P802.1Qdt Status Update and New Proposal Discussion

Lily Lv

Paul Congdon

Mick Seaman

James McIntosh

P802.1Qdt Status Update

Draft D0.2 is Available

- **Qdt draft has started in the Security TG**
  - Qdt includes 2 new functions:
    - Automatic PFC headroom measurement
    - MACsec protection on PFC frames
  - P802.1 Qdt draft 0.2 is available:

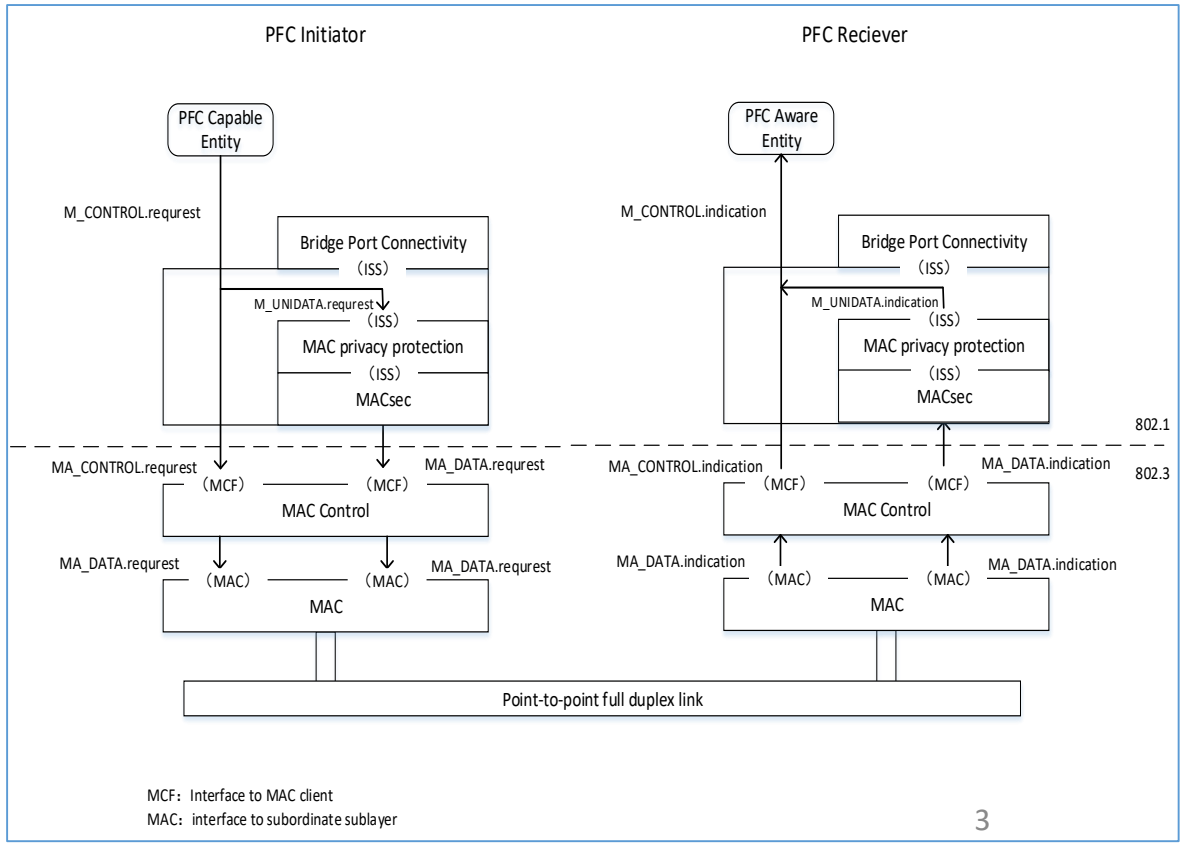
    https://www.ieee802.org/1/files/private/dt-drafts/d0/802-1Qdt-d0-2.pdf
- **D0.2 builds the overall framework of specification**
  - It specifies new functions in Clause 36 - Priority-based Flow Control (PFC)
  - It updates PFC management, covering Clause 12 - Bridge management, Clause 48 - YANG Data Models, Annex D - IEEE 802.1 Organizationally Specific TLVs
  - It updates other relevant clauses, e.g. Clause 1 - Overview, Clause 3 – Definitions, etc.

PFC headroom

Delay Value = 2 x (Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1 + 2 x (Max Frame) + (PFC Frame)

Medium delay | Internal processing delay | Fixed delay

MACsec protection



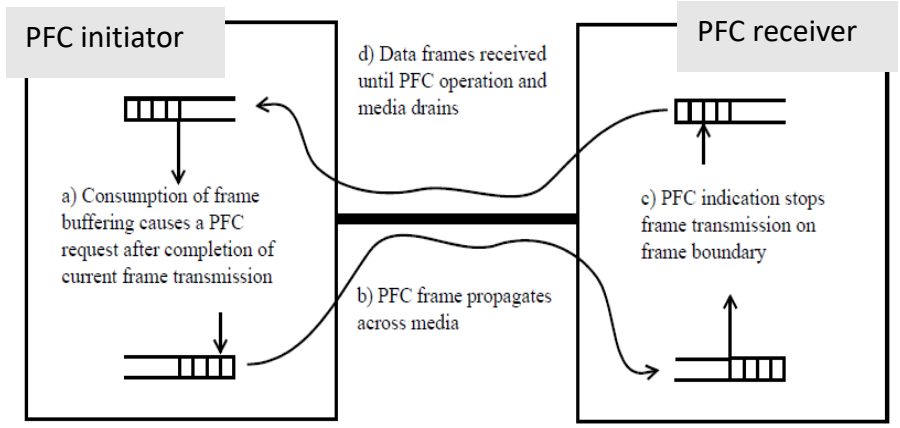
Re-thinking PFC Headroom Measurement

- The current approach for PFC headroom measurement aroused much discussion in the Security TG (see subsequent slides).
  - The conclusion is to propose a new way to obtain the PFC headroom measurement instead of re-using PTP.
  - Present the new proposal in TSN for broader discussion.

PFC Headroom Measurement Discussion

Take Another Look at PFC Headroom

What is actually required for PFC headroom is knowledge of the maximum amount of data that will be received after the decision to send a PFC has been made at PFC initiator.



This includes

- Data that the PFC receiver has already transmitted at that decision time
- Data sent during the time from decision time to PFC transmit including PFC transmit delay, PFC recognition and processing by PFC receiver, and time to halt transmission at PFC receiver.

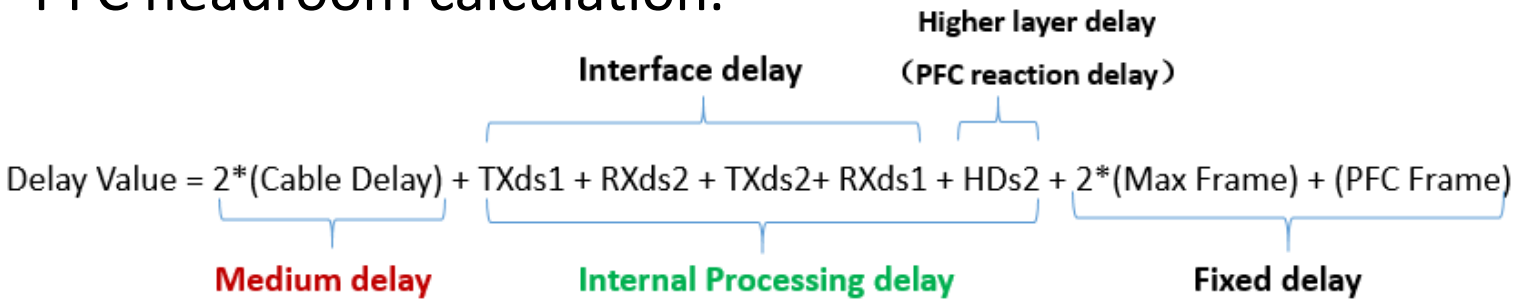
Delay Value = 2*(Cable Delay) + TXds1 + RXds2 + TXds2+ RXds1 + HDs2 + 2*(Max Frame) + (PFC Frame)

This is effectively the "**round trip** " from the PFC initiator's PFC transmitting back to the PFC initiator's reception of data controlled by the PFC.

It is actually independent of delay symmetry on the physical communication medium.

Current Method of PFC Headroom Measurement

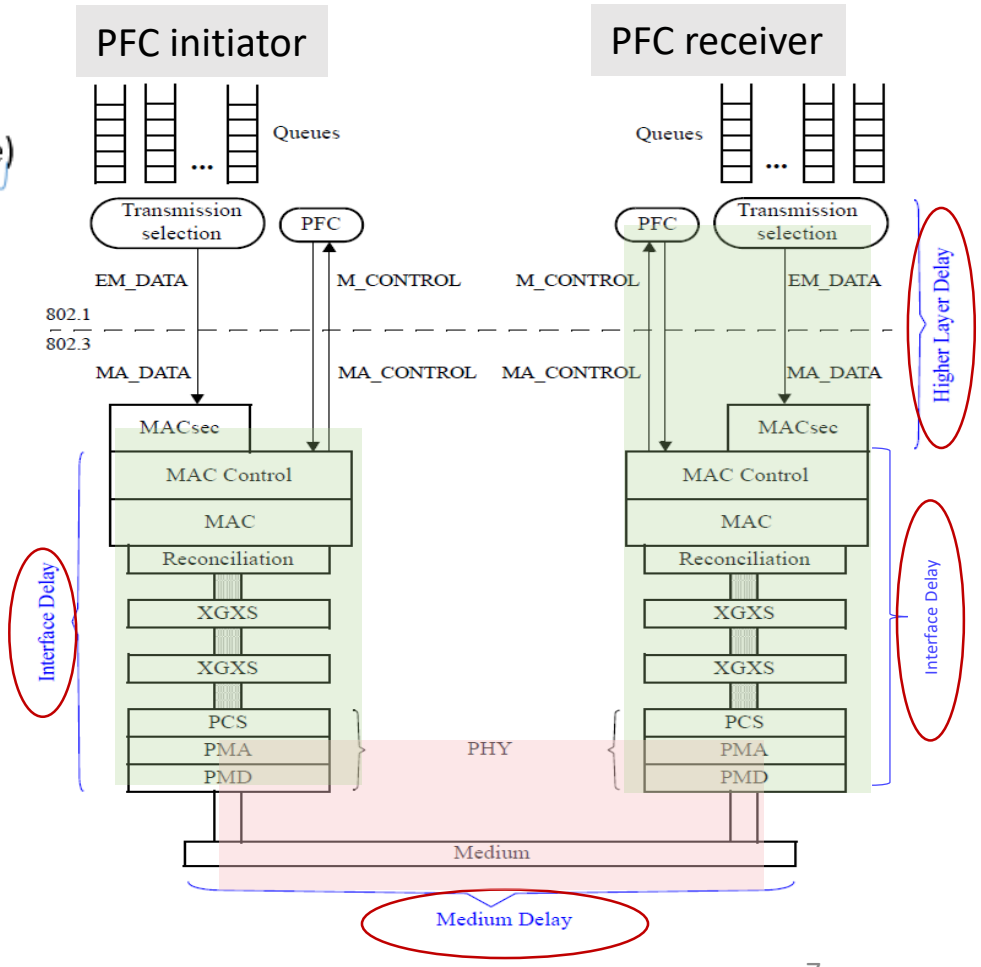
Use the existing PTP protocol and enhance the existing DCBX protocol to support automated PFC headroom calculation.



Cut the "round-trip" into medium delay and internal processing delay.

- **Medium delay**
  - Reuse PTP protocol to measure round-trip link delay
- **Internal Processing delay (implementation known value shared by DCBX)**
  - Both PFC initiator and PFC receiver know its own internal processing delay (interface delay + higher layer delay).
  - Define separate mechanism using DCBX to convey PFC receiver internal processing delay to PFC initiator.
  - PFC initiator calculates the total internal processing delay.

Finally, sum up medium delay, internal processing delay and fixed delay to get "round trip" delay as PFC headroom.



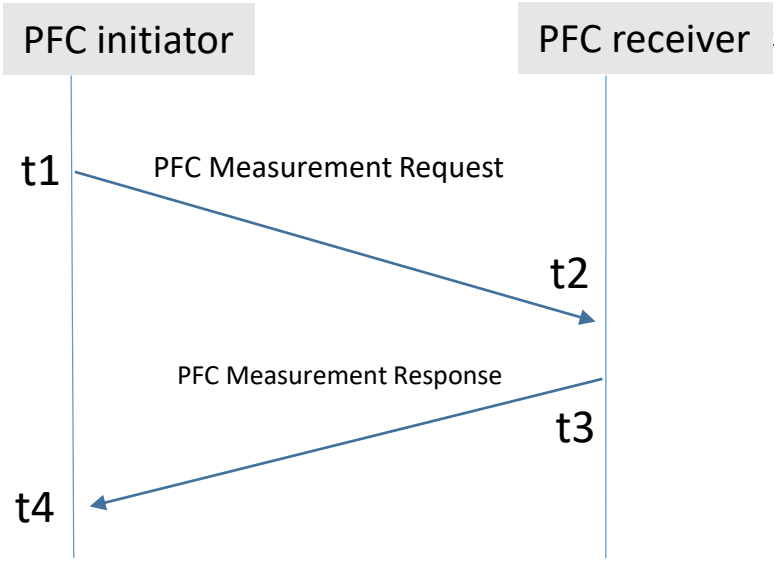
Concerns with the Current Method

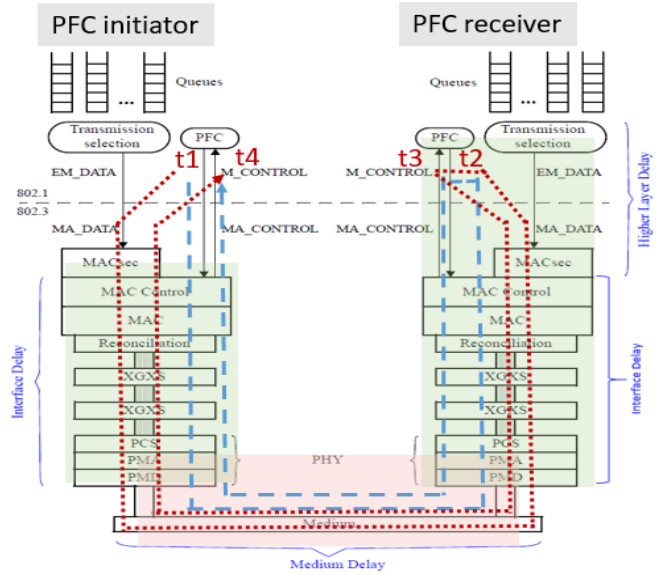
Requiring PTP for PFC Headroom measurement increases implementation/standard development complexity.

- PTP objectives and PFC Headroom measurement objectives are not aligned.
  - PTP was not designed to support PFC. It intends to precisely measure point to point cable delay. There are discussion on timestamp point, trying to make it as close to medium as possible.
  - However, PFC headroom measurement includes roundtrip delay from the point above the MAC to another point above the MAC on the other end.
  - Re-using PTP adds implementation restrictions/limitations to DC switches.
    - Hardware is required to be capable to get timestamp at lower point.
- 1588 would need to be a normative reference in .1Q.
  - In most cases, 1588 is not required/supported in DC environments.
  - Qdt only needs part of 1588.
    - Inventing subset of 1588 is not desirable.
    - Qdt becomes dependent on 1588 and must track updates.

New PFC Headroom Measurement proposal (1/3)

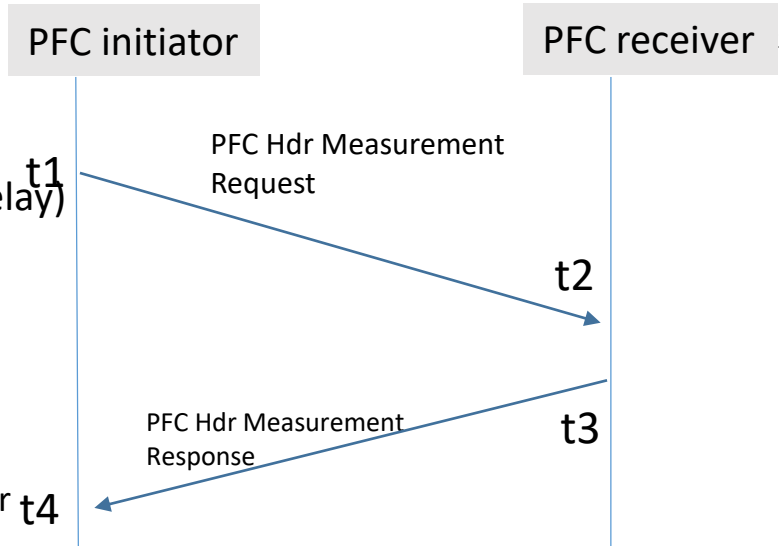
- Specify a new request-response measurement to measure the “round trip delay”.
  - Measure t1 (the timestamp of sending request) and t4 (the timestamp of receiving response)
    - Both t1 and t4 are timestamps above MAC on PFC initiator.
    - No strict requirement on the hardware.
  - DV = t4-t1
    - DV = t4 – t1 – (t3 – t2) + HD ≈ t4 – t1
      - t3-t2 is the time to generate PFC headroom measurement Response.
      - HD is Higher layer delay (PFC reaction delay).
      - (t3 – t2) is similar as HD. But if the implementation is different, it should be accounted for. The response message can carry the compensation value, or DCBX can carry it.
  - Design both request and response measurement frames as MAC data frames
    - If they are designed as MAC control frame, 802.3 would need to be involved. Better to minimize other standards dependency/involvement.

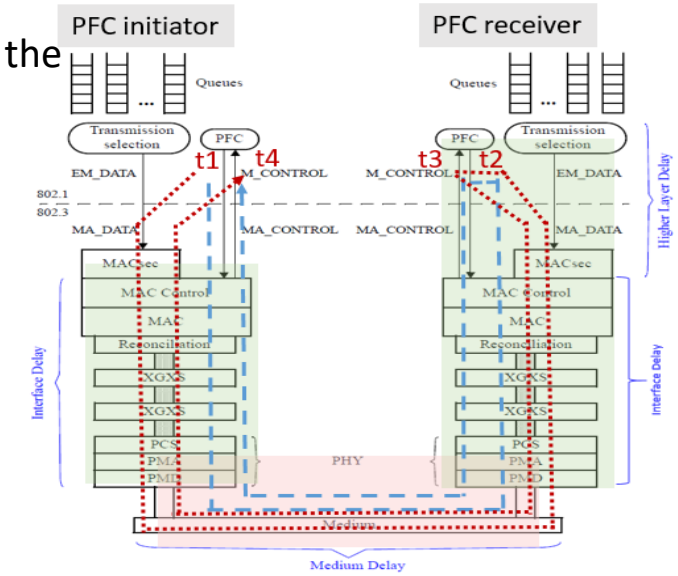




New PFC Headroom Measurement proposal (2/3)

- Accuracy analysis
  - t4-t1 accuracy
    - Assuming 100 meters, t4-t1 is on the micro-second level (1us + internal processing delay)
    - Clock measuring t4-t1 local to initiator independent of other clocks in the systems. Inaccuracy within tens of nano seconds will have little impact on PFC headroom measurement.
      - 1ns inaccuracy (100Gbps) leads to 100 bits mismatch
      - In practice, buffers to store the packet are usually allocated in chunks (e.g. 160 byte). Buffer chunk size could accommodate the inaccuracy.
    - Any t3-t2 adjustment is similarly local to PFC receiver (independent of other clocks in the systems).
    - PFC initiator can adjust for its own different reference points if necessary.

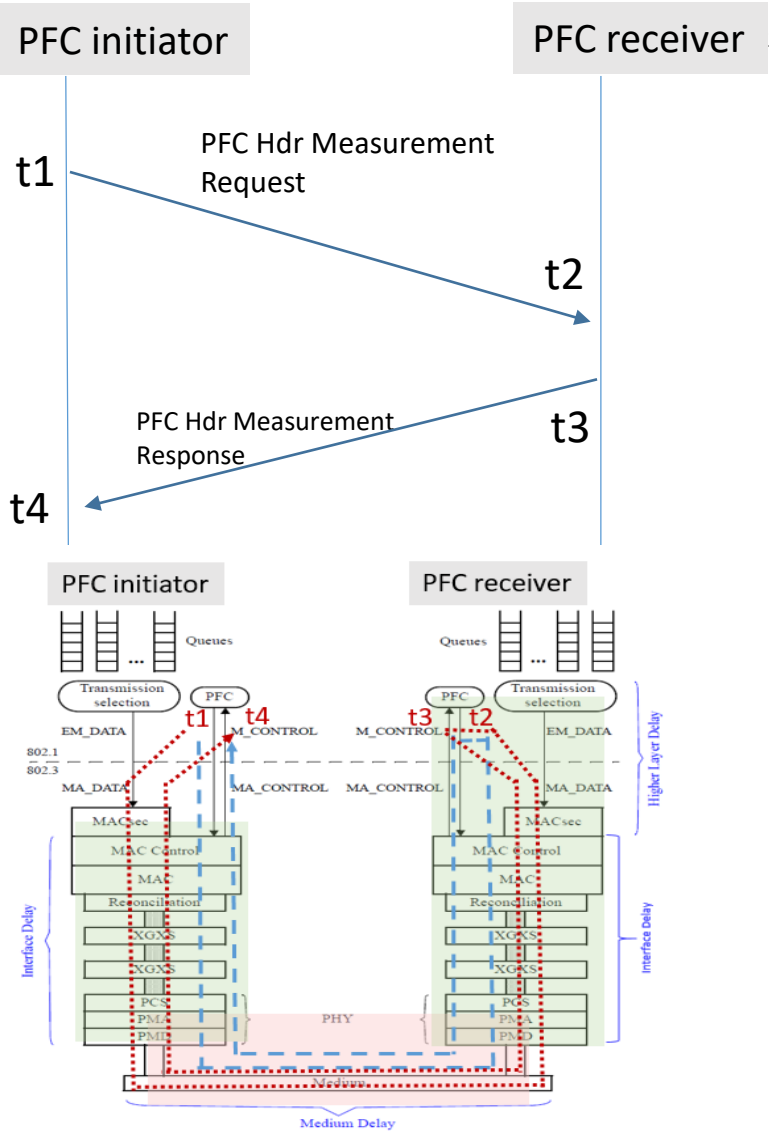




	Fixed Delay	Internal Processing Delay（802.3, no MACsec）	Medium Delay	Headroom（t4-t1）	t4-t1 mismatch		
100G,500m	32992	203776	500000	92KB	10 ns	0.125KB	0.1%
					100 ns	1.25KB	1%
100G,100m	32992	203776	100000	42KB	10 ns	0.125KB	0.3%
					100 ns	1.25KB	3%
100G,20m	32992	203776	20000	32KB	10 ns	0.125KB	0.4%
					100 ns	1.25KB	4%

New PFC Headroom Measurement proposal (3/3)

- Accuracy analysis
  - MAC control frame processing delay vs. MAC data frame processing
    - Traditional PFC is MAC control frame. The internal path of MAC control frame is different from MAC data frame. But the processing time difference is likely trivial.
    - MAC data frame measurement is more conservative than a MAC control frame measurement, but the difference will not waste unnecessary buffer memory.
    - PFC initiator can set its own adjustment parameter to accommodate the difference.
    - PFC receiver can convey the adjustment parameter to the PFC initiator via DCBX.



Benefits of the New Proposal

- The "round trip delay" is a better measure of the headroom as it avoids the need to add in a worst-case estimate of the internal transmission and processing delays in the systems.
- The "round trip delay” measurement would naturally account for any additional internal delays, such as the use of MACsec to protect the PFC frames themselves
- Benefits for standards development:
  - .1Q does not need to refer to 1588 as a normative reference.
  - .1Qdt has the flexibility to progress independently of other standards.
- Benefits for implementation:
  - No strict requirements on hardware.
  - Data center switches do not need to comply with 1588 or decide which relevant parts of the standard to use.

New Proposal Impact On Qdt PAR & CSD

- 'PTP' mentioned in current PAR

5.2.b **Scope of the project:** This amendment specifies procedures and managed objects for automated Priority-based Flow Control (PFC) headroom calculation and Media Access Control Security (MACsec) protection of PFC frames, using the existing Precision Time Protocol (PTP) and enhancements to the Data Center Bridging Capability Exchange protocol (DCBX).

- 'PTP' mentioned in current CSD

1.2.4 Technical Feasibility

b) Proven similar technology via testing, modeling, simulation, etc.

The proposed project enables peer nodes to advertise the new capability through the Data Center Bridging Capability Exchange (DCBX, specified in IEEE Std 802.1Q) mechanism which is widely deployed today using "Link Layer Discovery Protocol (LLDP, specified in IEEE Std 802.1AB). Roundtrip delay measurements for participating systems are based on the existing Precision Time Protocol (PTP, specified in IEEE Std 1588) delay measurement mechanism.

1.2.5 Economic Feasibility

c) Consideration of installation costs.

A modest reduction in installation cost of new equipment is expected.

There are no incremental installation costs relative to the existing PTP and DCBX that will be used by the proposed standard.

PAR&CSD need to be updated if new proposal is adopted.

Summary & Next Steps

Summary & Next steps

- .1Qdt draft work has started.
- The development of standard draft is limited until a PFC headroom measurement method can be decided.
- A new method of PFC headroom measurement method has been proposed by the Security TG.
- Modify PAR&CSD if new proposal is adopted.
- Question: Shall we produce a new draft with the new method, or do we need further discussion?

Backup Slides

	Fixed Delay	Internal Processing Delay（802.3, no MACsec）	Medium Delay	Headroom（t4-t1）	t4-t1 mismatch		
100G,500m	32992	203776	500000	92KB	10 ns	0.125KB	0.1%
					100 ns	1.25KB	1%
100G,100m	32992	203776	100000	42KB	10 ns	0.125KB	0.3%
					100 ns	1.25KB	3%
100G,20m	32992	203776	20000	32KB	10 ns	0.125KB	0.4%
					100 ns	1.25KB	4%

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