PFC Headroom and Enhancements Project Proposal

Paul Congdon, Lily Lv (Huawei)

Mick Seaman (Independent)

History

- Initial proposal focus on the problem of automatic configuration of PFC Headroom in data centers within TSN TG
- 802.1 WG process adjustment to ‘vet’ new work within Nendica – technical options narrowed
- Motion to authorize development of PAR and CSD in July was ‘tabled’ to address operation with MACSec
- Subsequent collaboration and presentations to propose technical solutions for MACSec interworking and PFC forwarding.
- Nendica consensus is that further discussion should be brought to the 802.1 WG or a Task Group based on the PAR description in 802.1-21-0052 (rev 00 or newer).

- Lots of previous presentations.
  - 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
  - https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-and-PTP-0602-v03.pdf - Adaptive PFC Headroom and PTP
  - https://www.ieee802.org/1/files/public/docs2021/cz-finn-pfc-headroom-0629-v01.pdf - Determining Priority Flow Control Headroom
  - https://www.ieee802.org/1/files/public/docs2021/new-lv-PFC-Headroom-Project-Proposal-0721-v01.pdf - PFC Headroom Measurement and Calculation Project Proposal
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0048-00-ICne-pfc-headroom-with-macsec.pdf - Incorporating MACSec into PFC Headroom Calculation
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0050-00-ICne-pfc-enhancements-project-proposal.pdf - PFC Enhancements Project Proposal
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0052-00-ICne-pfc-enhancements-next-steps.pdf - PFC Enhancements Project Proposal

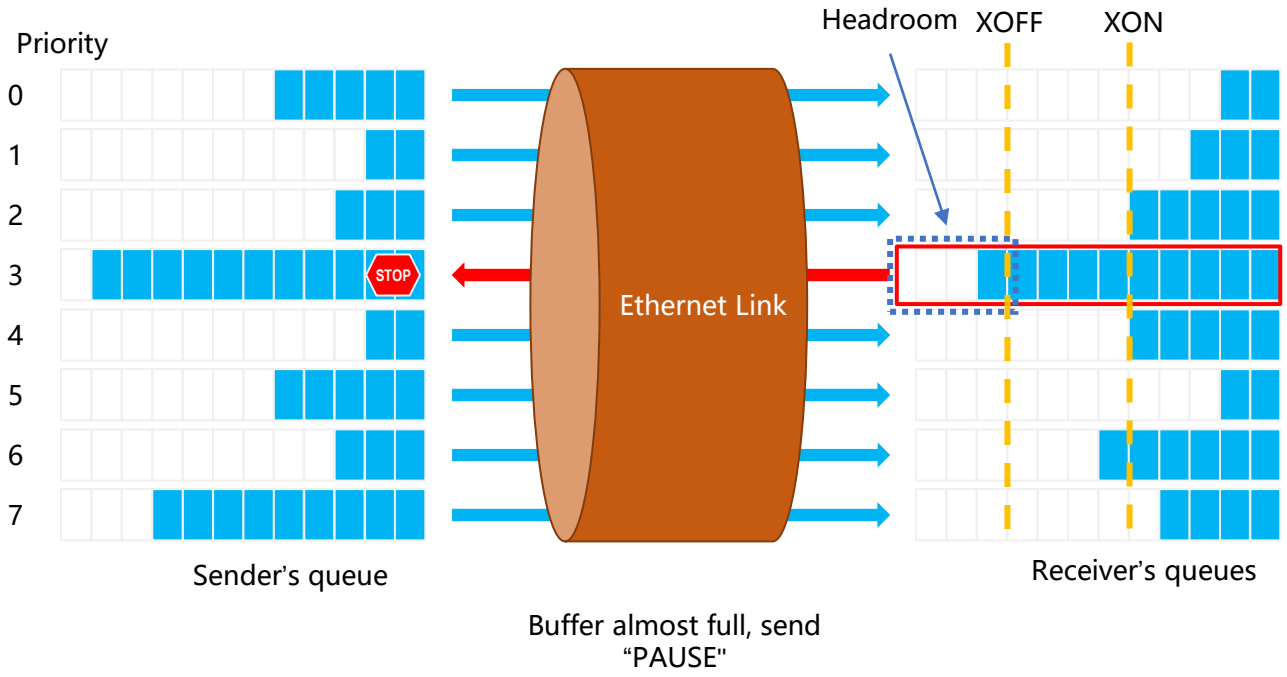
Motivation

- PFC is used in low-latency Ethernet data center networks to avoid packet loss.
- Deploying PFC today can be difficult
  - Manual configuration is complex and is different for each vendor solution
  - Consistent settings across a large-scale data center network is tedious
  - Vendor provided default values waste buffer resource, and do not work in certain circumstances (e.g. long distance data center interconnection)
- A standard is needed to specify any wire protocols (e.g. capability exchange) and a headroom measurement mechanism.
- Inconsistent and unclear specification of PFC and MACSec operation

See: https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-and-PTP-0602-v03.pdf

Adaptive PFC Headroom Calculation

Objective: Automatically calculate minimum PFC buffer requirements (i.e. headroom) for lossless operation, without user intervention.



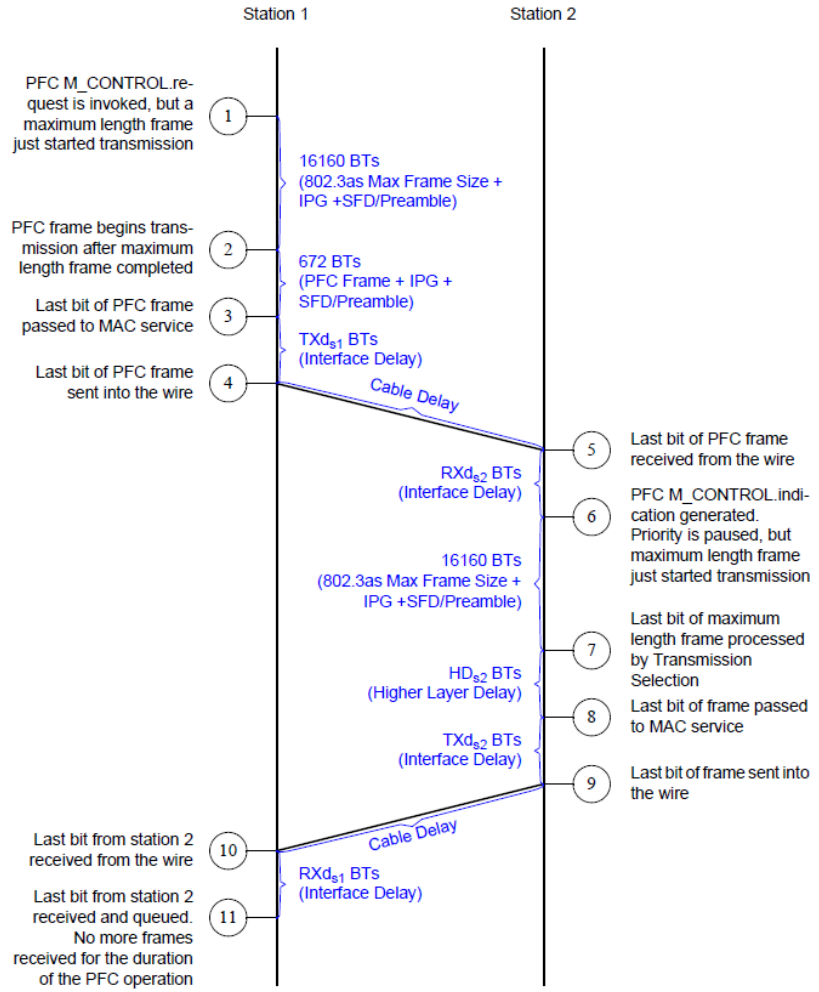
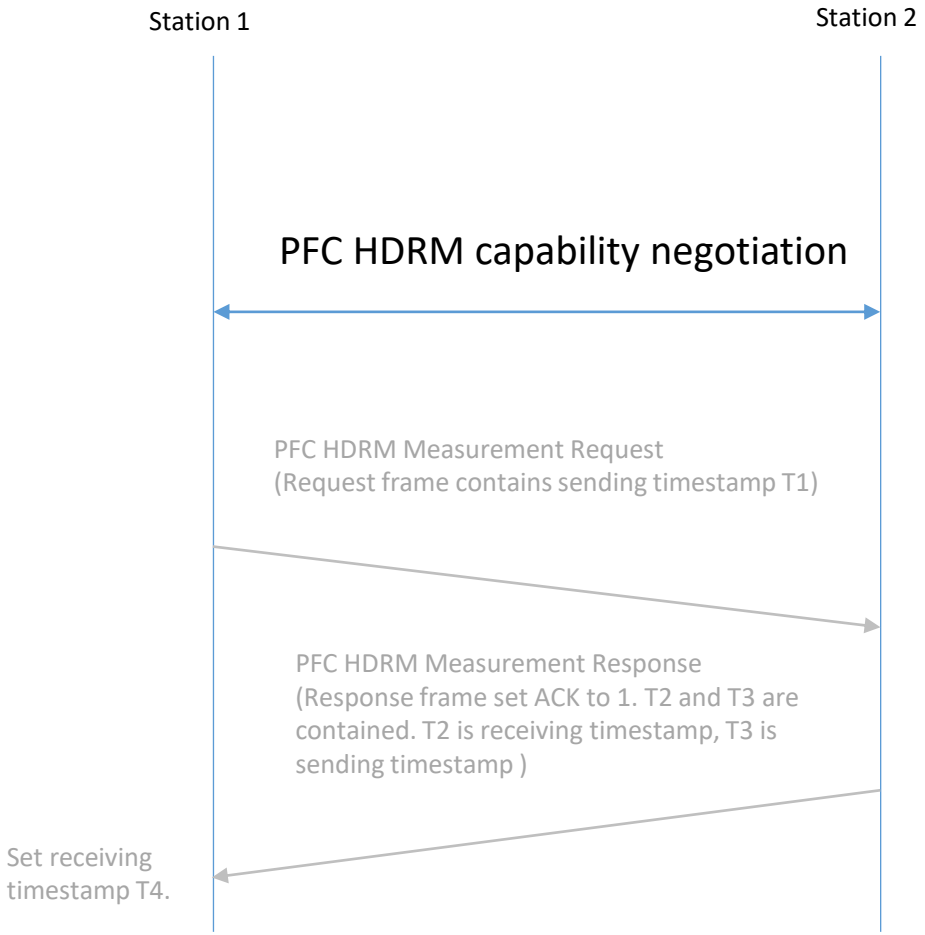
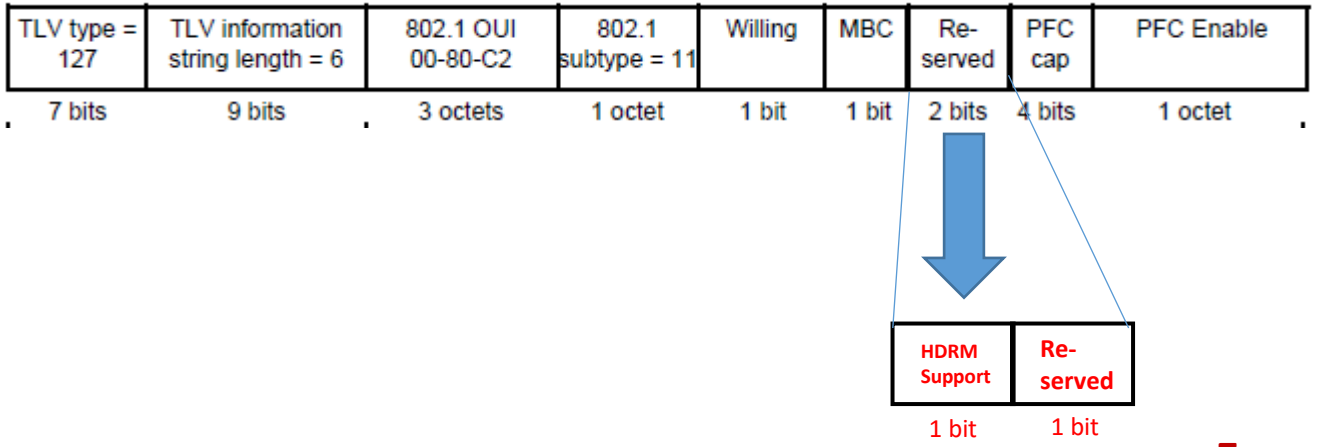


Figure N-3—Worst-case delay (802.1Q-2018)

Negotiating PFC Headroom Capability



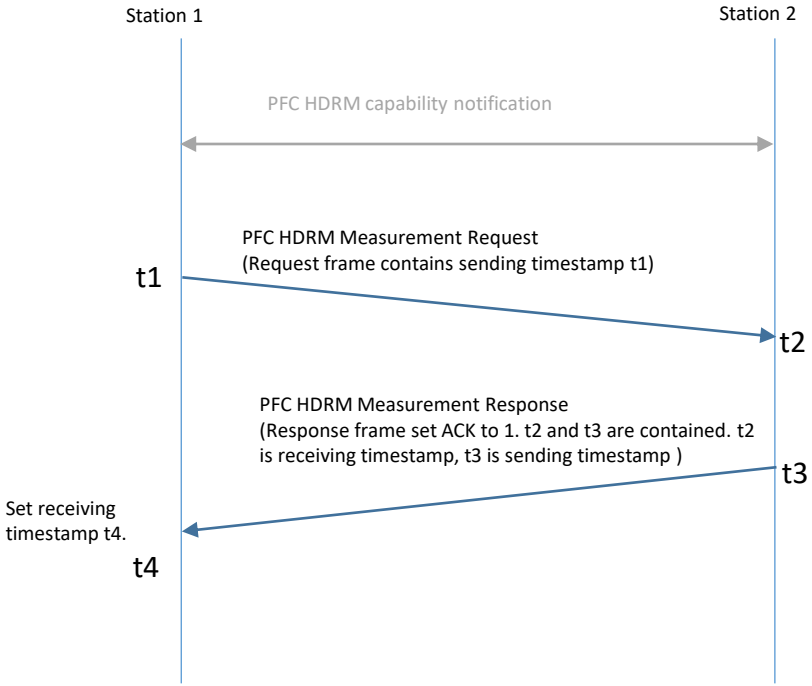
- Capability negotiation
  - Augment DCBX by extending PFC configuration TLV
  - DCBX uses LLDP with updated PFC configuration TLV to exchange HDRM capability
  - If both support PFC HDRM and PFC is enabled, initiate PFC HDRM Measurement Request, otherwise, stop the procedure.



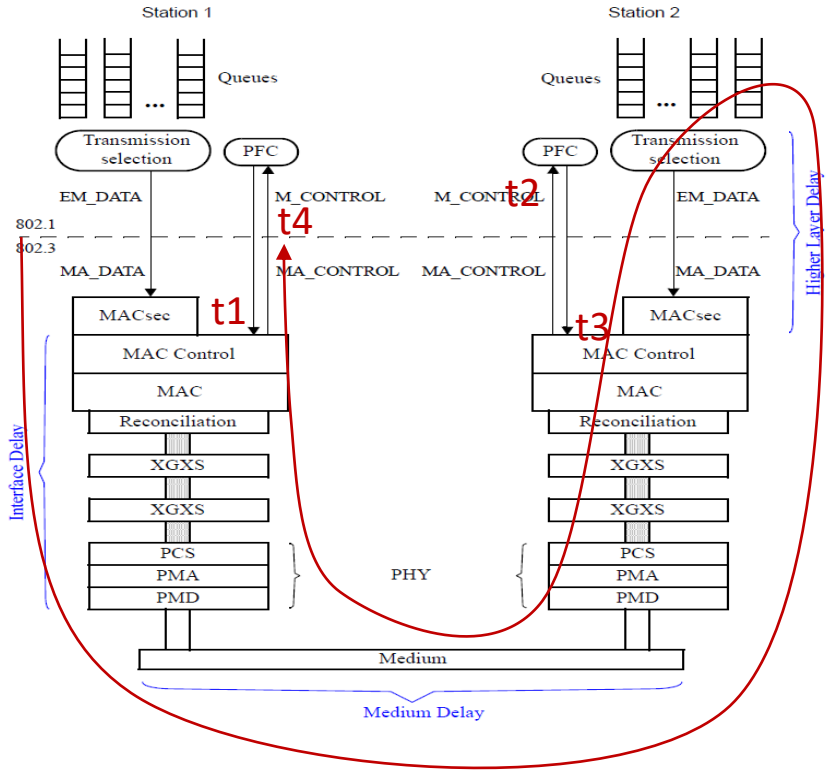
Example

Measuring PFC Headroom

- The delay measurement procedure can reuse 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



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



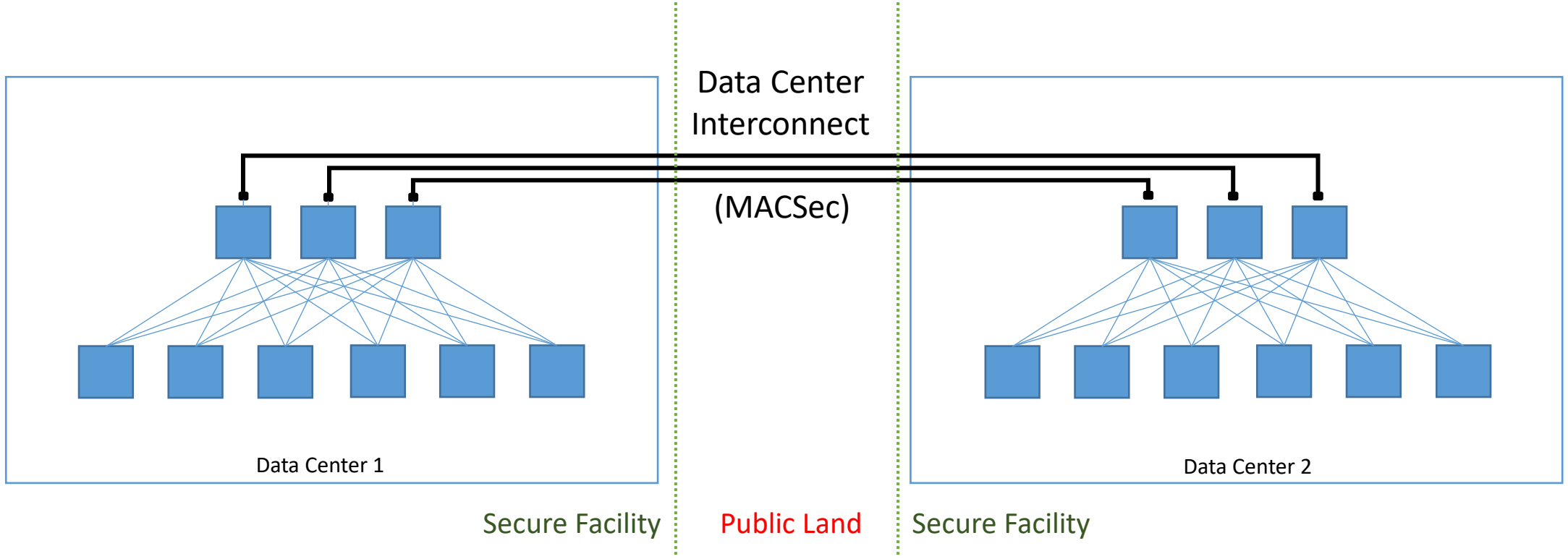
High Layer Delay(HD): the time needed for a queue to go into paused state after the reception of a PFC M_CONTROL.indication that paused its priority

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

Issues with PFC interoperation with MACSec

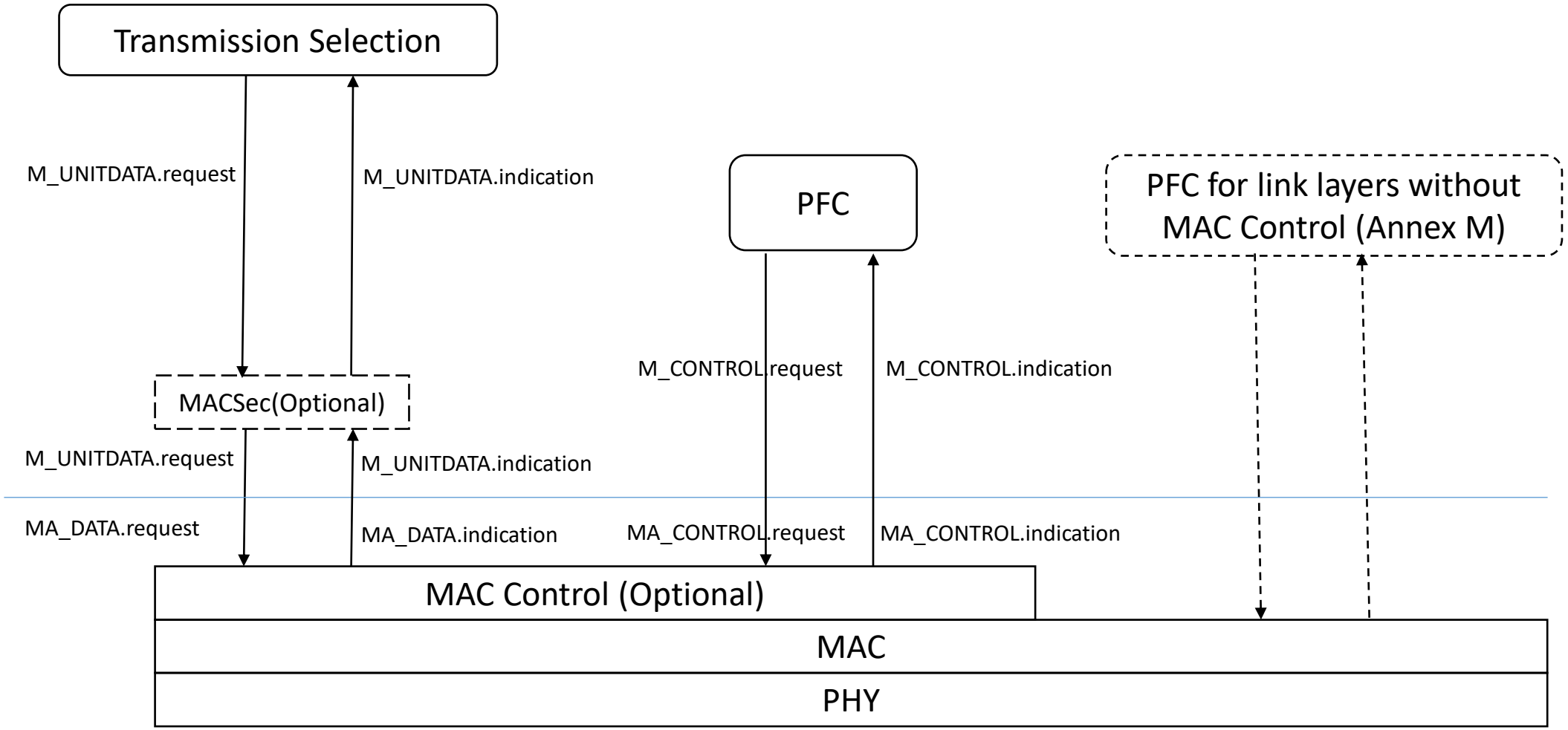
- Implementations in the field have interoperability issues
- Confusion in our specifications about interworking with MACSec
- Inadvertently or purposely forwarding PFC is problematic
- Desired use-case to MACSec protect PFC frames.

A Use Case To Consider with MACSec



NOTE: The RDMA protocol over Ethernet (RoCEv2) necessitates the use PFC to avoid frame loss
It is desirable to protect PFC frames when they traverse data center interconnect links

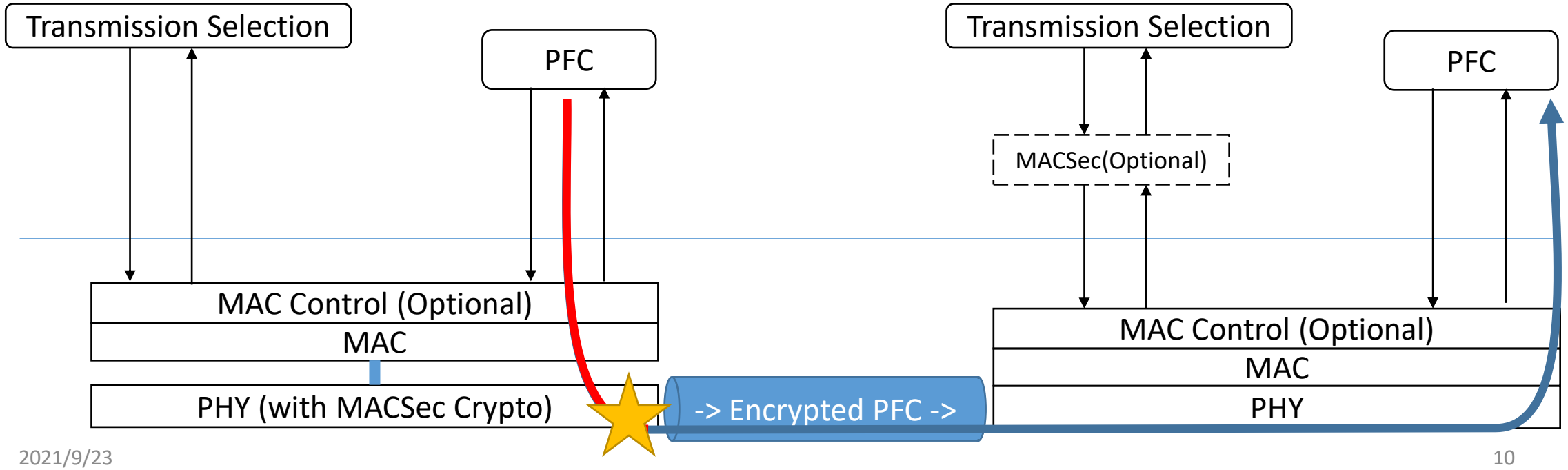
Current Protocol Layers



NOTE: Figure indicates that PFC Frames are not encrypted

Interoperability issue in the field

- Early implementations of MACSec were implemented external to the MAC (i.e. within a PHY as a ‘bump in the wire’).
  - These early implementations encrypt everything coming out of the MAC
  - These early implementations were never compliant with 802.1AE
  - These early implementations do not run MKA and may suffer outages



Ethernet Data Encryption (EDE-M)

Specifies 'Bump in the Wire' Behavior

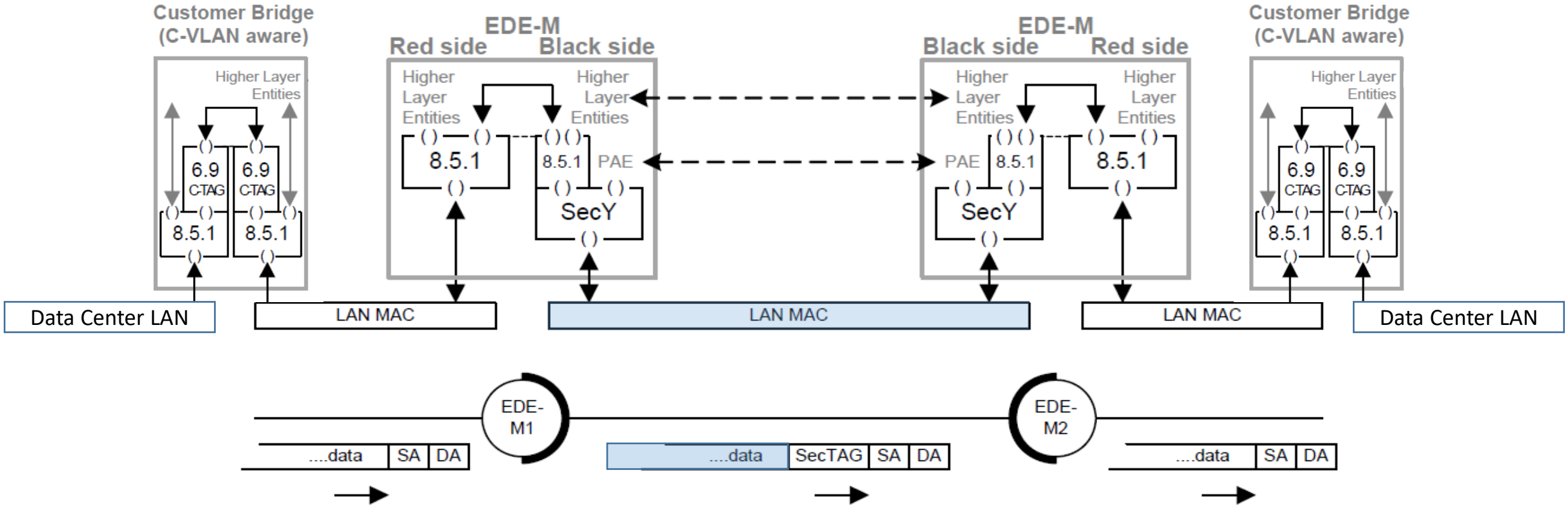
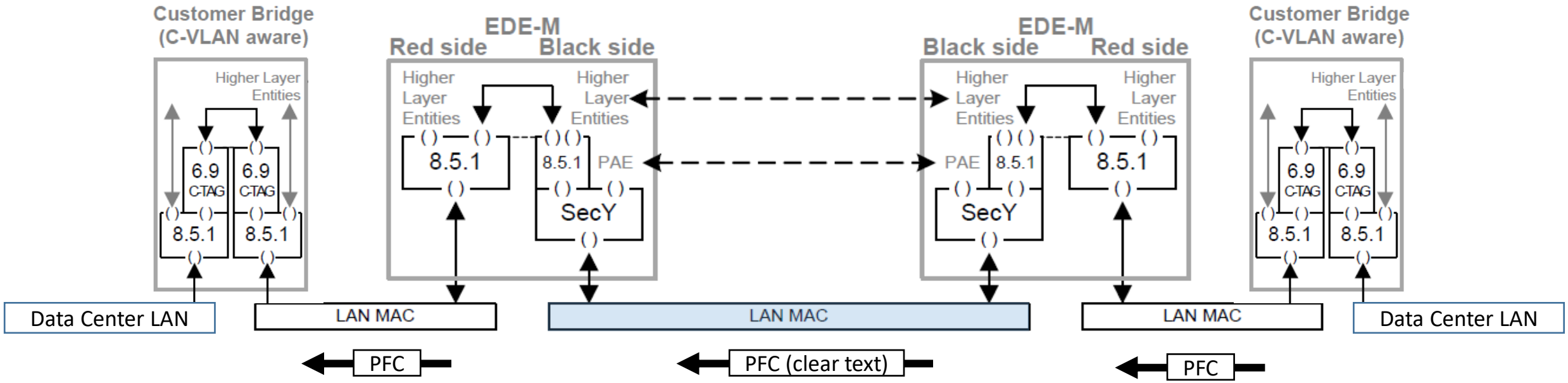


Figure 15-1—EDE-Ms connected by a point-to-point LAN

EDE-M Conformance:

a) Comprise a VLAN-unaware MAC Bridge as specified by IEEE Std 802.1Q (5.14 of IEEE Std 802.1Q-2018) with the constraints and exceptions specified in this standard.

No PFC Forwarding, but backward propagation



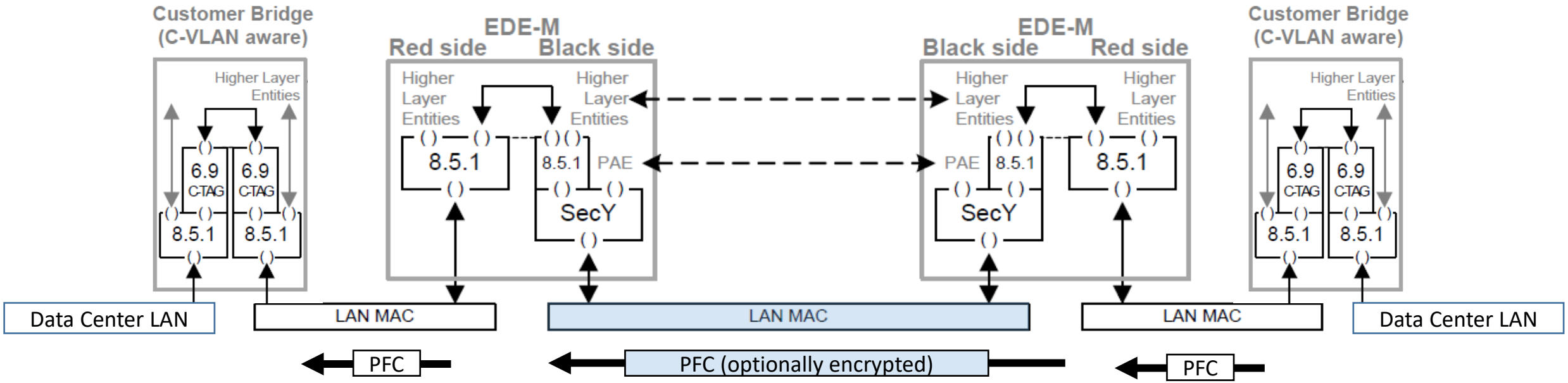
802.1 Reserved Addresses are not forwarded:

a) PFC uses reserved address 01-80-C2-00-00-01

b) C-VLAN Bridges, MAC Bridges, TPMRs and EDEs do not forward 01-80-C2-00-00-01

Desired Future State

- Allow an option for protecting/securing PFC frames
- Do not change the backward propagation model of PFC
- Include MACSec delay into the PFC Headroom calculation



PFC over PBNs

Use Case: distributed data centers connected by a service provider (PBN), rather than dark fiber. The envisaged need PFC from one Provider Edge Port (PEP) to another across the PBN

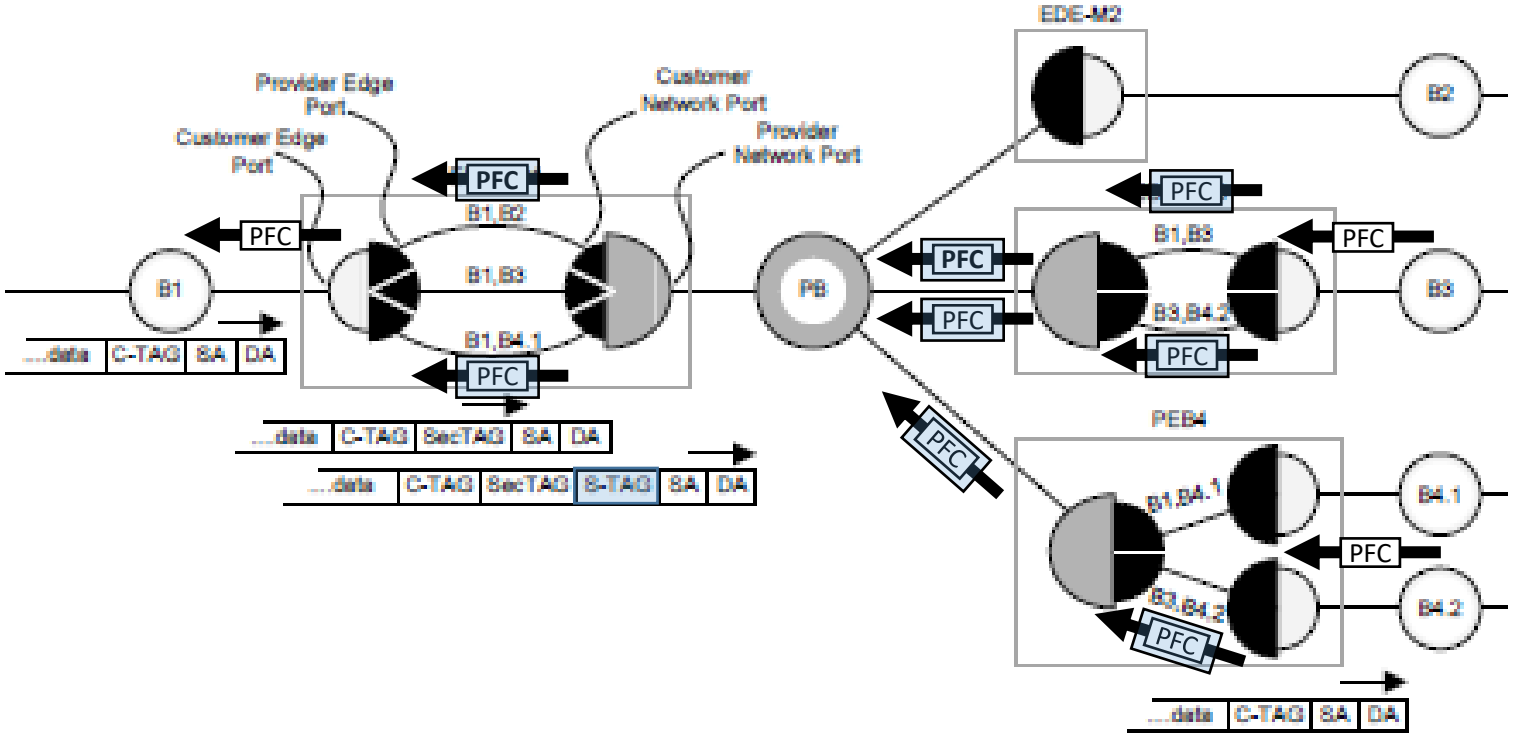


Figure 15-6—Example network with an EDE-CS

Imagine similar architecture (with or without MACSec), but encapsulating PFC

Challenges for PFC over PBNs

- Requires a new destination address for PFC frames
  - Allow forwarding through S-Components and TPMRs
- PBN can support multiple services per-port
  - Who should receive PFC frames when ingress buffer fills at the customer edge port?
  - PFC is not service specific, it is per-port
  - Increased congestion spreading if multiple services are ‘paused’
- Calculation of headroom is impossible
  - PFC frames would be subject to queuing delay
  - Lossless mode of operation can not be guaranteed

Recommendation: Do not support forwarding of PFC frames over PBN – as is the case today. Rely on other congestion control mechanisms.

Proposed Technical Solution:

Reuse PTP Measurement Procedure

- PTP/802.1AS supports peer-to-peer delay link measurement
- IEEE P802.3cx improves PTP timestamping accuracy
- The procedure can be reused in PFC headroom delay measurement

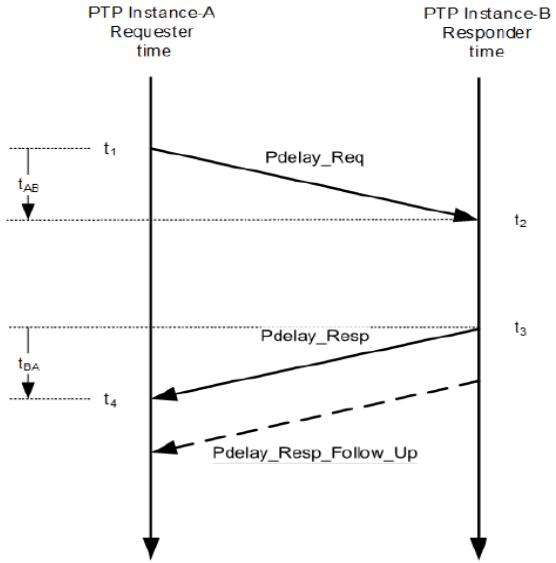
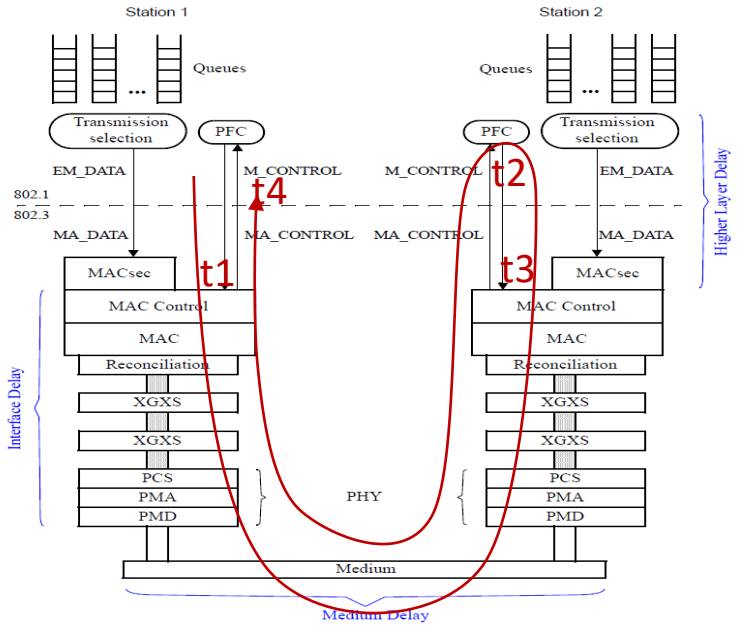


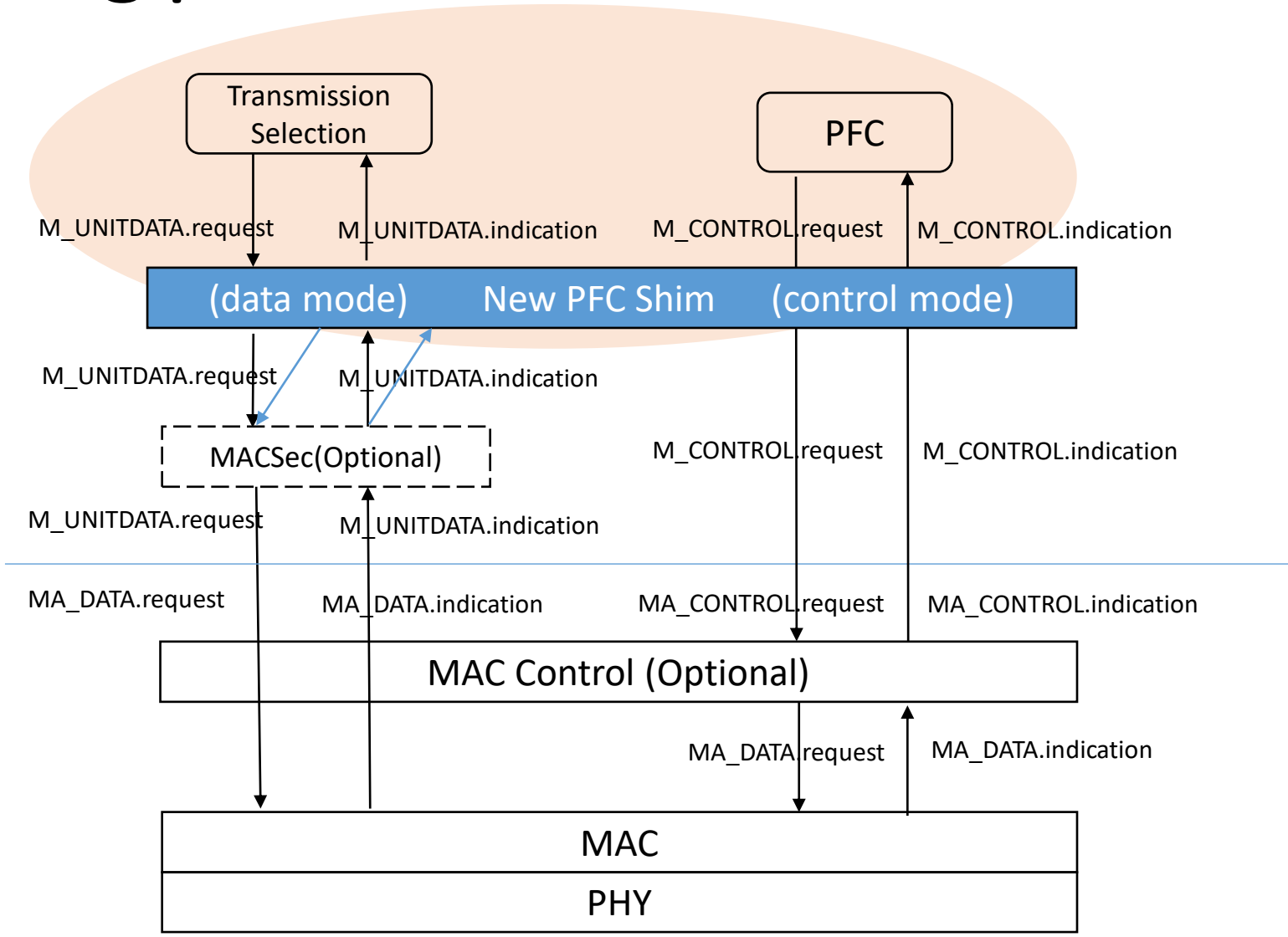
Figure 42—Peer-to-peer delay link measurement



- Proposed solution for PFC delay measurement:
  - Reuse PTP protocol to measure link delay
  - Define separate mechanism (using LLDP) to convey peer node(far-end) internal processing delay.
  - Sum up link delay, peer node processing delay and near-end processing delay to calculate PFC headroom.

New PFC shim enabling protected PFC frames

- Minimal (or no) impact to current PFC implementations
- Shim passes through existing MAC Control interface in ‘control mode’ (with no delay)
- Shim configured to generate and consume PFC frames if ‘data mode' is desired
- Internal delay calculation depends on Shim configuration
  - NOTE: MACsec delay is bounded and can be small and fixed.



Proposed Scope of Work

- Amendment to 802.1Q with limited changes are needed to support the PFC configuration mechanism and address errors and omissions
  - Update DCBX to discover the capability and auto-enable the feature. Address MBC inconsistency
  - LLDP + Pdelay procedure to measure delay
  - State machines and protocol description
  - Updates to DCBX MIBs and YANG
  - Enhanced descriptions in Annex M & N
  - Define PFC shim layer in 802.1 Q to allow MACSec protection of PFC frames.
  - Document the PFC propagation model as opposed to allowing PFC frames to be ‘forwarded’ transparently (e.g. through a PBN).

See: https://www.ieee802.org/1/files/public/docs2021/new-congdon-a-pfc-h-Q-changes-0521-v01.pdf
https://mentor.ieee.org/802.1/dcn/21/1-21-0048-00-ICne-pfc-headroom-with-macsec.pdf

List of likely impacted 802.1Q clauses

	Effort Estimation
• 1.3 Introduction	Small
• 5.4.1.7 DCBX Bridge requirements	Small
• 5.11 System requirements for Priority-based Flow Control (PFC)	Small
• 6.7.1 Support of the ISS by IEEE Std 802.3 (Ethernet)	Small
• 36. Priority-based Flow Control (PFC)	Large
• 36.1.3.3 Timing considerations (MACSec Bypass Control (MBC))	Small
• 38. Data Center Bridging eXchange protocol (DCBX)	Small
• D.2.10 Priority-based Flow Control Configuration TLV	Medium
• D.5 IEEE 802.1/LLDP extension MIB	Small
• D.6 IEEE 802.1/LLDP extension YANG	Small
• Annex M - Support for PFC in link layers without MAC Control	Small
• Annex N - Buffer requirements for PFC	Medium

Proposed Next Step

- Contributions to consider content for PAR and CSD (shared as a contribution within Nendica and/or 802.1WG/TSN)
- Provide a motion to develop a PAR and CSD for pre-circulation at the November plenary.

Backup

802.1Q MACSec Bypass Control is inconsistent

- While Figure 36-1 clearly indicates PFC frames are not encrypted
- 36.1.3.3 introduces MACSec Bypass Control (MBC)
- The DCBX Priority-based Flow Control Configuration TLV includes MBC as defined in D.2.10.4.
  - “The MACsec Bypass Capability Bit. If set to zero, the sending station is capable of bypassing MACsec processing when MACsec is disabled. If set to one, the sending station is not capable of bypassing MACsec processing when MACsec is disabled (see Clause 36).”

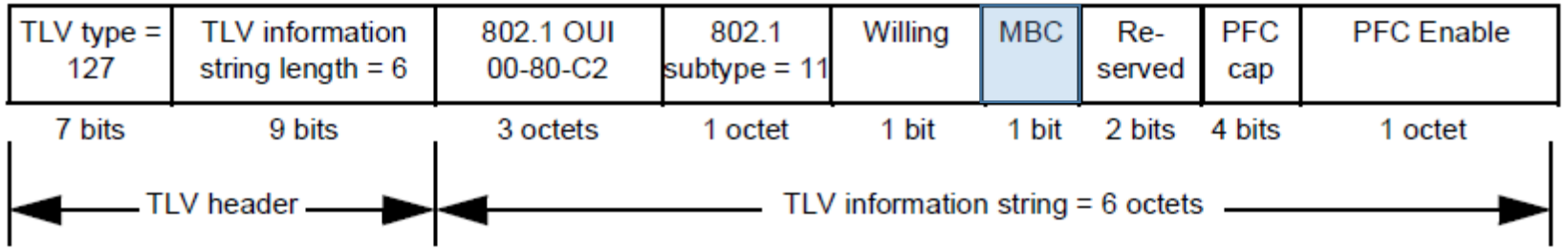


Figure D-10—Priority-based Flow Control Configuration TLV format

NOTE: IEEE Std 802.1AE has no concept of “MACSec Bypass”. Unprotected frames could use Uncontrolled Port