PFC Enhancements Project Proposal

Paul Congdon, Lily Lv (Huawei)

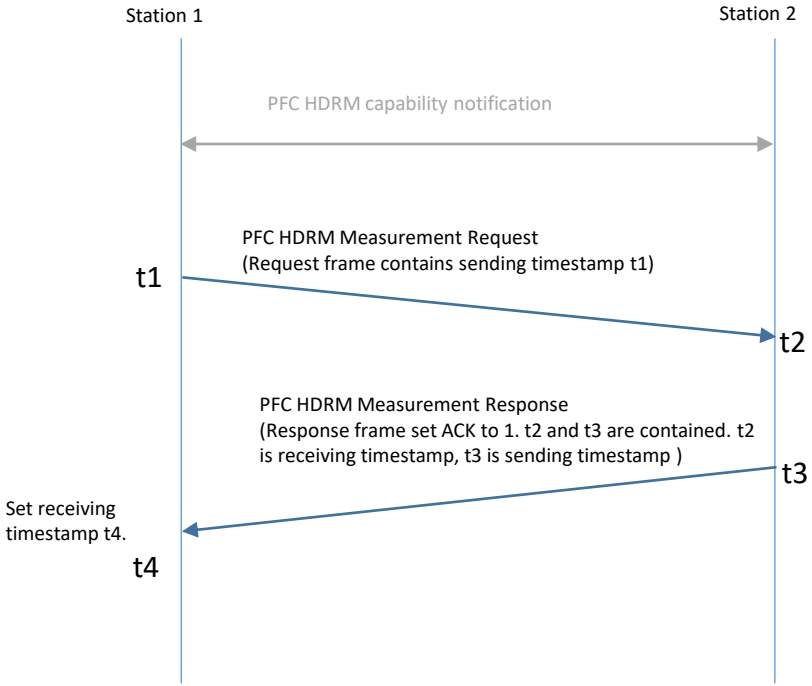
Mick Seaman (Independent)

Background

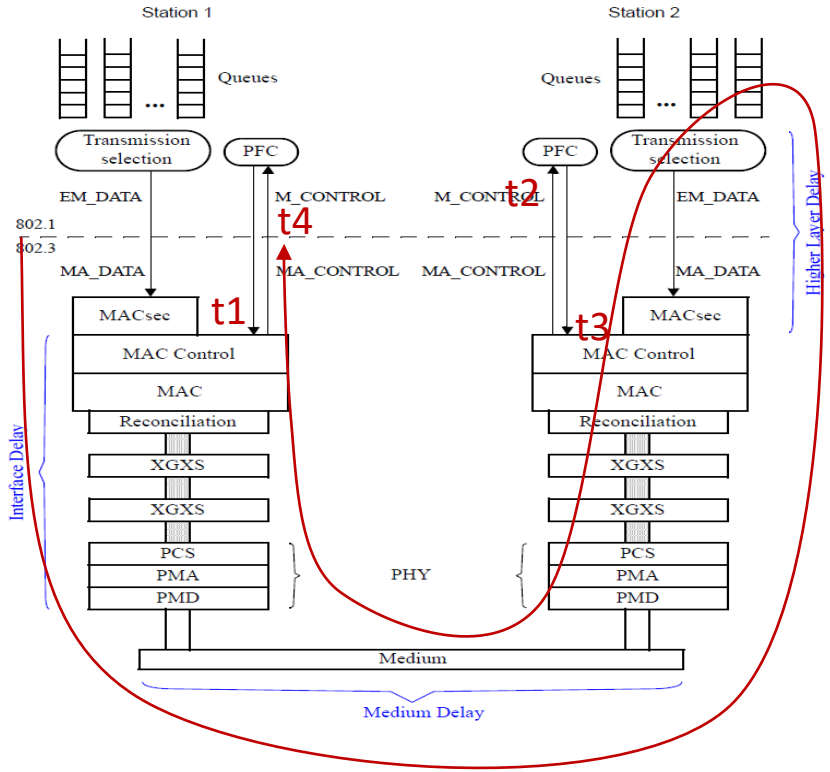
- Adaptive PFC headroom contribution proposes a new mechanism to automatically determine the amount of memory needed for PFC headroom – Lots of previous presentations.
  - https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-0121-v02.pdf Adaptive PFC Headroom
  - https://www.ieee802.org/1/files/public/docs2021/new-congdon-a-pfc-h-Q-changes-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
- 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.
- Additionally, there are unclarities and inconsistencies with respect to PFC and MACSec joint operation
- Additionally, clarify issues related to ‘forwarding’ PFC frames (e.g. across PBNs)

Background

- Proposed mechanism of adaptive 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

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

Need for The Project

- 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 usage (e.g MACSec operation, issues related to ‘forwarding’ PFC frames)

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

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

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

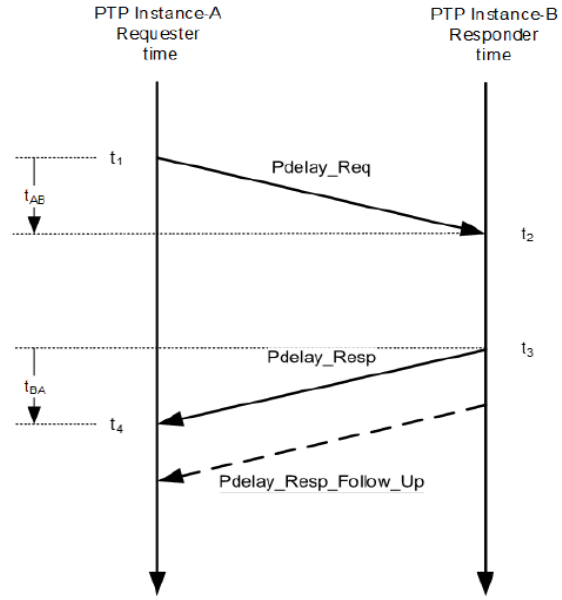
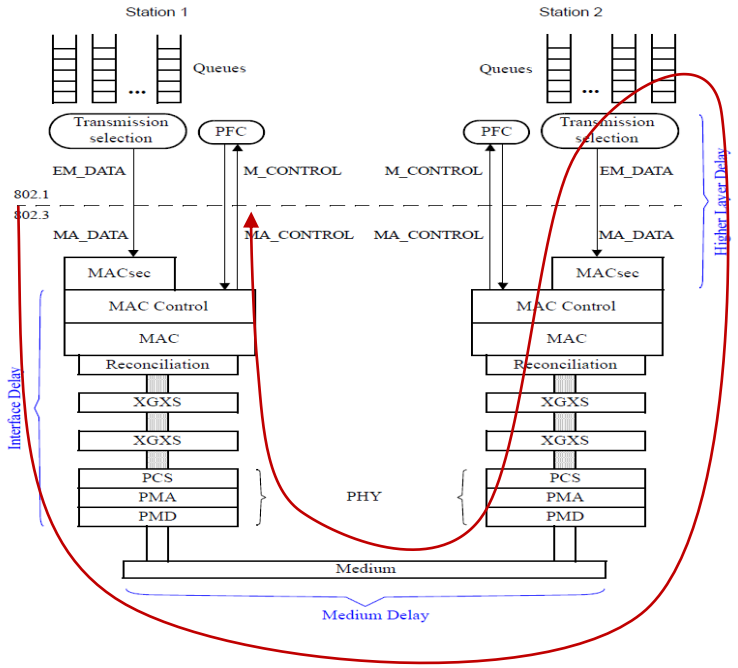


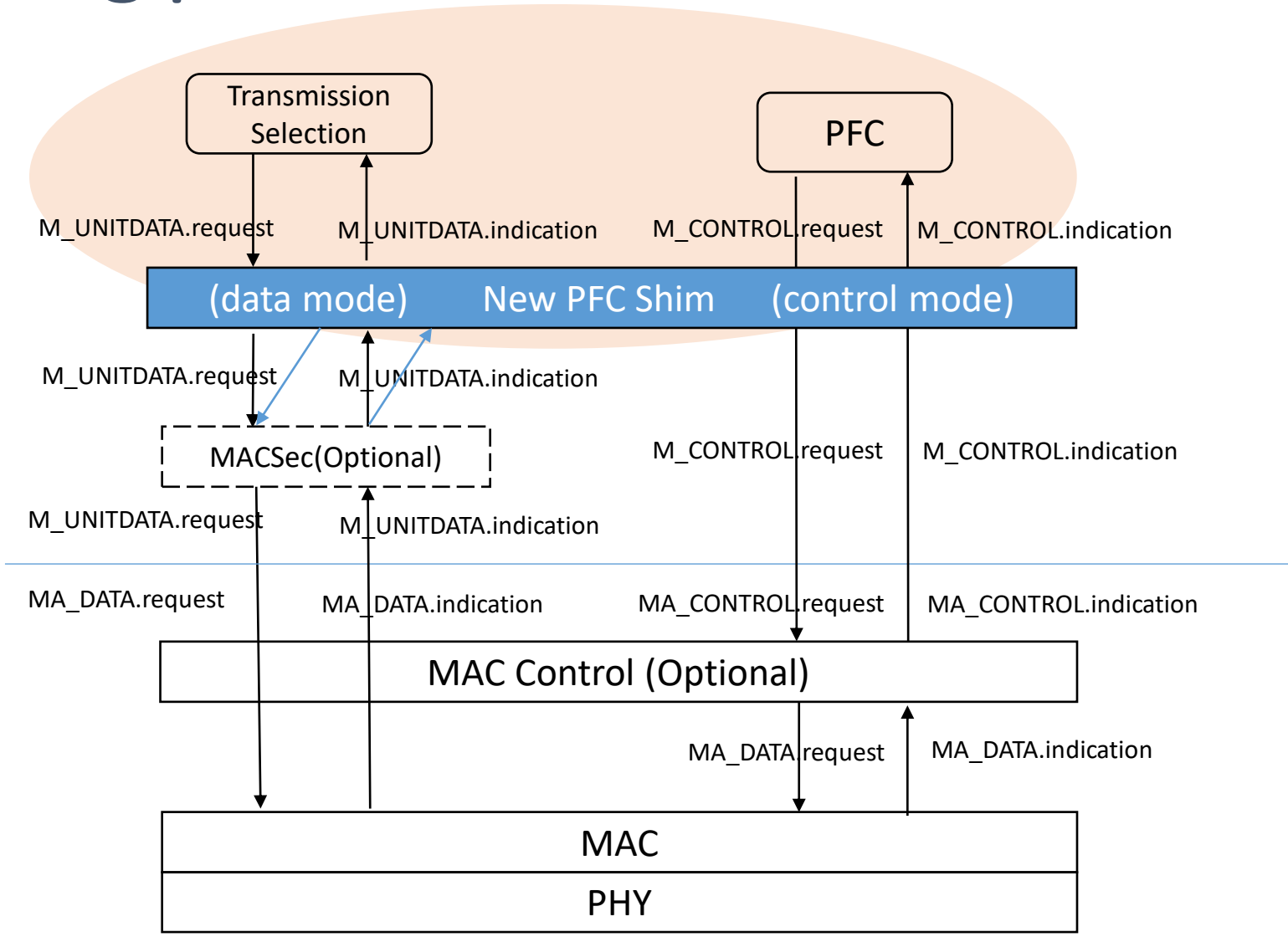
Figure 42—Peer-to-peer delay link measurement



- 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



Market Potential

- The data center market continues to grow very fast:
  - New high-performance applications (AI/ML).
  - Desire to converge high-performance computing and high-performance storage on Ethernet.
  - Desire to scale HPC networks as a Cloud Service.
  - A consolidated network saves operational and equipment costs, fueling more growth.
- RDMA over Converged Ethernet (RoCEv2) is widely used:
  - Requires low latency.
  - Requires lossless operation to avoid retransmission.
  - Is deployed within data centers and across data center interconnections.
  - Data center interconnects require PFC frame protection using MACSec
- Automating the configuration of PFC makes Ethernet technology more applicable for data center environments.

Technical Feasibility

- Proposed mechanism includes 2 major parts
  - Capability and configuration notification
  - PFC delay measurement
- Capability and configuration notification can be supported by extension of DCBX
- PFC delay measurement considers roundtrip delay between participating systems, which can be based on PTP peer-to-peer delay measurement mechanism.
- Both DCBX and PTP are mature technologies, which are currently available in production.

Proposed Next Step

- Straw poll in Nendica to propose a project within IEEE 802.1WG
- Continue drafting proposed text for PAR and CSD (shared as a contribution within Nendica)
- Provide a motion to develop a PAR and CSD for pre-circulation at the November plenary.