Headroom Measurement Protocol Design

Lily Lv (Huawei)

Fei Chen (Huawei)

To-Do List

- ✓ **Timestamp point clarification**
- ✓ **DCBX：PFC Configuration TLV format design**
  - ➢ PFC configuration TLV defines Capability (round-trip, PTP-based)
  - ➢ PFC informational TLV defines compensation value of PTP-based method
- Protocol design of request-response measurement
- Managed objects
  - ➢ The effort, implementation cost, and purpose of statistic gathering and retention requires careful consideration

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.

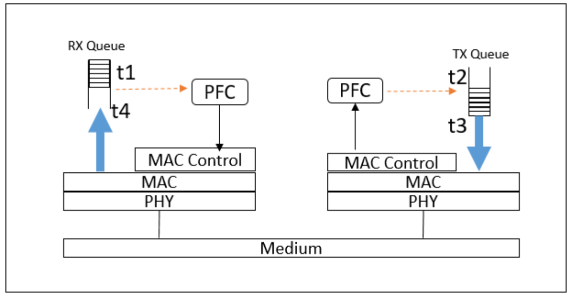
Timestamp Points

- Specify measurement timestamp points
- Non-MACsec and MACsec use the same definition of measurement timestamp points
- Headroom calculation for Non-MACsec and MACsec are different

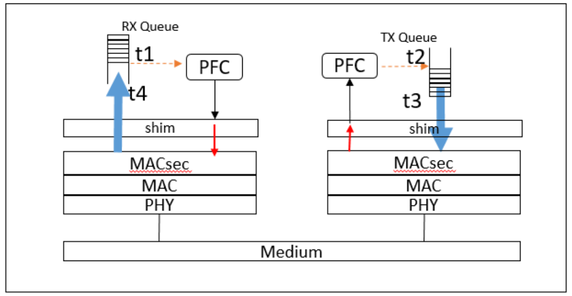
Timestamp Points

PFC timestamp points

Non-MACsec



MACsec

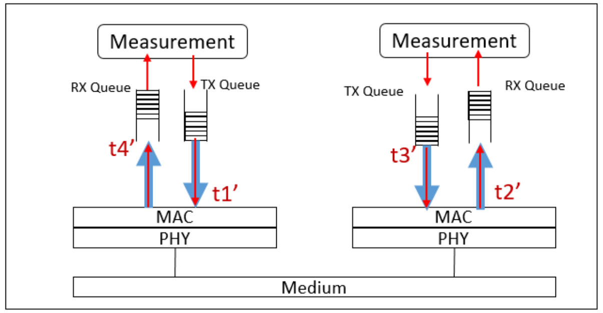


PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)

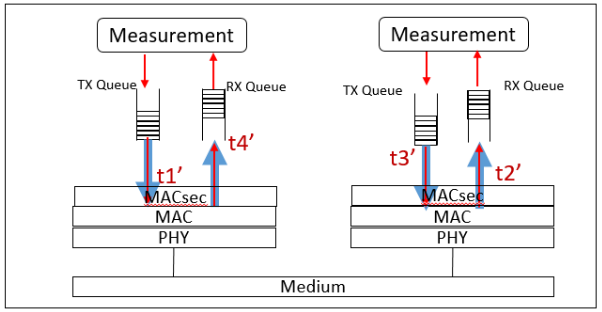
- t1: RX queue is above threshold and invokes signal to PFC module
- t2: PFC M_CONTROL.indicaton generated. Priority is paused, but max length frame just started transmission
- t3: last bit of maximum length frame processed by transmission selection
- t4: last bit of frame received and queued

Measurement timestamp points

Non-MACsec



MACsec



PFC Headroom = (t2'+ PFC reaction delay + r_tx_shim layer delay) – (t1'-PFC invocation delay-PFC frame - l_tx_shim layer delay) +t4'-t3' + 2*(Max Frame)

- t1': last bit of measurement req frame passed to MAC service
- t2': last bit of req frame is passed from MAC service
- t3': last bit of measurement resp frame processed by transmission selection
- t4': last bit of measurement resp frame received and queued

This is to be specified in Qdt.

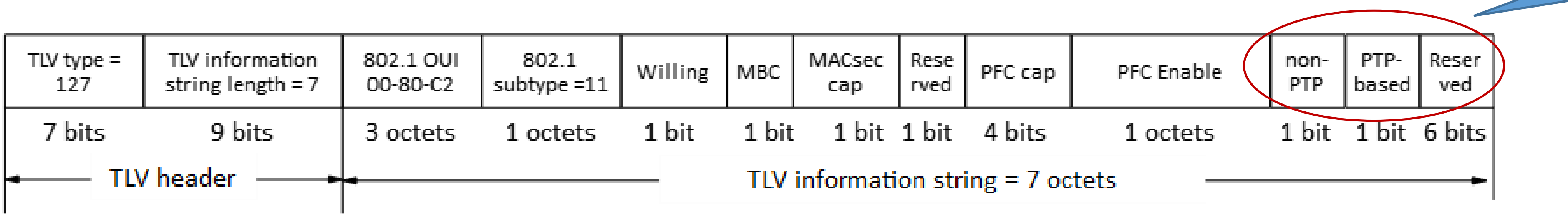
DCBX Design

- PTP-based measurement requires new informational TLV
- Measurement capability is reflected in PFC configuration TLV

PFC Configuration TLV format design

- Proposal :
  - ➢ PFC configuration TLV only includes ‘capability’

Each bit indicates one capability.

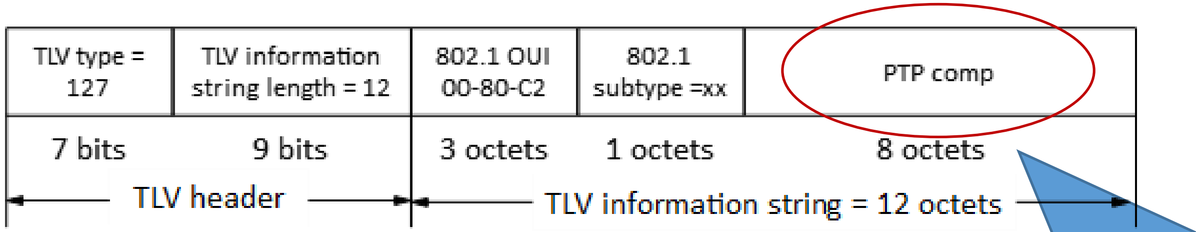


TLV type = 127	TLV information string length = 7	802.1 OUI 00-80-C2	802.1 subtype =11	Willing	MBC	MACsec cap	Reserved	PFC cap	PFC Enable	non-PTP	PTP-based	Reserved
7 bits	9 bits	3 octets	1 octets	1 bit	1 bit	1 bit	1 bit	4 bits	1 octets	1 bit	1 bit	6 bits
TLV header		TLV information string = 7 octets										

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**



TLV type = 127	TLV information string length = 12	802.1 OUI 00-80-C2	802.1 subtype =xx	PTP comp
7 bits	9 bits	3 octets	1 octets	8 octets
TLV header		TLV information string = 12 octets		

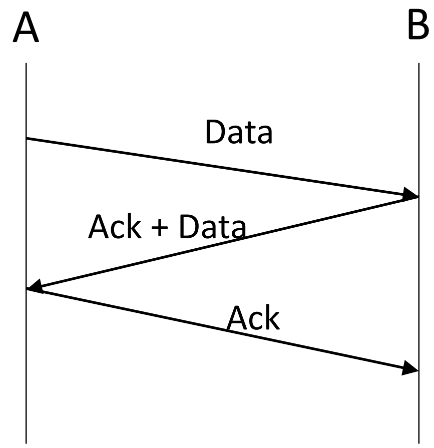
Compensation value for PTP-based measurement

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

Measurement Protocol Design

Benefit of Piggybacking Roundtrip Measurement

Piggybacking for TCP acknowledgement



Advantage：

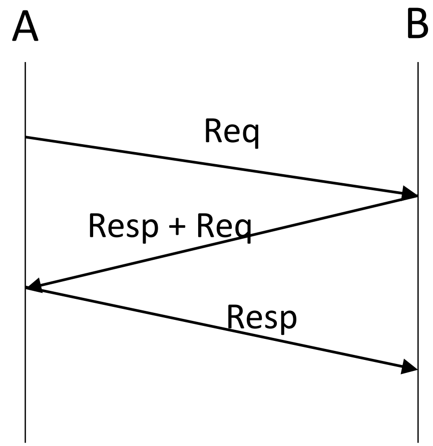
- Better utilize network bandwidth for full-duplex communication

Disadvantage:

- Delay in the transfer of the ACK

Set a counter on host 'B' to control the waiting time for data

Piggybacking for roundtrip measurement



Advantage：

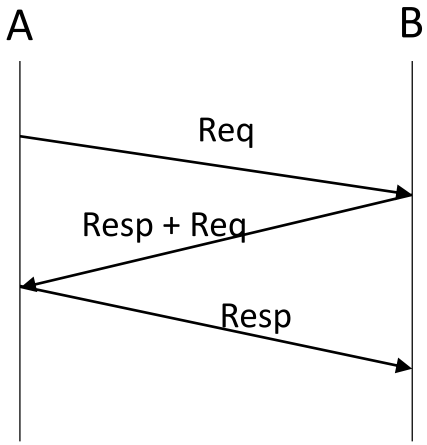
- Complete the measurement faster in query lost case

Disadvantage:

- Waste network bandwidth in some cases

Minimize bandwidth waste by optimizing the mechanism?

Piggybacking Roundtrip Measurement Mechanism（1/2）



Assumption:

- Auto calculated headroom --- successfully take roundtrip measurement at N times, and take the average value as headroom
  n: the number of roundtrip measurement
- Request message sending interval is no more than T, but no less than t
  req_timer: timer for request message sending interval
- req(i): the i^{th} request message
- resp(j): response message corresponding to the j^{th} request message

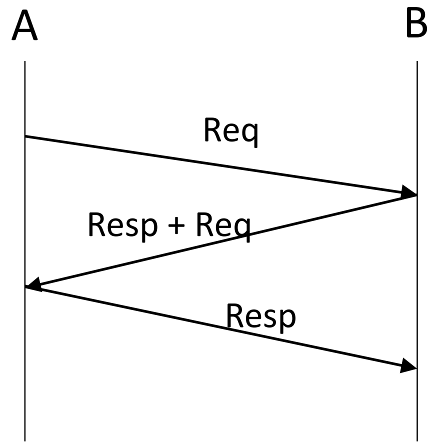
Processing:

- Initial stage： set n=0, req_timer=0, i=0; send req(i)

Cases entering initial stage:

- Node just started
- Link port status changed
- Manually trigger the auto headroom calculation
- Vendor specified measurement cycle
- ……

Piggybacking Roundtrip Measurement Mechanism（2/2）



Processing:

- **If req_timer increases to T**, send req(i+1); set req_timer=0
- **If receive req(k),**
  - If n==N, send resp(k)
  - If n<N,
    - If req_timer<t, send resp(k)
    - If req_timer>=t, send resp(k)+req(i+1); set req_timer=0
- **If receive resp(j),**
  - Finish a single time roundtrip measurement; set n=n+1
  - If n<N, continue increasing req_timer (until req_timer=T to send req(i+1))
  - If n=N, calculate headroom by averaging N times roundtrip measurement results, auto calculated headroom successful.
- **If receive resp(j)+req(k)，**
  - Finish a single time roundtrip measurement; set n=n+1
  - If n=N, calculate headroom by averaging N times roundtrip measurement results, auto calculated headroom successful; send resp(k)
  - If n<N,
    - If req_timer<t, send resp(k)
    - If req_timer>=t, send resp(k)+req(i+1); set req_timer=0

Exception handling:

- In order to avoid request message flooding the network, define a boundary of sent request messages M.
  - If i>M but n<N, system should stop the measurement, and report auto headroom calculation failed.

Measurement Message Format

	Octet	Length
PDU Ethertype	1	2
Version	3	4 bits
Subtype	3	4 bits
Headroom Measurement PDU	4	65-529

Re-use CI Ethertype 89-A2

Subtype 0, CIM

Subtype 1, Headroom Measurement Message

Measurement PDU

	Octet	Length
Version	1	4 bits
Reserved	1	2 bits
Req/Resp/Resp+Req	1	2 bits
Length	2	1
t1	3	8
t2	11	8
t3	19	8
t4	27	8
PSN	35	1
p_t1(optional)	36	8
p_PSN(optional)	44	1

Reduce

Request	Response	Resp+Req
Version	Version	Version
Reserved	Reserved	Reserved
0	1	2
36	36	45
t1	t1	t1
t2	t2	t2
t3	t3	t3
t4	t4	t4
PSN	PSN	PSN
		p_t1
		p_PSN

Other Explanation for the Mechanism

- Measurement time will not exceed T*M
- Node only maintains PSN(i) for req message, no additional status need to be stored.
- Do not wait for a response message to be received before sending the next request message. Do not set timer for response message timeout.
- It may send redundant request messages, but the effect can be controlled by t, T and M.