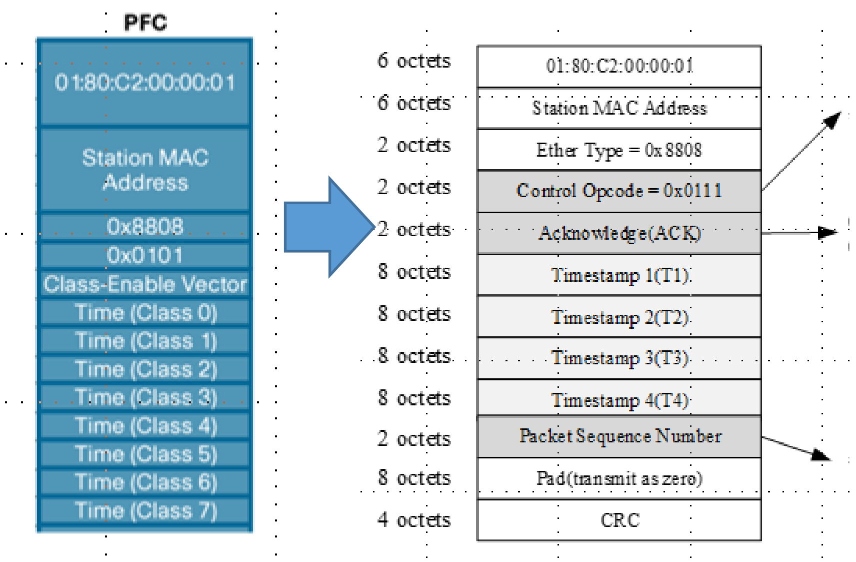
t2

PFC HDRM Measurement Response
(Response frame set ACK to 1. t2 and t3 are contained. t2 is receiving timestamp, t3 is sending timestamp)

t3

Set receiving timestamp t4.

t4



PFC

01:80:C2:00:00:01
Station MAC Address
0x8808
0x0101
Class-Enable Vector
Time (Class 0)
Time (Class 1)
Time (Class 2)
Time (Class 3)
Time (Class 4)
Time (Class 5)
Time (Class 6)
Time (Class 7)

PFC frame format

6 octets	01:80:C2:00:00:01
6 octets	Station MAC Address
2 octets	Ether Type = 0x 8808
2 octets	Control Opcode = 0x0111
2 octets	Acknowledge(ACK)
8 octets	Timestamp 1(T1)
8 octets	Timestamp 2(T2)
8 octets	Timestamp 3(T3)
8 octets	Timestamp 4(T4)
2 octets	Packet Sequence Number
8 octets	Pad(transmit as zero)
4 octets	CRC

Measurement frame format

identify Measurement frame

0x0000： measurement request
0x0001: measurement response

Packet sequence number

Proposals for Implementation(3/3)

- Option 3: reuse PTP protocol but define separate mechanism to get peer node internal processing delay
  - Reuse PTP protocol to measure cable delay
    - Pdelay_Resp/Pdelay_Resp_Follow_Up does not have reserved payload fields to carry more information
  - Develop additional procedure to request peer node internal processing delay
    - Peer node directly fill internal processing delay value in response message without measurement.
- Pros:
  - Reuse PTP delay measurement mechanism without any changes.
- Cons:
  - A separate procedure to get peer node internal processing delay.
  - Internal processing delay is not based on measurement, may introduce inaccuracy.

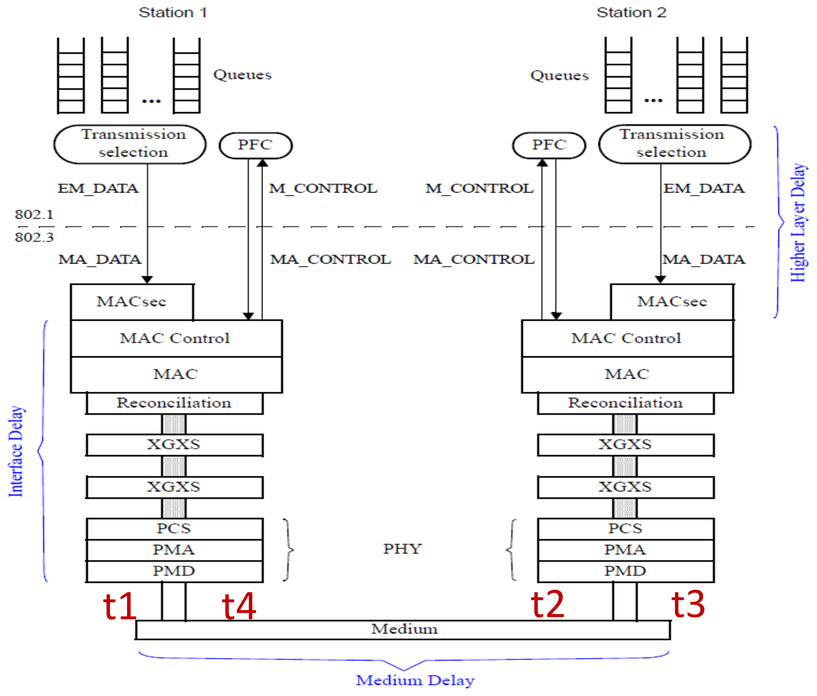
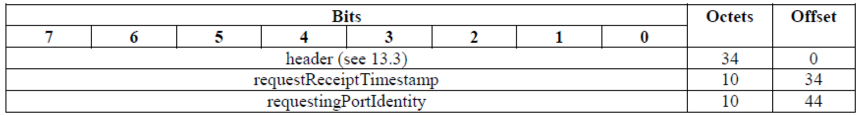


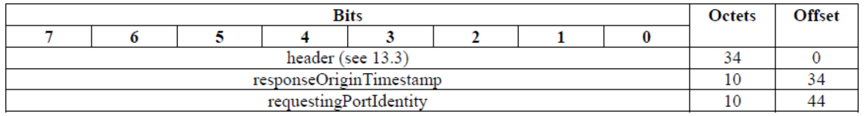
Figure N-2—Delay model (802.1Q-2018)

Table 48—Pdelay_Resp message fields



Bits								Octets	Offset
7	6	5	4	3	2	1	0		
header (see 13.3)								34	0
requestReceiptTimestamp								10	34
requestingPortIdentity								10	44

Table 49—Pdelay_Resp_Follow_Up message fields



Bits								Octets	Offset
7	6	5	4	3	2	1	0		
header (see 13.3)								34	0
responseOriginTimestamp								10	34
requestingPortIdentity								10	44

Summary & Next Steps

- Adaptive PFC headroom addresses the headroom configuration issue.
- PFC headroom measurement is technically feasible.
- 3 ways proposed to standardize PFC headroom measurement. Which one to choose could be further compared and decided when project starts.
- Next steps
  - Draft PAR & CSD to initiate a new project as amendment of 802.1Qbb(PFC)

Backup

PFC Environment Assumptions

- PFC is mainly used in datacenter network.
- Datacenter network is a different environment from typical TSN environment.
  - Higher link speed, could be 100Gbps or above.
    - Higher speed is more sensitive to delay.
  - Inter-Datacenter links can be as long as tens of kilometers.
    - Longer link put more pressure on buffer size.
  - PTP is NOT common in the datacenter
  - The delay measurement must cover not only link delay, but also **internal processing delay** of stations (including interface delay and higher layer delay).
    - Internal processing delay can be larger than link delay
    - Internal processing delay is hundreds of nanoseconds level or above, depending on implementation.
    - 802.3 defines maximum values.

Sublayer	25GbE(ns)	100GbE(ns)
RS, MAC and MAC control	327.68	245.76
BASE-R PCS	143.36	353.28
BASE-R PMA	163.84	92.16

		Fixed Delay	Internal Processing Delay	Medium Delay	Buffer size/queue	Queue number (5.8MB headroom)
Default Value	100G,500m	32992	203776	500000	92KB	32 ports * 2 queues
Real link distance	100G,20m	32992	203776	20000	32KB	32 ports * 5 queues
Real internal process delay	100G,20m	32992	203776*(3/4)=152832	20000	26KB	32 ports * 7 queues

		Fixed Delay	Internal Processing Delay	Medium Delay	Buffer size/queue	Queue number (23.552MB headroom)
Default Value	100G,500m	32992	203776	500000	92KB	128 ports * 2 queues
Real link distance	100G,20m	32992	203776	20000	32KB	128 ports * 5 queues
Real internal process delay	100G,20m	32992	203776*(3/4)=152832	20000	26KB	128 ports * 7 queues

		Fixed Delay	Internal Processing Delay	Medium Delay	Buffer size/queue	Queue number (5.28MB headroom)
Default Value	100G,200m	32992	203776	200000	55KB	48 ports * 2 queues
Real link distance	100G,20m	32992	203776	20000	32KB	48 ports * 3 queues
Real internal process delay	100G,20m	32992	203776*(3/4)=152832	20000	26KB	48 ports * 4 queues

		Cable delay (bit times)
100G Base-R	10m	5000 (0.6KB)
	10km	5 000 000 (625KB)

0.6KB for 10m estimation error (DCN case); 625KB for 10km estimation error (DCI case)

		ID + HD (bit times)
100G Base-R	802.3 max value	132 608
	Test value	100 000

Default settings may increase actual needs by 33%

PFC Delay Model

- PFC delay is RTT delay, from PFC pause frame is issued inside of station 1 until media drains.
- PFC delay consists of interface delay, medium delay and higher layer delay
  - Interface delay: the sum of MAC Control, MAC/RS, PCS, PMA, and PMD delays
  - Higher layer delay: the time needed for a queue to go into paused state after the reception of a PFC M_CONTROL.indication that paused its priority

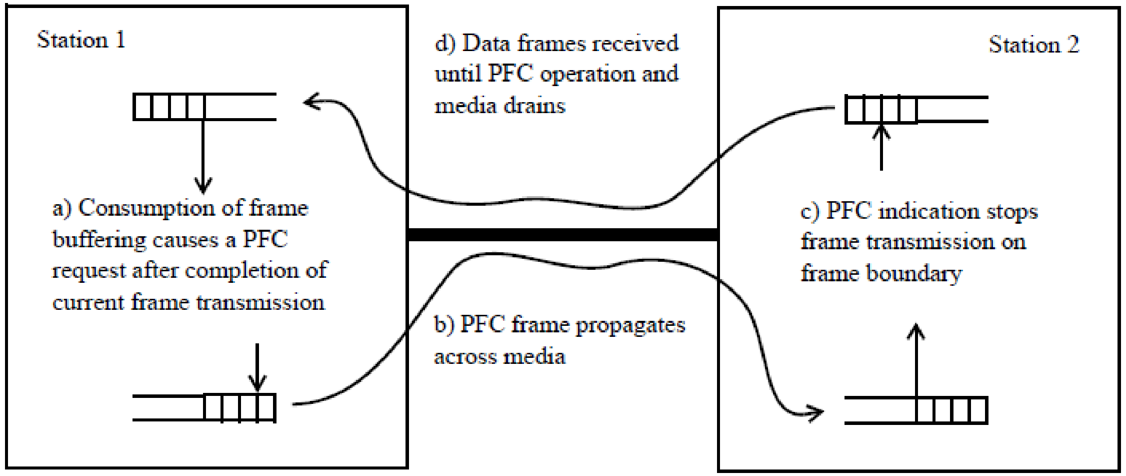
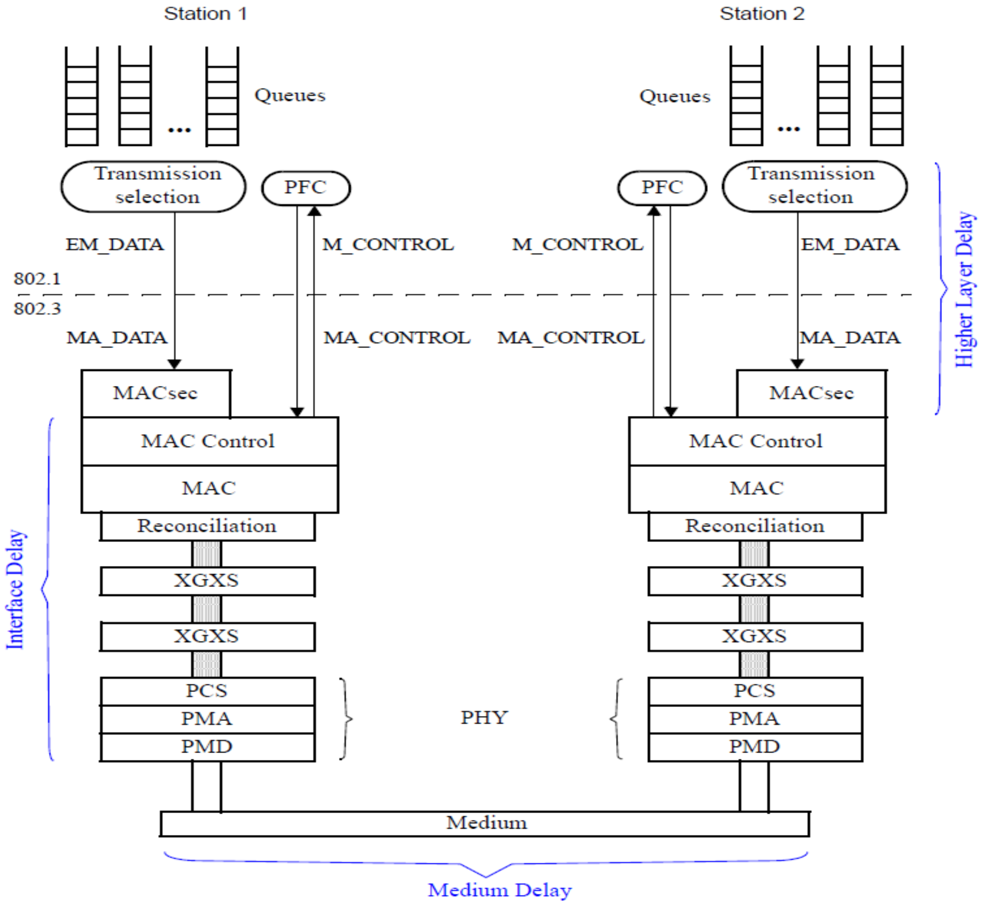


Figure N-1—PFC delays



Recap: Delay Measurement Mechanism in PTP and in 802.1AS

- PTP supports peer-to-peer delay link measurement
  - **It has one-step and two-step mechanisms**
  - One-step:
    - <meanLinkDelay> = [(t4 − t1) − correctedPdelayRespCorrectionField>]/2
    - correctedPdelayRespCorrectionField = t3-t2, **does not support sub-ns**
  - Two-step:
    - <meanLinkDelay> = [(t4 − t1) − (responseOriginTimestamp − requestReceiptTimestamp) − <correctedPdelayRespCorrectionField> − correctionField of Pdelay_Resp_Follow_Up]/2

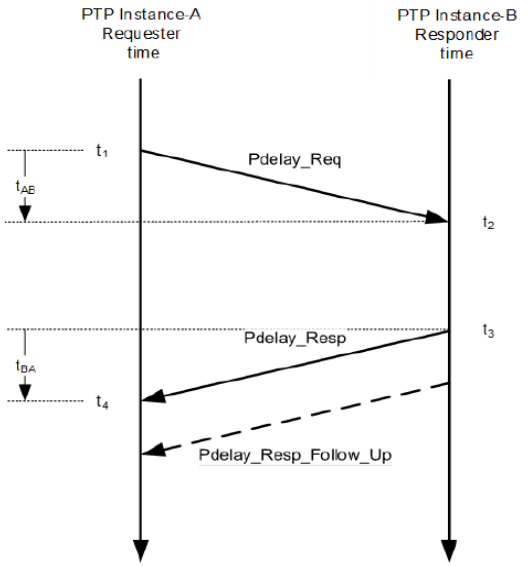
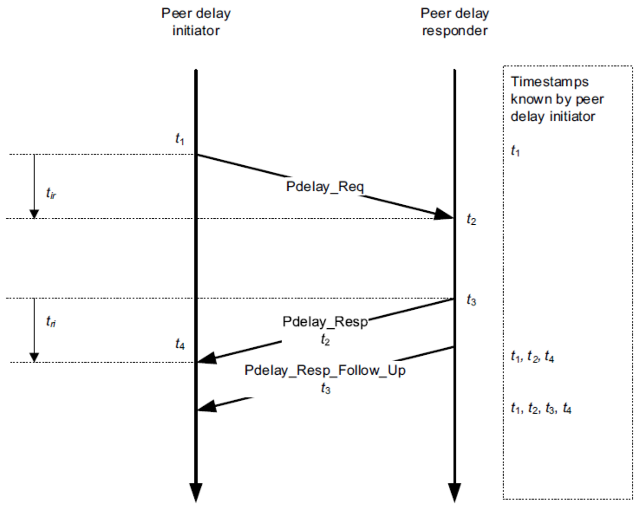


Figure 42—Peer-to-peer delay link measurement



igure 11-1—Propagation delay measurement using peer-to-peer delay mechanism

- 802.1AS follows PTP to measure propagation delay
  - Considering accuracy(sub-ns) and implementation complexity(compatibility, hardware capability), it chooses **two-step mechanism**.
    - “**The mechanism is the same as the peer-to-peer delay mechanism described in IEEE Std 1588-2019**, specialized to a two-step PTP Port and sending the requestReceiptTimestamp and the responseOriginTimestamp separately [see 11.4.2 of IEEE Std 1588-2019, item (c)(8)]. ”

Recap: Delay Measurement Timestamp Point in PTP and in 802.1AS

1588 (PTP)

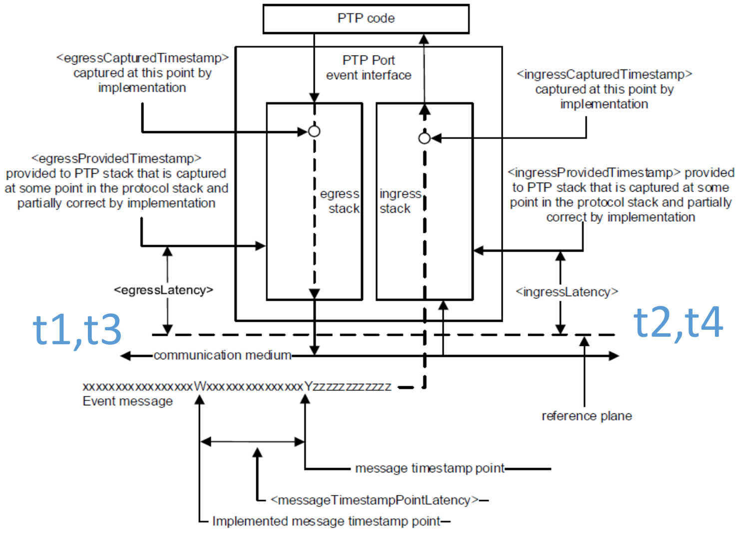


Figure 26—Definition of latency constants

ProvidedTimestamp = CapturedTimestamp +/- implementation-specific correction
messageTimestamp = ProvidedTimestamp +/- egress/ingress Latency

802.1AS

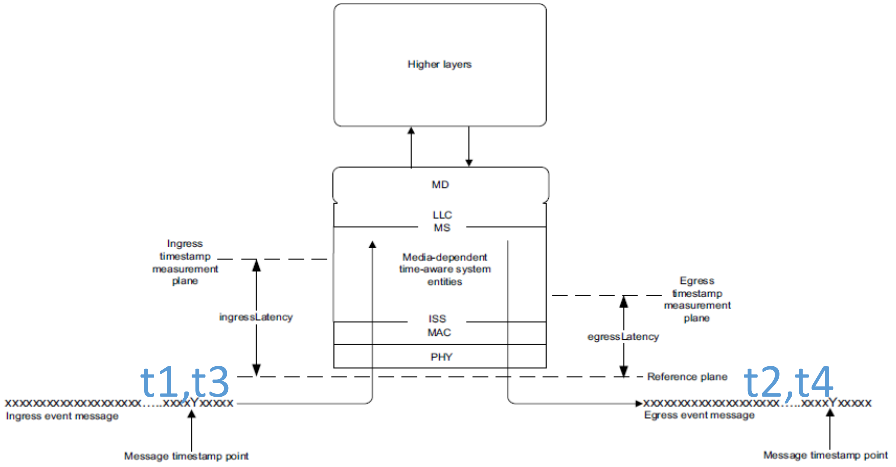


Figure 8-2—Definition of message timestamp point, reference plane, timestamp measurement plane, and latency constants

messageTimestamp = MeasuredTimestamp +/- egress/ingress Latency
“The timestamp measurement plane, and therefore the time offset of this plane from the reference plane, is likely to be different for inbound and outbound event messages”

802.3

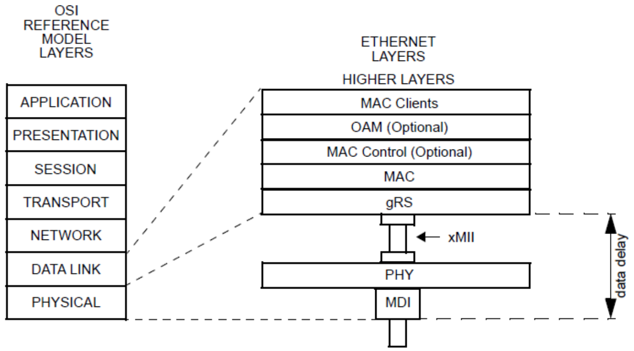


Figure 90–3—Data delay measurement

802.3 supports time sync by putting measurement timestamp point at xMII and providing PHY data delay(managed objects) as egress/ingress Latency.

- **Message timestamp point is at reference plane.** Correction is needed if implementation captured timestamp point is not message timestamp point.
- **Reference plane is between PTP instant and network.** For 802.1AS, it is between PHY and medium.
- **t1~t4 have same reference plane.**

Timestamp Point Analysis of PFC Headroom Measurement

- The delay includes time interval between point ① to point ⑪， not only cable delay, but also internal processing delay
  - $\text{Delay Value} = 2*(\text{Cable Delay}) + \underbrace{\text{TXds1} + \text{RXds2} + \text{HDs2} + \text{TXds2} + \text{RXds1}}_{\text{internal processing delay}} + \underbrace{2*(\text{Max Frame}) + (\text{PFC Frame})}_{\text{fixed value}}$
- Cable delay can reuse IEEE1588 or 802.1AS, but how about internal processing delay?
  - 2*(cable delay) = t4 – t1 – (t3 – t2)

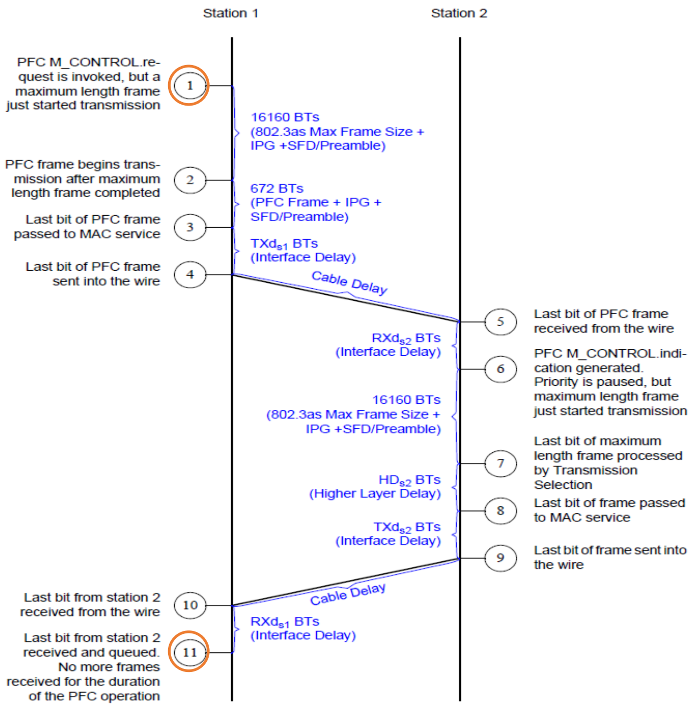
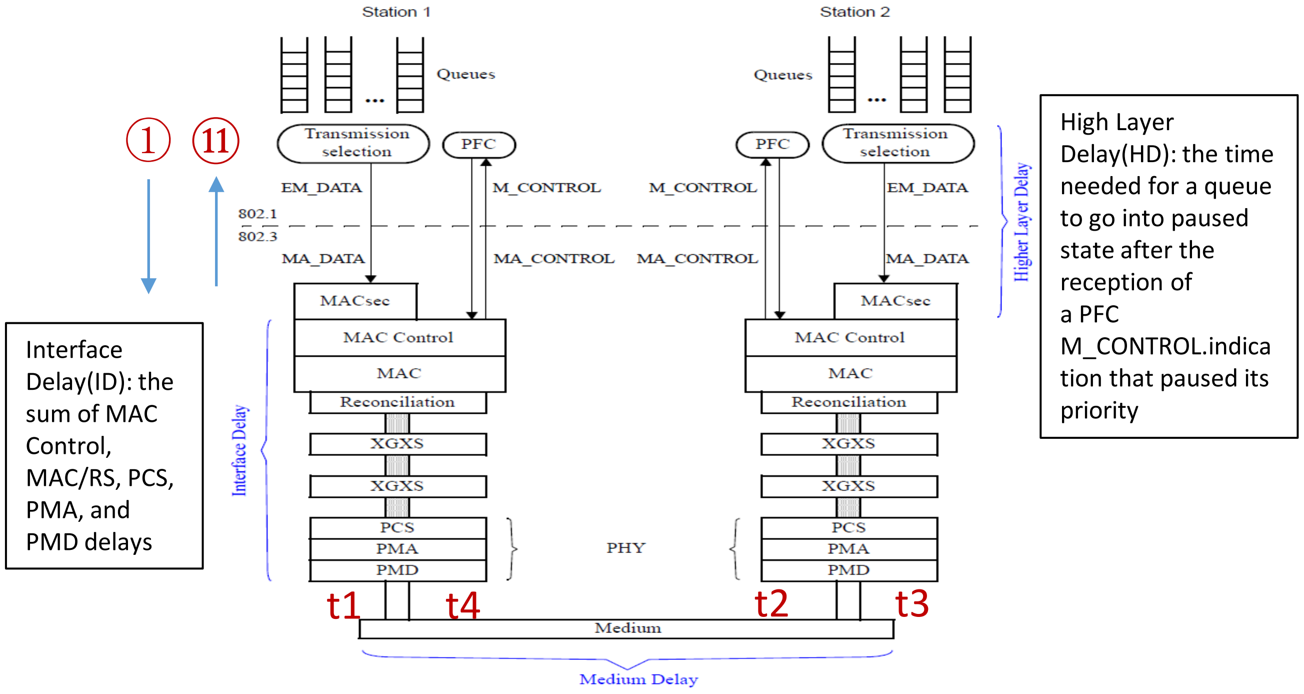


Figure N-3—Worst-case delay (802.1Q-2018)



Interface Delay(ID): the sum of MAC Control, MAC/RS, PCS, PMA, and PMD delays

High Layer Delay(HD): the time needed for a queue to go into paused state after the reception of a PFC M_CONTROL.indication that paused its priority

Figure N-2—Delay model (802.1Q-2018)

One-step or two-step in PFC Headroom Measurement Does Not Matter

- All 3 options does not care one-step or two-step mechanism for PFC headroom measurement.
  - Two-step is ok.
  - One-step could also be supported.
    - nanosecond level is accurate enough for headroom calculation
    - Implementation feasible
      - New function for PFC, no standard compatible issue as 802.1AS
      - Timestamp point does not need low level(PHY/MAC) support, so no stringent requirement on hardware