Adaptive PFC Headroom and PTP

Lily Lv (Huawei)

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
  - Reserve accurate buffer size as headroom to efficiently use buffer
    - Higher link speed (100Gbps or above in datacenter) is sensitive to delay
    - DCI links can be as long as tens of kilometers

		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)		

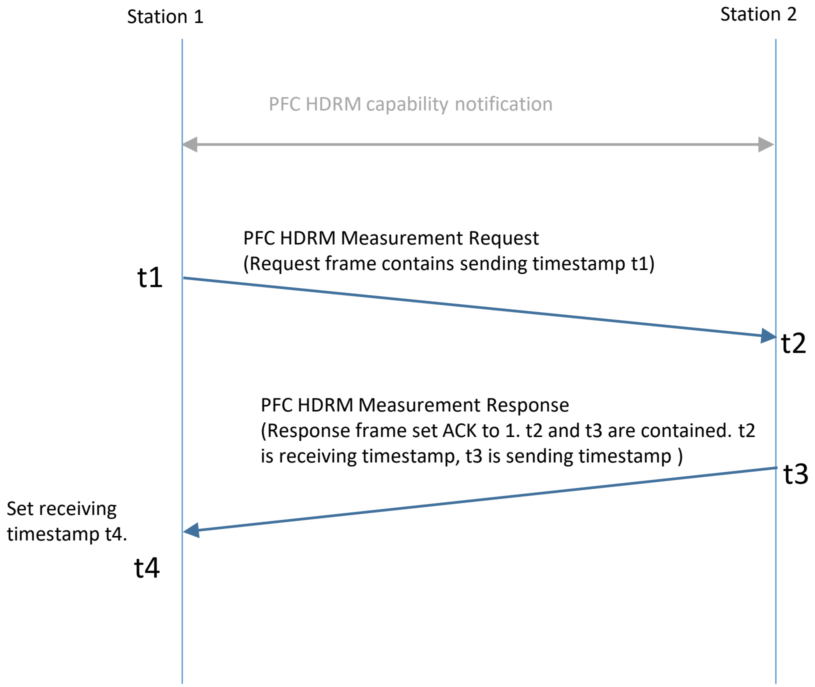
    - Implementation-specific delay could be better than max value defined in standard

		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%		

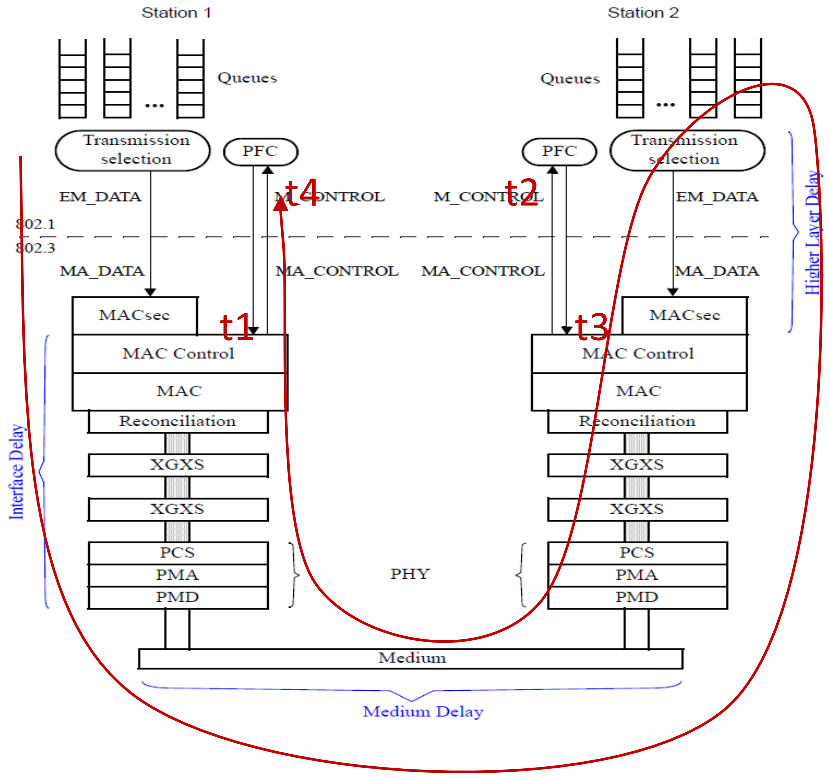
  - Automatic configuration to reduce heavy manual work
    - Headroom varies with link speed, link distance, vendor implementation etc.
    - Manual calculation for each port is time consuming task for engineers.

Background

- Proposed mechanism of adaptive PFC headroom
  - The delay measurement procedure is similar to PTP, to measure roundtrip delay, from PFC pause frame is issued inside of station 1 until media drains
  - 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.



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

- Feedbacks on the proposal were received
  - Q1: What is the measurement resolution requirement?
  - Q2: Can we leverage the existing protocol in 802.1AS or IEEE1588? (Implementation feasibility)

Q1: What is the measurement resolution requirement?

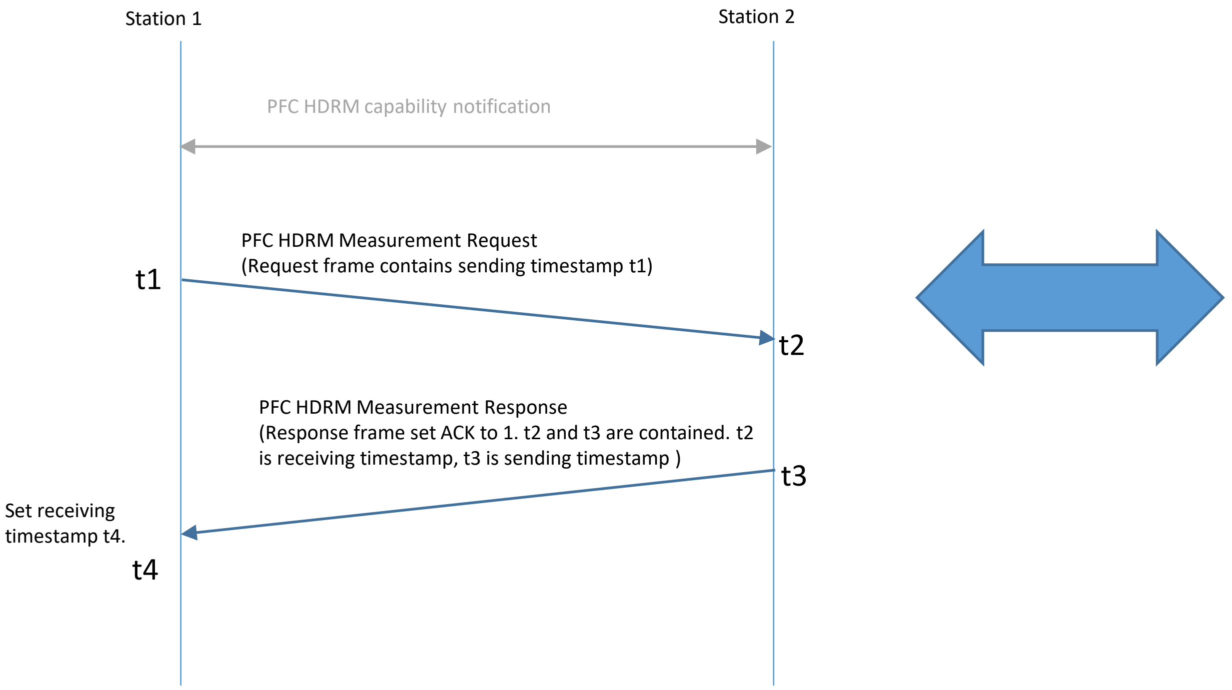
Time Accuracy Analysis of PFC Headroom Measurement

- 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
  - Assume 125MHz clock, timestamp resolution is 8ns
    - (t4-t1) is the roundtrip delay, including link delay and station internal processing delay. It is above micro-seconds.

Q2: Can we leverage the existing protocol in 802.1AS or IEEE1588? (Implementation feasibility)

Reuse PTP Measurement Procedure

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



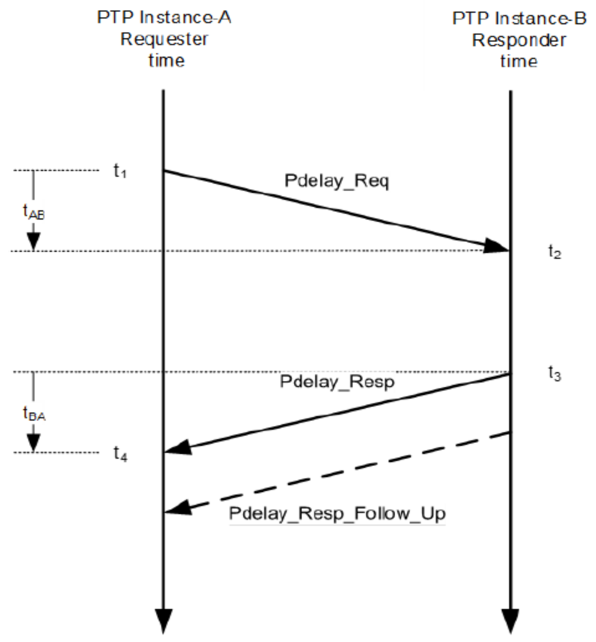


Figure 42—Peer-to-peer delay link measurement

Redefine Timestamp Points

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

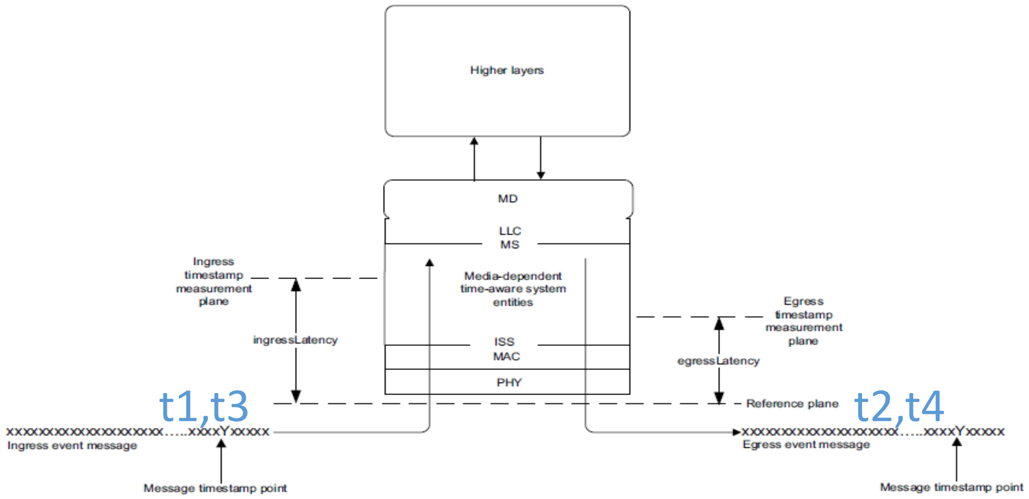
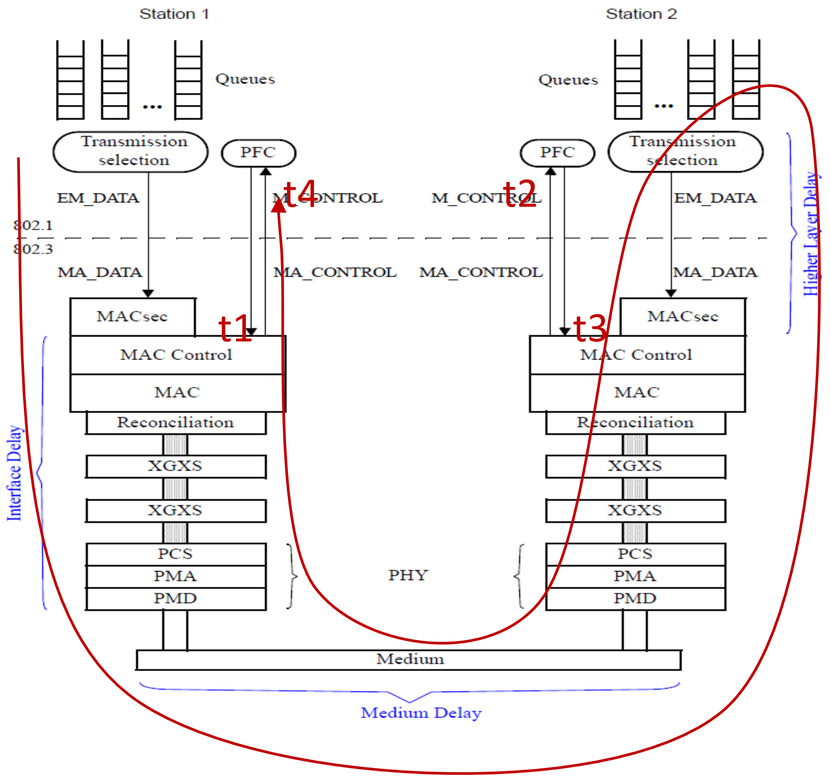


Figure 8-2—Definition of message timestamp point, reference plane, timestamp measurement plane, and latency constants

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 802.1AS, reference plane(s) for message timestamp points need to be redefined.
  - t1~t4 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 protocol but define separate mechanism to get peer node HD and ID
    - Reuse IEEE1588 or 802.1AS PTP protocol to measure cable delay
        - Pdelay_Resp/Pdelay_Resp_Follow_Up does not have reserved payload fields to carry more information
    - Develop new procedure and new message to request peer node HD and ID
        - Peer node directly fill HD, ID value in new defined response message without measurement.
- Pros:
    - Reuse IEEE1588 or 802.1AS delay measurement mechanism without any changes.
- Cons:
    - Switches have common understanding of HD and ID. Additional procedure to get HD, ID.
    - HD/ID value is not based on measurement, may introduce inaccuracy.

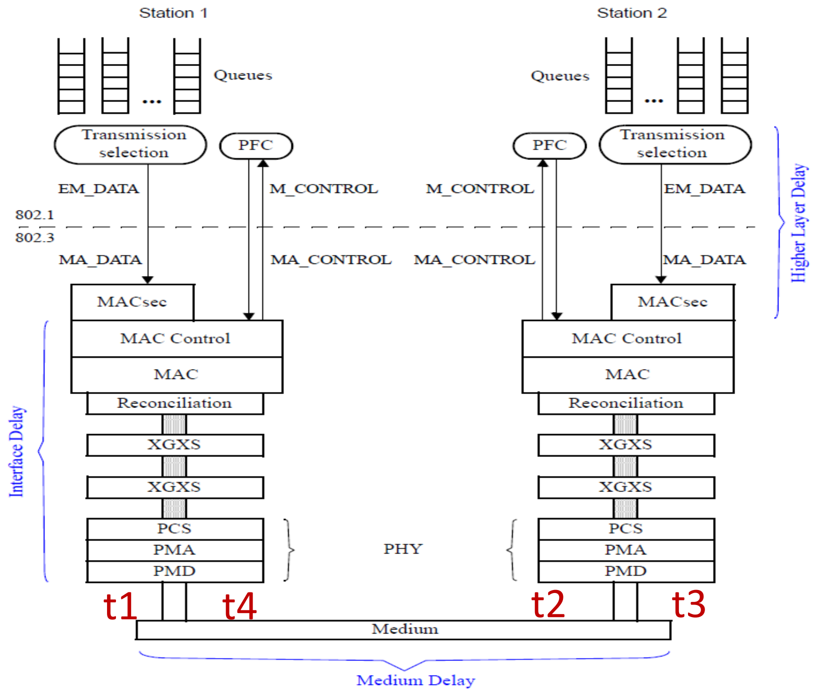


Figure N-2—Delay model (802.1Q-2018)

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

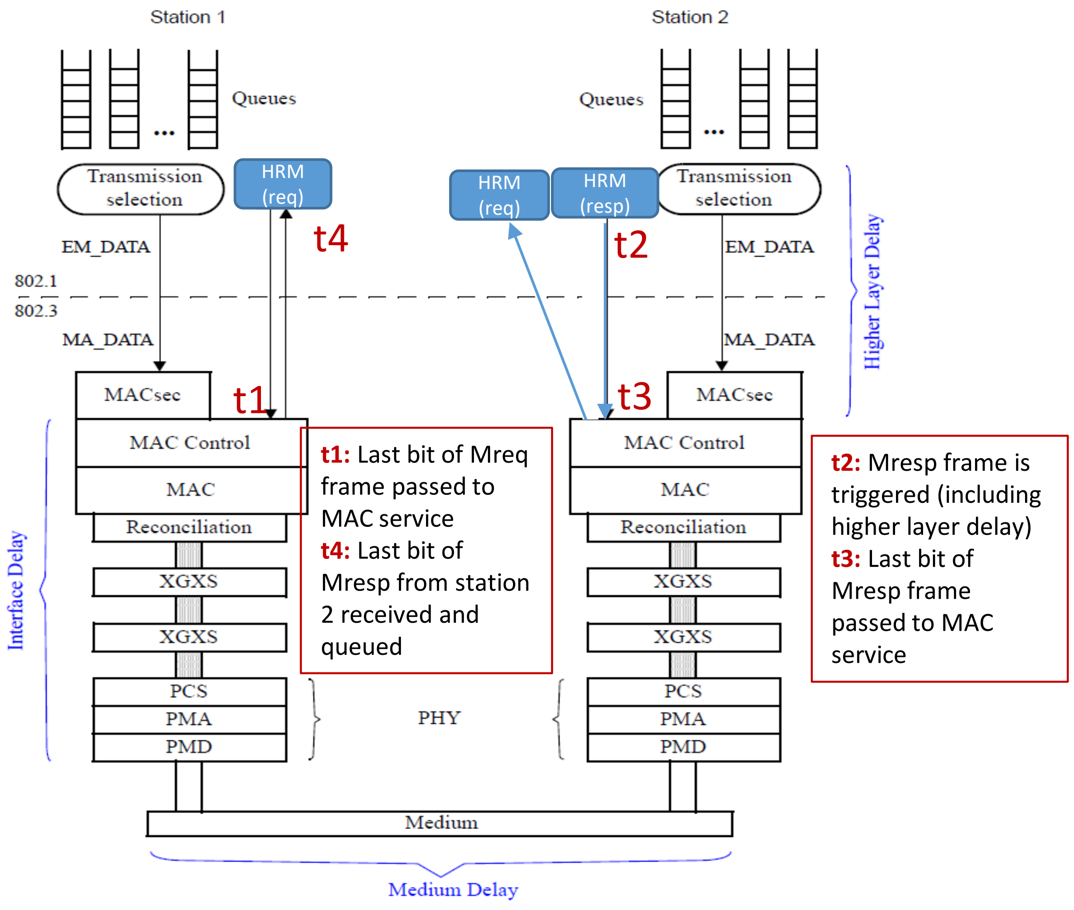
Table 49—Pdelay_Resp_Follow_Up message fields

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

Proposals for Implementation(2/3)

- Option 2: reuse PTP mechanism but change reference plane, including internal processing delay in the measurement
  - Neither of IEEE1588 and 802.1AS PTP reference plane can be used
    - IEEE1588 PTP reference plane is general, between PTP instant and network.
    - 802.1AS redefines PTP reference plane between PHY and medium.
  - PFC headroom measurement expects reference plane 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
    - Message timestamp calculation is the same as 802.1AS, egress/ingress latency can be different for different timestamp point.

      messageTimestamp = MeasuredTimestamp +/- egress/ingress Latency
- Pros:
  - Little changes to PTP delay measurement, but with measured HD/ID, more accurate for headroom calculation.
- Cons:
  - Need to redefine reference plane.

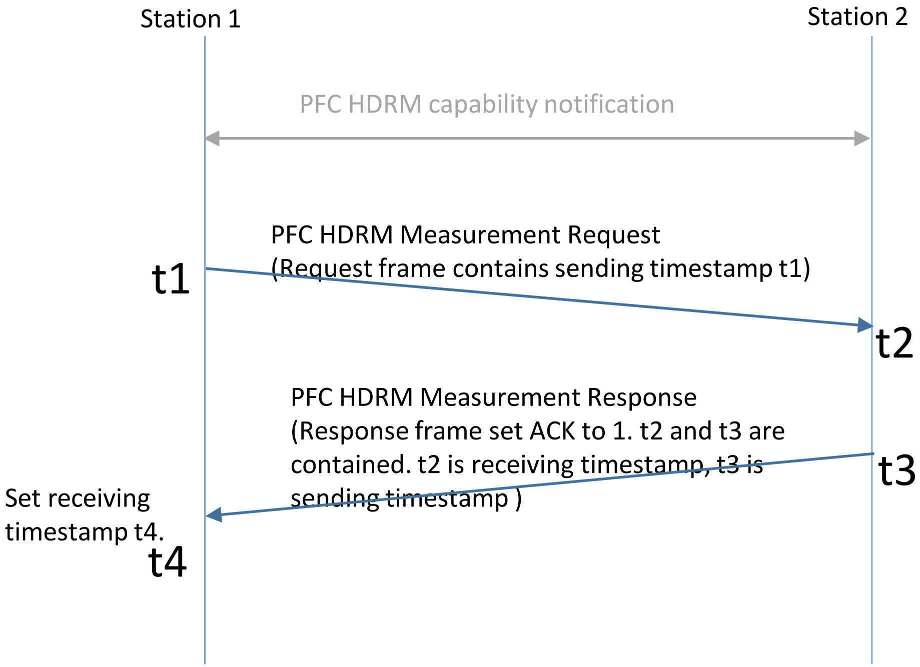


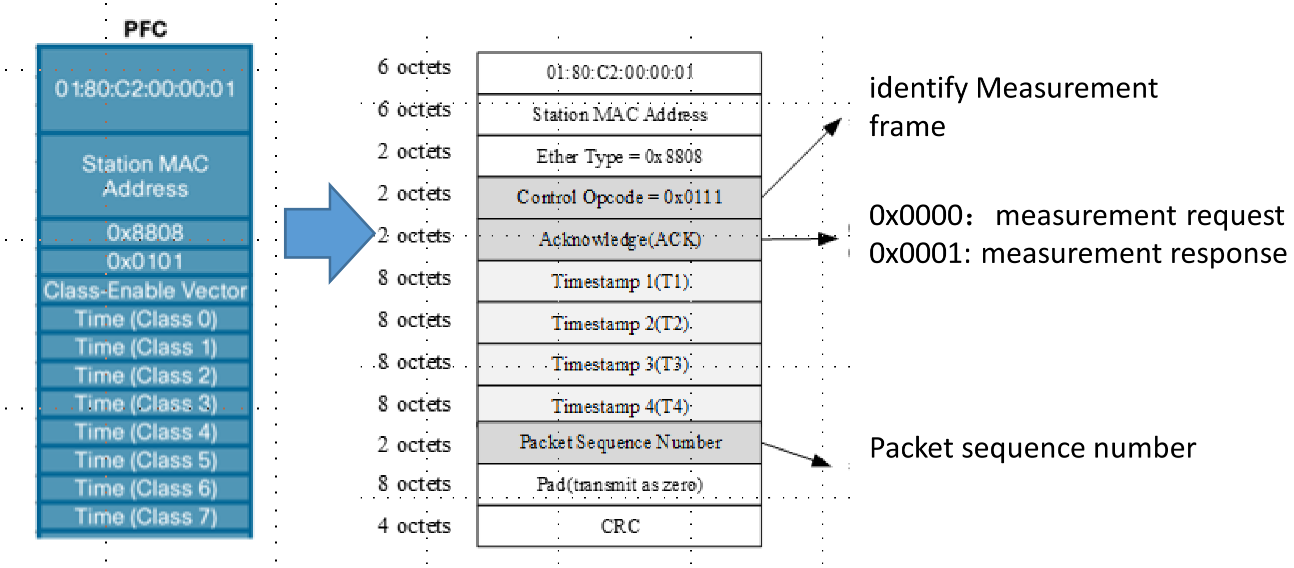
DV = 2*(Max Frame) + (PFC Frame) + 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1

t4-t1-(t3-t2)

Proposals for Implementation(3/3)

- Option 3: design MAC control frame as delay measurement message
  - Internal processing delay(ID) for MAC control frame and MAC data frame may have difference.
    - PFC frame is MAC control frame, while PTP delay measurement frame is MAC data frame.
  - Measurement mechanism and reference plane is the same as option2, but design MAC control frame as the interactive messages.
- Pros:
  - More like PFC delay procedure, can be more accurate
  - Implementation friendly, do not change time sync module.
- Cons:
  - New design of message format
  - Need to redefine reference plane.





PFC frame format

Measurement frame format

Summary & Next Step

- Adaptive PFC headroom benefits buffer usage efficiency and manual work reduction.
- 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 step
  - 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

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

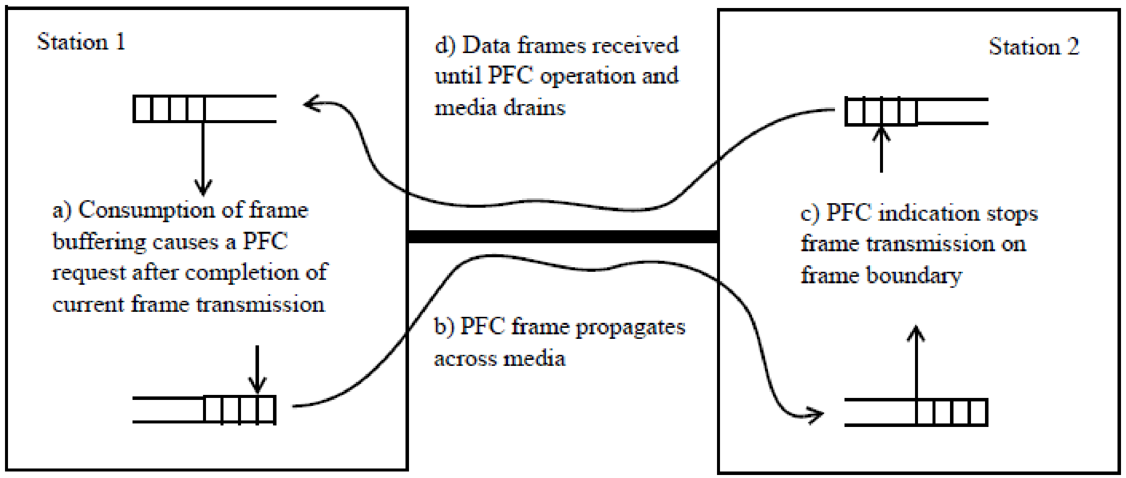
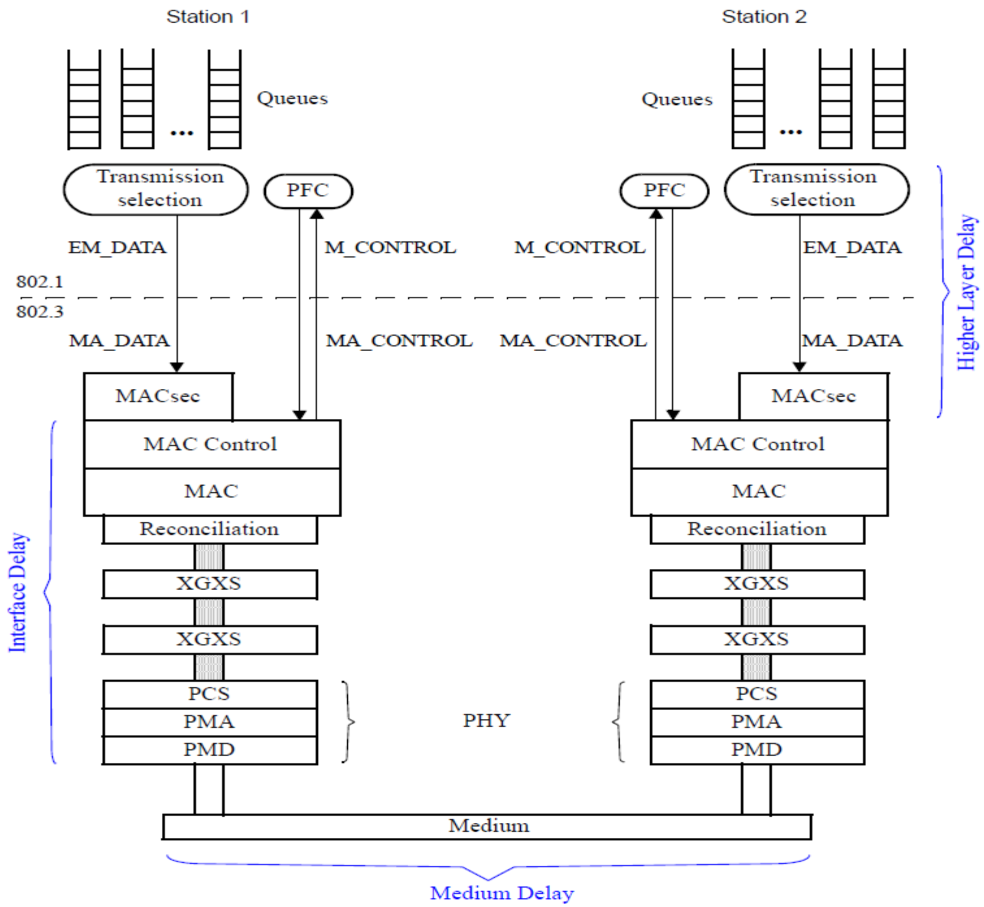


Figure N-1—PFC delays



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

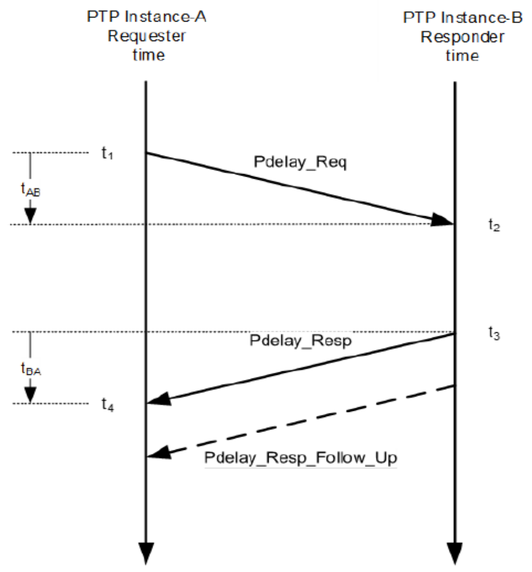
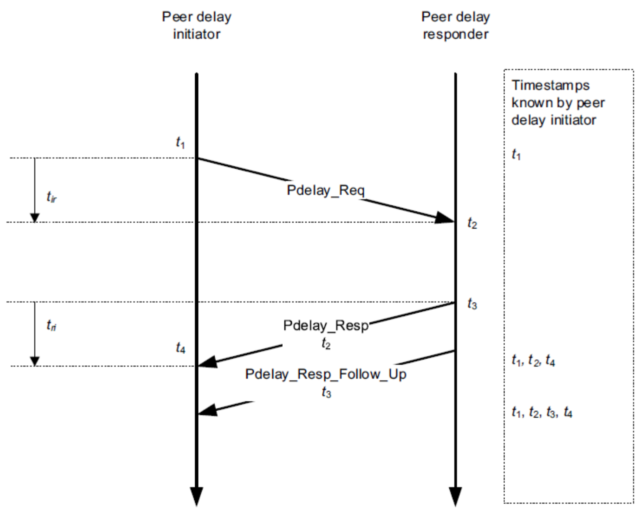


Figure 42—Peer-to-peer delay link measurement



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)

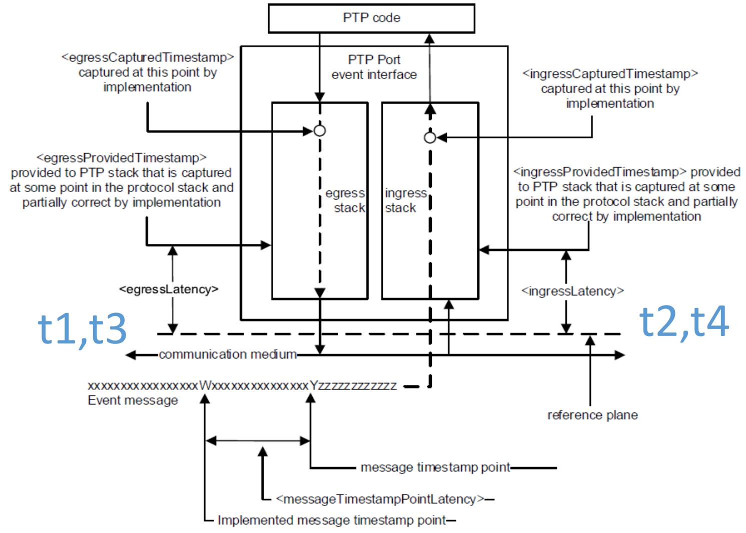


Figure 26—Definition of latency constants

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

802.1AS

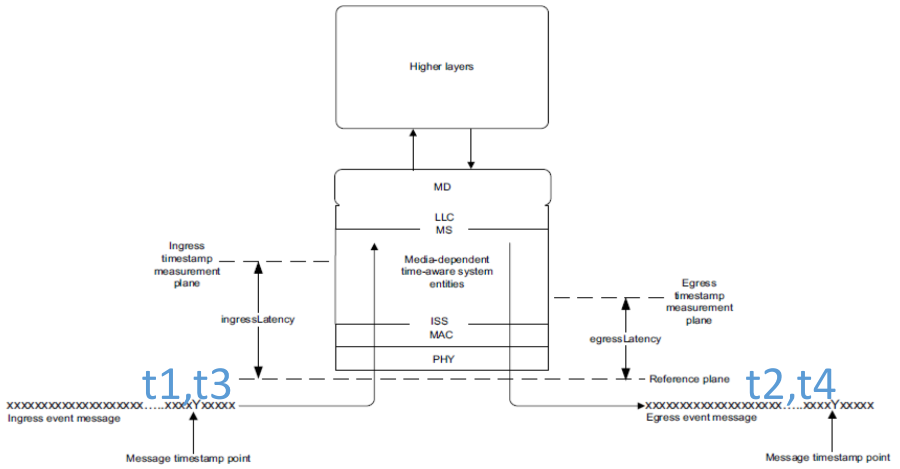


Figure 8-2—Definition of message timestamp point, reference plane, timestamp measurement plane, and latency constants

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

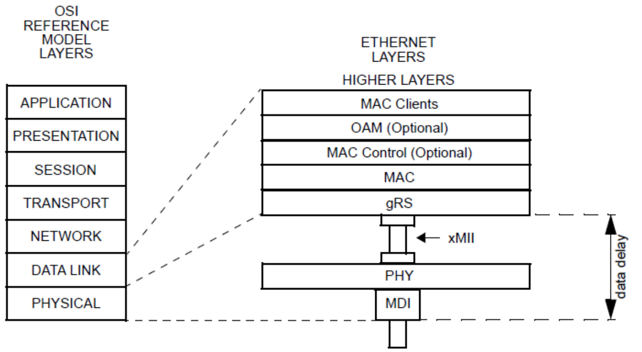


Figure 90–3—Data delay measurement

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
  - Delay Value = 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1 + 2*(Max Frame) + (PFC Frame)
    - TXds1 + RXds2 + HDs2 + TXds2 + RXds1: internal processing delay
    - 2*(Max Frame) + (PFC Frame): fixed value
- Cable delay can reuse IEEE1588 or 802.1AS, but how about internal processing delay?
  - 2*(cable delay) = t4 – t1 – (t3 – t2)

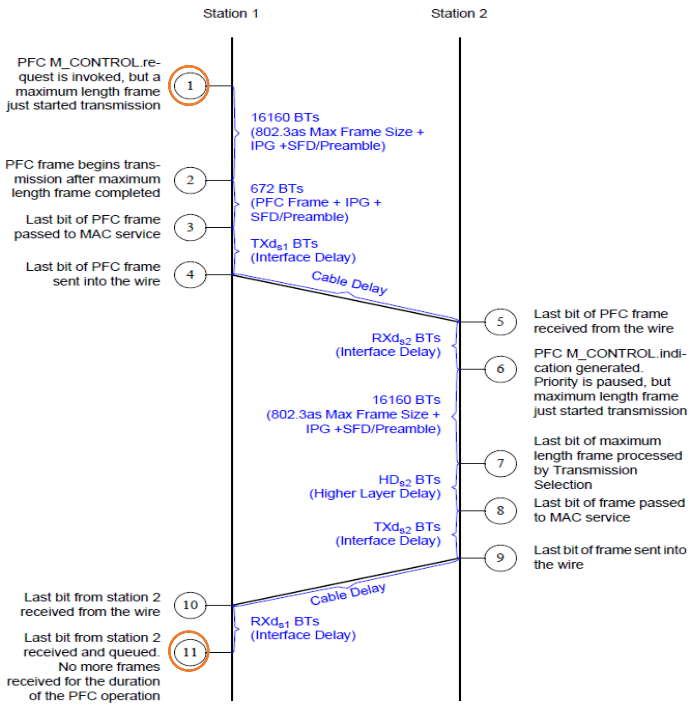
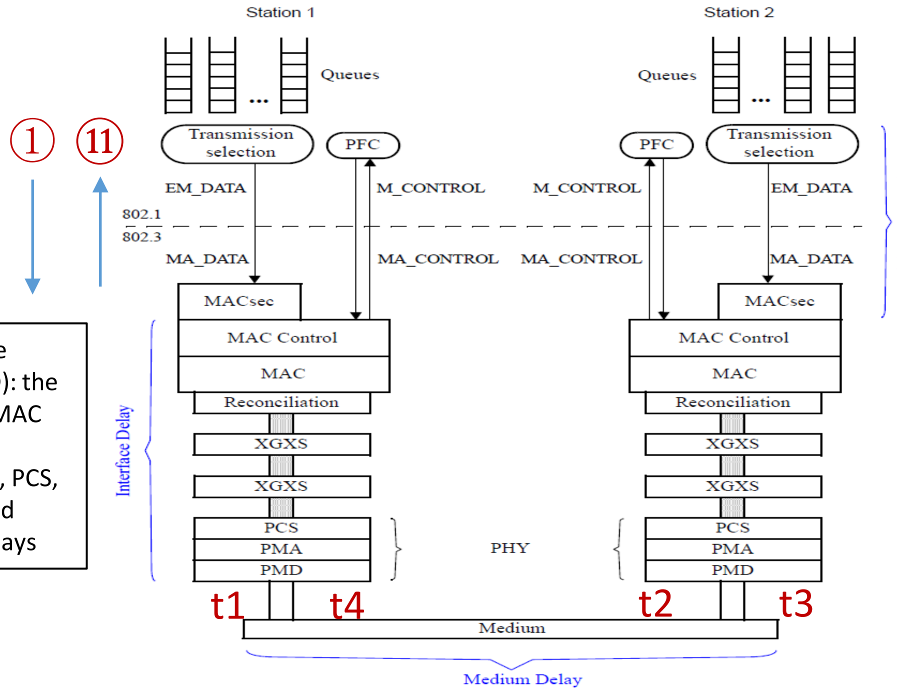


Figure N-3—Worst-case delay (802.1Q-2018)

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