Headroom Measurement Protocol Design

Lily Lv (Huawei)

Fei Chen (Huawei)

Liang Guo (CAICT)

Jie Li (CAICT)

To-Do List

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

- **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 ?**
- 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

Qcz definition

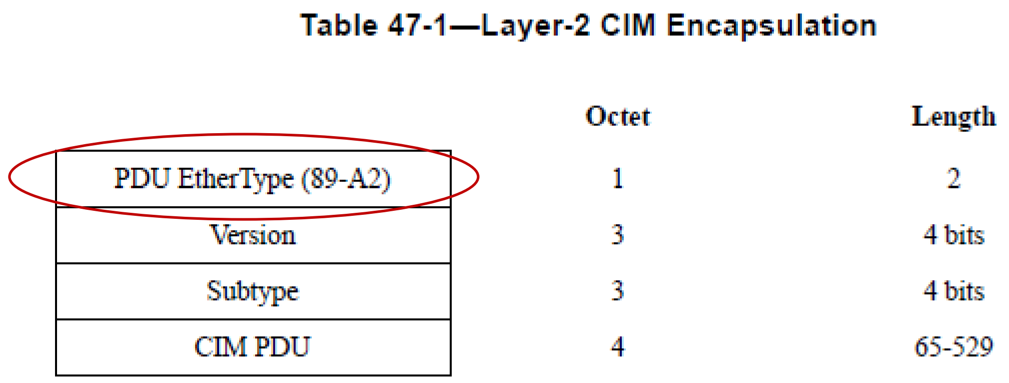


Table 47-1—Layer-2 CIM Encapsulation

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

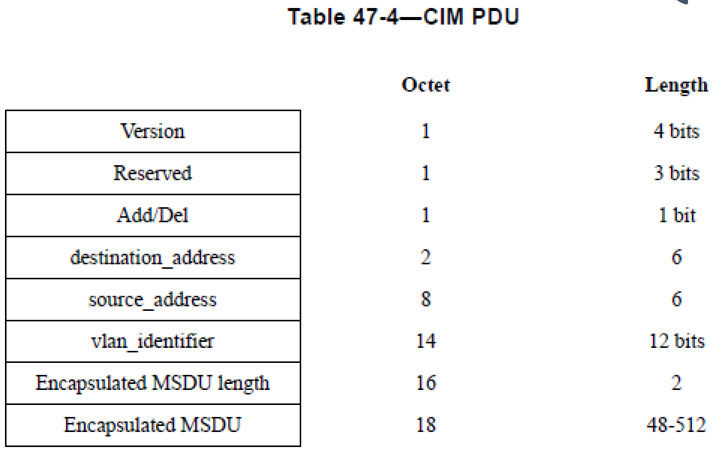


Table 47-4—CIM PDU

	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

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.

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

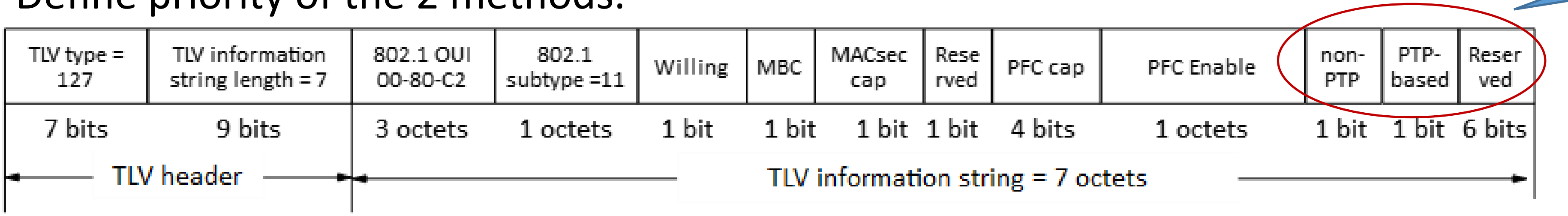
Question:

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.



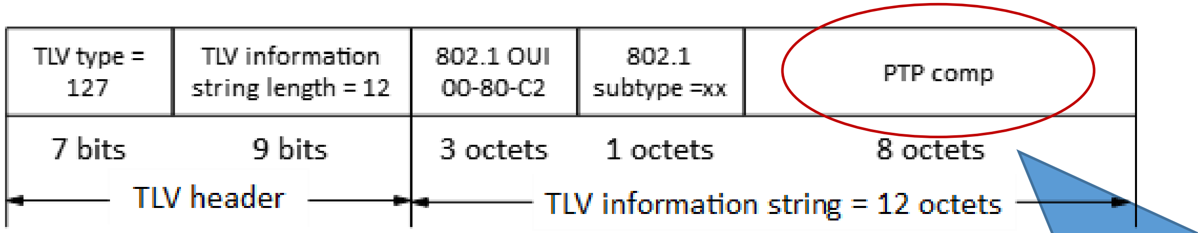
TLV type = 127	TLV information string length = 7	802.1 OUI 00-80-C2	802.1 subtype =11	Willing	MBC	MACsec cap	Reserved	PFC cap	PFC Enable	non-PTP	PTP-based	Reserved
7 bits	9 bits	3 octets	1 octets	1 bit	1 bit	1 bit	1 bit	4 bits	1 octets	1 bit	1 bit	6 bits
TLV header		TLV information string = 7 octets										

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**



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

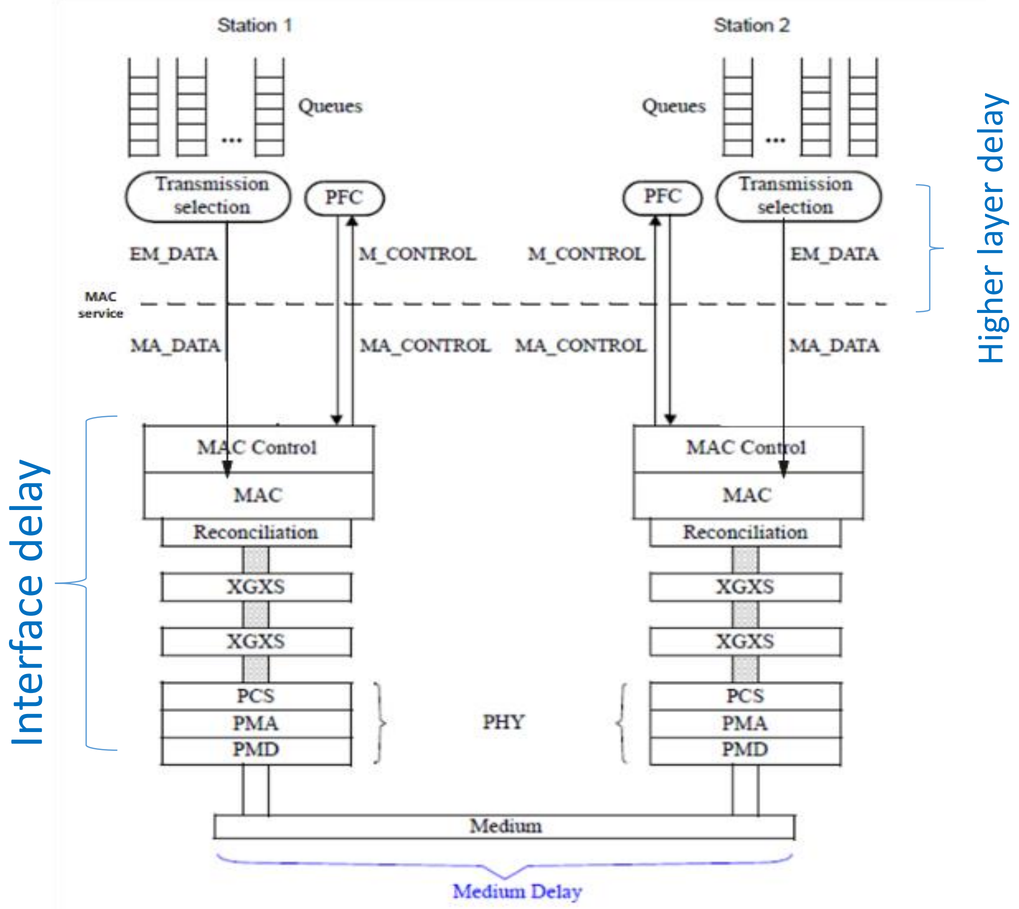
DCBX informational attributes: "Informational attributes are exchanged via LLDP without any participation in a DCBX state machine."

Timestamp Point Clarification (1/2)

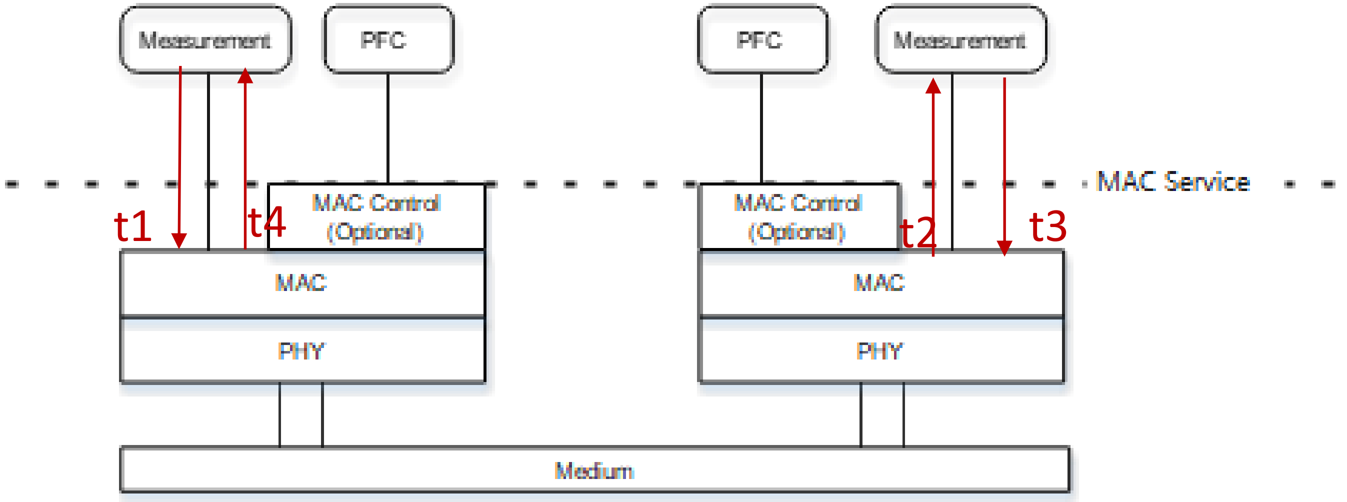
Without MACsec

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



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

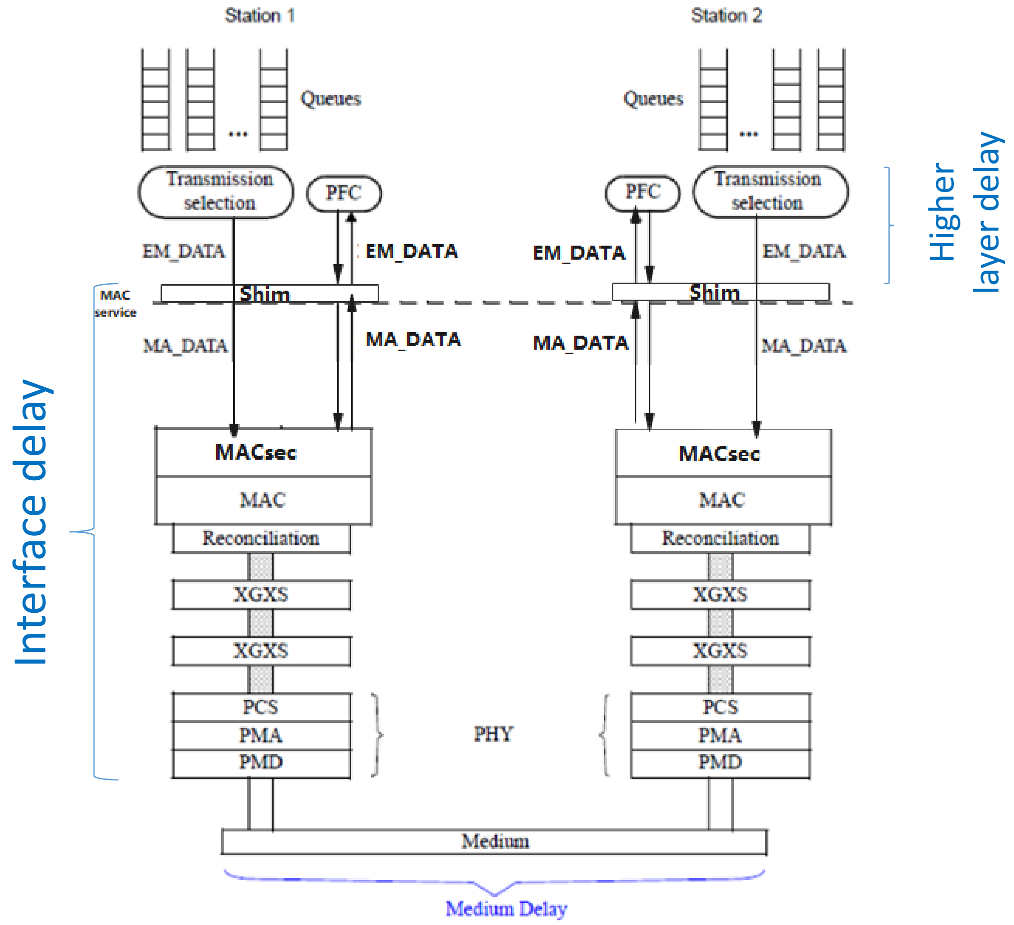
Roundtrip delay = t4 – (t1 – (MAC control processing time))
– (t3 – (t2 + (MAC control processing time))
+ (PFC reaction time)
≈ **t4 – t1 – (t3 – t2)**

Timestamp Point Clarification (2/2)

Roundtrip delay

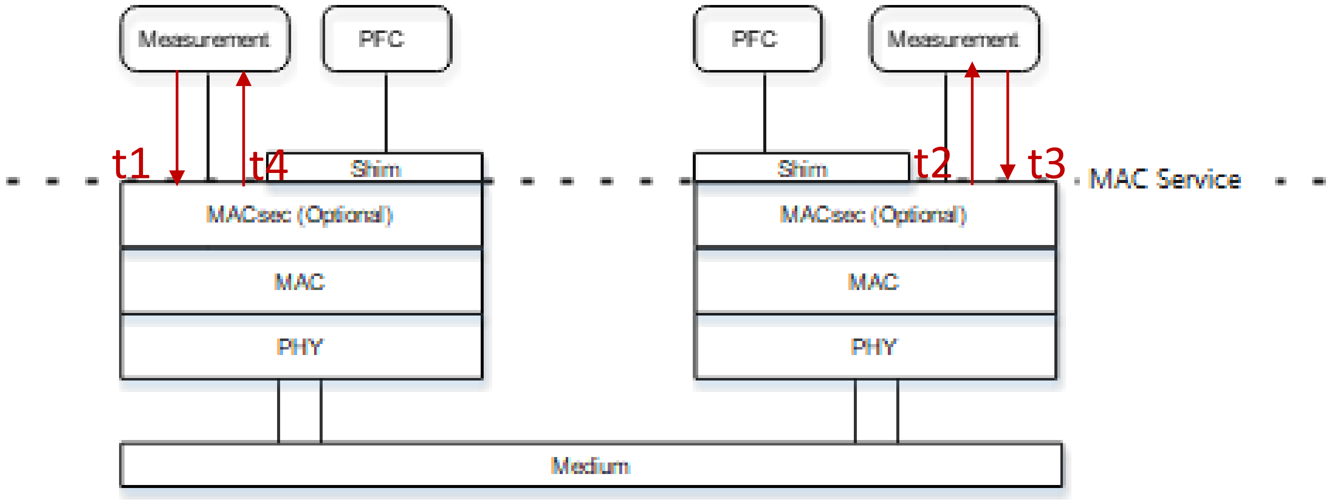
Delay Value = 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1

+ 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

- 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

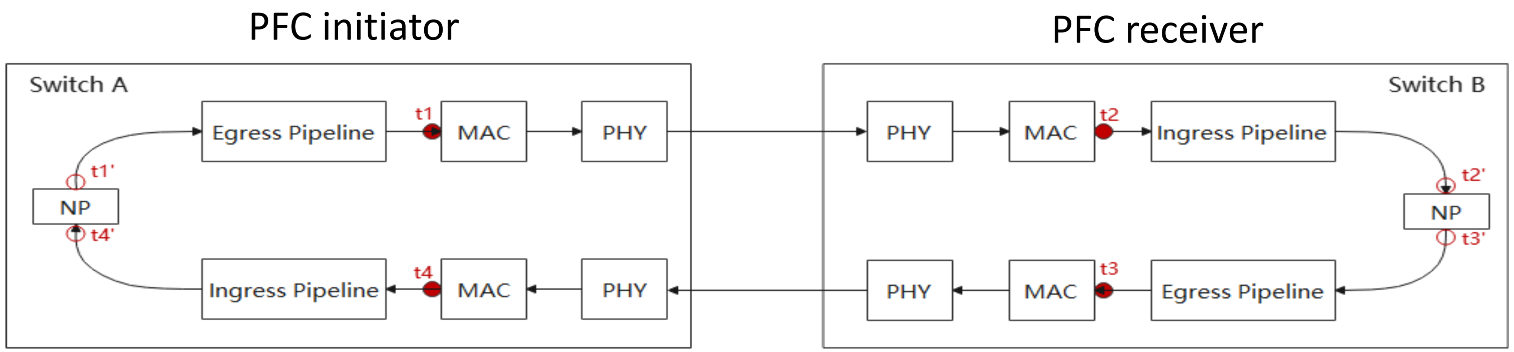
- Local clock frequency drift analysis

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

Implementation example:

t1 = t1’ + ePP delay
t4 = t4’ – iPP delay
t1 = t1’ + ePP delay
t4 = t4’ – iPP delay



Test:

1) 10 meters case: RTT_{10m}= (t4’ - t1’) – (t3’ – t2’) = 20,200 ns
2) 500 meters case: RTT_{500m} = (t4’ - t1’) – (t3’ – t2’) = 25,055 ns

Result：

RTT_{500m} - RTT_{10m} = 4,855ns → 4.954ns/m ≈ 5ns/m (fiber propagation latency)

Protocol Design of Request-Response Measurement

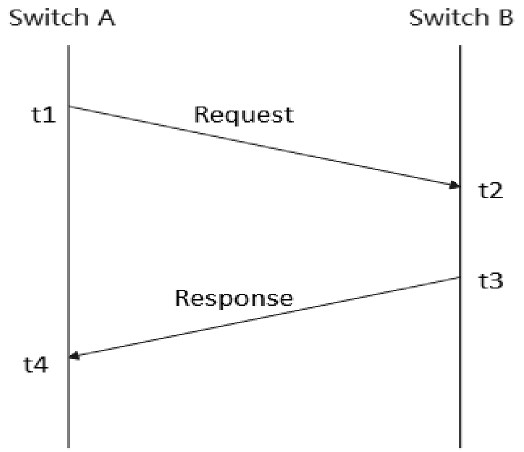
Protocol design consideration:

- Avoid to design a complex new protocol
- Keep the fixed and same size of all measurement messages 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
Headroom Measurement PDU	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



Packet design

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

Procedure:

- Switch A sends Request message.
  - 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)

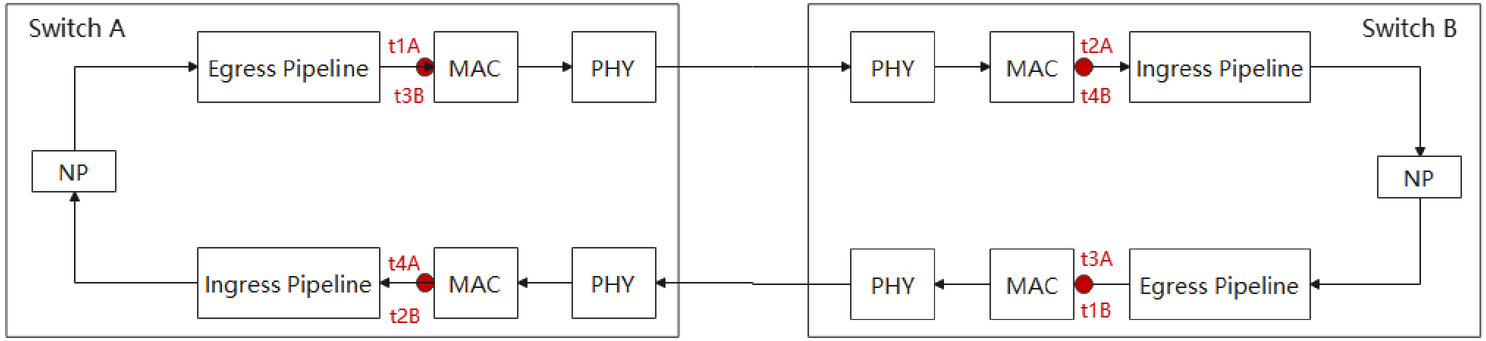
Protocol Design of Request-Response Measurement

Option 2:

Headroom Measurement PDU

	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

Packet design



Procedure:

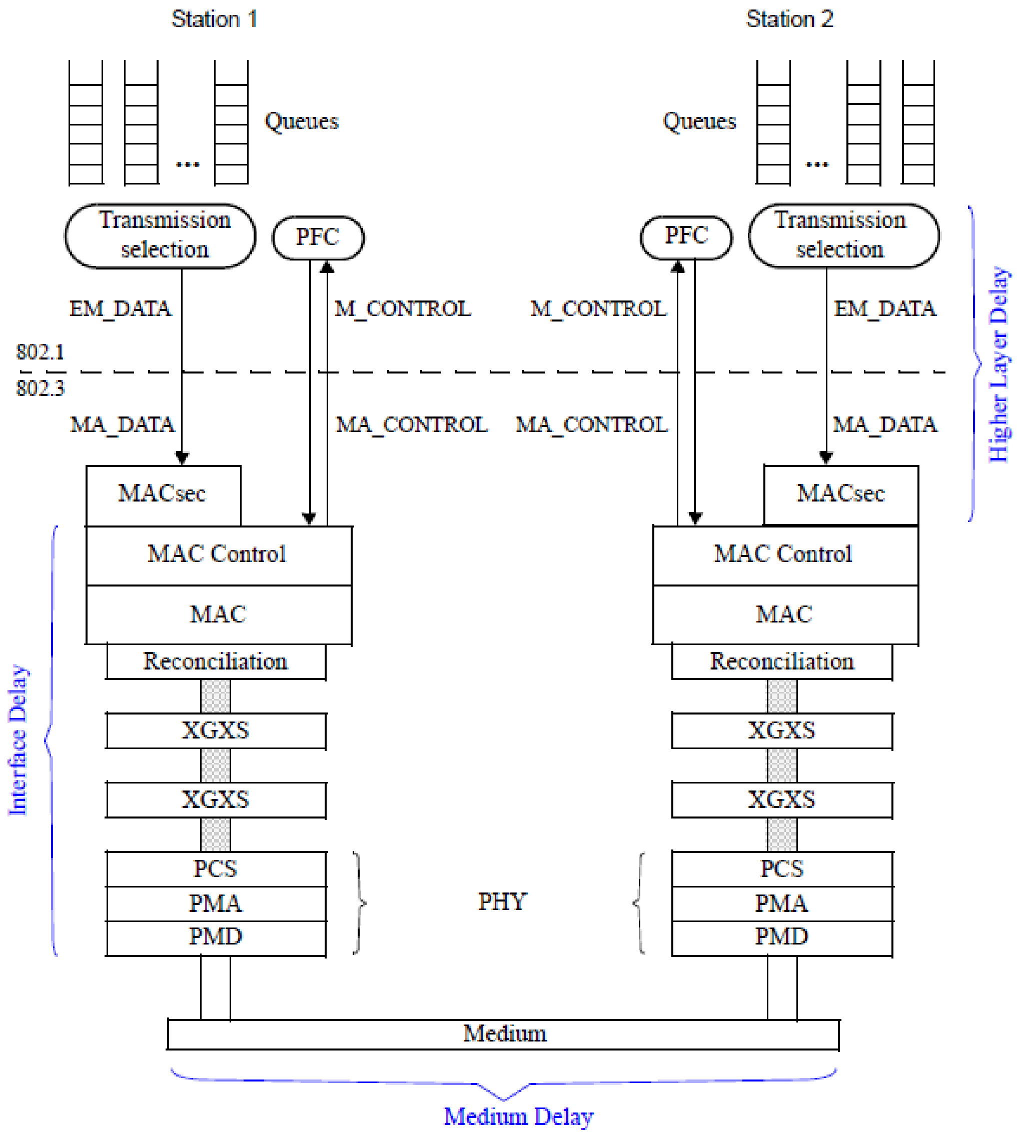
- The difference from option 1 is that, switch B send back a new generated measurement packet with t2A,t3A as well as t1B.
- After switch A receiving the measurement packet, it generates another packet filling in t2B and t3B.

* 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

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:

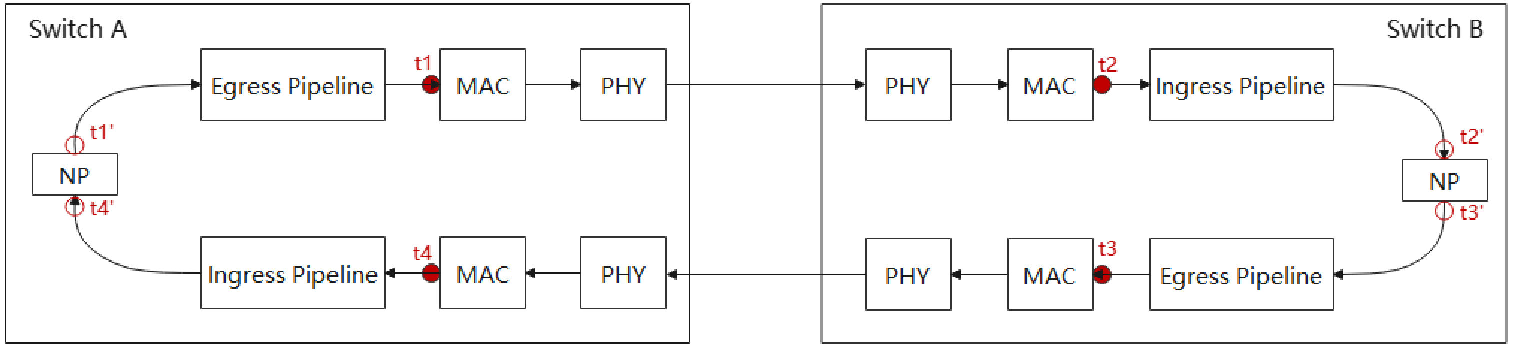
t1: last bit of measurement request message passed to MAC service
t4: last bit of measurement request message passed from MAC service
t2: last bit of measurement request message passed from MAC service
t3: last bit of measurement request message passed to MAC service

Implementation example:

t1 = t1’ + ePP delay
t4 = t4’ – iPP delay
t1 = t1’ + ePP delay
t4 = t4’ – iPP delay

PFC initiator

PFC receiver


Switch A
Egress Pipeline
t1
MAC
PHY
t1'
NP
t4'
Ingress Pipeline
t4
MAC
PHY
Switch B
PHY
MAC
t2
Ingress Pipeline
t2'
NP
t3'
PHY
MAC
t3
Egress Pipeline