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.

![Diagram showing priority levels and headroom calculation](image)

Figure N-3—Worst-case delay (802.1Q-2018)
Proposal for Adaptive PFC Headroom

- 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.

**Example**

```
+-----------------+-----------------+
<table>
<thead>
<tr>
<th>TLV type</th>
<th>TLV information</th>
</tr>
</thead>
<tbody>
<tr>
<td>127</td>
<td>string length = 6</td>
</tr>
</tbody>
</table>
+-----------+-----------------+---+----+-------+----+----+----+----+----+----+
| 802.1 CUI | 802.1 subtype | Willing | MBC | Reserved | PFC cap | PFC Enable |
| 00-80-C2 | 11           | 1       | 1   | 1       | 1           | 1           |
+-----------+-----------------+---+----+-------+----+----+----+----+----+----+
```

Set receiving timestamp T4.
Proposal for Adaptive PFC Headroom (2/4)

- 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

![Diagram of the process]

**Example**

- PFC frame format:
  - 6 octets: Station MAC Address
  - 2 octets: Ether Type = 0x8808
  - 2 octets: Control Opcode = 0x0011
  - 2 octets: Acknowledged (ACK)
  - 8 octets: Timestamp T1
  - 8 octets: Timestamp T2
  - 8 octets: Timestamp T3
  - 8 octets: Timestamp T4
  - 8 octets: Packet Sequence Number
  - 4 octets: Pad (transmit as zero)
  - 4 octets: CRC

- Measurement frame format:
  - 0x0000: measurement request
  - 0x0001: measurement response
Proposal for Adaptive PFC Headroom (4/4)

- Phase 3: Headroom calculation
  - \( X = \text{Port speed} \times (T4 - T1 - (T3 - T2)) \)
  - \( DV = X + 2 \times (\text{Max Frame}) + (\text{PFC Frame}) \)
  - Headroom = \( DV \times \alpha \)
    - \( \alpha \) 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

<table>
<thead>
<tr>
<th>Clause</th>
<th>Effort Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3 Introduction</td>
<td>Small</td>
</tr>
<tr>
<td>5.4.1.7 DCBX Bridge requirements</td>
<td>Small</td>
</tr>
<tr>
<td>5.11 System requirements for Priority-based Flow Control (PFC)</td>
<td>Small</td>
</tr>
<tr>
<td>6.7.1 Support of the ISS by IEEE Std 802.3 (Ethernet)</td>
<td>Small</td>
</tr>
<tr>
<td>36. Priority-based Flow Control (PFC)</td>
<td>Large</td>
</tr>
<tr>
<td>38. Data Center Bridging eXchange protocol (DCBX)</td>
<td>Small</td>
</tr>
<tr>
<td>D.2.10 Priority-based Flow Control Configuration TLV</td>
<td>Medium</td>
</tr>
<tr>
<td>D.5 IEEE 802.1/LLDP extension MIB</td>
<td>Small</td>
</tr>
<tr>
<td>D.6 IEEE 802.1/LLDP extension YANG</td>
<td>Small</td>
</tr>
<tr>
<td>Annex M - Support for PFC in link layers without MAC Control</td>
<td>Small</td>
</tr>
<tr>
<td>Annex N - Buffer requirements for PFC</td>
<td>Medium</td>
</tr>
</tbody>
</table>
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
Proposition for Adaptive PFC Headroom

802.1Q PFC

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

Example of Adaptive headroom

DV = 2*(Max Frame) + (PFC Frame) + X

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