Consideration of Adaptive PFC Headroom in 802.1Q

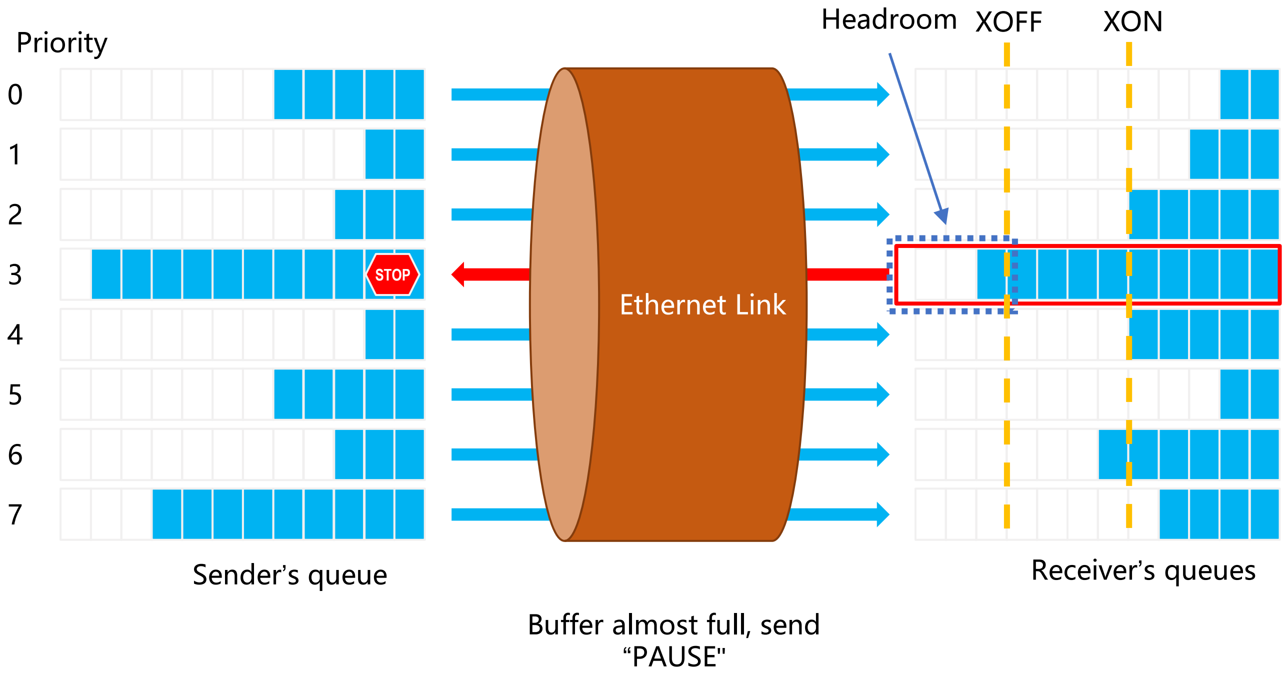
Paul Congdon

Outline

- Brief review of proposal
- Proposed scope of work
- Considerations on 802.1Q

Adaptive PFC Headroom Calculation

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



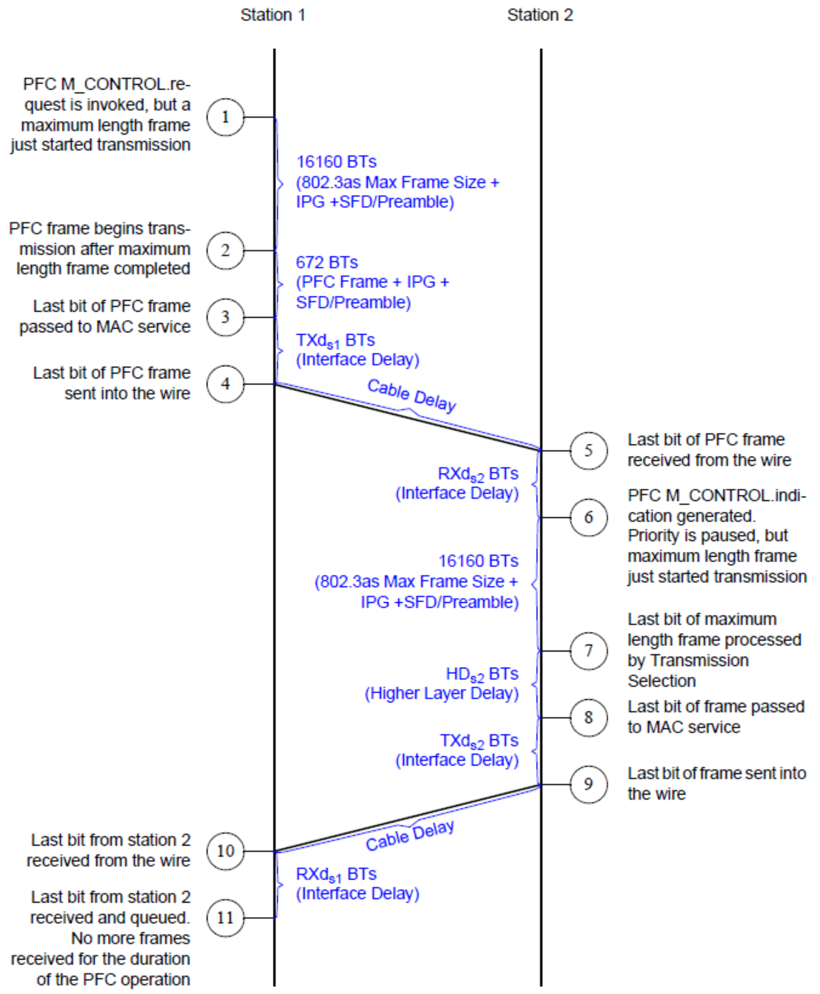
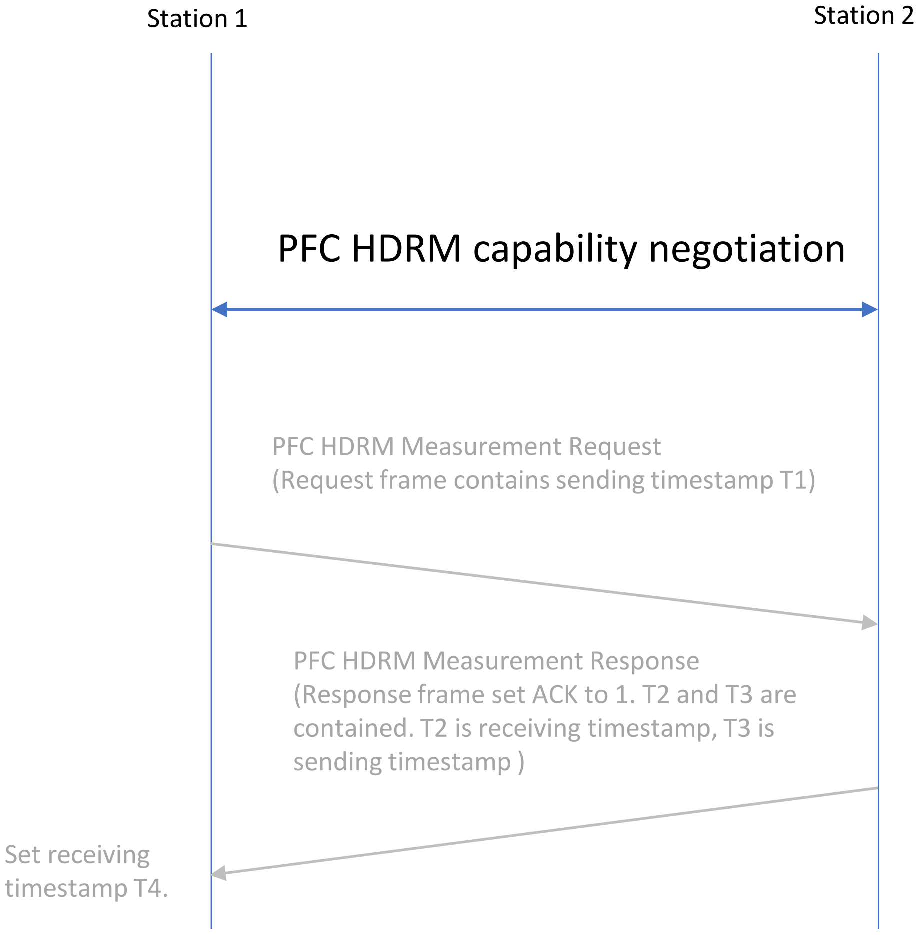

Station 1
Station 2
1
PFC M_CONTROL.request is invoked, but a maximum length frame just started transmission
16160 BTs (802.3as Max Frame Size + IPG +SFD/Preamble)
2
PFC frame begins transmission after maximum length frame completed
672 BTs (PFC Frame + IPG + SFD/Preamble)
3
Last bit of PFC frame passed to MAC service
TXd$_{s1}$ BTs (Interface Delay)
4
Last bit of PFC frame sent into the wire
Cable Delay
5
Last bit of PFC frame received from the wire
RXd$_{s2}$ BTs (Interface Delay)
6
PFC M_CONTROL.indication generated. Priority is paused, but maximum length frame just started transmission
16160 BTs (802.3as Max Frame Size + IPG +SFD/Preamble)
7
Last bit of maximum length frame processed by Transmission Selection
HD$_{s2}$ BTs (Higher Layer Delay)
8
Last bit of frame passed to MAC service
TXd$_{s2}$ BTs (Interface Delay)
9
Last bit of frame sent into the wire
Cable Delay
10
Last bit from station 2 received from the wire
RXd$_{s1}$ BTs (Interface Delay)
11
Last bit from station 2 received and queued. No more frames received for the duration of the PFC operation

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

Proposal for Adaptive PFC Headroom



Station 1

Station 2

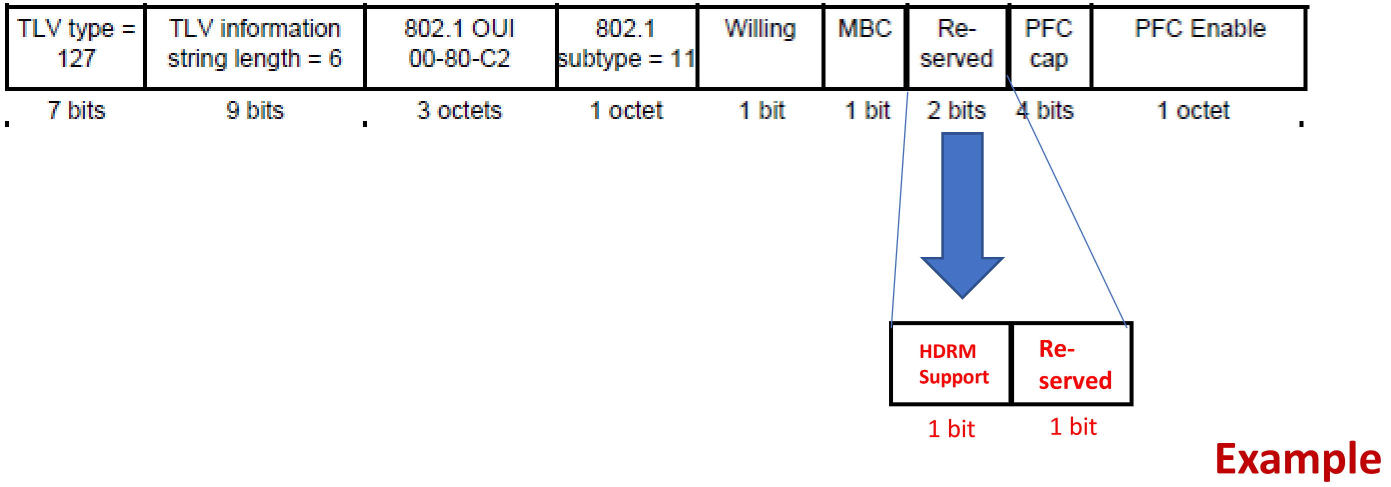
PFC HDRM capability negotiation

PFC HDRM Measurement Request
(Request frame contains sending timestamp T1)

PFC HDRM Measurement Response
(Response frame set ACK to 1. T2 and T3 are contained. T2 is receiving timestamp, T3 is sending timestamp)

Set receiving timestamp T4.

- Phase 1: 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.

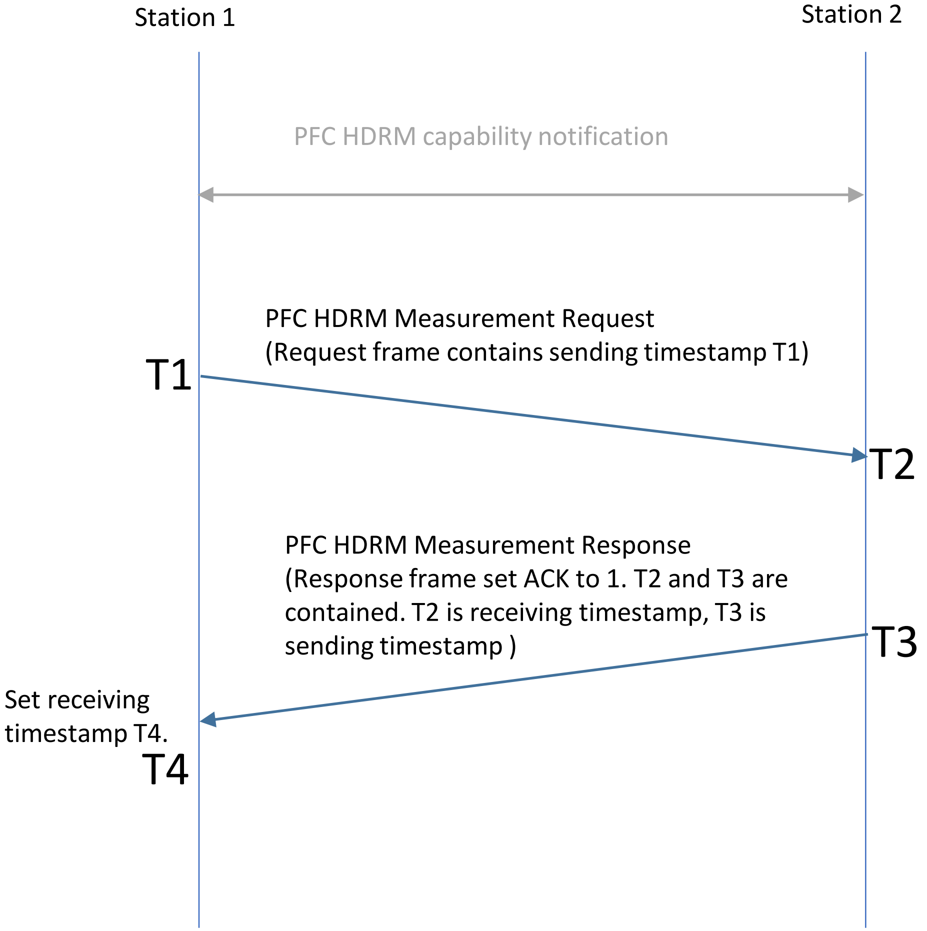


TLV type = 127	TLV information string length = 6	802.1 OUI 00-80-C2	802.1 subtype = 11	Willing	MBC	Re-served	PFC cap	PFC Enable
7 bits	9 bits	3 octets	1 octet	1 bit	1 bit	2 bits	4 bits	1 octet

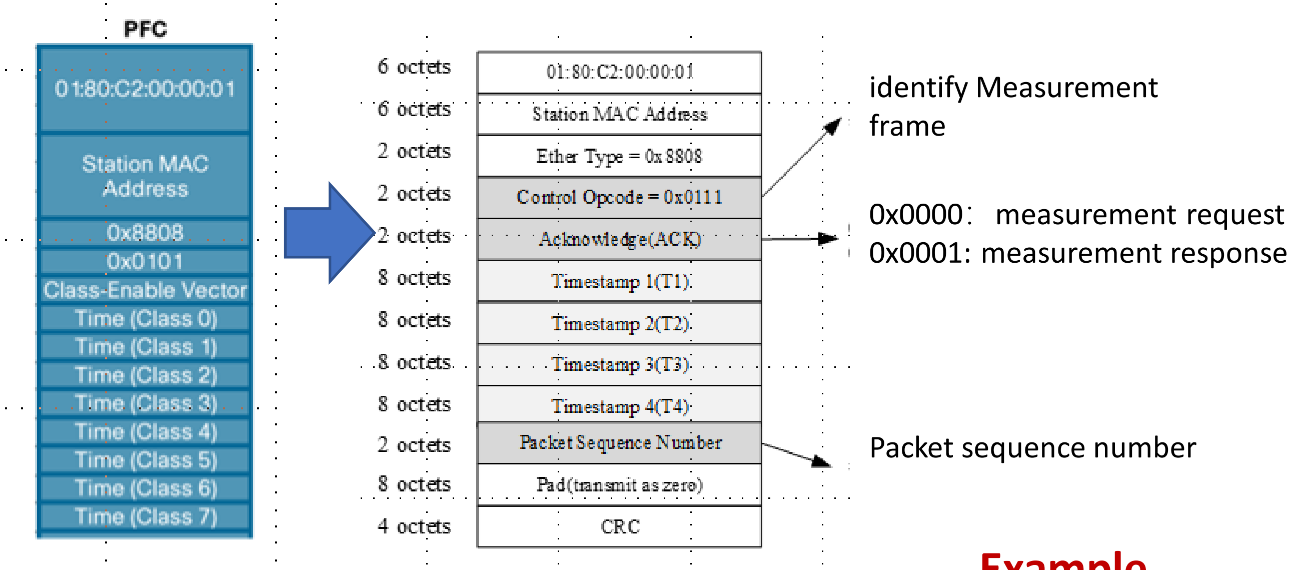
HDRM Support	Re-served
1 bit	1 bit

Example

Proposal for Adaptive PFC Headroom (2/4)


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

- Phase 2: Delay Measurement
  - Measurement request is sent from station 1 to station 2 with sending timestamp T1
  - Measurement response is sent from station 2 to station 1 with receiving timestamp T2 and sending timestamp T3
  - Station 1 set receiving timestamp T4
  - Measurement request and response frame is a new MAC control frame



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 = 0x8808
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

identify Measurement frame

0x0000: measurement request
0x0001: measurement response

Packet sequence number

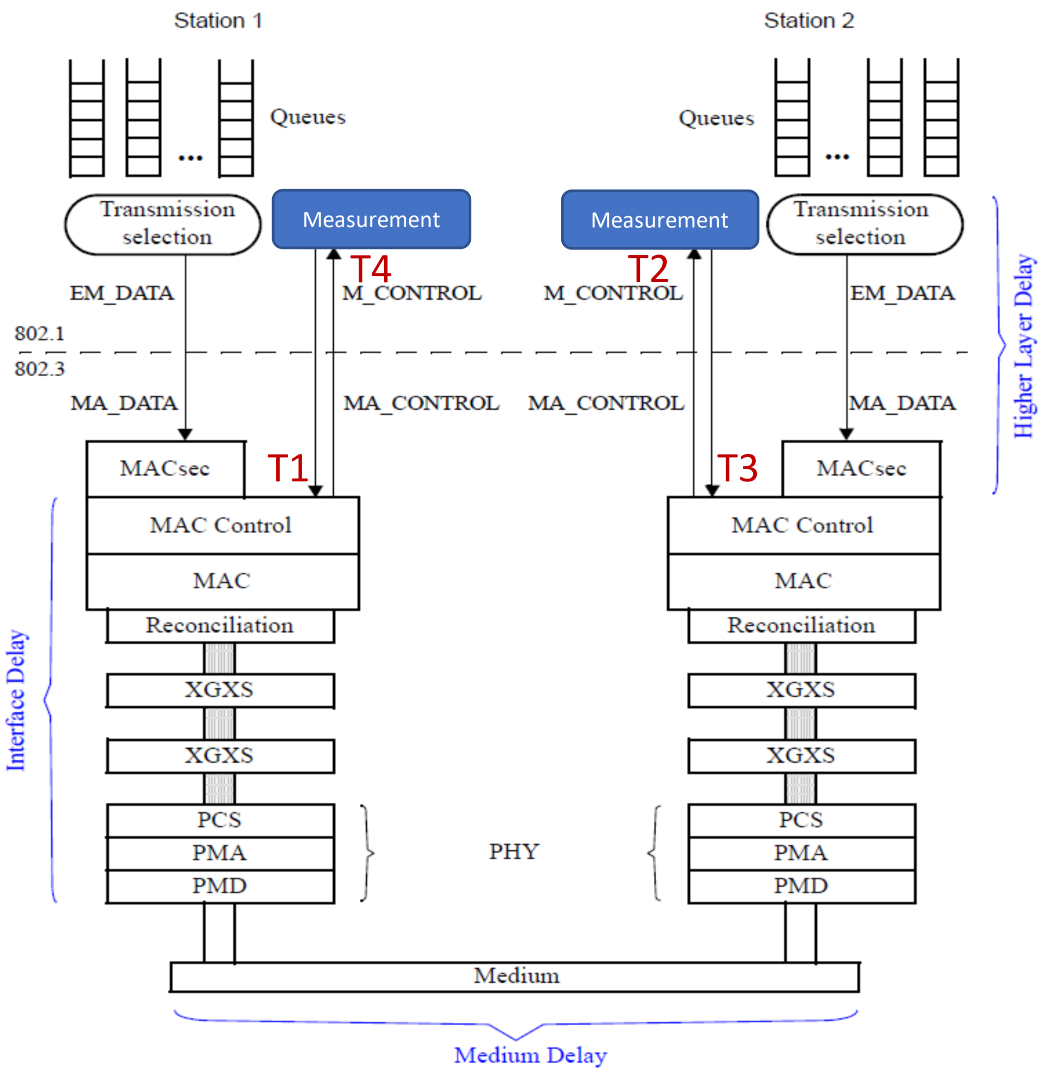
Example

Measurement frame format

Proposal for Adaptive PFC Headroom (4/4)

Example

- Phase 3: Headroom calculation
  - X = Port speed * (T4-T1-(T3-T2))
  - DV = X + 2*(Max Frame) + (PFC Frame)
  - Headroom = DV * alpha
    - alpha is implementation dependent, considering internal buffer chunk size



Scope of work summary

- Consider an amendment to 802.1Q that does the following:
  - specifies protocols, procedures and managed objects that support the automatic configuration of PFC buffer requirements.
- Specific functionality includes:
  - Update DCBX to discover the capability and auto-enable the feature
  - Specify timestamp points
  - New M_CONTROL primitive to measure delay and interworking with 802.3
  - State machines and protocol description
  - Updates to DCBX MIBs and YANG
  - Enhanced descriptions in Annex M & N

List of 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
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

Next Steps

- Discussion?
- What else is needed before asking for authorization to draft a PAR and CSD?

Backup

1.3 Introduction

- This standard specifies protocols, procedures, and managed objects to support Priority-based Flow Control (PFC). These allow a Virtual Bridged Network, or a portion thereof, to enable flow control per traffic class on IEEE 802 point-to-point full-duplex links. To this end, it

  bh) Defines a means for a system to inhibit transmission of data frames on certain priorities from the remote system on the link.

  bi) Defines a means for two participating systems to automatically calculate the minimum buffer requirements to assure lossless operation.

5.4.1.7 DCBX Bridge requirements

- A device supporting DCBX shall

  a) Support Link Layer Discovery Protocol (LLDP) transmit and receive mode (IEEE Std 802.1AB).

  b) Support the DCBX ETS Configuration Type, Length, Value (TLV) (D.2.8).

  c) Support the ETS Recommendation TLV (D.2.9).

  d) Support the Priority-based Flow Control Configuration TLV (D.2.10).

  e) Support the Application Priority TLV (D.2.11).

  f) Support the asymmetric and symmetric DCBX state machines (38.4).

  g) Support the Application VLAN TLV (D.2.14).

- A device supporting DCBX may

  a) Support automatic PFC buffer requirement configuration (x.x.x)

5.11 System requirements for Priority-based Flow Control (PFC)

- A system that conforms to the provisions of this standard for PFC may

  g) Support enabling PFC on up to eight priorities per port.

  h) Support the IEEE8021-PFC-MIB (17.7.17).

  i) Support automatic configuration of PFC buffer requirements for lossless operation.

6.7.1 Support of the ISS by IEEE Std 802.3 (Ethernet)

- Update description of mapping PFC M_CONTROL.requests to the MAC control interface associated with the express MAC (eMAC) to include the measurement M_CONTROL.request interface.

36. Priority-based Flow Control (PFC)

- Updates to the overview section to describe adaptive PFC headroom calculation
- New clause 36.3 to specify adaptive PFC headroom
  - Specify new M_CONTROL primitives
  - Specify protocol state machines
  - Include architectural diagrams for timestamps
  - Other considerations?

38. Data Center Bridging eXchange protocol (DCBX)

- Update 38.2 Goals to include buffer calculation for PFC
- Augmenting the PFC capability negotiation using Symmetric Attribute Passing

D.2.10 Priority-based Flow Control Configuration TLV

- Define one of the two ‘reserved’ bits as follows:
  - ABC capable – auto-buffer calculation capability is supported

- If a device is ABC capable and PFC is enabled on at least one traffic class, the measurement process and the automatic headroom calculation will be enabled.

D.5 IEEE 802.1/LLDP extension MIB
D.6 IEEE 802.1/LLDP extension YANG

- Update the 802.1 Extension MIB for the PFC TLV with new bits
- Update the 802.1 Extension YANG for the PFC TLV with new bits

Annex M - Support for PFC in link layers without MAC Control

- Updates for new primitive measurement PDU

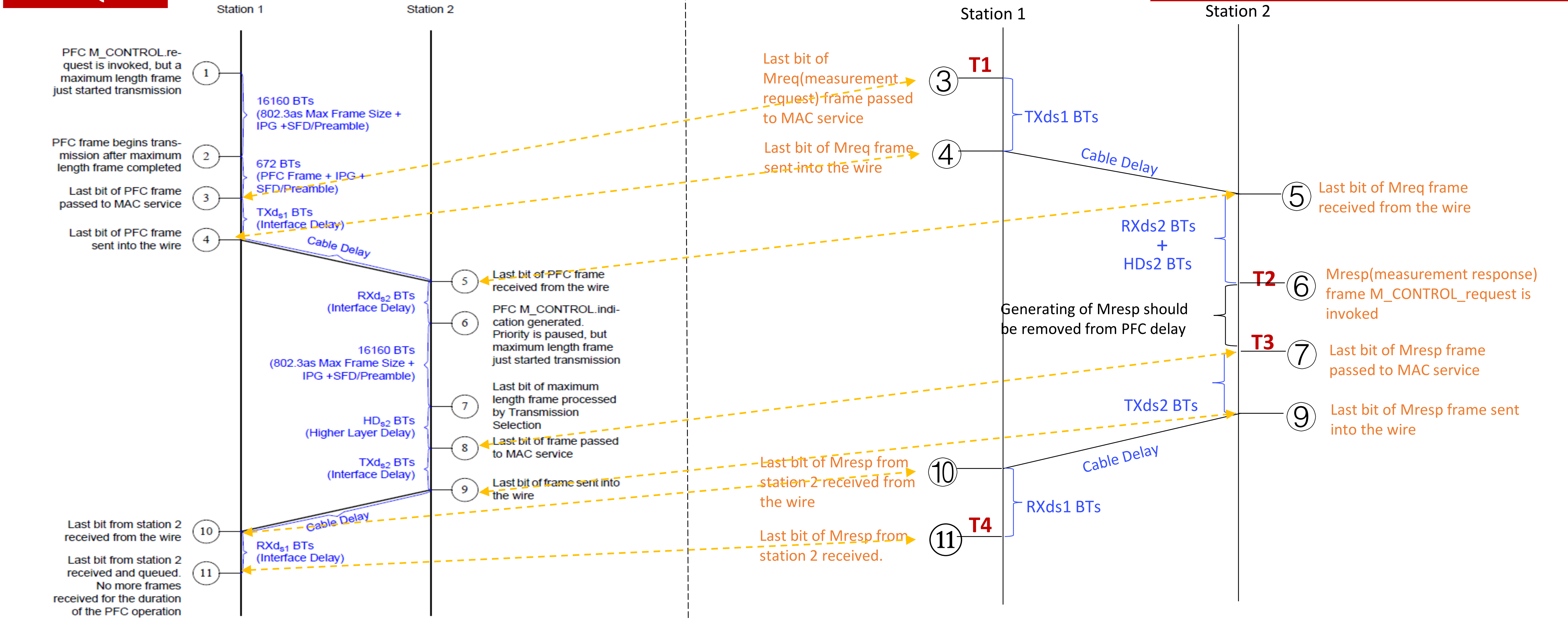
Annex N - Buffer requirements for PFC

- New N.7 subclause showing an informative example

Proposal for Adaptive PFC Headroom

802.1Q PFC

Example of Adaptive headroom


Station 1
Station 2
PFC M_CONTROL.request is invoked, but a maximum length frame just started transmission
1
16160 BTs (802.3as Max Frame Size + IPG +SFD/Preamble)
PFC frame begins transmission after maximum length frame completed
2
672 BTs (PFC Frame + IPG + SFD/Preamble)
Last bit of PFC frame passed to MAC service
3
TXd_{s1} BTs (Interface Delay)
Last bit of PFC frame sent into the wire
4
Cable Delay
5
Last bit of PFC frame received from the wire
RXd_{s2} BTs (Interface Delay)
6
PFC M_CONTROL.indication generated. Priority is paused, but maximum length frame just started transmission
16160 BTs (802.3as Max Frame Size + IPG +SFD/Preamble)
7
Last bit of maximum length frame processed by Transmission Selection
HD_{s2} BTs (Higher Layer Delay)
8
Last bit of frame passed to MAC service
TXd_{s2} BTs (Interface Delay)
9
Last bit of frame sent into the wire
Cable Delay
10
Last bit from station 2 received from the wire
RXd_{s1} BTs (Interface Delay)
11
Last bit from station 2 received and queued. No more frames received for the duration of the PFC operation
Station 1
Station 2
Last bit of Mreq(measurement request) frame passed to MAC service
3
T1
TXds1 BTs
Last bit of Mreq frame sent into the wire
4
Cable Delay
5
Last bit of Mreq frame received from the wire
RXds2 BTs
+
HDs2 BTs
T2
6
Mresp(measurement response) frame M_CONTROL_request is invoked
Generating of Mresp should be removed from PFC delay
T3
7
Last bit of Mresp frame passed to MAC service
TXds2 BTs
9
Last bit of Mresp frame sent into the wire
Cable Delay
10
Last bit of Mresp from station 2 received from the wire
RXds1 BTs
T4
11
Last bit of Mresp from station 2 received.

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

=

X = (T4-T1- (T3-T2)) * Speed = 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1
DV = 2*(Max Frame) + (PFC Frame) + X