

Headroom Measurement Protocol Design

Lily Lv (Huawei)
Fei Chen (Huawei)

Liang Guo (CAICT)
Jie Li (CAICT)

To-Do List

- **Timestamp point clarification**
 - **Will (t3-t2) be impacted (variably) by queue delay?**
 - **further specify t1, t4**
- **Timestamp accuracy**
 - **What is the accuracy of t1, t4?**
- **Protocol design of request-response measurement**
 - **After DCBX or could be before DCBX?**
 - **Request-> request + response -> response ?**
- **Headroom and buffer allocation strategy**
 - **Headroom for each priority?**
- **Managed objects**
 - **The effort, implementation cost, and purpose of statistic gathering and retention requires careful consideration**

DONE:

- ✓ **Ethertype for Qdt**
 - Reuse Qcz (CI) Ethertype 89-A2
- ✓ **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

Done: Ethertype for Qdt

Reuse Qcz (CI) Ethertype 89-A2

Qcz definition

Table 47-1—Layer-2 CIM Encapsulation

	Octet	Length
PDU EtherType (89-A2)	1	2
Version	3	4 bits
Subtype	3	4 bits
CIM PDU	4	65-529

Subtype:

This field, 4 bits in length, shall be transmitted with the value 0 to indicate an encapsulated CIM PDU. The Subtype field occupies the least significant 4 bits of the first octet of the layer-2 CIM Encapsulation.

Table 47-4—CIM PDU

	Octet	Length
Version	1	4 bits
Reserved	1	3 bits
Add/Del	1	1 bit
destination_address	2	6
source_address	8	6
vlan_identifier	14	12 bits
Encapsulated MSDU length	16	2
Encapsulated MSDU	18	48-512

Qdt proposal

	Octet	Length
PDU Ethertype(89-A2)	1	2
Version	3	4 bits
Subtype	3	4 bits
Headroom Measurement PDU	4	65-529

Subtype 0, CIM

Subtype 1, Headroom Measurement Message

Question:

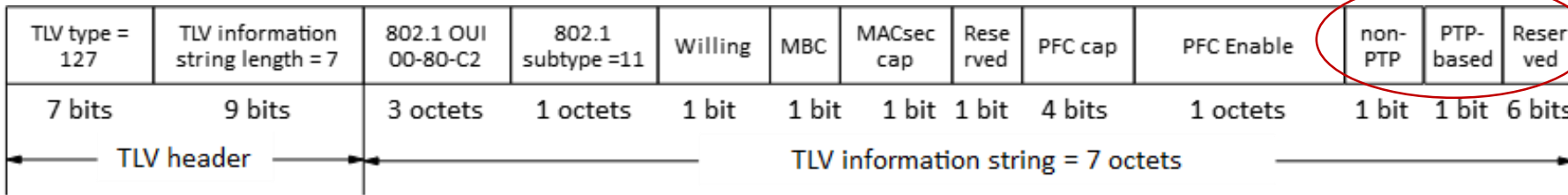
Is “65-529” too big for headroom measurement PDU?

Done: PFC Configuration TLV format design

- Proposal :

- PFC configuration TLV only includes 'capability'

Define priority of the 2 methods.

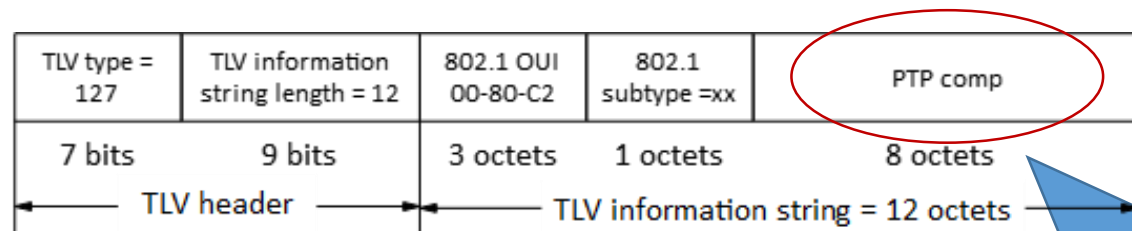


Each bit indicates one capability.

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**



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

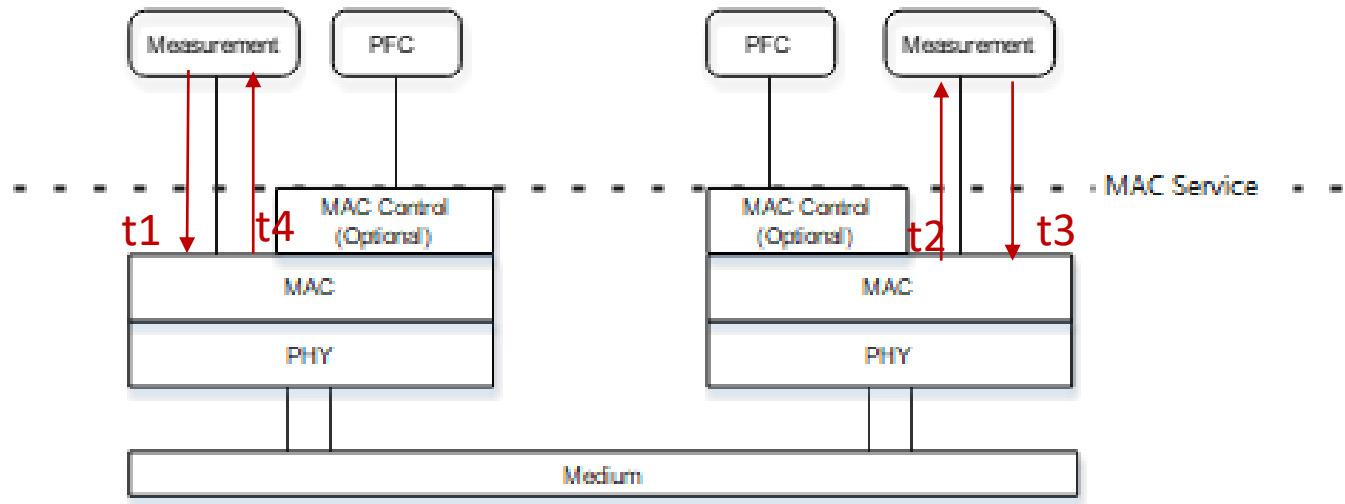
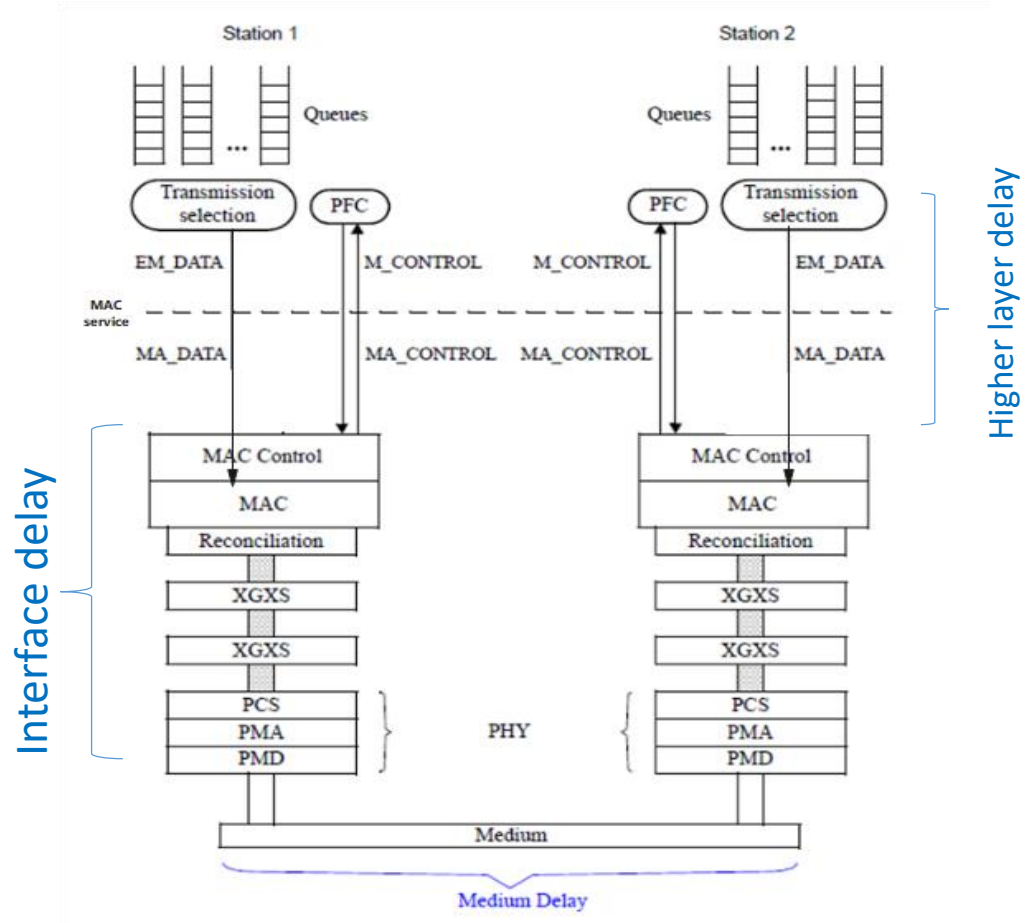
Compensation value for PTP-based measurement

Timestamp Point Clarification (1/2)

Without MACsec

Roundtrip delay

$$\text{Delay Value} = 2 * (\text{Cable Delay}) + \text{TXds1} + \text{RXds2} + \text{HDs2} + \text{TXds2} + \text{RXds1} + 2 * (\text{Max Frame}) + (\text{PFC Frame})$$



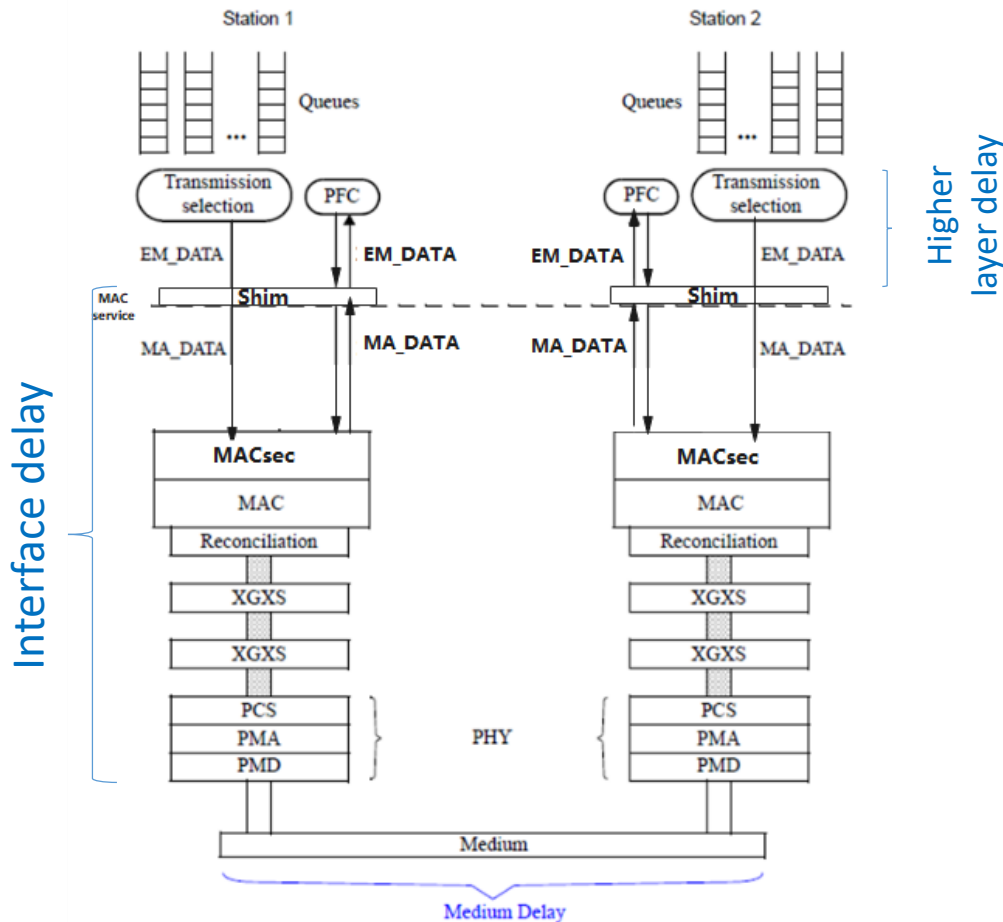
- t1:** last bit of measurement request message passed to MAC service
- t4:** last bit of measurement request message passed from MAC service
- t2:** last bit of measurement request message passed from MAC service
- t3:** last bit of measurement request message passed to MAC service

$$\begin{aligned} \text{Roundtrip delay} &= t4 - (t1 - (\text{MAC control processing time})) \\ &\quad - (t3 - (t2 + (\text{MAC control processing time}))) \\ &\quad + (\text{PFC reaction time}) \\ &\approx t4 - t1 - (t3 - t2) \end{aligned}$$

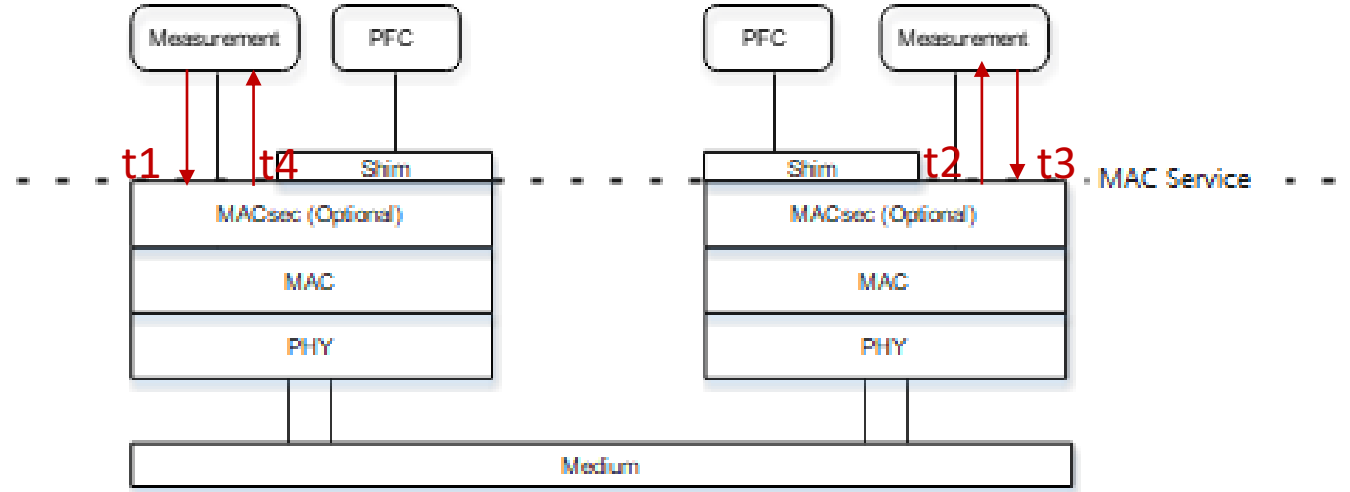
Timestamp Point Clarification (2/2)

Roundtrip delay

$$\text{Delay Value} = 2 * (\text{Cable Delay}) + \text{TXds1} + \text{RXds2} + \text{HDs2} + \text{TXds2} + \text{RXds1} + 2 * (\text{Max Frame}) + (\text{PFC Frame})$$



With MACsec



- 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

$$\begin{aligned} \text{Roundtrip delay} &= t4 - (t1 - (\text{shim processing time})) \\ &\quad - (t3 - (t2 + (\text{shim processing time}))) \\ &\quad + (\text{PFC reaction time}) \\ &\approx t4 - t1 - (t3 - t2) \end{aligned}$$

Timestamp Accuracy

- We do not require peer nodes to be synchronized.
- The longer the cable length is, the higher tolerance of timestamp inaccuracy is.

	Fixed Delay	Internal Processing Delay (802.3, no MACsec)	Medium Delay	Headroom (t4-t1)	t4-t1 mismatch		
100G,500m	32992	203776	500000	92KB	10 ns	0.125KB	0.1%
					100 ns	1.25KB	1%
100G,100m	32992	203776	100000	42KB	10 ns	0.125KB	0.3%
					100 ns	1.25KB	3%
100G,20m	32992	203776	20000	32KB	10 ns	0.125KB	0.4%
					100 ns	1.25KB	4%

- The factors impacting timestamp accuracy
 - Local clock frequency drift
 - Captured timestamp point

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.**

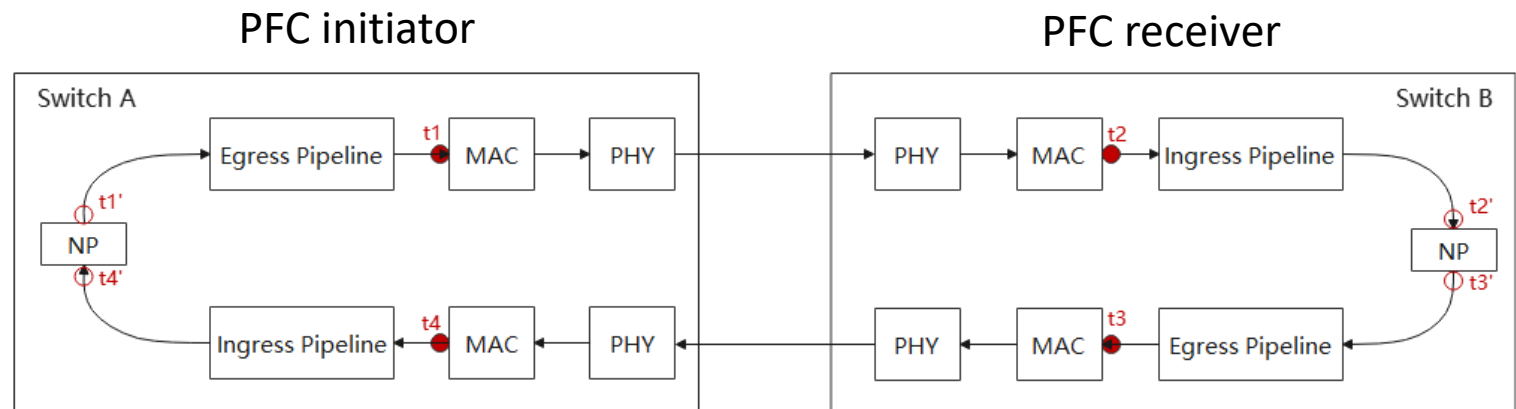
- Captured timestamp point analysis

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:

$$\begin{aligned}t1 &= t1' + \text{ePP delay} \\t4 &= t4' - \text{iPP delay} \\t1 &= t1' + \text{ePP delay} \\t4 &= t4' - \text{iPP delay}\end{aligned}$$



Protocol Design of Request-Response Measurement

Protocol design consideration:

- Avoid to design a complex new protocol
- Keep the fixed and same size of all measurement packets to increase accuracy
- Add less state on switch to decrease implementation complexity

Protocol Design of Request-Response Measurement

Option 1:

	Octet	Length
PDU Ethertype(89-A2)	1	2
Version	3	4 bits
Subtype (0001)	3	4 bits
Version	4	4 bits
Reserved	4	2 bits
Req/Resp	4	2 bits
Timestamp 1 (t1)	5	8
Timestamp 2 (t2)	13	8
Timestamp 3 (t3)	21	8
Timestamp 4 (t4)	29	8

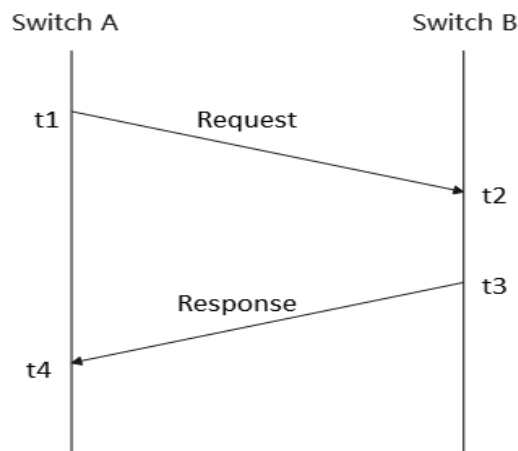
Headroom
Measurement
PDU

Packet design

- Req/Resp: 2 types of measurement packet, request and response.

Procedure:

- Switch A sends Request packet.
 - Triggering condition could be port status/configuration changes.
 - Request packet includes t1. Other timestamp fields are NULL.
- Switch B generate Response packet after receiving request packet.
 - Response packet includes t1,t2 and t3. t4 field is NULL.
- Switch B sends response message back to switch A.
- Switch A receives response message, capturing timestamp t4
- Switch A calculates roundtrip measurement by $t4 - t1 - (t3 - t2)$



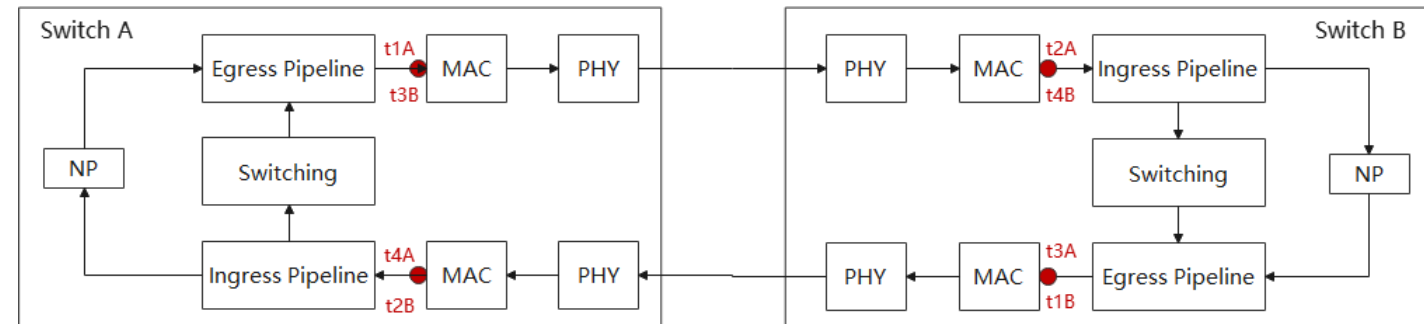
Protocol Design of Request-Response Measurement

Option 2:

	Octet	Length
PDU Ethertype(89-A2)	1	2
Version	3	4 bits
Subtype (0001)	3	4 bits
Version	4	4 bits
Reserved	4	2 bits
Timestamp 1 (t1A)	5	8
Timestamp 2 (t2A)	13	8
Timestamp 3 (t3A)	21	8
Timestamp 4 (t4A)	29	8
Timestamp 5 (t1B)	37	8
Timestamp 6 (t2B)	45	8
Timestamp 7 (t3B)	53	8
Timestamp 8 (t4B)	61	8

Headroom
Measurement
PDU

Packet design



Procedure:

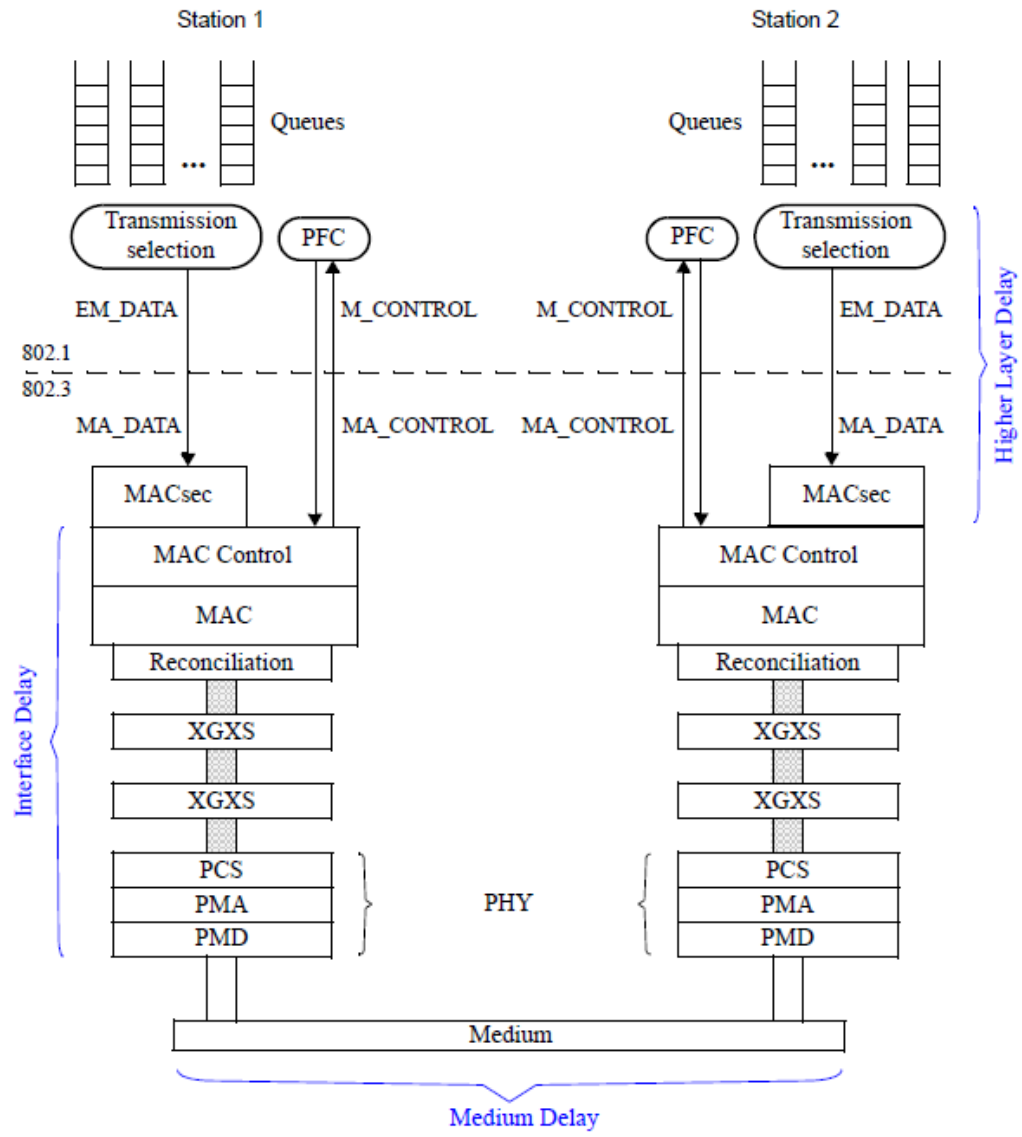
- The difference from option 1 is that, switch B send back a new generated measurement packet with t2A,t3A as well as t1B.
- After switch A receiving the measurement packet, it generates another packet filling in t2B and t3B.

* There might be smarter design to make the procedure work with smaller size measurement packet, but that will add on complexity of implementation.

Protocol Design of Request-Response Measurement

	Pros	Cons
Option 1 (preferred)	Simple logic, easy to implement	Potential waste of bandwidth
Option 2 (not preferred)	Potential benefit on saving bandwidth	Complex state machine design

Thanks



802.1Q Figure N-2—Delay model