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**

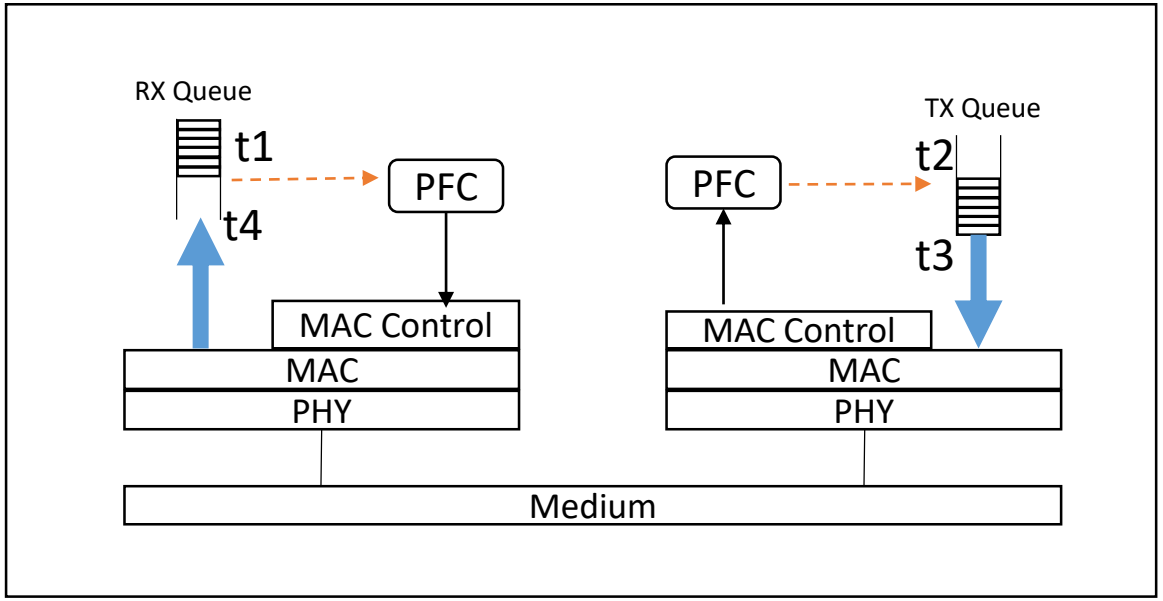
- Reuse Qcz (CI) Ethertype 89-A2

✓ **Timestamp accuracy**

- Describe accuracy by number of pause quantas or number of maximum length frames, instead of number of nans seconds.

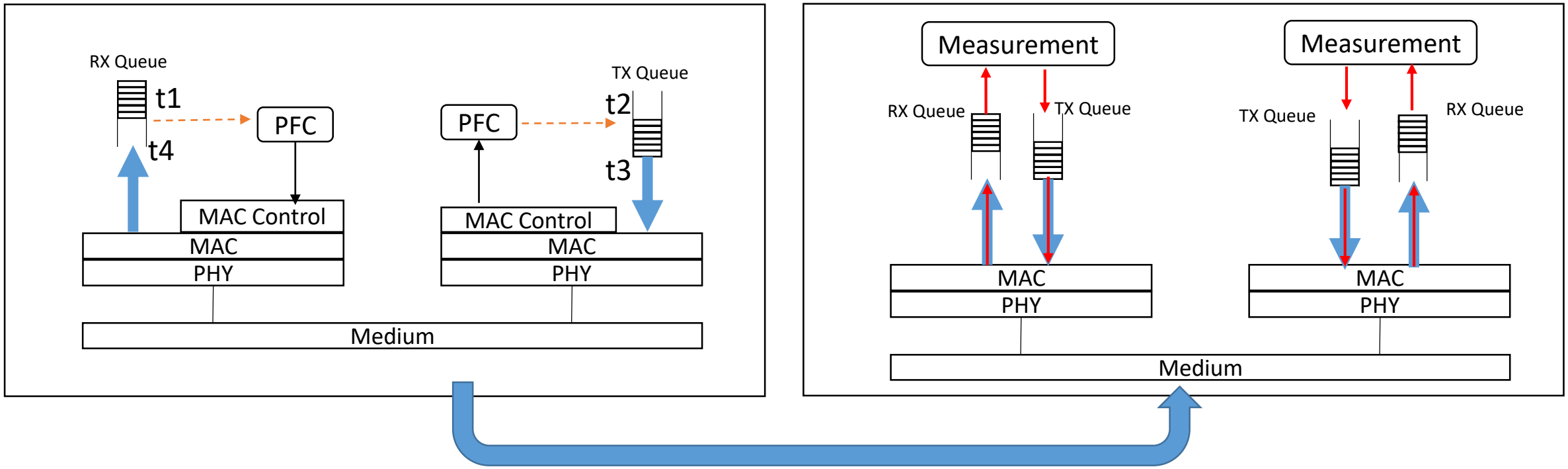
Timestamp Points Clarification

PFC Timestamp Points (Non-MACsec)



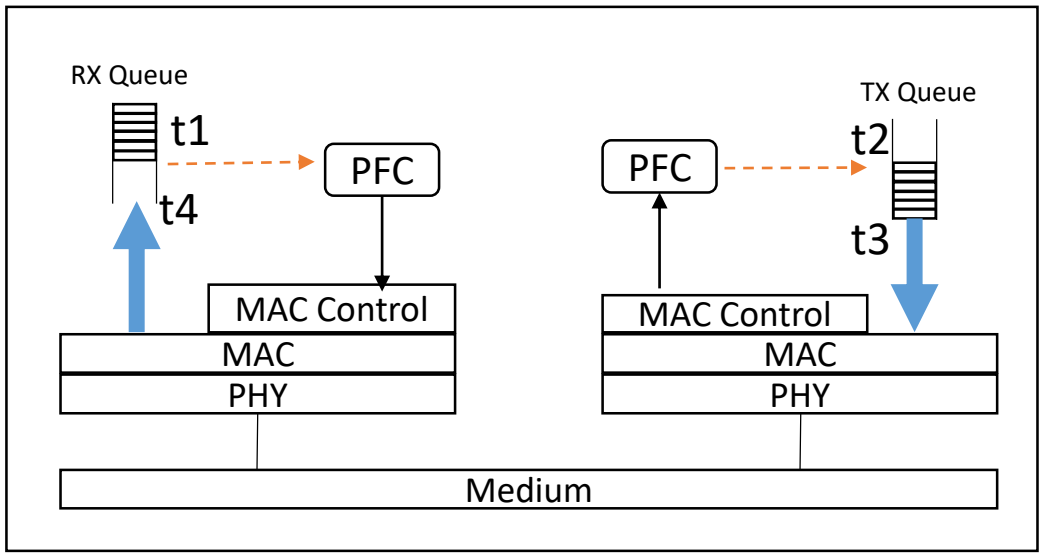
- PFC Headroom = t2-t1+ t4-t3
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue

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)



802.1Q

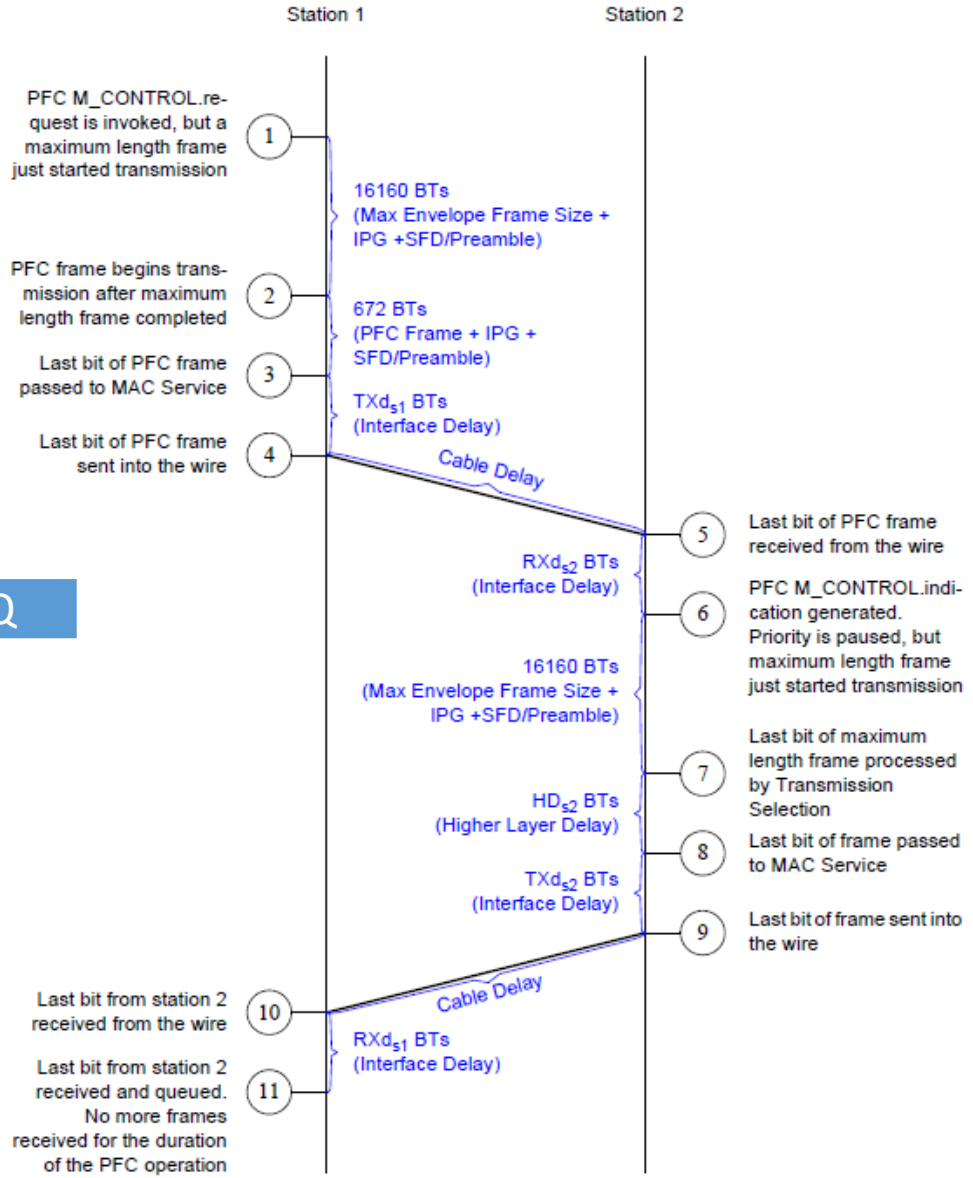


Figure N-3—Worst-case delay

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)

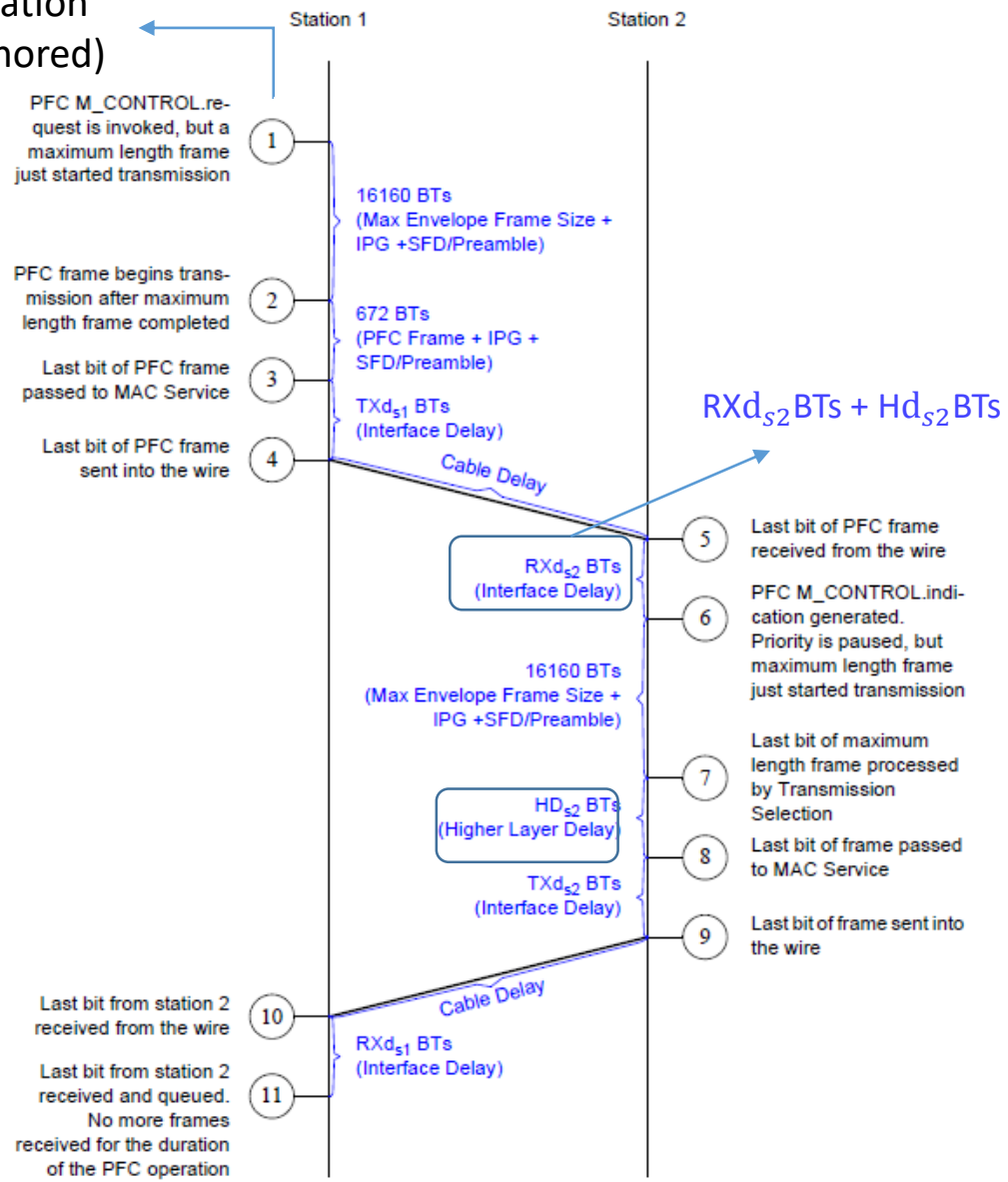
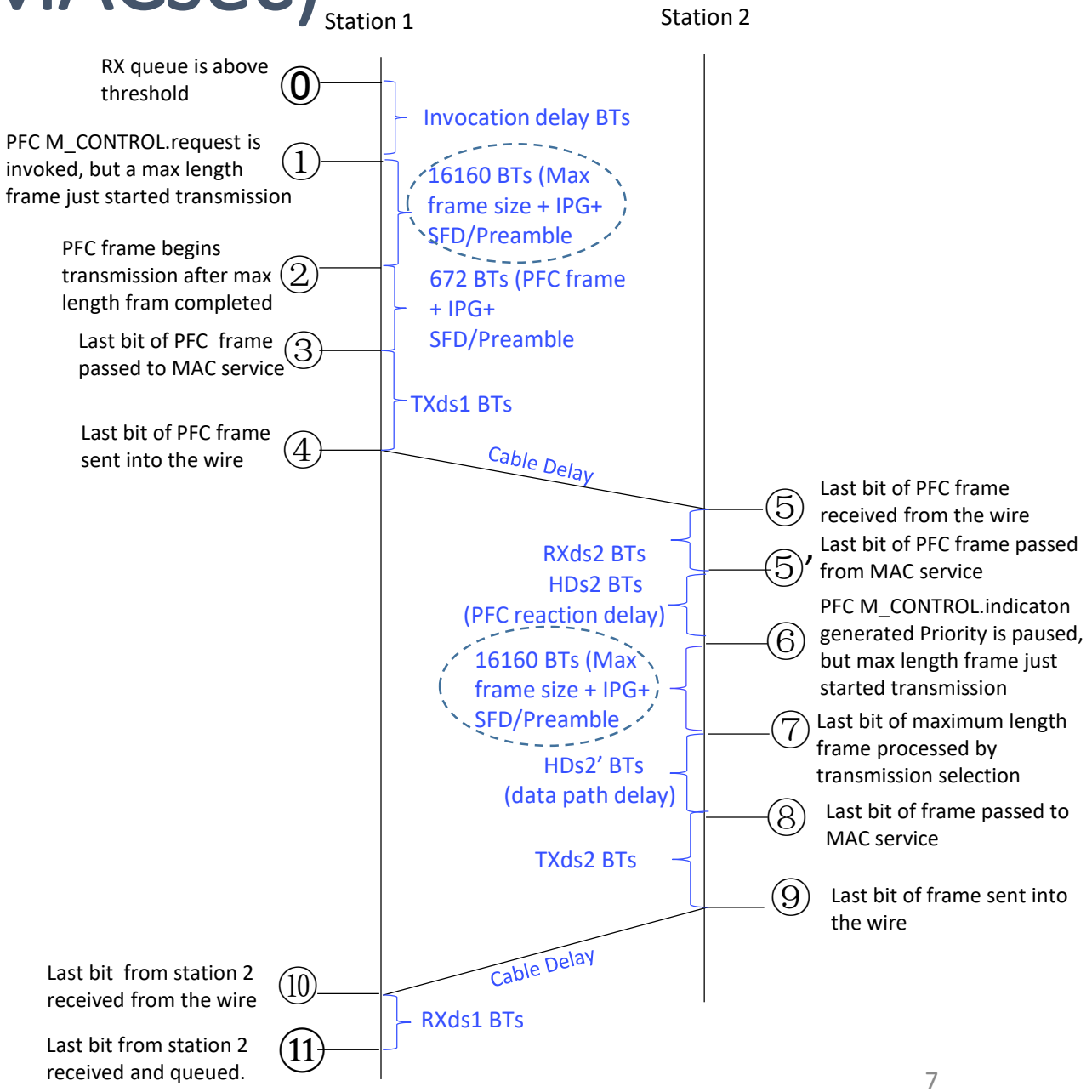
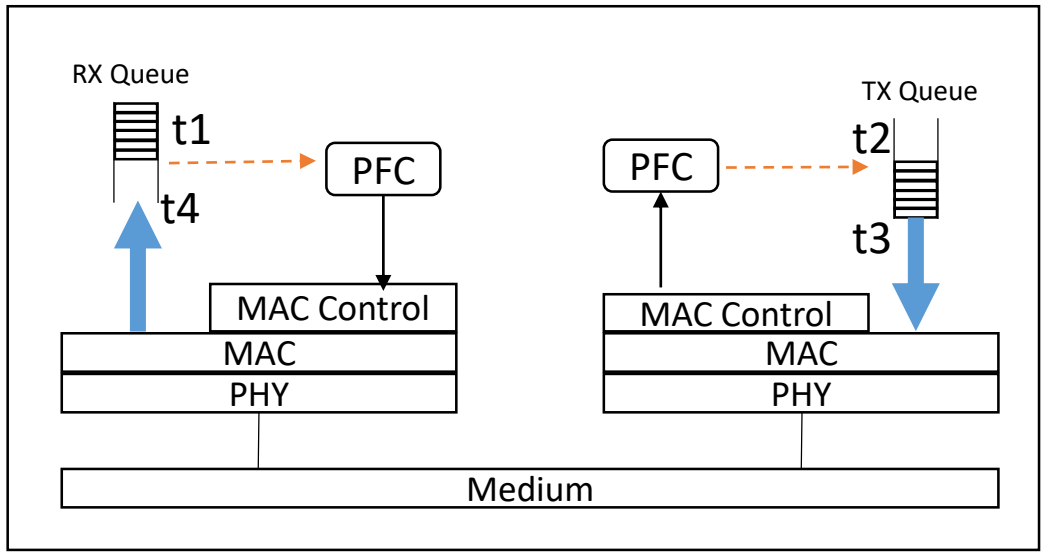
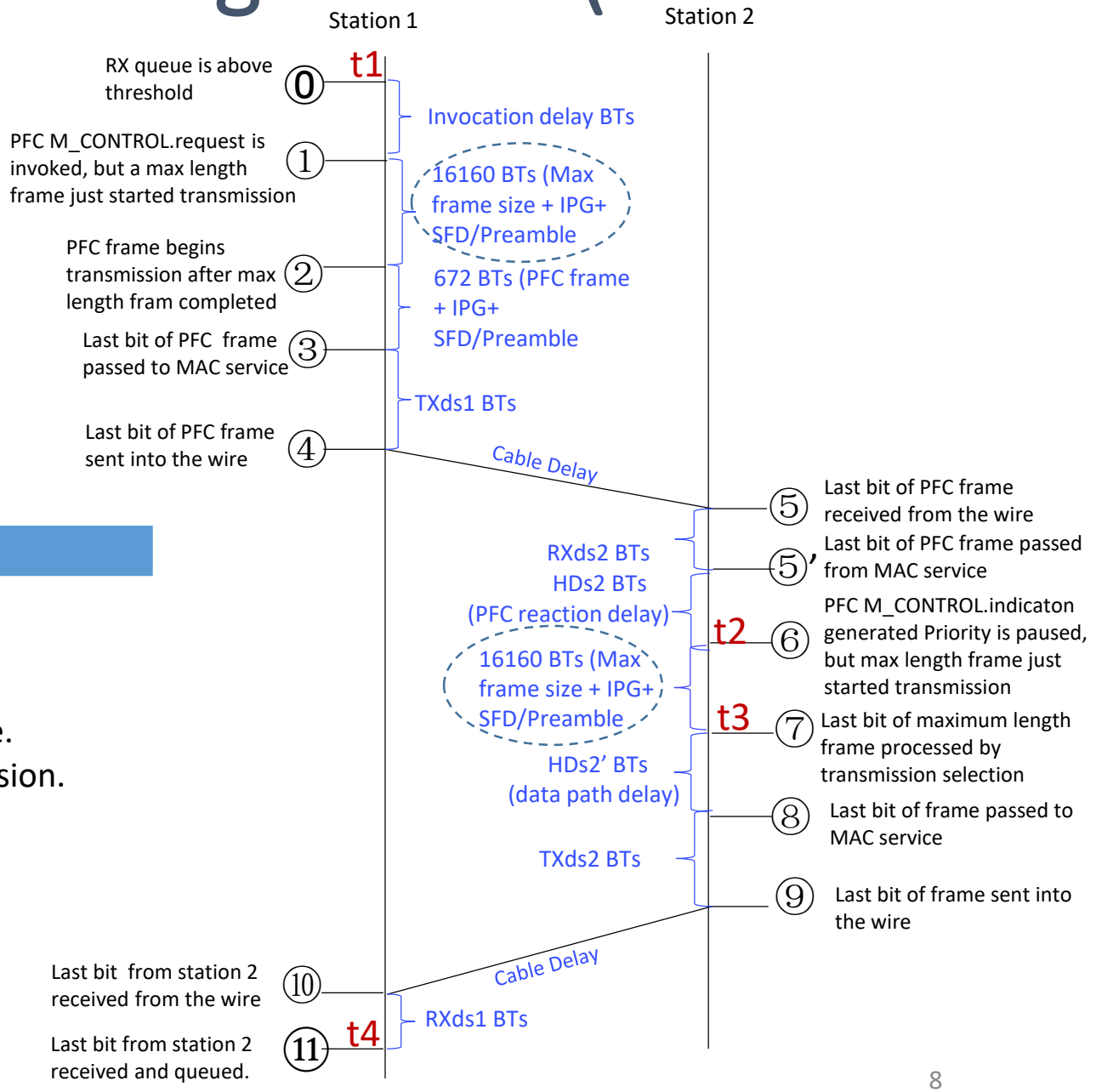


Figure N-3—Worst-case delay



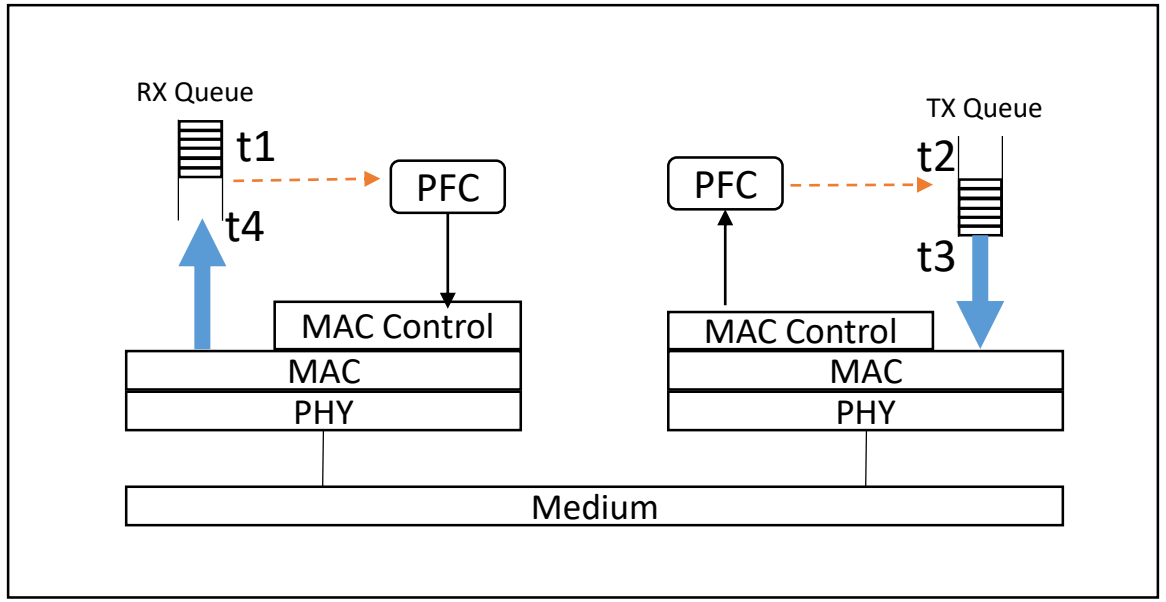
PFC Timestamp Points in New Figure N-3 (Non-MACsec)





- 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

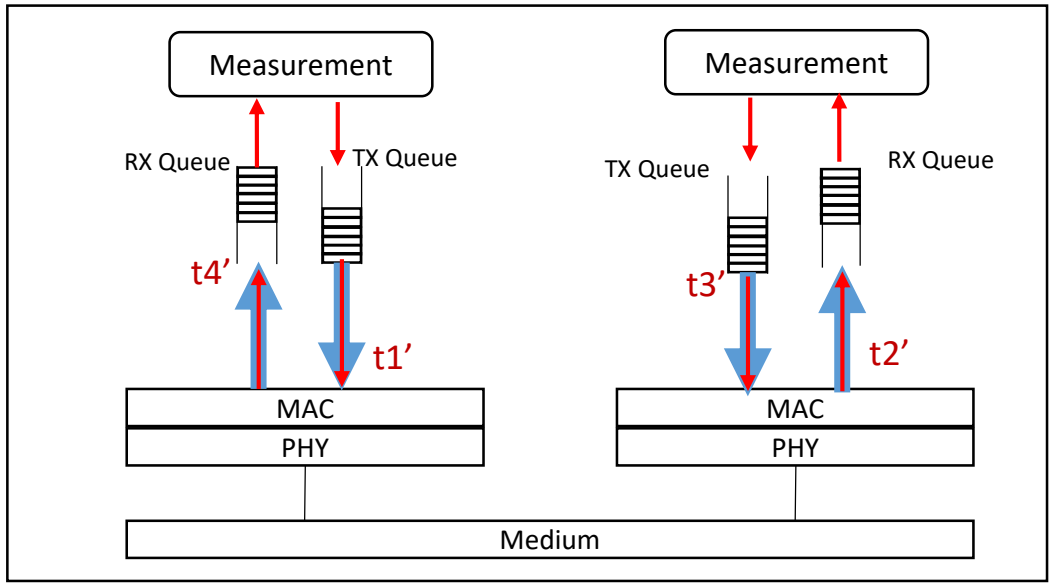
PFC Headroom Calculation (Non-MACsec)

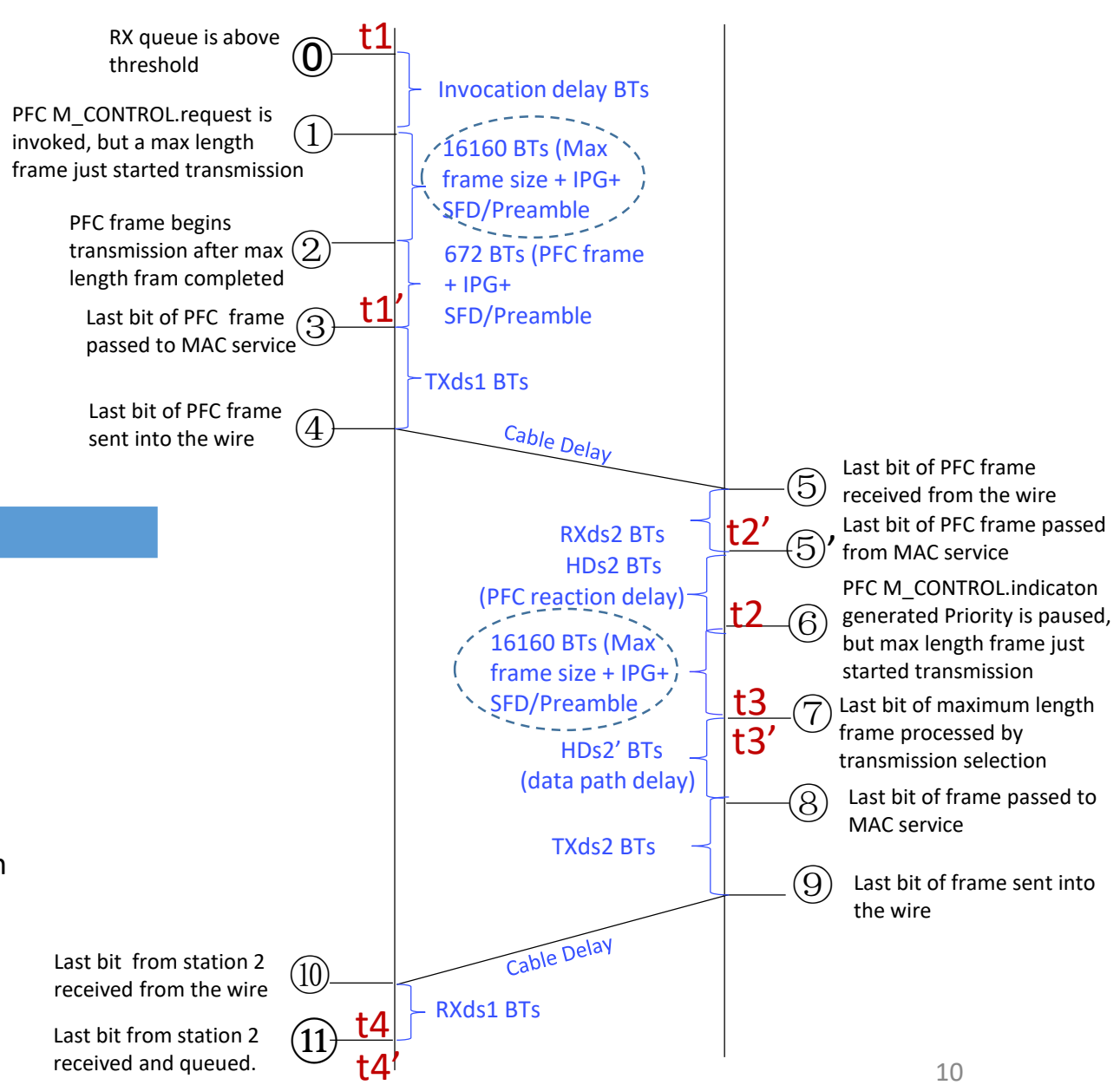


- PFC Headroom = t2-t1+ t4-t3
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue

- 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.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

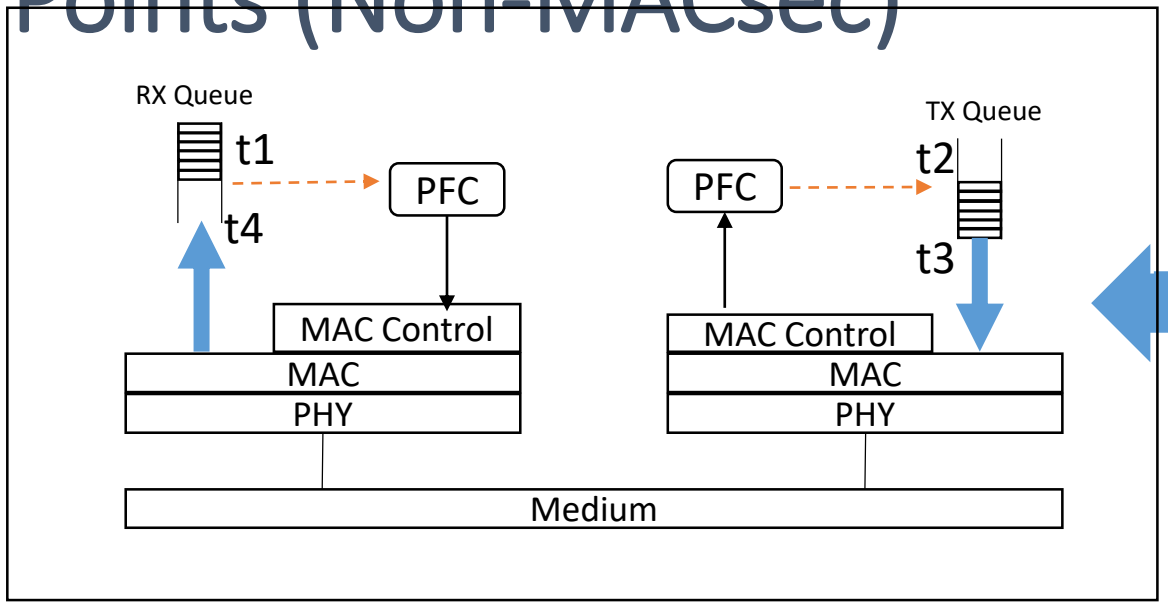
Measurement Timestamp Points (Non-MACsec)

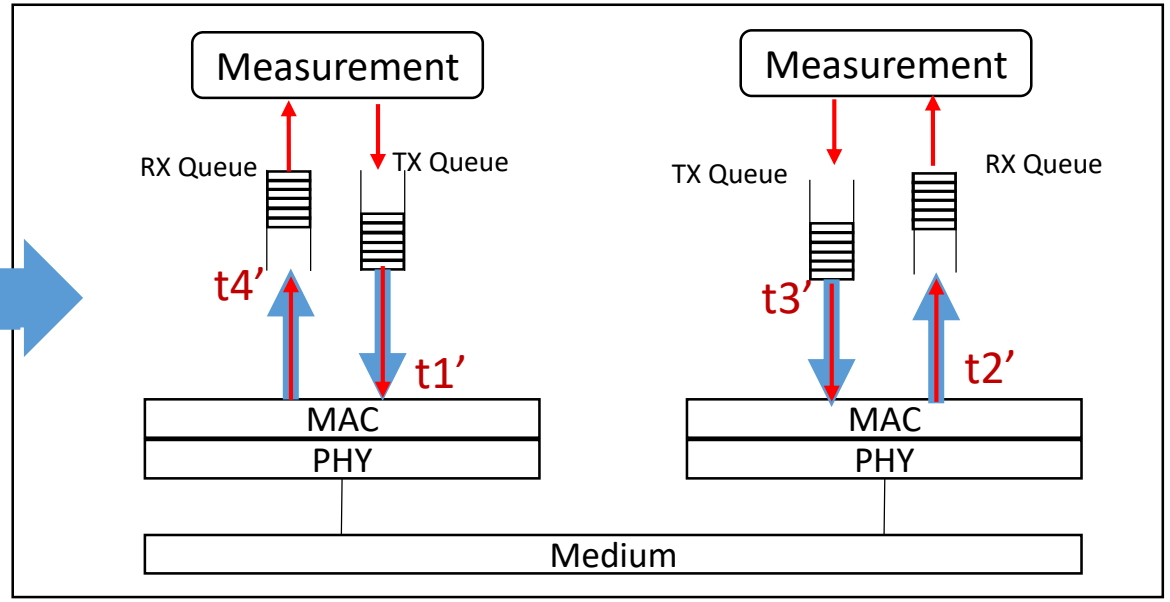




- t1': last bit of measurement req frame passed to MAC service
  - t1' – PFC invocation delay – PFC frame = t1
- t2': last bit of measurement req frame received from MAC service
  - t2' + PFC reaction delay= t2
- t3': last bit of measurement resp frame processed by transmission selection
  - t3' = t3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

PFC Headroom Calculation by Measurement Timestamp Points (Non-MACsec)

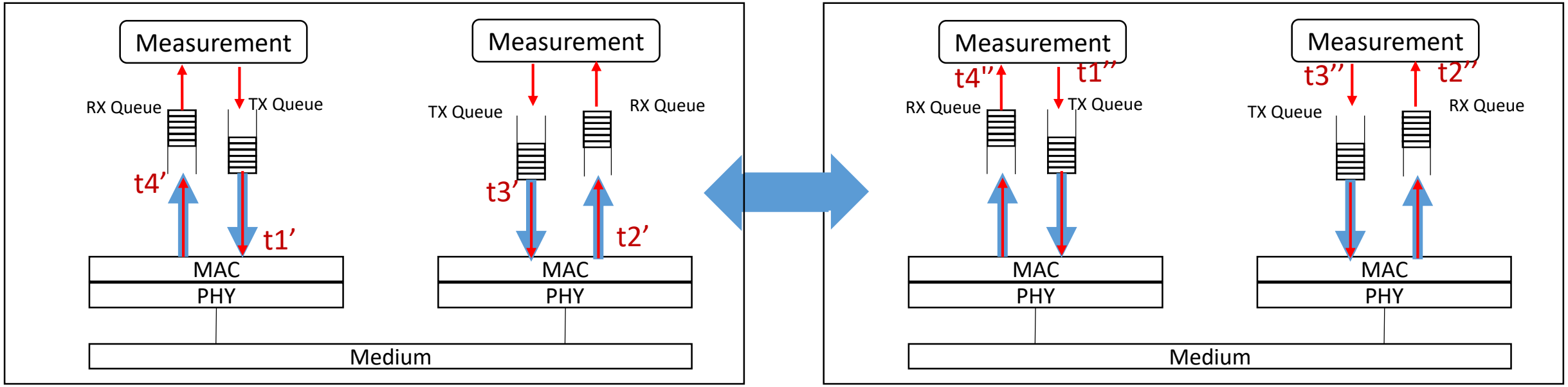




- **PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)**
  - t1: RX queue is above threshold and invokes signal to PFC module
  - t2: PFC M_CONTROL.indicaton generated. Priority is paused, but max length frame just started transmission
  - t3: last bit of maximum length frame processed by transmission selection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

- **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

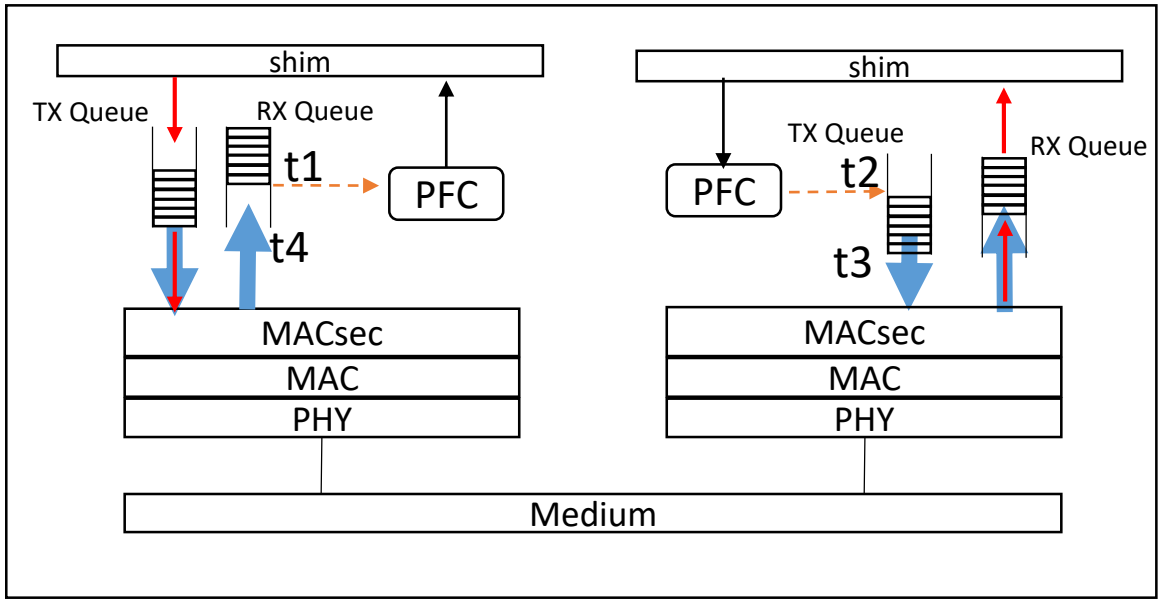
Implementation Example (Non-MACsec)



- 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'' + l_tx 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'

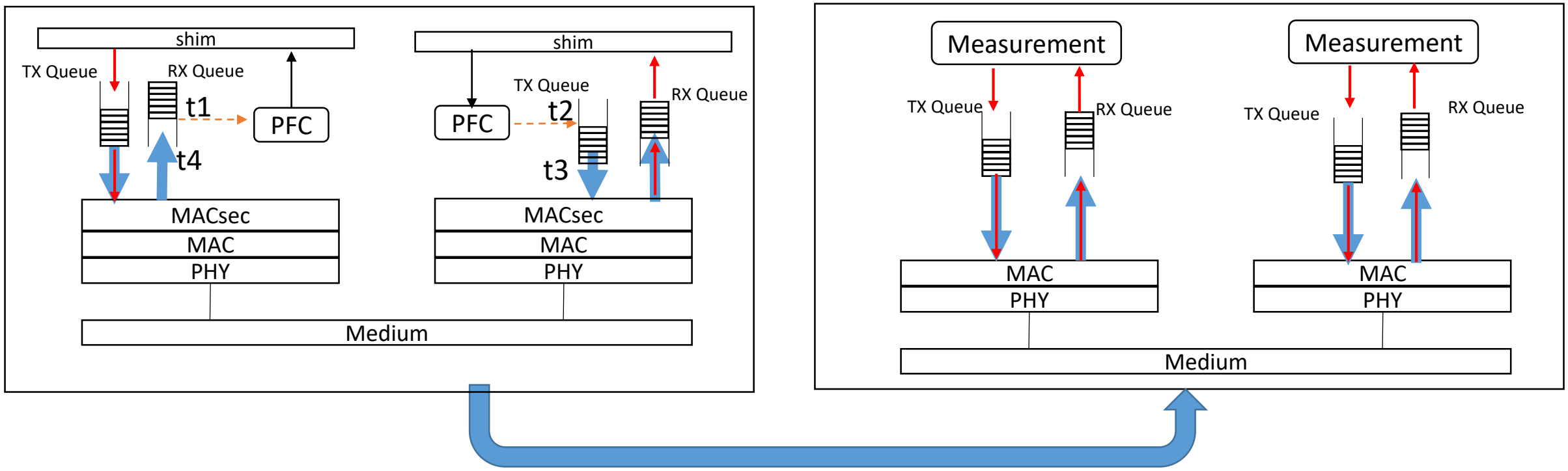
Note: r_tx data path delay and l_rx data path delay are not full data path delay.

PFC Timestamp Points (MACsec)



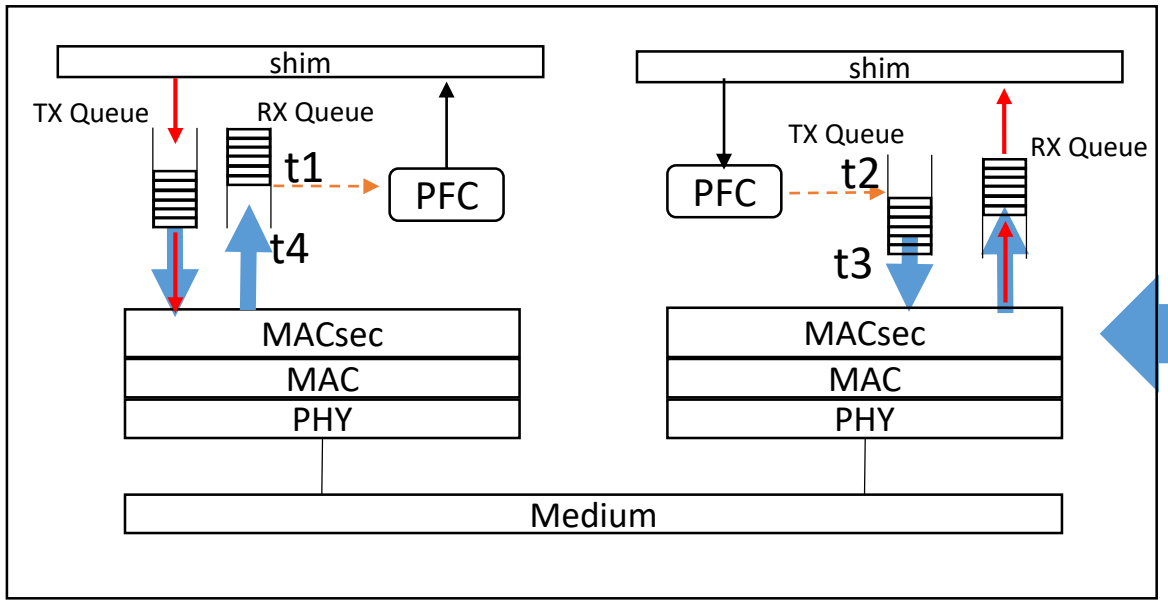
- PFC Headroom = t2-t1+ t4-t3
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue

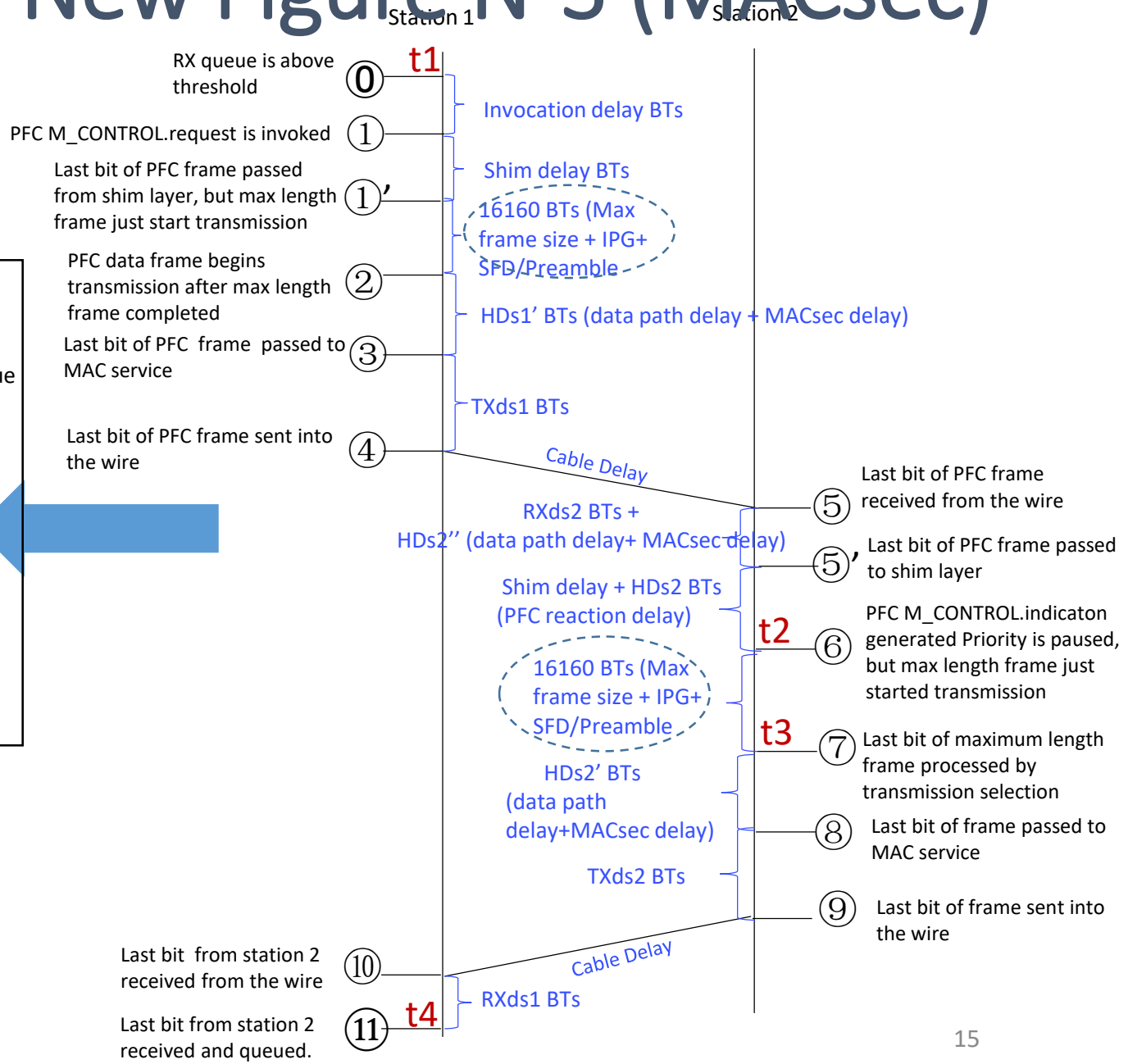
PFC Timestamp Points (MACsec)



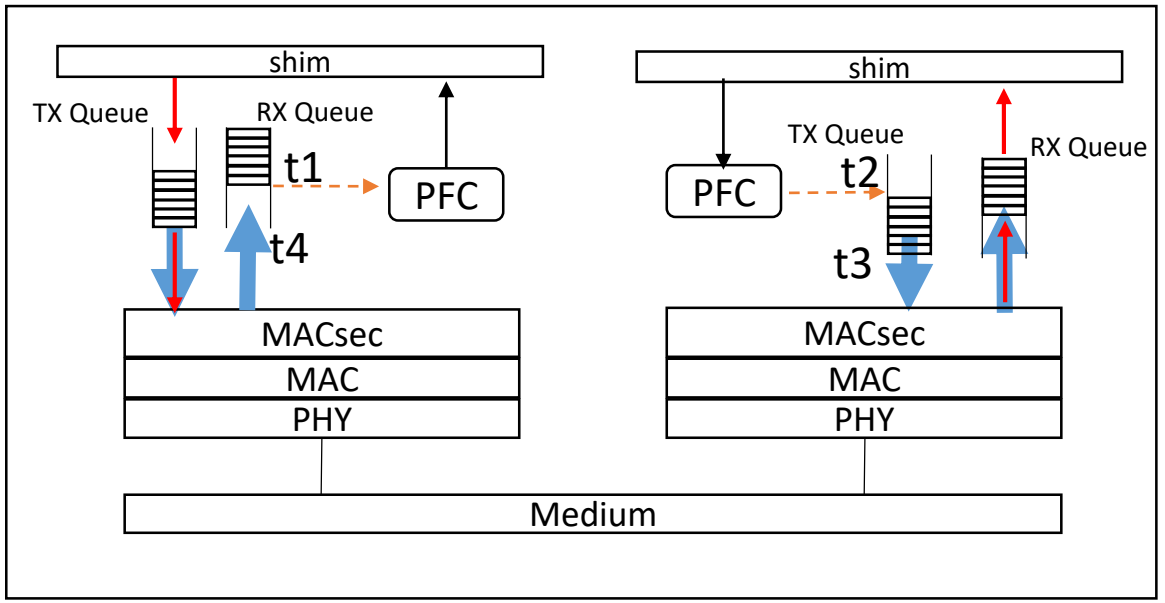
- Different procedure
  - PFC invocation-> traffic stop vs. Measurement request-> measurement response
- ~~MAC control frame vs. MAC data frame~~
  - ~~PFC pause frame traverses on 'quick path' ------ > no data path delay~~
- PFC pause frame waits at most 1 MAC data frame to be sent ------ > t2-t1 is 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)





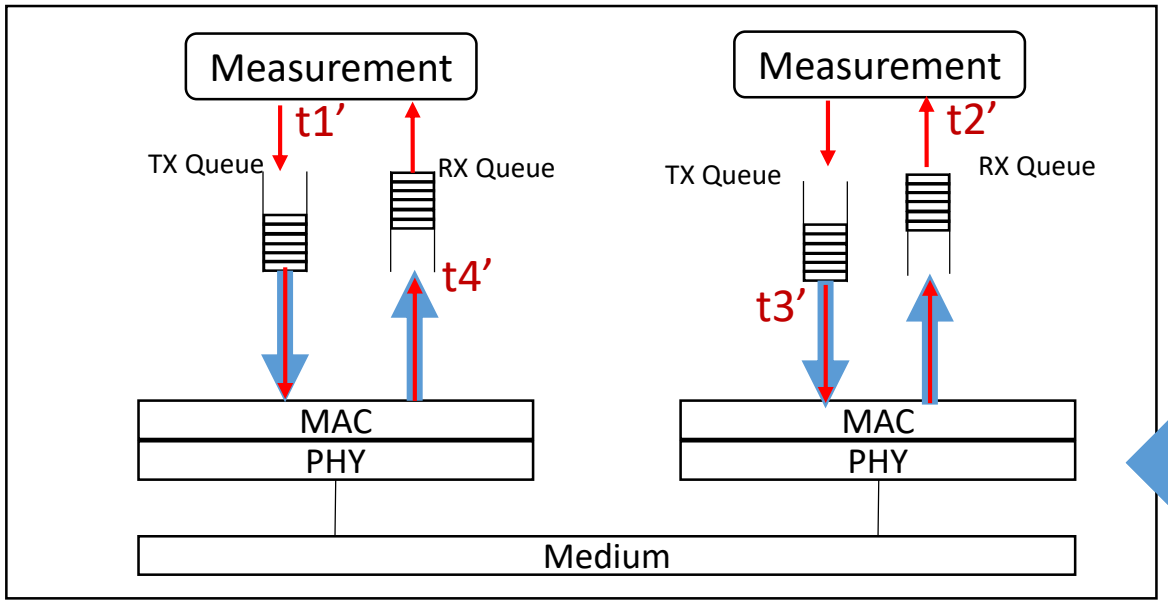
PFC Headroom Calculation (MACsec)



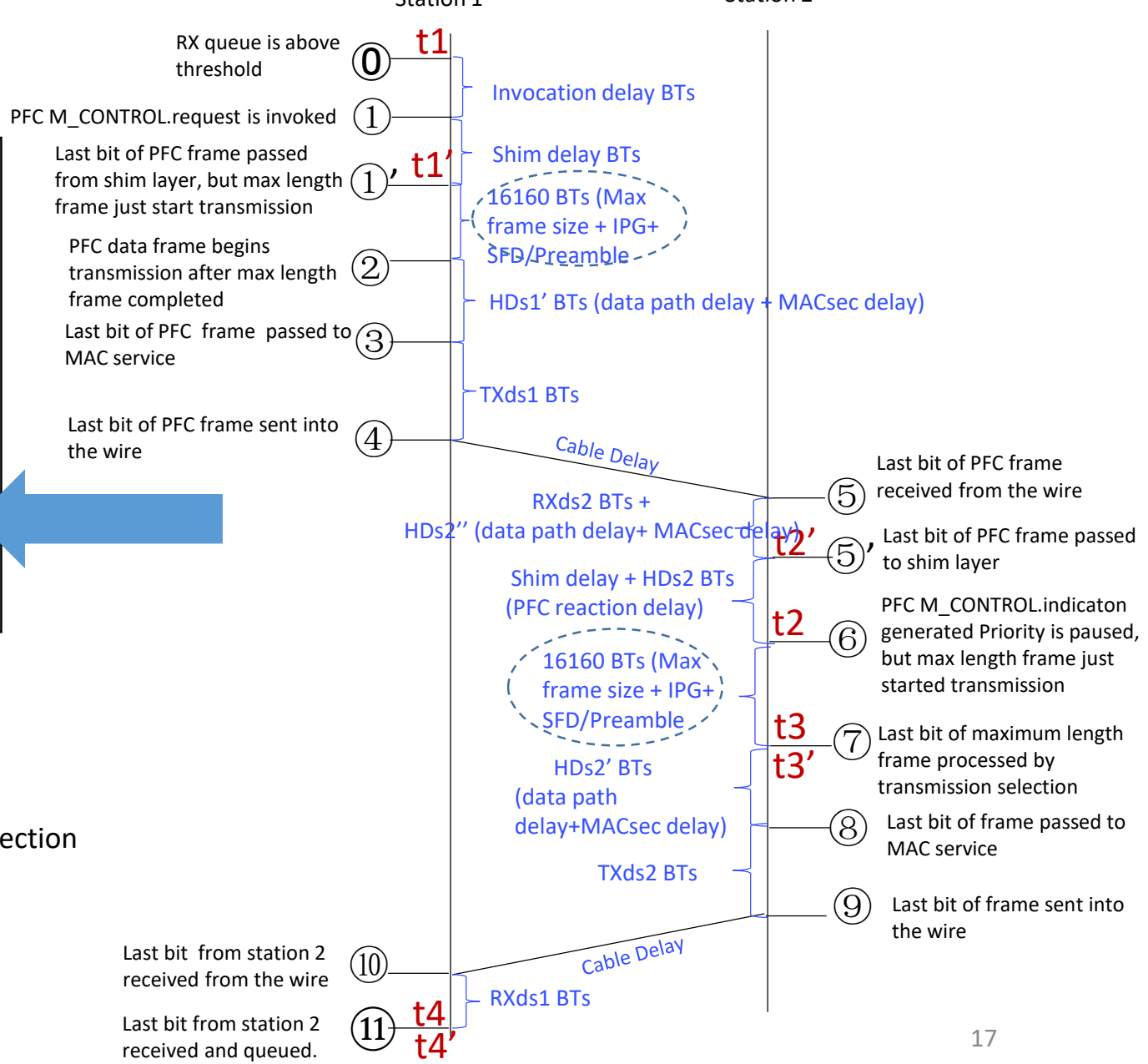
- PFC Headroom = t2-t1+ t4-t3
  - t1: RX queue is above threshold and invokes signal to PFC module.
  - t2: TX queue receives signal from PFC module and stops transmission.
  - t3: last packet is sent after TX queue is stopped
  - t4: last packet is received by RX queue

- 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.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

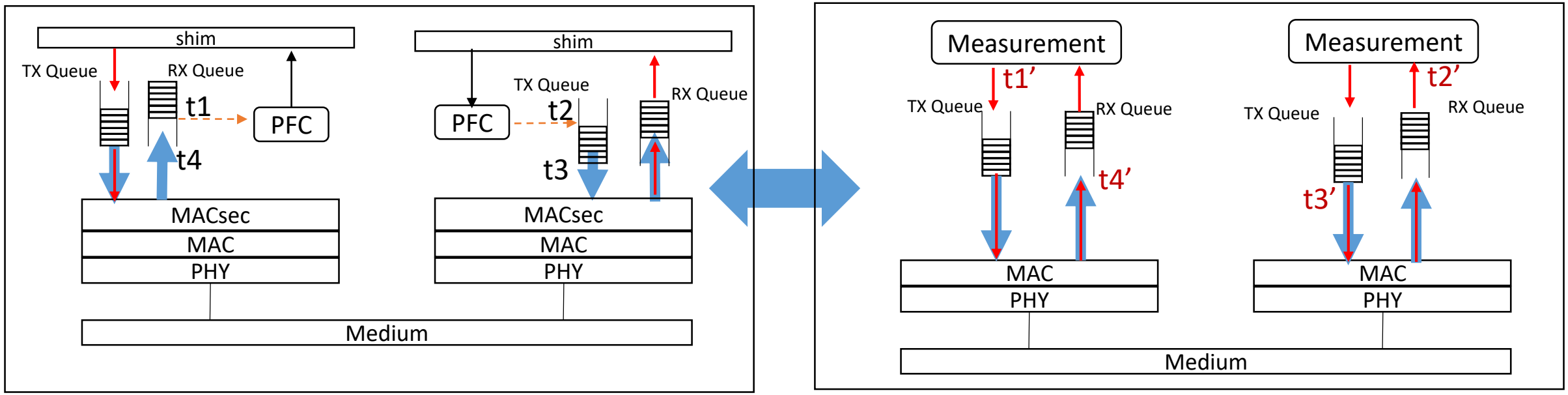
Measurement Timestamp Points (Non-MACsec)



- 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 (MACsec)

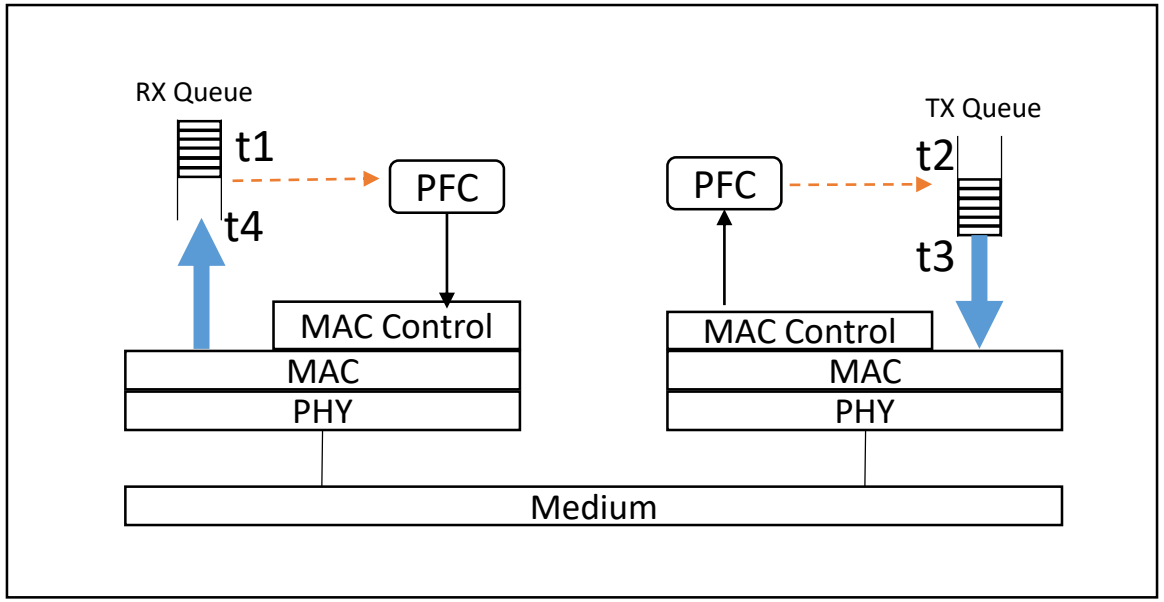


- PFC Headroom = t2-t1+ t4-t3 + 2*(Max Frame)
  - t1: RX queue is above threshold and invokes signal to PFC module
  - t2: PFC M_CONTROL.indicaton generated. Priority is paused, but max length frame just started transmission
  - t3: last bit of maximum length frame processed by transmission selection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

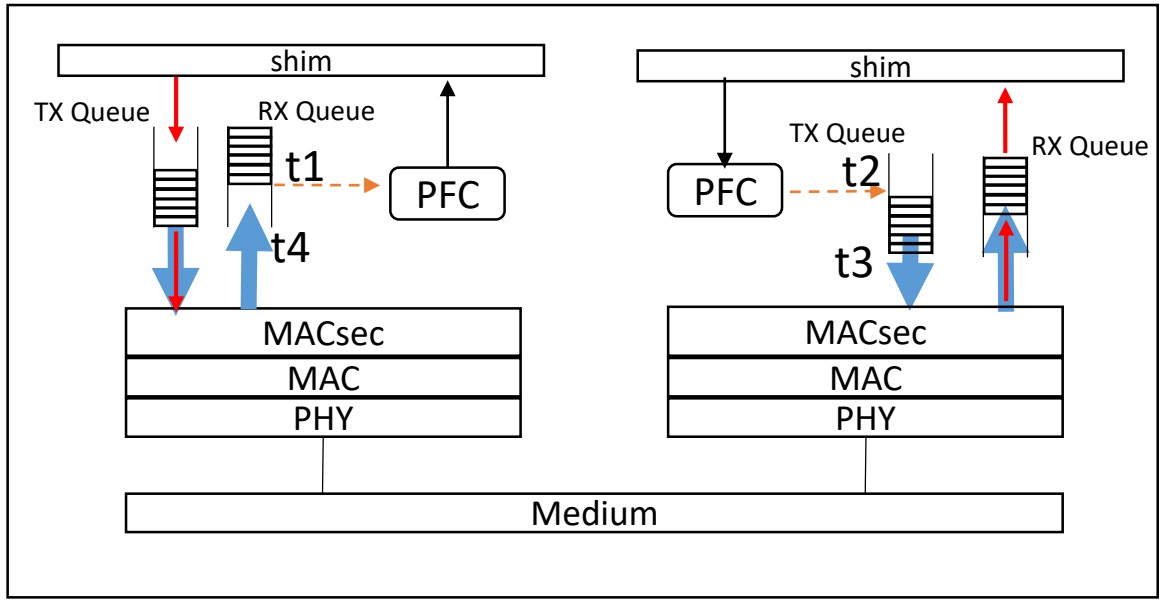
- **PFC Headroom = (t2' + r_rx_shim layer delay + PFC reaction delay) –(t1' – PFC invocation delay – l_tx_shim layer delay) +t4'-t3' + 2*(Max Frame)**
  - t1': last bit of req frame is passed from measurement module
    - t1' – PFC invocation delay – l_tx_shim layer delay = t1
  - t2': last bit of req 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



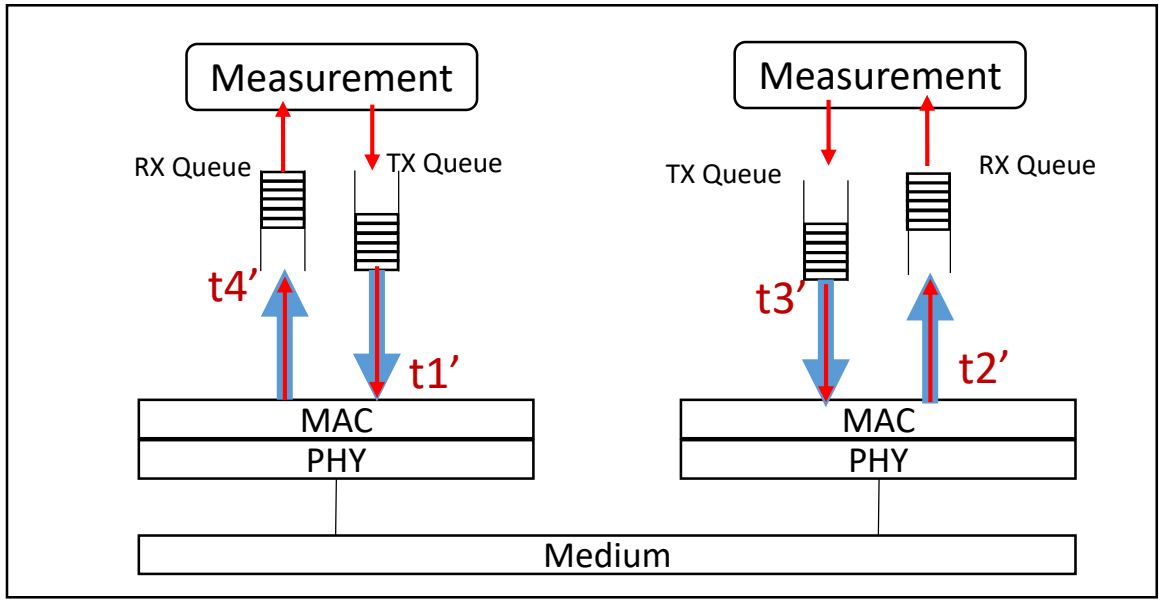
MACsec



- 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

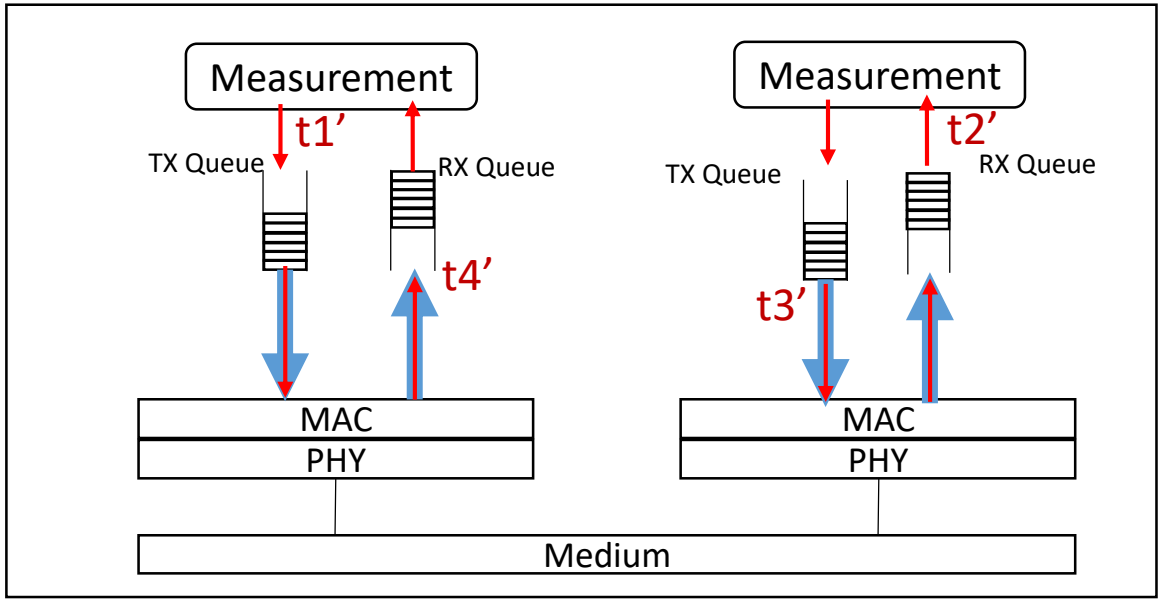
Summary of Measurement Timestamp Points

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

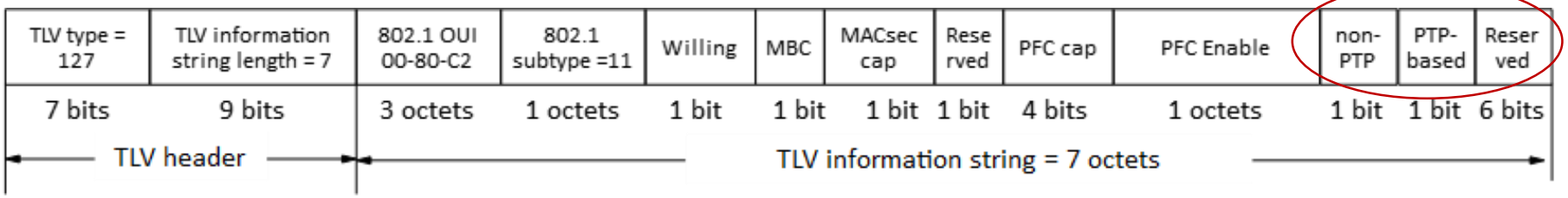
Should the timestamp points（t1’, t2’） converged?

Thanks

Done: PFC Configuration TLV format design

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

    Define priority of the 2 methods.

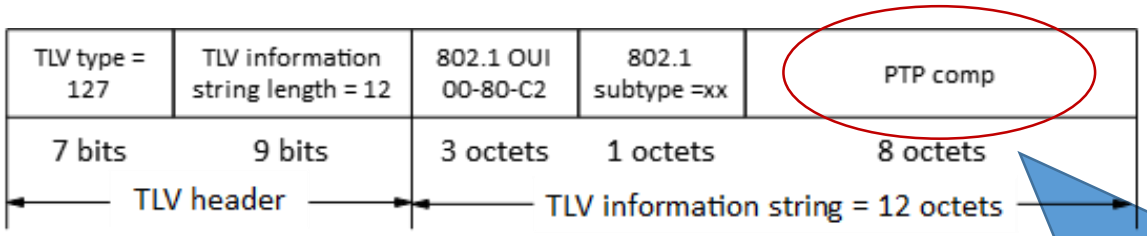


Each bit indicates one capability.

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**



Compensation value for PTP-based measurement

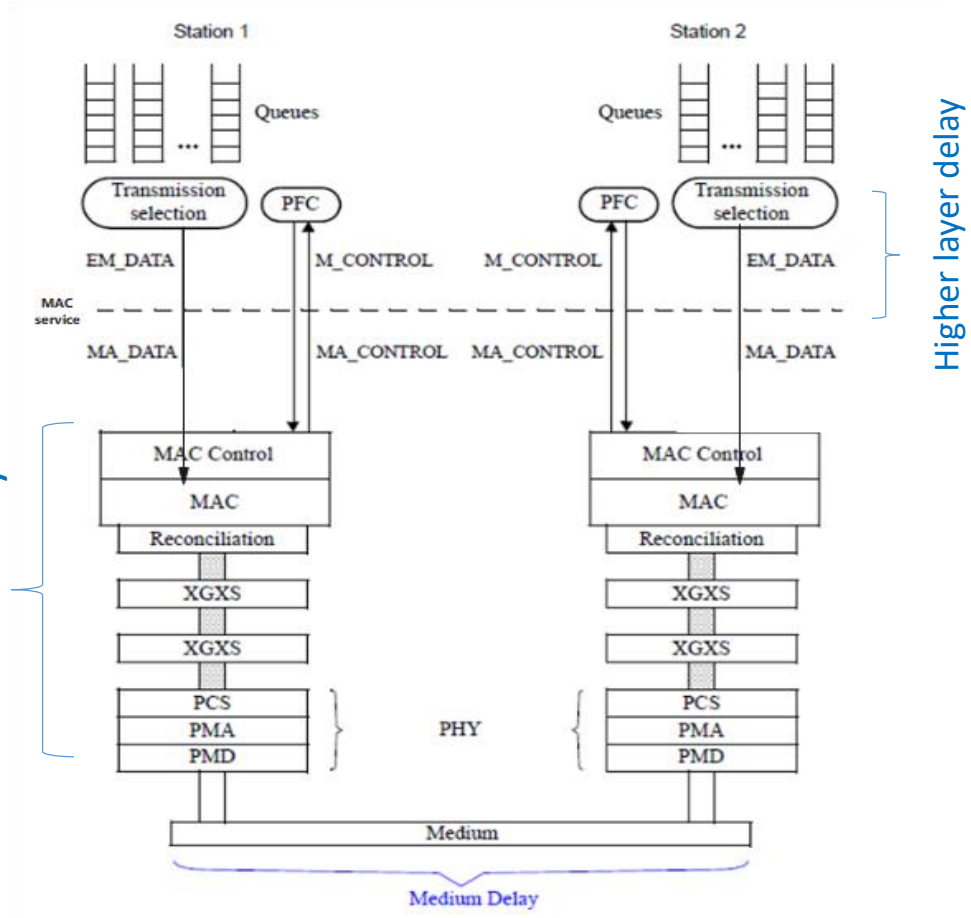
DCBX informational attributes: “Informational attributes are exchanged via LLDP without any participation in a DCBX state machine.”

Timestamp Point Clarification (1/2)

Roundtrip delay

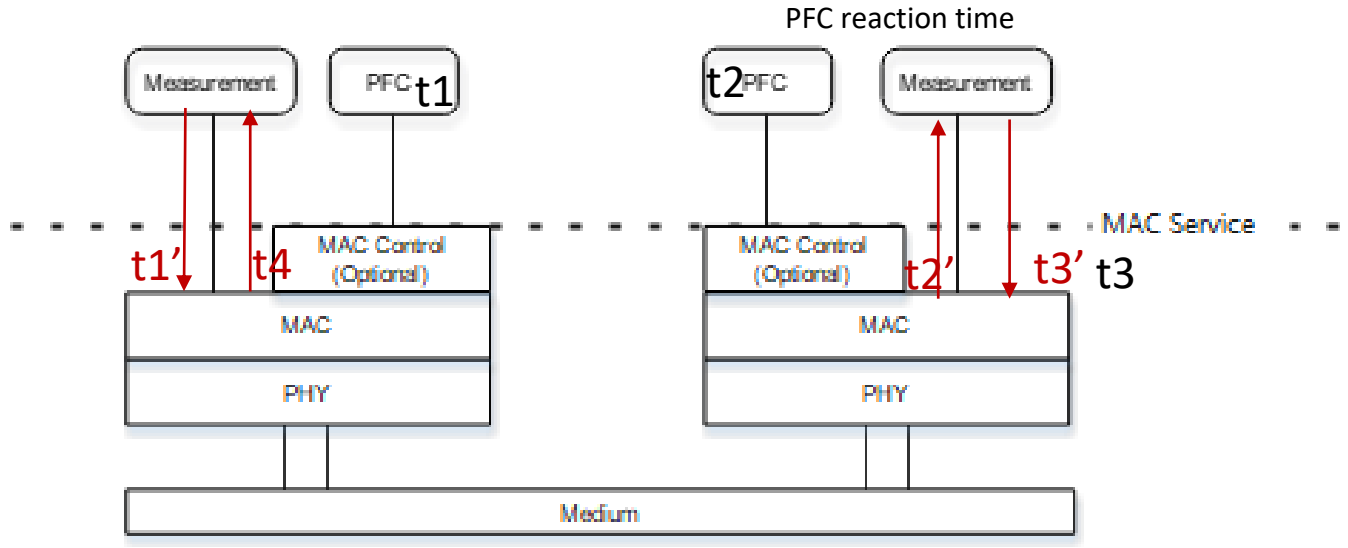
Delay Value = 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1

+ 2*(Max Frame) + (PFC Frame)



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

Without MACsec



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

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

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

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

Roundtrip delay = t4 – (t1' – (MAC control processing time))

– (t3 – (t2 + (MAC control processing time))

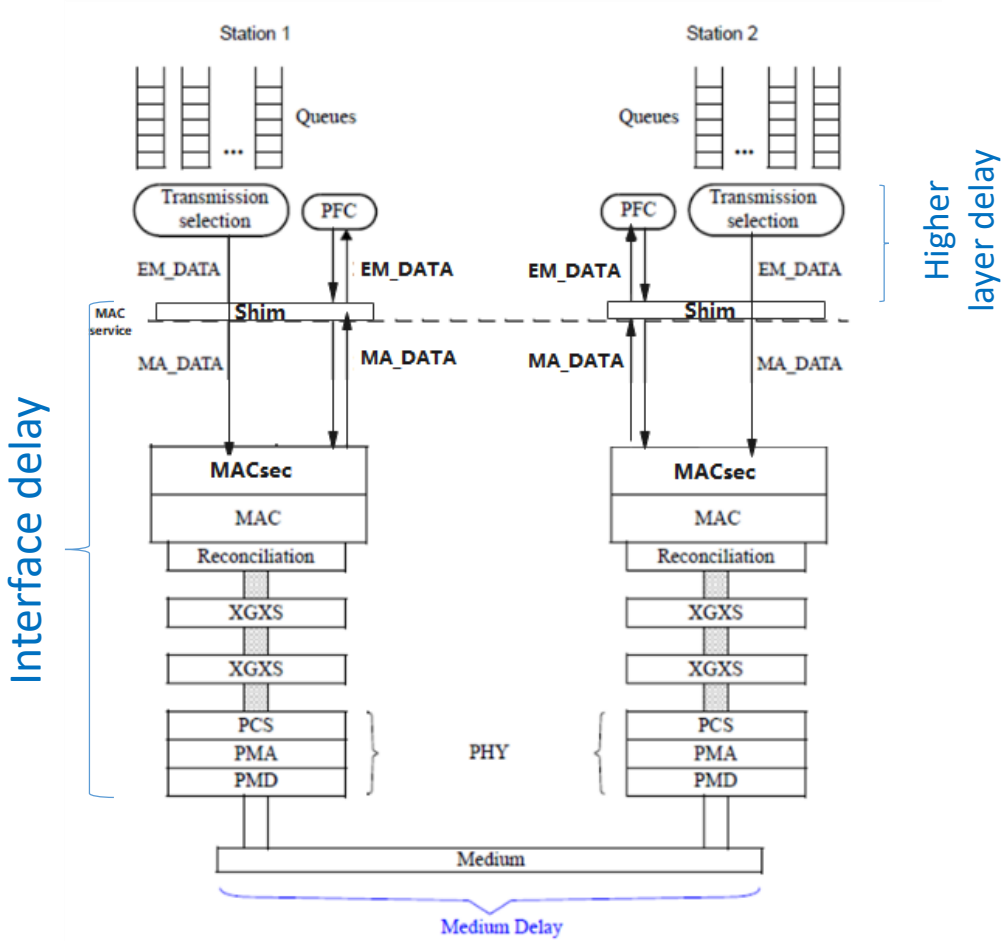
+ (PFC reaction time)

≈ t4 – t1 – (t3 – t2)

Timestamp Point Clarification (2/2)

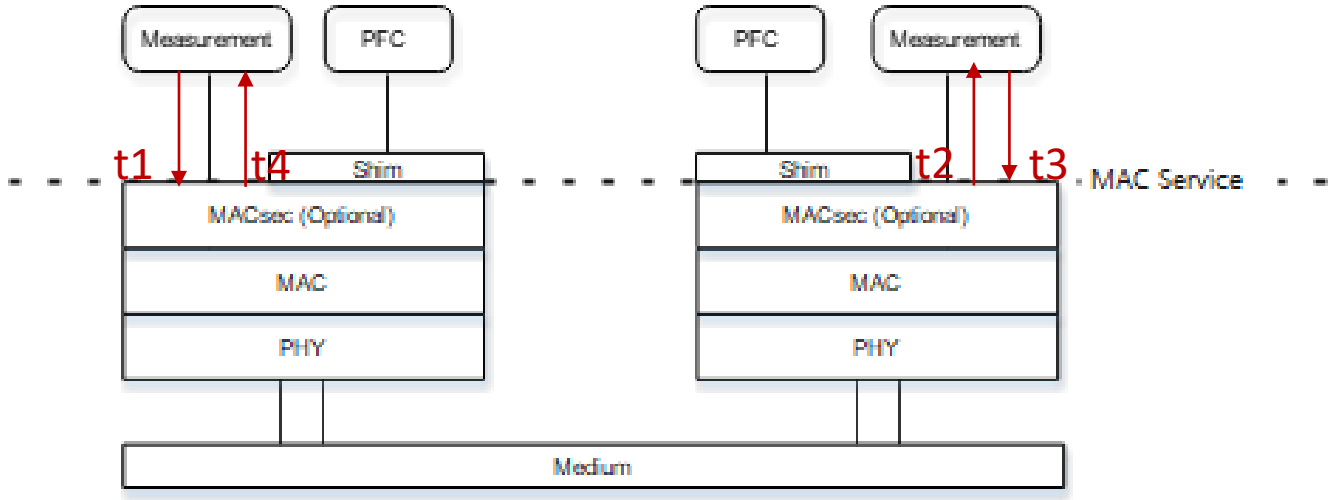
Roundtrip delay

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



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

With MACsec



t1: last bit of measurement request message passed to MAC service
t4: last bit of measurement response message passed from MAC service

t2: last bit of measurement request message passed from MAC service
t3: last bit of measurement response message passed to MAC service

Roundtrip delay = t4 – (t1 – (shim processing time))
– (t3 – (t2 + (shim processing time))
+ (PFC reaction time)
≈ t4 – t1 – (t3 – t2)

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