Annex N

(informative)

Buffer requirements for PFC

N.1 Overview

To ensure that data frames are not lost due to lack of receive buffer space, receivers must ensure that a PFC M_CONTROL.request primitive is invoked while there is sufficient receive buffer to absorb the data that can continue to be received during the time needed by the remote system to react to the PFC operation. The PFC headroom (see 36.1.1) is the minimum buffer size that needs to remain available on receiver. It helps implementation to allocate buffer for PFC-enabled priorities. But the ~~The~~ precise calculation of this buffer requirement and buffer allocation are ~~is~~ highly implementation dependent. ~~This annex provides an example of how it can be calculated based on a hypothetical delay model. Setting the PFCLinkDelayAllowance (see 12.22.6) to less than the round-trip delay value can result in frames loss.~~

This annex explains delay model of PFC headroom, and provides an example of buffer allocation based on the PFC headroom calculation.

~~Figure N-1 provides an high-level view of the various delays to consider:~~

a) ~~Processing and queuing delay of the PFC request~~
b) ~~Propagation delay of the PFC frame across the media~~
c) ~~Response time to the PFC indication at the far end~~
d) ~~Propagation delay across the media on the return path~~

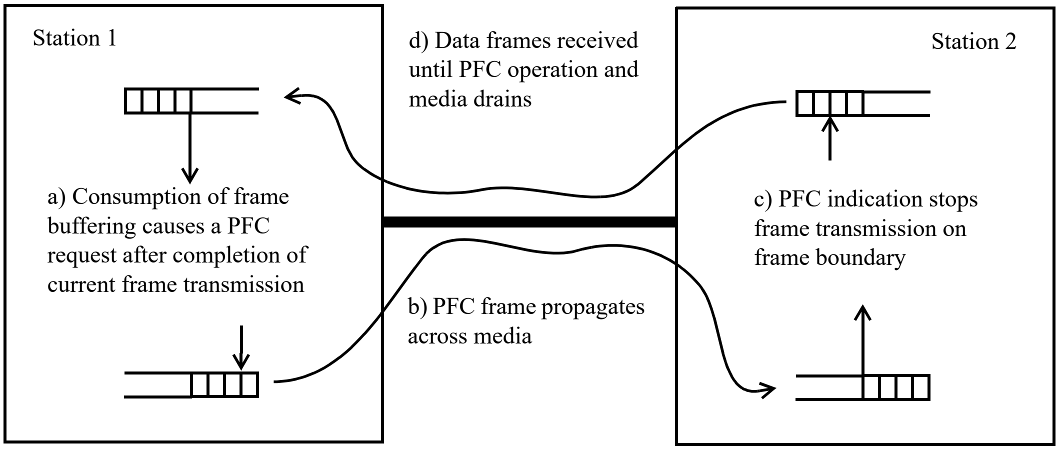
Station 1
d) Data frames received until PFC operation and media drains
Station 2
a) Consumption of frame buffering causes a PFC request after completion of current frame transmission
c) PFC indication stops frame transmission on frame boundary
b) PFC frame propagates across media

Figure N-1—PFC delays

N.2 Delay model of PFC headroom

PFC headroom calculation considers various delays accumulated, from item a) to l) (see 36.1.1)

PFC frame transmission in PFC initiator station B:

a) B’s reception processing to calculate the remaining buffering following frame receipt.
b) B’s PFC Initiator to initiate PFC following that buffering calculation and PFC frame encoded ready for transmission.
c) Any prior in-progress frame transmission by B (possibly of a maximum sized frame) to complete.
d) First bit of PFC frame sent to MAC service.
e) Last bit of PFC frame sent on the physical link.

PFC frame transmission across link from B to A:

f) The link delay for transmission from B to A.

PFC frame reception in PFC receiver station A (including PFC taking action):

g) PFC frame reception since the last bit of PFC frame received on link,including frame validation, by A’s interface stack.
h) A’s PFC Receiver to decode the PFC frame and halt transmission selection for specified priorities.

User data transmission in PFC receiver station A:

i) Any in-progress frame transmission by A (possibly of a maximum sized frame) to complete.
j) Last frame sent on the physical link.

User data transmission across link from A to B:

k) The link delay for transmission from A to B.

User data reception in PFC initiator station A:

l) Reception delays associated with B’s interface stack, reception processing, and buffering.

Figure N-2 shows how to model the various delays between two stations connected by a point-to-point full-duplex IEEE 802.3 link.

~~The main delay components shown in Figure N-2 are as follows:~~

When calculating PFC headroom, the delays are classified into 3 main components shown in Figure N-1, as below:

a) ~~**PFC transmission delay:** the time needed by a station to request transmission of a PFC frame after a PFC M_CONTROL.request has been invoked (e.g., because a maximum length data frame can be transmitted).~~
b) ~~**Interface Delay (ID):** the sum of MAC Control, MAC/RS, PCS, PMA, and PMD delays, including item e), g), k). Interface Delay is dependent on the MAC and physical layer in use.~~
c) ~~**Cable Delay:** the number of bits in flight stored in the transmission medium. This delay value is dependent on the selected technology and on the medium length.~~
d) ~~**Higher Layer Delay (HD):** the time needed for a queue to go into paused state after the reception of a PFC M_CONTROL.indication that paused its priority. A substantial portion of this delay component is implementation specific.~~

a) **Internal processing delay (ID):** the time spent on frame processing within PFC initiator station and PFC receiver station, such as interface stack delay, and buffering delay etc, queue status change delay, assuming no prior in-progress frame transmission. ID includes item a) b) d) e) g) h) j) l).
b) **Link delay (LD):** the time spent on physical link between PFC initiator station and PFC receiver station. LD includes item f) and item k).
c) **Worst-case delay (WD):** the additional time needed for a maximum sized frame transmission before the PFC frame transmission at PFC initiator station, and the additional time needed for a

maximum sized frame transmission after the PFC frame taking effect at PFC receiver station. WD includes item c) and item i). It is not shown in Figure N-2, but shall be considered by PFC headroom calculation.

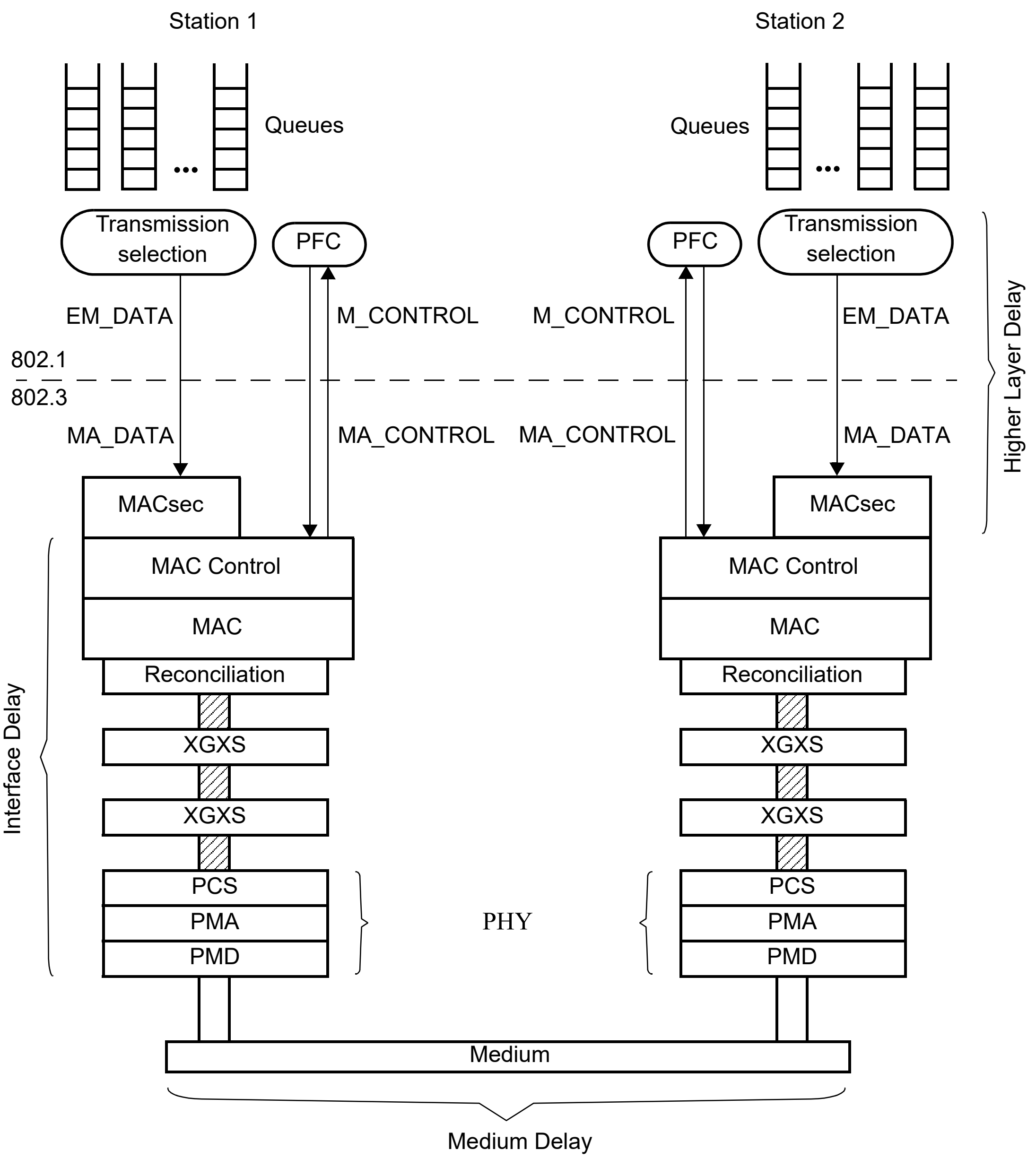

Figure N-2—~~Delay model~~

The total delay value of PFC headroom is the sum of ID, LD and WD.

When calculating PFC headroom using link delays (see 36.8.1), 1588 measures LD. ID is based on peer notification and local knowledge. WD depends on size of maximum frame and MACsec capability of user data.

When calculating PFC headroom using measurement protocol (see 36.9), ID + LD is obtained by running the protocol. Then by adding WD, the total delay is got. Keeping the measurement requests and responses the same MACsec capability as PFC frames increases the measurement accuracy.

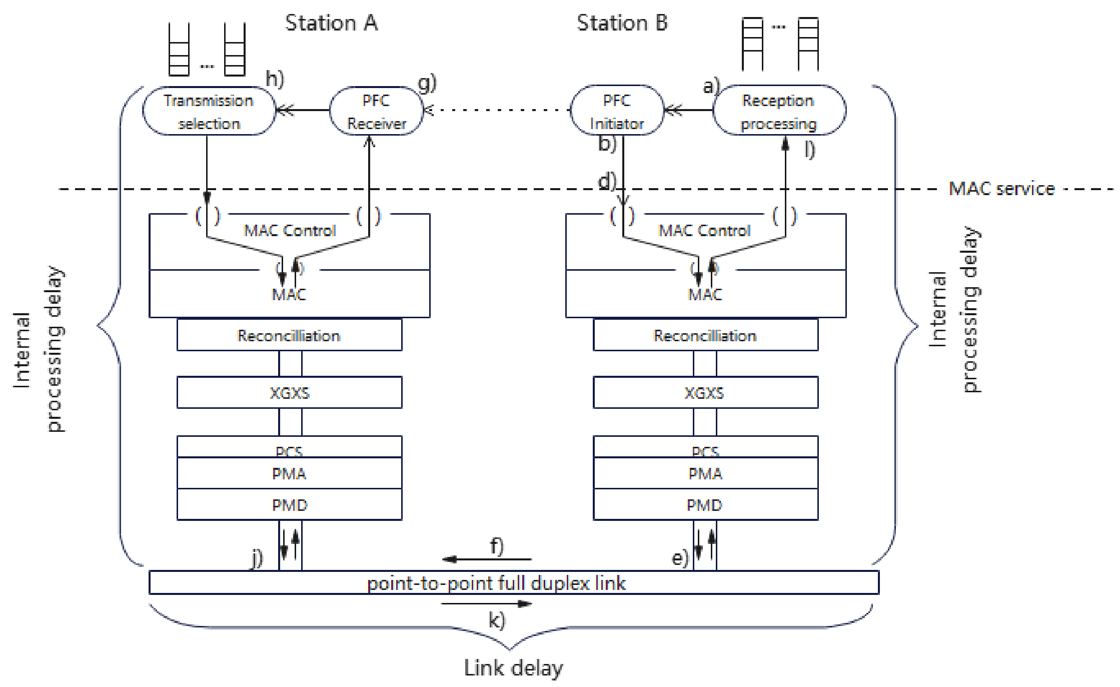

Figure N-3—Delay model

Figure N-4 shows a possible worst-case delay example where MACsec is disabled for PFC frame and user data.

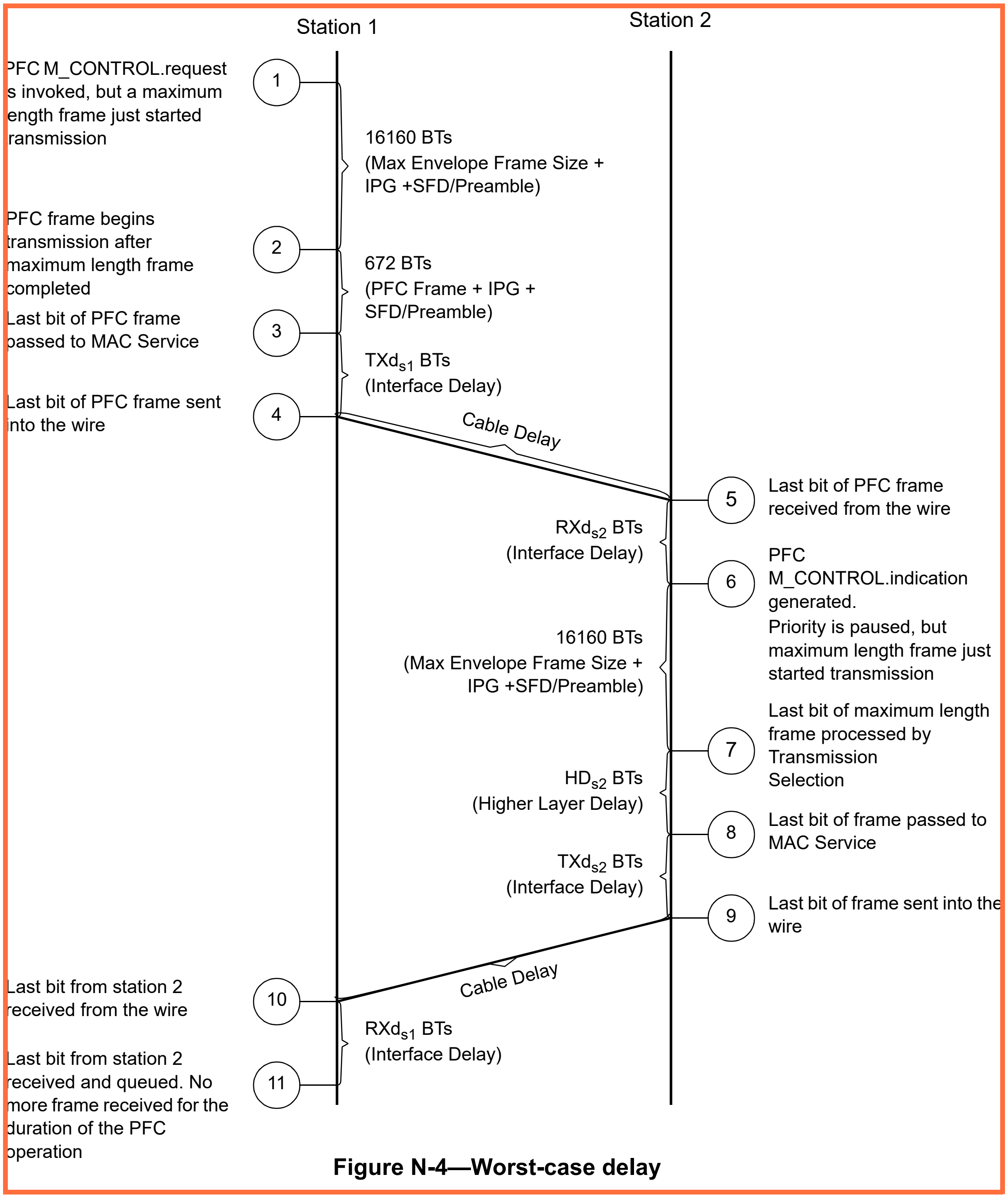
Station 1
Station 2
PFC M_CONTROL.request s invoked, but a maximum ength frame just started ransmission
1
16160 BTs
(Max Envelope Frame Size + IPG +SFD/Preamble)
PFC frame begins transmission after maximum length frame completed
2
672 BTs
(PFC Frame + IPG + SFD/Preamble)
Last bit of PFC frame passed to MAC Service
3
TXd$_{s1}$ BTs
(Interface Delay)
Last bit of PFC frame sent into the wire
4
Cable Delay
5
Last bit of PFC frame received from the wire
RXd$_{s2}$ BTs
(Interface Delay)
6
PFC M_CONTROL.indication generated.
Priority is paused, but maximum length frame just started transmission
16160 BTs
(Max Envelope Frame Size + IPG +SFD/Preamble)
7
Last bit of maximum length frame processed by Transmission Selection
HD$_{s2}$ BTs
(Higher Layer Delay)
8
Last bit of frame passed to MAC Service
TXd$_{s2}$ BTs
(Interface Delay)
9
Last bit of frame sent into the wire
Cable Delay
Last bit from station 2 received from the wire
10
RXd$_{s1}$ BTs
(Interface Delay)
Last bit from station 2 received and queued. No more frame received for the duration of the PFC operation
11

Figure N-4—Worst-case delay

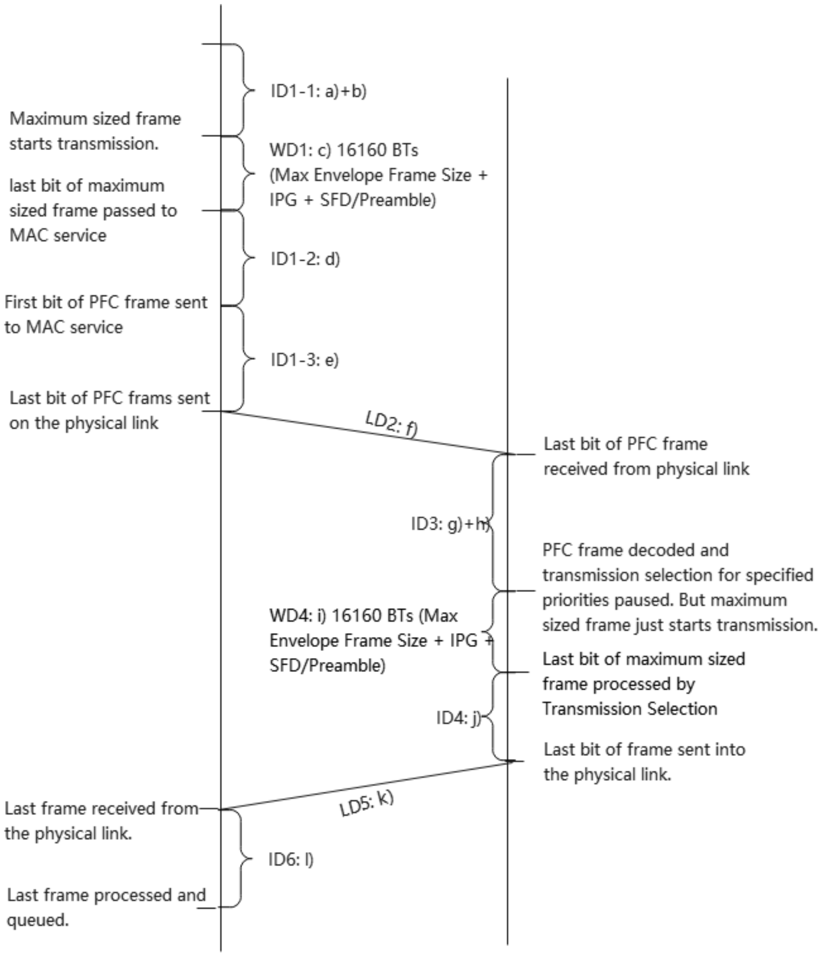
ID1-1: a)+b)
Maximum sized frame starts transmission.
WD1: c) 16160 BTs (Max Envelope Frame Size + IPG + SFD/Preamble)
last bit of maximum sized frame passed to MAC service
ID1-2: d)
First bit of PFC frame sent to MAC service
ID1-3: e)
Last bit of PFC frams sent on the physical link
LD2: f)
Last bit of PFC frame received from physical link
ID3: g)+h)
PFC frame decoded and transmission selection for specified priorities paused. But maximum sized frame just starts transmission.
WD4: i) 16160 BTs (Max Envelope Frame Size + IPG + SFD/Preamble)
Last bit of maximum sized frame processed by Transmission Selection
ID4: j)
Last bit of frame sent into the physical link.
LD5: k)
Last frame received from the physical link.
ID6: l)
Last frame processed and queued.

Figure N-5—Worst-case delay

The total Delay Value (DV) is the sum of all delays shown in Figure N-4:

~~$DV = 2 \times (Max\ Frame) + (PFC\ Frame) + 2 \times (Cable\ Delay) + TXd_{s1} + RXd_{s2} + HD_{s2} + TXd_{s2} + RXd_{s1}$~~

$DV = ID_1 + WD_1 + LD_2 + ID_3 + WD_4 + ID_4 + LD_5 + ID_6$

$ID_1 = ID_{1-1} + ID_{1-2} + ID_{1-3}$

It is the round-trip delay from PFC initiation to last frame reception and buffered, plus 2 maximum sized frames represented by bit times.

~~For any given station the Interface Delay includes both transmit and receive paths (i.e., ID = TXd + RXd). Therefore:~~

~~$DV = 2 \times (Max\ Frame) + (PFC\ Frame) + 2 \times (Cable\ Delay) + ID_{s1} + ID_{s2} + HD_{s2}$~~

~~Usually the peer stations connected by a point-to-point link use the same technology, therefore $ID_{s1} = ID_{s2}$:~~

~~$DV = 2 \times (Max\ Frame) + (PFC\ Frame) + 2 \times (Cable\ Delay) + 2 \times ID + HD_{s2}$~~

N.3 ~~Interface Delay~~ Internal Processing Delay

The Internal Processing Delay is implementation dependent. It comprises frame processing delays above MAC service which is between MAC control client and transmission selection, as well as MAC and PHY layer interface delays.

Example of processing delays above MAC service are MACsec and entering pause state delays.

For link speeds of up to 10Gb/s, MACsec constrains each of the transmit delay and the receive delay to a maximum of 19 360 bit times (see 36.1.3.3).

This standard defines a queue shall go into paused state in no more than 614.4 ns (see 36.1.3.3). This delay is equivalent to 6144 bit times at the speed of 10Gb/s.

IEEE 802.3 defines different interfaces delay constraints for different MAC and PHY. Table N-1 shows the delay constraints for some IEEE 802.3 interfaces.

~~The Interface Delay comprises all delay components below the MAC Control Client, excluding the cable delay. Table N-1 shows the Interface Delay constraints for some IEEE 802.3 interfaces.~~

Table N-1—IEEE 802.3 Interface Delays

Sublayer	Maximum RTT (bit times)	Maximum RTT (pause quanta)	Reference (subclause of IEEE Std 802.3-2018 [B14])
10G MAC Control, MAC, and RS	8192	16	46.1.4
XGXS and XAUI	2048	4	48.5
10GBASE-X PCS	2048	4	49.2.15
10GBASE-R PCS	3584	7	50.3.7
LX4 PMD	512	1	53.2
CX4 PMD	512	1	54.3
Serial PMA and PMD	512	1	52.2
10GBASE-T	25 600	50	55.11

N.4 ~~Cable Delay~~ Link Delay

The ~~Cable~~ Link Delay is the propagation delay over the transmission medium and can be approximated by the following equation:

~~Cable~~ Link Delay = Medium Length × $\frac{1}{\mathrm{BT} \times \upsilon}$

where υ is the signal propagation speed in the medium and BT is the bit time of the medium.

N.5 ~~Higher Layer Delay~~ Worst-case Delay

The Worst-case Delay comprises 2 parts.

At PFC initiator station, it is assumed a maximum sized frame just start transmission from Transmission Selection when PFC is invoked. PFC frame has to wait until this in-progress frame complete transmission.

At PFC receiver station, it is assumed queue is paused but a maximum sized frame just starts transmission. Thus, bit times of the maximum sized frame is added into the total delay.

~~The Higher Layer Delay comprises the delay components between the MAC Control Client and the port Transmission Selection. Example of these delays are MACsec and implementation specific delays.~~

~~For link speeds of up to 10Gb/s, MACsec constrains each of the transmit delay and the receive delay to a maximum of 19 360 bit times (see 36.1.3.3).~~

~~This standard constrains the implementation specific delays to be less that 614.4 ns (see 36.1.3.3). This delay is equivalent to 6144 bit times at the speed of 10Gb/s.~~

N.6 Buffer allocation ~~Computation~~ example

A station needs to be capable of buffering DV bit times of data to ensure no frame loss due to congestion. The worst case is with a 10GBASE-T PHY. Assuming MACsec is not supported, this results in the following:

- PFC frame generation: 200 bit times;
- Maximum envelope frame size: 2000 octets, 16 160 bit times;
- PFC frame size: 64 octets, 672 bit times;
- XGMII MAC/RS and XAUI interface: 8192 + 2 × 2048 = 12 288 bit times;
- 10GBASE-T Delay: 25 600 bit times;
- 100 meters Cat6 cable: 5556 bit times (computed assuming υ = 0.6 × c, where c is the speed of the light in meters per second);
- Entering paused state ~~HD~~ = 6144 bit times

The total Delay Value in this scenario results as follows:

~~DV = 2 × (Max Frame) + (PFC Frame) + 2 × (Cable Delay) + 2 × ID + HD$_{s2}$~~

~~DV = 2 × (16 160) + (672) + 2 × (5556) + 2 × (25 600) + 2 × (12 288) + 6144= 126 024 bit times~~

$DV = ID_1 + WD_1 + LD_2 + ID_3 + WD_4 + ID_4 + LD_5 + ID_6$

DV = (200) + (16 160) + (672 + (12 288 + 25 600) /2)+ (5556) + ((12 288 + 25 600) /2 + 6144) + (16 160) + ((12 288 + 25 600) /2) + (5556) + ((12 288 + 25 600) /2) = 126 224 bit times

For this case, the amount of buffering needed to ensure no frame loss due to congestion results to be ~~126 024~~ 126 224 bit times, roughly equivalent to ~~15.5~~ 15.4 kB. 30.8kB is allocated to PFC enabled priority queue. XON/XOFF threshold is set to 15.4kB. So PFC guarantees no frame loss and no throughput loss.

If MACsec is used for user data, WD1 and ID4, each ~~the High Layer Delay~~ is incremented by 19 360 bit times; therefore, the total Delay Value results as follows:

~~DV = 2 × (16 160) + (672) + 2 × (5556) + 2 × (25 600) + 2 × (12 288) + 25 504 = 145 384 bit times~~

DV = 126 224 + 19 360 + 19 360 = 164 944 bit times

For this case, the amount of buffering needed to ensure no frame loss due to congestion results to be ~~145 384~~ 164 944 bit times, roughly equivalent to ~~18~~ 20 kB. Similar as non-MACsec case, 40kB is allocated to PFC enabled priority queue. XON/XOFF threshold is set to 20kB.