PFC Delay Value Constraint Model

Mark Gravel
mark.gravel@hp.com
USA 916.785.5955

1 Introduction

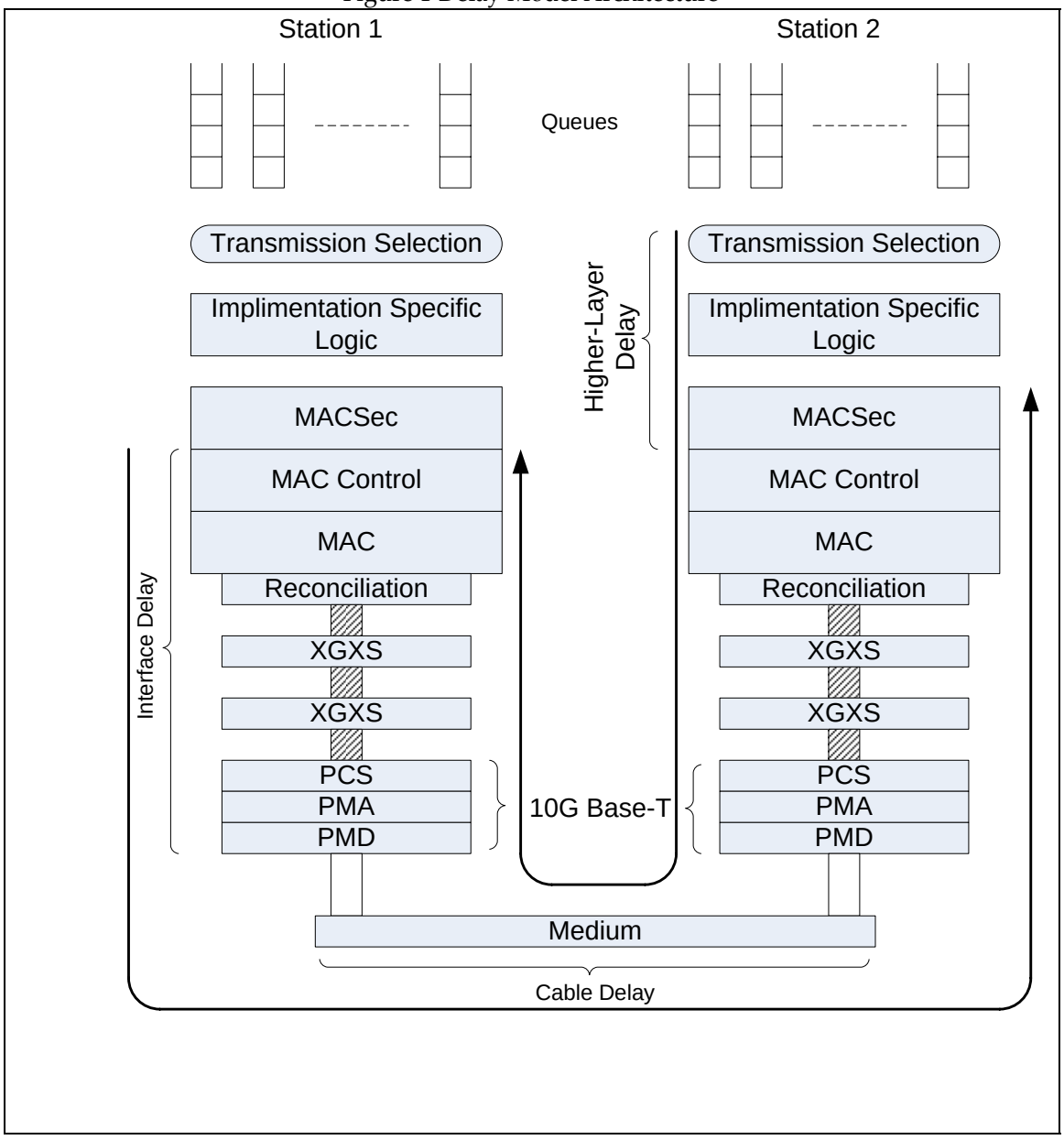
Proper PFC operation dictates that there must be an upper bound on the propagation delays through the network in order to constrain the amount of reserve buffer space necessary to guarantee lossless behavior. This document outlines a method for modeling peer-to-peer delay constraints using the Delay Constraint Reference Model.

2 Delay Constraint Reference Model

The total amount of reserved buffer space needed at the receiver to assure lossless behavior is the sum of two maximum length frames, one PFC frame, and peer-to-peer round trip delay. Overall peer-to-peer delay varies with implementation choices and the underlying Physical-layer technology connecting peer stations, as illustrated in Figure 1:

1. PFC transmission delay – a station that receives a PFC transmission request my have just committed to transmitting a maximum length frame. The assumption is that PFC frames are injected at the MAC Control sub-layer; however, this is somewhat implementation specific.
2. Interface Delay – sum of MAC Control, MAC/RS, PCS, PMA, and PMD delays. Interface Delay is technology dependent. A 10GBase-T physical device and XAUI interface is considered to have the worse case delay constraints to date. The worst case scenario is likely to change with the standardization 40G/100G interfaces.
3. Cable Delay – number of bits in flight stored in transmission medium, the exact delay value is contingent on the selected technology (i.e. Cu or fiber) and overall medium length.
4. Higher-layer Delay – amount of bits in flight between the output PFC queue and the MAC Control Client. A substantial portion of this delay component may be implementation specific.

Figure 1 Delay Model Architecture



The delay constraint model lumps MAC Control, MAC/RS, PCS, PMA, and PMD delay components into on single variable called Interface Delay (ID). All delay elements above the MAC Control Client are lumped into a single variable called Higher-layer Delay (HD). Both ID and HD delays are highly technology and implementation dependent.

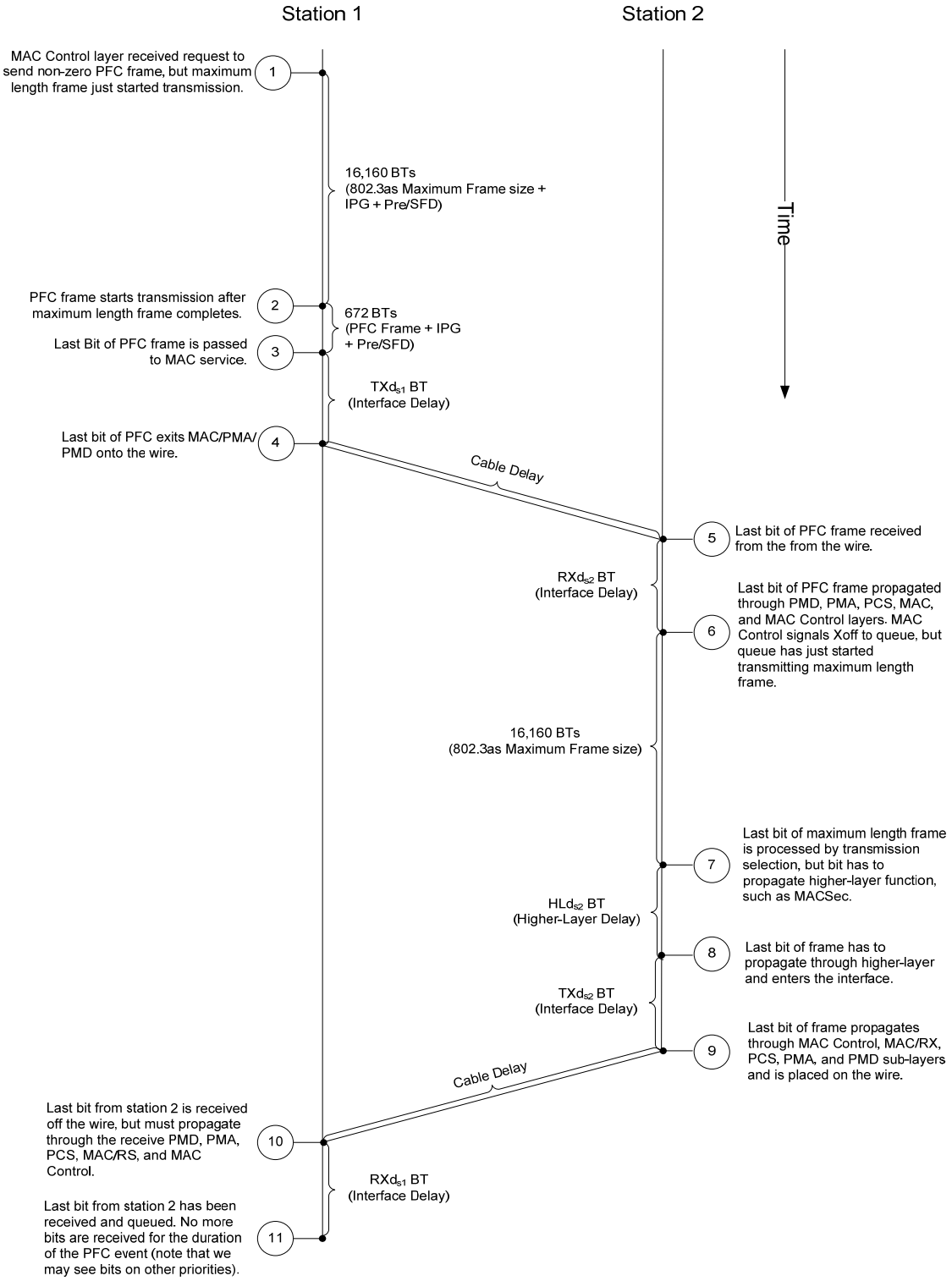
Figure 2 illustrates chronologically all delay elements that must be taken into account when determining the amount of reserve buffer space necessary to assure lossless behavior within a PFC enabled Priority Group. These delay elements include, but are not limited to the following:

1. Priority Group N (PG[n]) exceeds established PFC-Xoff threshold and sends a PFC frame transmit request to MAC Control sub-layer. However, the MAC Control has just

accepted a maximum length data frame for transmission. MAC Control must finish transmitting this maximum length frame before starting PFC frame transmission.

2. Upon completion of the current MA_DATA.request, the MAC Control sub-layer services the MA_CONTROL.request and starts PFC frame transmission.
3. MA_CONTROL.request is complete with last bit of PFC frame sent to the MAC.
4. Last bit of PFC frame must propagate through MAC Control, MAC/RS, PMA, and PMD sub-layers before it is placed on the wire.
5. Last bit of PFC frame propagates through the medium and enters the receiving PHY.
6. Last bit of PFC frame must propagate the PMA, PMD, and MAC/RS sub-layers, and it is seen by the MAC Control sub-layer. MAC control sub-layer sends PG[n] (Priority Group n) Xoff indication to transmission selection process (TSP); however, TSP has just committed sending a maximum length frame from PG[n]. The port must finish transmitting the current frame before ceasing traffic on PG[n].
7. Last bit of maximum length frame must now propagate through any higher-layer delays, such as implementation specific buffering or MACSec SecY processing.
8. Last bit of maximum length frame now propagates through MAC Control, MAC/RS, PCS, PMA, and PMD of the transmitter.
9. Last bit of maximum length frame is placed on medium after N bits of transmit interface delay.
10. Last bit of maximum length frame exits medium, and it must now propagate the receive interface (i.e. PMD, PMA, PCS, MAC/RS, and MAC Control).
11. Last bit in maximum length packet is received from the port for PG[n]. No more bits are received for the duration of the PFC event.

Figure 2 Delay Constraint Model



3 Delay Value Equation Derivation

Delay Value (DV) is the sum of all delay elements between the MAC Client and peer transmission selection process; it is represented with the following equation:

$$DV = 2*(\max\_frame\_size) + (PFC\_frame\_size) + 2*(cable\_delay) + TXd_{s1} + RXd_{s2} + TXd_{s2} + RXd_{s1} + HD_{s2}$$

Noting that for any given station's Interface Delay (i.e. MAC Control, MAC/RS, PCS, PMA, and PMD), the delay model includes both transmit and receive paths, so the following simplification can be made:

$ID_{s1} = TXd_{s1} + RXd_{s1}$, and likewise

$ID_{s2} = TXd_{s2} + RXd_{s2}$

Therefore,

$$DV = 2*(\max\_frame\_size) + (PFC\_frame\_size) + 2*(cable\_delay) + ID_{s1} + ID_{s2} + HD_{s2}$$

The DV equation is further simplified if the MAC Control, MAC/RS, PCS, PMA, and PMD technologies are identical in the peer stations:

$ID = ID_{s1} + ID_{s2}$

Therefore,

$$DV = 2*(\max\_frame\_size) + (PFC\_frame\_size) + 2*(cable\_delay) + 2*(ID) + HD_{s2}$$

4 Interface Delay

Interface Delay consist of all delay elements existing below the MAC Control Client, such as the MAC control layer, MAC Reconciliation layer, PCS, PMA, and PMD sub-layers, excluding Cable Delay. Ostensibly, Interface Delay will vary widely for any given MAC and physical layer technology (e.g. 1GE10GE) or will become available in near future (e.g. 40GE/100GE).

Interface Delay Constraints for some existing IEEE 802.3 interfaces are outlined in Table 1. Worst case Interface Delay to date consists of a 10GBase-T PHY with a XAUI interface to an ASIC – 37,888 bit times per station (10G MAC, 2*XAUI, and 10GBase-T).

Table 1 Existing 802.3 Interface Delay Constraints

Sub-layer	**Maximum RTT (bit time)**	**Maximum RTT (pause quanta)**	**Reference** IEEE 802.3, clause
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

Source: IEEE802.3an, Clause 44.3.

5 Cable Delay

Cable Delay consists of the propagation delay of the underlying media that must be taken into consideration when calculating the end-to-end delay constraint of a PFC capable device. Cable Delay can be approximated with the following calculation:

$$CD = MediaLength(meters) * \left[\frac{1}{BT * v} \right]$$, where

$v \equiv$ Signal velocity in medium (n *c) - n is the scalar for non-free space wave propagation and c the speed of light in free space (300,000,000 ms).

$BT \equiv$ Bit-Time of media.

Note that cable velocity (v) varies depending on the delivery medium (i.e. Cu vs. Fiber). Moreover, velocity in Cu medium varies noticeably with temperature.

6 Higher-layer Delay Constraints

Higher-layer delay is the sum of all the delay elements existing between the MAC Control Client interface and the port transmission selection process.

Delay elements may include, but are not limited to, the following:

- IEEE 802.3AE – MAC Security
- Implementation specific delays, such as buffering and pipelining present in a real world memory system implementation.

Table 2 Higher-layer Delays

Sub-layer	Maximum Delay Constraint (bit time)	Maximum Delay Constraint (pause quanta)	Reference
MACSec – SecY Transmit Delay	17024	33.25	IEEE 802.3AE, table 10-1
MACSec – SecY Receive Delay	17024	33.25	IEEE 802.3AE, table 10-1
Memory/Interface Pipelining	16160	31.25	Propose one 802.3as maximum frame size + Pre/sfd + IPG.

7 Delay Value Calculation Example

The receiving station must be capable of buffering DV Bit-times worth of data to assure lossless behavior. This example DV calculation assumes the following:

- 802.3as Maximum frame – 2000 Octets – 16,160 bit times.
- PFC Frame size – 64 Octets or 672 bit times.
- AISC Interface (XGMII MAC/RS and XAUI) – (8192 + 2x2048 = 12,288) bit times.
- PHY Interface (XAUI interface to 10GBase-T PHY) – 25,600.
- 100 meters, category 6 cable – 5556 bit-times.

As outlined in EIA-568-B, standard for building telecommunications cabling, maximum, worst case Cat 6 propagation delay is 555ns. This propagation delay corresponds to a propagation velocity of 0.60**c* (*c*=speed of light in meters per second) and a Cable Delay of 5,556 Bit-times.

The total Delay Value for this scenario:

$$DV = 2*(\max\_frame\_size) + (PFC\_frame\_size) + 2*(cable\_delay) + 2*(ID) + HD$$

DV= 2(16,160) + (672) +2*(5556) + 2*(37,888) + (33,184)*
DV=153,064 BT
The total Delay Value Constraint is roughly 19.1K Bytes.

8 Conclusion

Delay Values vary widely depending on interface technology. To date, 10G Base-T represents the worse case delay constraint requiring a Delay Valve of roughly 19.1K Bytes of reserved packet buffer per PFC enabled queue. Developing technologies, such as 40G and 100G Ethernet, are expected to displace 10G Base-T as having worse case delay constraints.