Headroom Measurement Protocol Design

Lily Lv (Huawei)
Fei Chen (Huawei)
To-Do List

- **Timestamp point clarification**
  - Need model/figure with labeled time points
- Protocol design of request-response measurement
- Managed objects
  - The effort, implementation cost, and purpose of statistic gathering and retention requires careful consideration
- DCBX: PFC Configuration TLV format design ----> more generic way???
  - PFC configuration TLV defines Capability (round-trip, PTP-based)
  - PFC informational TLV defines compensation value of PTP-based method

Conclusions:

- ✓ Ethertype for Qdt
  - Reuse Qcz (CI) Ethertype 89-A2
- ✓ Timestamp accuracy
  - Describe accuracy by number of pause quantas or number of maximum length frames, instead of number of nans seconds.
Timestamp Points Clarification
PFC Timestamp Points (Non-MACsec)

- PFC Headroom = t2-t1 + t4-t3
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue
PFC Timestamp Points (Non-MACsec)

- Different procedure
  - PFC invocation -> traffic stop vs. Measurement request -> measurement response
  - MAC control frame vs. MAC data frame
    - PFC pause frame takes the ‘quick path’ ----> no data path delay, ‘quick path’ delay can be ignored
  - PFC pause frame waits at most 1 MAC data frame to be sent ----> t2-t1 is variable, consider worst case
  - After PFC is taken action, at most one more MAC data frame is sent ----> t4-t3 is variable, consider worst case
Figure N-3 Helps to Define Measurement Timestamp Points (Non-MACsec)

Delay Value = 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1 + 2*(Max Frame) + (PFC Frame)
Updated Figure N-3 (Non-MACsec) (PFC invocation delay is ignored)

1. PFC M_CONTROL request is invoked, but a maximum length frame just started transmission
2. PFC frame begins transmission after maximum length frame completed
3. Last bit of PFC frame passed to MAC service
4. Last bit of PFC frame sent into the wire
5. Last bit of PFC frame received from the wire
6. Last bit of PFC frame received from the wire
7. Last bit of PFC frame received from the wire
8. Last bit of PFC frame received from the wire
9. Last bit of PFC frame received from the wire
10. Last bit from station 2 received from the wire
11. Last bit from station 2 received from the wire
12. Last bit from station 2 received from the wire
PFC Timestamp Points in New Figure N-3 (Non-MACsec)

- t1: RX queue is above threshold and invokes signal to PFC module.
- t2: TX queue receives signal from PFC module and stops transmission.
- t3: last packet is sent after TX queue is stopped
- t4: last packet is received by RX queue
PFC Headroom Calculation (Non-MACsec)

- PFC Headroom = t2-t1+ t4-t3
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue

- PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)
  - t1: RX queue is above threshold and invokes signal to PFC module
  - t2: PFC M_CONTROL.indication generated. Priority is paused, but max length frame just started transmission
  - t3: last bit of maximum length frame processed by transmission selection
  - t4: last bit of frame received and queued
Measurement Timestamp Points (Non-MACsec)

- **t1**: last bit of measurement req frame passed to MAC service
  - t1' – PFC invocation delay – PFC frame = t1
- **t2**: last bit of measurement req frame received from MAC service
  - t2' + PFC reaction delay = t2
- **t3**: last bit of measurement resp frame processed by transmission selection
  - t3' = t3
- **t4**: last bit of measurement resp frame received and queued
  - t4' = t4

- **t1'**: last bit of measurement req frame passed to MAC service
  - t1' – PFC invocation delay – PFC frame = t1
- **t2'**: last bit of measurement req frame received from MAC service
  - t2' + PFC reaction delay = t2
- **t3'**: last bit of measurement resp frame processed by transmission selection
  - t3' = t3
- **t4**: last bit of measurement resp frame received and queued
  - t4' = t4
**PFC Headroom Calculation by Measurement Timestamp Points (Non-MACsec)**

- **PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)**
  - t1: RX queue is above threshold and invokes signal to PFC module
  - t2: PFC M_CONTROL.indication generated. Priority is paused, but max length frame just started transmission
  - t3: last bit of maximum length frame processed by transmission selection
  - t4: last bit of frame received and queued

- **PFC Headroom = (t2’+ PFC reaction delay) –(t1’-PFC invocation delay – PFC frame) +t4’-t3’ + 2*(Max Frame)**
  - t1’: last bit of measurement req frame passed to MAC service
    - t1’ – PFC invocation delay – PFC frame = t1
  - t2’: last bit of measurement req frame received from MAC service
    - t2’ + PFC reaction delay= t2
  - t3’: last bit of measurement resp frame processed by transmission selection
    - t3’ = t3
  - t4’: last bit of measurement resp frame received and queued
    - t4’ = t4
Implementation Example (Non-MACsec)

- PFC Headroom = (t2'+ PFC reaction delay) –(t1’-PFC invocation delay-PFC frame) +t4’-t3’ + 2*(Max Frame)
  = (t2’ – r_rx data path delay + PFC reaction time) – (t1’’ + lTx data path delay – PFC invocation delay-PFC frame)
  +(t4’’ – l_rx data path delay )-(t3’’ + r_tx data path delay ) + 2*(Max Frame)
- t1’’: last bit of req frame is generated by measurement module
  - t1’’ + l_tx data path delay = t1’
- t2’’: last bit of req frame is received by measurement module
  - t2’’ – r_rx data path delay = t2’
- t3’’: last bit of resp frame is generated by measurement module
  - t3’’ + r_tx data path delay = t3’
- t4’: last bit of resp frame is received by measurement module
  - t4’’ – l_rx data path delay = t4’

Note:  r_tx data path delay and l_rx data path delay are not full data path delay.
PFC Timestamp Points (MACsec)

- PFC Headroom = t2-t1+ t4-t3
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue
PFC Timestamp Points (MACsec)

- Different procedure
  - PFC invocation -> traffic stop vs. Measurement request -> measurement response
- MAC control frame vs. MAC data frame
  - PFC pause frame traverses on ‘quick path’ -> no data path delay
- PFC pause frame waits at most 1 MAC data frame to be sent -> t2-t1 is variable
- After PFC is taken action, at most one more MAC data frame is sent -> t4-t3 is variable
PFC Timestamp Points in New Figure N-3 (MACsec)

- **Station 1**
  - **t1**: Invocation delay BTs
  - **t2**: Shim delay BTs
  - **t3**: 16160 BTs (Max frame size + IPG + SFD/Preamble)
  - **t4**: HDs1' BTs (data path delay + MACsec delay)
  - **t5**: Last bit of PFC frame received from the wire
  - **t6**: Last bit of PFC frame passed to shim layer
  - **t7**: PFC M_CONTROL.indication generated Priority is paused, but max length frame just started transmission
  - **t8**: Last bit of maximum length frame processed by transmission selection
  - **t9**: Last bit of frame passed to MAC service
  - **t10**: Last bit of frame sent into the wire
  - **t11**: RXds1 BTs

- **Station 2**
  - **t1**: Invocation delay BTs
  - **t2**: Shim delay BTs
  - **t3**: 16160 BTs (Max frame size + IPG + SFD/Preamble)
  - **t4**: HDs2' BTs (data path delay + MACsec delay)
  - **t5**: Last bit of PFC frame received from the wire
  - **t6**: Last bit of PFC frame passed to shim layer
  - **t7**: PFC M_CONTROL.indication generated Priority is paused, but max length frame just started transmission
  - **t8**: Last bit of maximum length frame processed by transmission selection
  - **t9**: Last bit of frame passed to MAC service
  - **t10**: Last bit of frame sent into the wire
  - **t11**: RXds1 BTs

**Diagram Explanation**
- **TX Queue**: Transmit queue
- **RX Queue**: Receive queue
- **Shim**: Shim layer
- **PFC**: Prioritization for Forwarding Control
- **MACsec**: MAC Security
- **PHY**: Physical Layer
- **Medium**: Medium for data transmission
- **RX Queue is above threshold**: Indicates when the receive queue exceeds its threshold.
- **PFC M_CONTROL.request is invoked**: Triggered by the PFC mechanism.
- **Last bit of PFC frame passed to shim layer, but max length frame just start transmission**: Transition point for data handling.
- **PFC data frame begins transmission after max length frame completed**: Sequential data transmission.
- **Last bit of PFC frame passed to MAC service**: Final transmission step.
- **Last bit of PFC frame sent into the wire**: Final data transmission.
PFC Headroom Calculation (MACsec)

- **PFC Headroom = t2-t1+ t4-t3**
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue

- **PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)**
  - t1: RX queue is above threshold and invokes signal to PFC module
  - t2: PFC M_CONTROL.indication generated. Priority is paused, but max length frame just started transmission
  - t3: last bit of maximum length frame processed by transmission selection
  - t4: last bit of frame received and queued

![Diagram](image)
- t1’: last bit of req frame is passed from measurement module
  - t1’ = PFC invocation delay + l_tx_shim layer delay = t1
- t2’: last bit of req frame is passed to measurement module
  - t2’ + r_rx_shim layer delay + PFC reaction delay = t2
- t3’: last bit of measurement resp frame processed by transmission selection
  - t3’ = t3
- t4’: last bit of measurement resp frame received and queued
  - t4’ = t4

**Measurement Timestamp Points (Non-MACsec)**

1. **t1**: Last bit of PFC frame sent into the wire
2. **t2**: Last bit of PFC frame passed from shim layer, but max length frame just start transmission
3. **t3**: Last bit of PFC frame passed to MAC service
4. **t4**: Last bit of PFC frame sent into the wire

**Diagram Notes**
- RX queue is above threshold
  - PFC M_CONTROL.request is invoked
- Last bit of PFC frame passed from shim layer, but max length frame just start transmission
- PFC data frame begins transmission after max length frame completed
- Last bit of PFC frame passed to MAC service
- Last bit of PFC frame sent into the wire

**Non-MACsec Delays**
- HDs1’ BTs (data path delay + MACsec delay)
- HDs2’ BTs (data path delay + MACsec delay)
**Measurement Timestamp Points (MACsec)**

- **PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)**
  - t1: RX queue is above threshold and invokes signal to PFC module
  - t2: PFC M_CONTROL.indication generated. Priority is paused, but max length frame just started transmission
  - t3: last bit of maximum length frame processed by transmission selection
  - t4: last bit of frame received and queued

- **PFC Headroom = (t2’ + r_rx_shim layer delay + PFC reaction delay) –(t1’ – PFC invocation delay – l_tx_shim layer delay) +t4’-t3’ + 2*(Max Frame)**
  - t1’: last bit of req frame is passed from measurement module
    - t1’ – PFC invocation delay – l_tx_shim layer delay = t1
  - t2’: last bit of req frame is passed to measurement module
    - t2’ + r_rx_shim layer delay + PFC reaction delay = t2
  - t3’: last bit of measurement resp frame processed by transmission selection
    - t3’ = t3
  - t4’: last bit of measurement resp frame received and queued
    - t4’ = t4
Summary of PFC Timestamp Points

Non-MACsec

- PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)
  - t1: RX queue is above threshold and invokes signal to PFC module
  - t2: PFC M_CONTROL.indication generated. Priority is paused, but max length frame just started transmission
  - t3: last bit of maximum length frame processed by transmission selection
  - t4: last bit of frame received and queued

MACsec
Summary of Measurement Timestamp Points

**Non-MACsec**

- PFC Headroom = (t2’ + PFC reaction delay) – (t1’ - PFC invocation delay - PFC frame) + t4’ - t3’ + 2*(Max Frame)
  - t1’: last bit of measurement req frame passed to MAC service
  - t2’: last bit of measurement req frame received and queued
  - t3’: last bit of measurement resp frame processed by transmission selection
  - t4’: last bit of measurement resp frame received and queued

**MACsec**

- PFC Headroom = (t2’ + r_rx_shim layer delay + PFC reaction delay) – (t1’ - PFC invocation delay - l_tx_shim layer delay) + t4’ - t3’ + 2*(Max Frame)
  - t1’: req frame is passed from measurement module
  - t2’: req frame is passed to measurement module
  - t3’: last bit of measurement resp frame processed by transmission selection
  - t4’: last bit of measurement resp frame received and queued

*Should the timestamp points (t1’, t2’) converged?*
Thanks
Done: PFC Configuration TLV format design

• Proposal:
  - PFC configuration TLV only includes ‘capability’
    Define priority of the 2 methods.

<table>
<thead>
<tr>
<th>TLV type</th>
<th>TLV information string length</th>
<th>802.1 OUI 00-80-C2</th>
<th>802.1 subtype =11</th>
<th>Willing</th>
<th>MBC</th>
<th>MACsec cap</th>
<th>Reserv ed</th>
<th>PFC cap</th>
<th>PFC Enable</th>
<th>non-PTP</th>
<th>PTP-based</th>
<th>Reserv ed</th>
</tr>
</thead>
<tbody>
<tr>
<td>127</td>
<td>7 octets</td>
<td>1 octets</td>
<td>1 bit</td>
<td>1 bit</td>
<td>1 bit</td>
<td>1 bit</td>
<td>1 bit</td>
<td>4 bits</td>
<td>1 octets</td>
<td>1 bit</td>
<td>1 bit</td>
<td>6 bits</td>
</tr>
<tr>
<td></td>
<td>TLV header</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If non-PTP and PTP-based are supported on both sides, each node choose its own preference.

- ‘PTP comp’ for PTP-based measurement passes to peer separately.

Define a new informational TLV - **PFC informational TLV**

DCBX informational attributes: “Informational attributes are exchanged via LLDP without any participation in a DCBX state machine.”

Compensation value for PTP-based measurement
Delay Value = 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1
+ 2*(Max Frame) + (PFC Frame)

Roundtrip delay = t4 – (t1' – (MAC control processing time)) – (t3 – (t2 + (MAC control processing time))) + (PFC reaction time)
≈ t4 – t1 – (t3 – t2)
Timestamp Point Clarification (2/2)

Roundtrip delay

$$\text{Delay Value} = 2 \times (\text{Cable Delay}) + \text{TXds1} + \text{RXds2} + \text{HDs2} + \text{TXds2} + \text{RXds1}$$

$$+ 2 \times (\text{Max Frame}) + (\text{PFC Frame})$$

$t1$: last bit of measurement request message passed to MAC service

$t2$: last bit of measurement request message passed from MAC service

$t3$: last bit of measurement response message passed to MAC service

$t4$: last bit of measurement response message passed from MAC service

$$\text{Roundtrip delay} = t4 - (t1 - (\text{shim processing time})) - (t3 - t2) - (t3 - t2) + \text{(PFC reaction time)}$$

$$\approx t4 - t1 - (t3 - t2)$$

Modified model based on 802.1Q Figure N-2—Delay model

With MACsec
Timestamp Accuracy

- Local clock frequency drift analysis

  Assume 5ppm oscillator, fiber cable 100Gbps and 10km link distance: 
  (t4-t1) is no more than 200us: 100us link delay plus internal processing delay 
  1ns time offset in 200us, can be ignored.

- Captured timestamp point analysis

  Expected timestamp point: 
  t1: last bit of measurement request message passed to MAC service 
  t4: last bit of measurement request message passed from MAC service 
  t2: last bit of measurement request message passed from MAC service 
  t3: last bit of measurement request message passed to MAC service

  Implementation example: 
  
  t1 = t1' + ePP delay 
  t4 = t4' – iPP delay 
  t1 = t1' + ePP delay 
  t4 = t4' – iPP delay