

# Priority-based Flow Control— Proposed Clause 36 changes

Mick Seaman

## Summary

The current organization, in IEEE Std 802.1Q–2022, Clause 36 Priority-based Flow Control assumes that the reader is already completely familiar with PFC operation and knows all the answers. It omits significant detail, which can only be had by looking at the 802.3 Standard (but does not provide detailed references). For example, the existing text goes to some lengths to avoid saying that MAC Control sends a frame (it's just as if a M\_CONTROL.request results in an M\_CONTROL.indication by magic—discussion of frame transmission appears only in Annex N covering PFC delays/headroom). These omissions make it hard to use the existing text as a starting point for P802.1Qdt. Reviewing our discussions during the July meeting, I believe that the current omissions are an obstacle to making progress. It's hard for a group to remain on the same page when the prime reference pages are missing basic information. As a further example: nowhere does the existing 802.1Q text point out that the MAC Control frame has an 88-08 EtherType, and that the interception of MAC Control frames is based purely on that EtherType and not on the DA. That information is essential to any discussion of MACsec protection of PFC frames.

Adding text to Clause 36 to say “start again, because here is what you need to know” will not yield a defensible result. At the same time, existing 802.3 MAC Control specific detail should be retained, placing that in the overall big picture. No change to existing conformance with respect to 802.3 MAC Control is intended. This note includes (in order) the following:

- Proposed replacement [Clause 36](#) text (not yet complete).
- Notes on PFC related issues in IEEE Std 802.1Q–2022 Clause 36 and elsewhere (but see below), and in IEEE Std 802.1AX–2020 and IEEE Std 802 (under revision).
- Relevant references to, and excerpts from, PFC relevant IEEE Stds 802.1Q–2022, 802.3–2022, 802.3.1–2013, and 802.1AX–2020.
- References to prior P802.1Qdt contributions. Much of the proposed text is based on these contributions and prior discussion, though that process is not complete. They need checking to complete that process, and make sure intentional differences have been discussed and resolved.
- A PDF post-it markup of IEEE Std 802.1Q–2022 Clause 36 (may be missing from this copy).

## 1 36. Priority-based Flow Control (PFC)

2 Priority-based Flow Control (PFC) allows a MAC Client to flow control the transmission of data frames by  
 3 a peer MAC Client attached to the same individual LAN.

4 This clause provides an overview of PFC operation (36.1) and further describes and specifies:

- 5 a) Network and system considerations and limitations for PFC use (36.2).
- 6 b) PFC operation with IEEE 802.3 MAC Control support (36.3).
- 7 c) PFC-capable interface stack operation with MACsec (36.4, 36.5), MAC Privacy protection (36.6),  
 8 and Link Aggregation (36.7).
- 9 d) The receive buffering (PFC headroom) required to avoid against frame loss (36.1.1, 36.8).
- 10 e) A PFC round-trip delay measurement protocol that supports automatic headroom calculation (36.9).
- 11 f) Management of PFC, including parameter exchanges using DCBX/LLDP, the headroom  
 12 measurement protocol, and MACsec Key Agreement (MKA) (36.10).
- 13 The encoding of DCBX/LLDP parameters is specified in Annex D.

14 The models of operation in this clause provide a basis for specifying the externally observable behavior of  
 15 PFC and are not intended to place additional constraints on implementations; these can adopt any internal  
 16 model of operation compatible with the externally observable behavior specified.

### 17 36.1 PFC overview

18 A station can initiate PFC on a point-to-point link to request its peer station to temporarily pause  
 19 transmission on a per-priority basis. This flow control attempts to eliminate or reduce frame loss resulting  
 20 from a temporary lack of receive buffering. The buffer shortage can be a result of inability to process frames  
 21 at unusually high reception rate or, in a bridge or router, congestion of one or more links to which frames are  
 22 to be forwarded. The PFC mechanism operates independently of the reason for its use (see W.2 for  
 23 additional discussion).

24 Each PFC-capable station's MAC Client interface stack is associated with a PFC Initiator, capable of  
 25 monitoring receive buffering, and a PFC Receiver capable of selectively pausing transmission selection of  
 26 frames of one or more priorities. Figure 36-1 provides an example of PFC use with IEEE 802.3 MACs that  
 27 include the optional MAC Control sublayer.

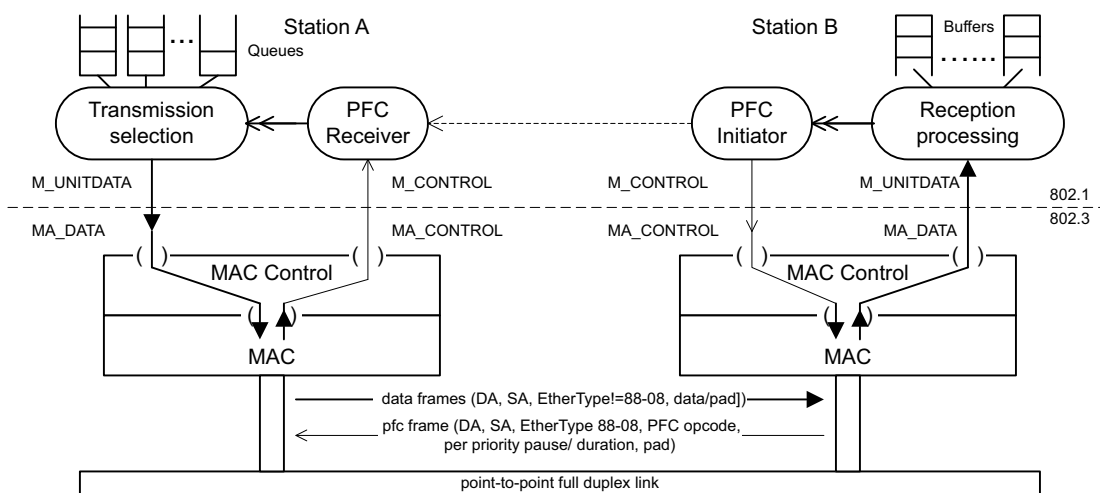


Figure 36-1—PFC example

1 In Figure 36-1, Station B reacts to a possible lack of buffers for receiving data frames. Its PFC Initiator  
2 makes a MAC Control request specifying the globally assigned IEEE MAC-specific Control Protocols  
3 group address 01-80-C2-00-00-01, the PFC opcode 01-01, the priorities for which transmission is to be  
4 paused, and for each priority the duration of the pause. The MAC Control request prompts MAC  
5 transmission of a frame with the specified destination MAC address, the station's individual source MAC  
6 address, and a Length/Type field with EtherType 88-08 followed by the PFC opcode and priority  
7 parameters.

8 NOTE—Each station does not need to know the other's individual MAC address to send and receive PFC frames. A  
9 point-to-point link connects only two stations, so the destination address can be a well-known multicast address  
10 provided that the frame is confined to the connecting link. Frames with the 01-80-C2-00-00-01 destination address are  
11 not forwarded by any Bridge (8.6.3).

12 If Station B's MAC supports preemption, the PFC is transmitted as an express frame (6.7.2).

13 Station A's MAC is configured to receive frames with the destination MAC address 01-80-C2-00-00-01.  
14 Valid frames received with that address together with any other valid frames the MAC has been configured  
15 to receive are passed to MAC Control. MAC Control passes each frame with a value of the 802.3  
16 Length/Type other than 88-08 directly to the MAC Client interface stack with an MA\_DATA.indication as  
17 shown for Station B. Each received frame with Length/Type 88-08 followed by the PFC opcode 01-01 is  
18 passed with an MA\_CONTROL.indication directly to the PFC Receiver which maintains a Priority\_Paused  
19 variable (TRUE or FALSE) for the MAC for each of the eight priorities. A frame of a given priority is not  
20 available for transmission selection by a Bridge's MAC Relay Entity's Forwarding Process (8.6.8) if  
21 transmission is paused for the MAC for that priority and MAC.

22 A Bridge's Forwarding Process queues frames forwarded for transmission on a Bridge Port on the basis of  
23 traffic class (8.6.6). Transmission selection can select frames from the queue in FIFO order (8.6.6, 8.6.8) so  
24 the reception of a PFC that pauses transmission for a given priority can pause transmission for frames of  
25 other priorities assigned to the same traffic class. A PFC Initiator does not rely on this possibility, but  
26 specifies pausing for each priority to be paused in PFC requests.

### 27 **36.1.1 PFC headroom**

28 After Station B initiates PFC, it can continue to receive frames with PFC-enabled priorities until it has  
29 received the last such frame transmitted by Station A before the latter's PFC Receiver has halted  
30 transmission selection. Station B might not be able to empty currently occupied buffers—transmission from  
31 those buffers to a further link might itself be halted, currently or imminently—so its reception processing  
32 can expect to make use of additional buffering during the cumulative time for:

- 33 a) B's reception processing to calculate the remaining buffering following frame receipt.
- 34 b) B's PFC Initiator to initiate PFC following that buffering calculation.
- 35 c) Encoding of the PFC frame and any other transmission delays associated with B's interface stack.
- 36 d) Any prior in-progress frame transmission by B (possibly of a maximum sized frame) to complete.
- 37 e) PFC frame transmission on the physical link.
- 38 f) The link delay for transmission from B to A.
- 39 g) PFC frame reception, including frame validation, by A's interface stack.
- 40 h) A's PFC Receiver to decode the PFC frame and halt transmission selection for specified priorities.
- 41 i) Any in-progress frame transmission by A (possibly of a maximum sized frame) to complete.
- 42 j) The link delay for transmission from A to B.
- 43 k) Reception delays associated with B's interface stack, reception processing, and buffering.

44 The PFC *headroom* is the buffering that needs to remain available to B's reception process before PFC is  
45 initiated to ensure that frames are not lost as a result of a shortage of buffers. If, when not PFC paused, data  
46 frames that would occupy those buffers can be transferred at full link rate from A's transmit buffers to those

1 monitored by B's reception process and PFC initiator, a) through k) are additive, with all delays being times  
2 during which additional bits can be encode in frames to be transmitted or buffered awaiting processing. In  
3 that case the PFC headroom is the link speed multiplied by that total, the round-trip time for PFC operation  
4 (from B's receipt and buffering of a frame that prompts PFC initiation, to B's receipt and buffering of the last  
5 frame transmitted before the PFC took effect).

6 NOTE 1—Direct use of MAC Control for PFC frame transmission and reception emphasizes the need for timely  
7 transmission and reception processing of MAC Control PFC frames. As part of bounding the buffer allocation required  
8 to avoid frame loss, IEEE Std 802.3 places timing requirements on that processing. For detailed specification of PFC  
9 operation with IEEE 802.3 MAC Control see 36.3. Annex N provides a detailed example of headroom calculation.

10 NOTE 2—The PFC frame can be transmitted as an express frame, but so could an in-progress frame [item d) above].

## 11 **36.2 Network and system considerations and limitations**

### 12 **36.2.1 Data center network protocol support**

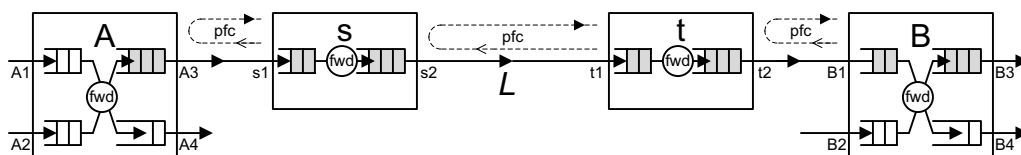
13 PFC can be used to support data center networks. Data center protocols can require very low frame loss  
14 without depending on end-to-end loss detection and retransmission, which can be less timely than required  
15 and are therefore not a focus of protocol design. Traffic patterns can be bursty and unpredictable at network  
16 design time. Arbitrary sets of traffic sources can have low long-term bandwidth requirements, while still  
17 needing to be able to access full network bandwidth without the delays inherent in making and releasing  
18 reservations. Intermediate systems can forward received frames from several links to a single link in excess  
19 of the latter's capacity for periods that can be too short to determine and signal appropriate transmission  
20 rates to the traffic sources. The number of links supported by any given intermediate system and their speed  
21 means that practical implementations have limited buffer capacity.

22 This bursty traffic can be supported by one or more PFC-enabled priorities. Other priorities can be assigned  
23 to frames for other protocols or flows whose traffic patterns are better known, are explicitly supported by  
24 bandwidth reservation or traffic shapers, or for whom frame loss is an explicit part of error recovery,  
25 congestion control, and fairness of network use by multiple flows (e.g. TCP).

### 26 **36.2.2 Hop-by-hop operation**

27 An intermediate system that receives a PFC frame on a given MAC, and pauses transmission, can find its  
28 own buffers filling as it continues to receive frames for transmission on that MAC from other system  
29 interfaces, requiring PFC transmission on those interfaces. This hop-by-hop back pressure flow control can  
30 propagate, through multiple intermediate systems to the source(s) of the excess traffic if their transmission is  
31 not slowed by other means or naturally exhausted. Less buffering needs to be allocated in each intermediate  
32 system than would be required by relying on signaling through successive intermediate systems to each of  
33 the current and potential sources of flows passing through the system.

34 Distributed data centers can use data center protocols over links are significantly longer than those typically  
35 found in an individual data center (e.g. 60 km as opposed to 100 metres) and introduce corresponding PFC  
36 headroom buffering requirements as consequence of the increased transmission delays. When a data center  
37 system connects to such a link is via a local intervening Bridge, its PFC headroom requirement is  
38 determined by the round-trip delay to that Bridge, as shown in Figure 36-2, and is unaffected by the length  
39 of the link between the data centers. This is true even if the intervening Bridges are Two-Port MAC Relays  
40 (TPMRs), which are transparent to the operation of some bridge-to-bridge protocols.

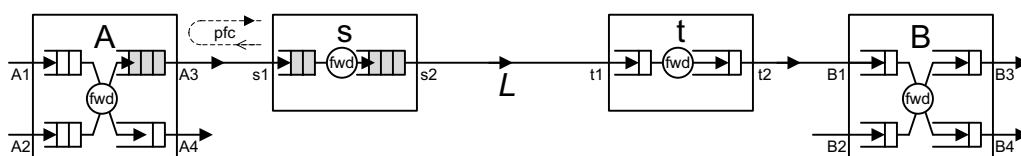


**Figure 36-2—PFC hop-by-hop flow control with TPMRs**

1 Figure 36-2 shows the buffering of user data frames, as they flow from data center switch A (bridge or  
 2 router) to data center switch B, passing through TPMRs *s* and *t*. Port B3 is congested, which has led to PFC  
 3 initiation on port B3 pausing transmission from port *t2*. The round-trip from B3’s PFC initiation to its last  
 4 reception of a PFC-enable priority data frame is indicated above the *t2*–B1 link. Following *t2*’s transmission  
 5 pause, *t*’s buffers filled, causing *t1* to initiate a pause on the *s2*–*t1* link. If the congestion at B3 persists, *s*  
 6 eventually initiate PFC at *s1*, applying back-pressure to A3, as shown.

7 NOTE 1—Frames, including PFC frames, destined to the well-known IEEE MAC-specific Control Protocols group  
 8 address are not forwarded by any Bridge (8.6.3). This example uses TPMRs to emphasize the fact that PFC operates  
 9 hop-by-hop for any frame forwarding device. The same would be true if *s* and *t* in Figure 36-2 were Provider Bridges.

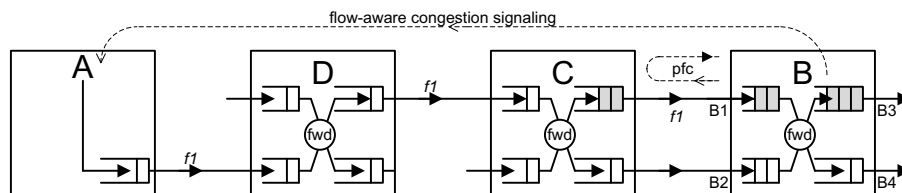
10 If the *s2*–*t1* link *L*’s data rate is less than that of the A3–*s1* link, congestion can arise at port *s2*, with PFC  
 11 initiation at *s1* back-pressuring A3, as shown in Figure 36-3



**Figure 36-3—PFC hop-by-hop flow control with link rate mismatch**

### 12 36.2.3 PFC and flow-aware congestion signaling

13 PFC can be used in conjunction with protocols that attribute congestion to individual flows and provide  
 14 feedback towards the source(s) of those flows, as shown in Figure 36-4 and Figure 36-5.



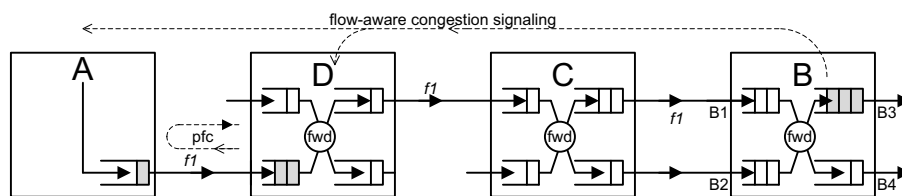
**Figure 36-4—Flow-aware congestion signaling with PFC loss prevention**

15 In Figure 36-4, B attributes the congestion at port B3 to flow *f1* with source A, and sends a message directly  
 16 to A requesting a flow rate reduction. The immediate effect of the congestion is to fill buffers allocated for  
 17 reception from B1, initiating a PFC to prevent loss until A’s rate reduction propagates to B1. PFC operation  
 18 depends only on buffer use and is independent of flow-aware signaling. While the latter takes longer to take  
 19 effect, it avoids the congestion spreading (36.2.4) that can accompany sustained use of PFC.

20 NOTE 1—A can be the true source of the flow, or an intermediate system, e.g., a router. The congestion notification  
 21 provided by QCN (Clause 30, 31, and 32) signals to the flow’s source MAC Address.

22 NOTE 2—Providing minimal buffering and relying on PFC to prevent loss prevention can affect flow-aware congestion  
 23 control performance and fairness. The QCN analysis in Clause 30 did not take PFC into account.

1

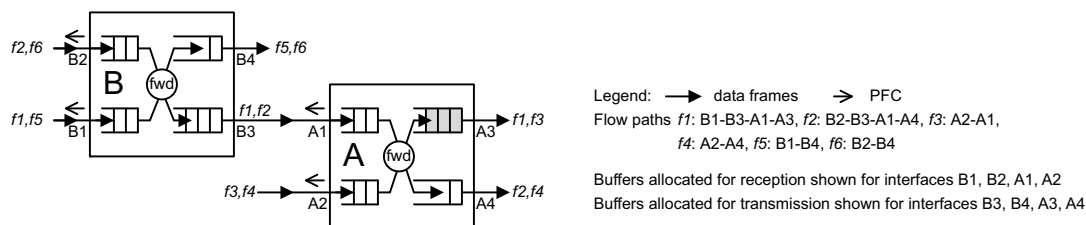


**Figure 36-5—Flow-aware congestion signaling with PFC back-pressure**

2 In Figure 36-5, B has sent a message to A requesting a rate reduction for flow  $f_1$ , but A does not implement  
 3 the congestion signaling protocol. If D intercepts that flow rate reduction message and reduces its own  
 4 transmission for  $f_1$  or other flows transmitted by A, D's buffers can fill, triggering PFC to pause flows with  
 5 PFC-enabled priorities. As in Figure 36-4, PFC operation depends only on buffer use and is independent of  
 6 flow-aware signaling and the details of D's interception of congestion signaling message (not specified by  
 7 this standard).

### 8 36.2.4 Congestion spreading

9 PFC's hop-by-hop back pressure flow control can cause congestion spreading, pausing any link that is used  
 10 by a flow that subsequently uses a paused link. Figure 36-6 provides an example.



**Figure 36-6—PFC congestion spreading**

11 In Figure 36-6, Bridge A's remaining buffer allocation for reception from MAC A1 or MAC A2 and  
 12 subsequent transmission by MAC A3 has been nearly exhausted by frames for flows  $f_1$  and  $f_3$ . Bridge A  
 13 initiates PFC for A1 and A2 to prevent subsequent frame loss, which in turn leads to near exhaustion of  
 14 Bridge B's buffering for frames received from B1 and B2 and transmission by B3, as B3's transmission is  
 15 paused for the priorities if all the flows shown. Consequently Bridge B initiates PFC for B1 and B2. The  
 16 result of the  $f_1$  and  $f_3$  transmission congestion at A3 is thus to congest transmission at B3, even though the  
 17 sum of  $f_1$  and  $f_2$ 's bandwidth requirements do not exceed that MAC's capability. Frames for flows  $f_2$  and  $f_4$   
 18 are delayed, even though they will not be transmitted by the MAC (A3) with flows in excess of transmission  
 19 bandwidth capability. Frames for flows  $f_5$  and  $f_6$  are delayed, even though they are not to be forwarded by a  
 20 system with any MAC that lacks the bandwidth to support the network flows.

### 1 36.2.5 Potential for deadlock and delay

2 PFC's hop-by-hop back pressure flow control can result in deadlock. Figure 36-7 provides an example.

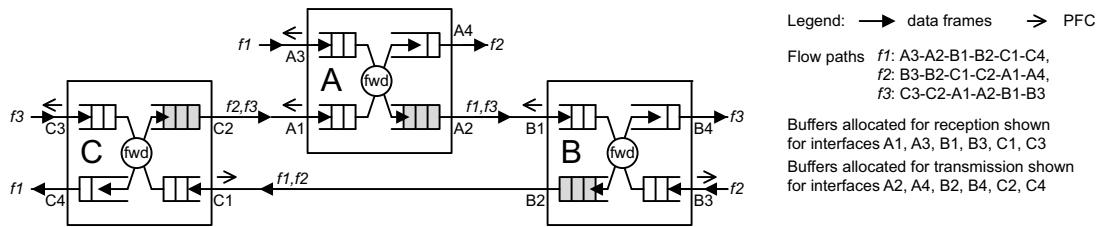


Figure 36-7—PFC deadlock example

3 In Figure 36-7, flow  $f_1$  traverses Bridges A, B, and C in that order; flow  $f_2$  traverses B, C, and A; and flow  $f_3$  traverses C, B, and A. While none of the flows loops in this set of Bridges (flow  $f_1$ , e.g., is received by MAC A3 and transmitted on C4), there is a circular buffer dependency as PFC operates per-priority and is independent of any particular flow. If flows  $f_1$  and  $f_3$  cause congestion at A2, A can initiate PFC for the link A1-C2, causing C (after received frames fill buffers for C2) to initiate PFC for C1-B2, and B in turn to initiate PFC for B1-A2. As A2's transmission is now blocked, A cannot let the PFC for A1 lapse without losing frames.

10 Circular buffer dependency is a necessary condition for PFC deadlock, and does not occur in some network topologies (a simple case is where all flows follow the same tree). However, even in networks whose intended topology is circular buffer dependency free, there remains the possibility of such a dependency during network reconfiguration as a consequence of link loss or addition. The operation of network configuration and management protocols should be independent of PFC operation (36.2.8). Each Bridge enforces a maximum Bridge transit delay (6.5.6), discarding frames queued for longer. That discard can suffice to remove a deadlock, if the network converges on a circular buffer dependency free topology.

### 17 36.2.6 PFC and MAC Security

18 User data frames and PFC frames can be MACsec protected (36.4, 36.5). Although MACsec does not defend against physical attack on a link or interference with the details of MAC operation, it can ensure that data and PFC frames were transmitted by an authenticated and authorized peer, reducing exposure to adversarial actions that can be less easy to detect than link failure.

22 Whether or not PFC frames are MACsec protected, it is important that a system that uses PFC does not provide a way (e.g., by inappropriate tunneling) for a distant adversary to transmit a PFC frame on a link.

24 MACsec peers can communicate over links that include intervening Bridges. Two Customer Bridges can, e.g., secure connectivity across a Provider Bridged Network. If one of those Customer Bridges protects a PFC frame with the same MACsec Secure Channel (SC), that frame will be discarded by the first Provider Bridge. Each Customer Bridge can secure connectivity (if desired, including PFC transmission) to its nearest Provider Bridge with a separate SC (see 11.7 of IEEE Std 802.1AE-2018).

29 NOTE 1—All PFC frames have MAC destination address 01-80-C2-00-00-01. Frames with that address are discarded by all Bridges (8.6.3). If they are integrity protected by the Customer Bridge to Customer Bridge SC, the Provider Bridge will not be able to identify them as PFC frames.

### 1 **36.2.7 PFC and MAC Privacy protection**

2 MAC Privacy protection can be applied to user data frames and PFC frames (36.6, IEEE Std 802.1AEdk).  
3 PFC transmission reflects a possible shortage of reception buffers, and can thus provide an adversary with  
4 information as to the real level of user traffic even when frame confidentiality has been augmented by the  
5 transmission of user data frames in a Privacy Channel. To reduce the privacy compromise, PFC frames can  
6 also be transmitted in Privacy Channel MPPDUs, at the possible cost of an increase in PFC headroom  
7 (36.1.1, 36.8) depending on MPPDU transmission intervals.

8 NOTE—Privacy Channels provides regular transmission of fixed sized MAC Privacy protection PDUs (MPPDUs),  
9 independent of the level of user traffic, encapsulating privacy protected frames. Privacy Frames provide address  
10 encapsulation and configurable for individual frames (see Clause 17 of IEEE Std 802.1AEdk). While an adversary will  
11 not be certain that short frames transmitted outside a Privacy Channel are PFCs, observations can be useful if their  
12 contribution to a probabilistic fingerprint of activity outweighs the cost of acquisition. The cost to an adversary of  
13 erroneous conclusions can be minimal (see IEEE Std 802E).

14 Since MPPDUs encapsulate MAC addresses, PFC frames shall only be transmitted in Privacy Channels or  
15 Privacy Frames if the supporting MACsec Secure Channel (SC) provides protection to, and only to, the  
16 nearest Bridge of any type. PFC frames extracted from received MPPDUs whose transmission is supported  
17 by an SC that protects frames passing through intermediate relay systems shall be discarded. To ensure that  
18 the SC has the intended scope, the address is also used by the peer PAEs to exchange EAPOL frames, which  
19 include MKA (MACsec Key Agreement) frames, should be the Nearest Bridge group address (8.6.3).

### 20 **36.2.8 Network configuration and management protocols**

21 Sound design requires that a system any system or network recover from erroneous conditions or state,  
22 however implausible, within known bounded time during which network configuration and management  
23 protocols operate correctly and the frames they transmit are correctly received. Timely and successful  
24 configuration and network management protocol operation is facilitated by the following:

- 25 a) Transmission is not subject to PFC, and not excessively delayed by transmission of other frames  
26 including high priority forwarded frames.
- 27 b) Reception, and delivery to the correct protocol processing and/or forwarding entities does not  
28 depend on the processing of frames subject to PFC.

29 NOTE 1—Use of FIFO ingress buffering by an interface provides an example of possible interaction between  
30 PFC controlled and other frames, if the ingress buffering is not separated by priority as shown in Figure W-5.

31 Satisfaction of these constraints can depend on network design and configuration choices, including the  
32 priority assigned to network configuration protocol and management frames and the use of VLAN tags to  
33 convey that priority between intermediate systems, including Bridges.

34 A Bridge shall meet the above constraints [a) and b)] for all interfaces for all network configuration and  
35 management protocol entities for which it transmits or receives frames.

36 Frames for the spanning tree protocols (RSTP, MSTP, Clause 13), and Shortest Path Bridging (SPB,  
37 Clause 27) including those for ISIS-SPB, are transmitted and received without a VLAN tag and addressed to  
38 the nearest peer (using, e.g., the Nearest Customer Bridge group address as the MAC destination address). In  
39 the common case where there are no intervening frame buffering or store and forward intermediate systems,  
40 correct interface implementation can be sufficient to satisfy a) and b) for peer protocol entity  
41 communication. Where one or more intervening intermediate systems (e.g., TPMRs or Provider Bridges) are  
42 present, the priority they assign to untagged frames needs to be one that provides a high probability of timely  
43 delivery in the presence of other flows and one that is not subject to PFC. Frames for other traffic flows can  
44 be VLAN-tagged by the configuration protocol peers to explicitly signal a different priority as part of  
45 satisfying this requirement. TPMRs, Provider Bridges, and Provider Backbone Bridges should not expedite  
46 frames for configuration protocols simply on the basis of their MAC destination address. Such expediting



1 can result in out of order delivery for MACsec protected frames, and discarding of subsequent data frames  
2 now outside the recipient's replay protection window.

3 NOTE 2—RSTP, MSTP, and SPB frames that are MACsec protected by their originating system Bridge component are  
4 not VLAN-tagged, before or after protection, by that component.

5 Frames for network management protocols (e.g., NETCONF over TLS) are commonly forwarded through  
6 intermediate systems before reaching their intended destinations. The priority assigned to those frames  
7 needs to be one not associated with PFC by those intermediate systems.

8 NOTE 3—Priority is a parameter both of the EISS, that adds VLAN tags to frames, and of the ISS (6.6,  
9 IEEE Std 802.1AC). The priority to be associated with a received frame that is to be forwarded by a Bridge can be  
10 derived from its VLAN tag (6.8, 6.9.4) if present or a default value (6.6, 6.7, 12.6.2.1, 6.9.4) in the absence of a VLAN  
11 tag, and can be further modified by flow classification and metering (8.6.5).

12 NOTE 4—Configuration and control frame priority can determine how those frames are transmitted by the originating  
13 interface stack, e.g. where MAC Security is used to protect integrity, confidentiality, or privacy (36.4, 36.5, 36.6).

### 14 36.2.9 Point-to-point operation

15 PFC is specified only for a pair of full duplex MACs (e.g., IEEE 802.3 MACs operating in point-to-point  
16 full-duplex mode) connected by a single point-to-point link.

## 17 36.3 Detailed specification of PFC operation with IEEE 802.3 MAC Control

### 18 36.3.1 PFC primitives

19 A MAC Client wishing to pause transmission of data frames on certain priorities from the remote system on  
20 the link generates an M\_CONTROL.request (11.4 of IEEE Std 802.1AC-2016; Annex 31D of  
21 IEEE Std 802.3-2022) specifying the following:

- 22 a) The globally assigned 48-bit multicast address 01-80-C2-00-00-01.
- 23 b) The PFC opcode (i.e., 01-01, as specified in Annex 31A of IEEE Std 802.3-2022).
- 24 and a request\_operand\_list with two operands as follows:
  - 25 c) priority\_enable\_vector: a 2-octet field, with the most significant octet being reserved (i.e., set to zero  
26 on transmission and ignored on receipt). Each bit of the least significant octet indicates if the  
27 corresponding field in the time\_vector parameter is valid. The bits of the least significant octet are  
28 named e[0] (the LSB) to e[7] (the MSB). Bit e[n] refers to priority n. For each e[n] bit set to one, the  
29 corresponding time[n] value is valid. For each e[n] bit set to zero, the corresponding time[n] value is  
30 invalid.
  - 31 d) time\_vector: a list of eight 2-octet fields, named time[0] to time[7]. The eight time[n] values are  
32 always present regardless of the value of the corresponding e[n] bit. Each time[n] field is a 2-octet,  
33 unsigned integer containing the length of time for which the receiving station is requested to inhibit  
34 transmission of data frames associated with priority n. The field is transmitted most significant octet  
35 first, and least significant octet second. The time[n] fields are transmitted sequentially, with time[0]  
36 transmitted first and time[7] transmitted last. Each time[n] value is measured in units of  
37 pause\_quanta, equal to the time required to transmit 512 bits of a frame at the data rate of the MAC.  
38 Each time[n] field can assume a value in the range of 0 to 65 535 pause\_quanta.

39 As a result of the processing of the PFC M\_CONTROL.request, the peering PFC station receives a PFC  
40 M\_CONTROL.indication with the same multicast address and PFC opcode, and an indication\_operand\_list  
41 with the operands specified for the M\_CONTROL.request.

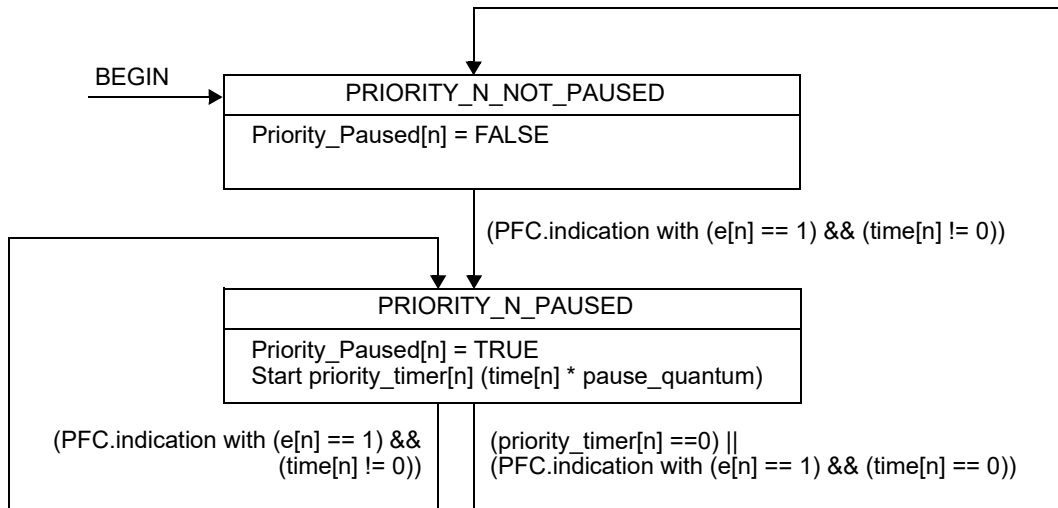
42 NOTE—IEEE Std 802.1AC maps M\_CONTROL.requests and M\_CONTROL.indications to and from the  
43 MA\_CONTROL.requests and MA\_CONTROL.indications specified by IEEE Std 802.3 respectively.

1 As specified in IEEE Std 802.3, when PFC is enabled on a port for at least one priority over an IEEE 802.3  
 2 link layer, the IEEE Std 802.3 PAUSE mechanism is not used for that port.

3 **36.3.2 Processing PFC M\_CONTROL.indications**

4 The PFC Receiver maintains and makes available to Transmission Selection the vector of the  
 5 Priority\_Paused[n] variables, indicating the state of each of the eight priorities. Each Priority\_Paused[n]  
 6 variable is a boolean. When Priority\_Paused[n] is FALSE, priority n is not in paused state. When  
 7 Priority\_Paused[n] is TRUE, priority n is in paused state.

8 Figure 36-8 shows the PFC state diagram for priority n. If PFC is not enabled for priority n, then the PFC  
 9 state diagram does not apply to priority n and Priority\_Paused[n] is FALSE.



**Figure 36-8—PFC Receiver state diagram for priority n**

10 Upon receipt of a PFC M\_CONTROL.indication, the PFC Receiver programs up to eight separate timers,  
 11 each associated with a different priority, depending on the priority\_enable\_vector. For each bit in the  
 12 priority\_enable\_vector that is set to one, the corresponding timer value is set to the corresponding time value  
 13 in the time\_vector parameter. Priority\_Paused[n] is set to TRUE when the corresponding timer value (i.e.,  
 14 priority\_timer[n]) is nonzero. Priority\_Paused[n] is set to FALSE when the corresponding timer value (i.e.,  
 15 priority\_timer[n]) counts down to zero. A time value of zero in the time\_vector parameter has the same  
 16 effect as the timer having counted down to zero. If PFC is not enabled for priority n and a PFC indication is  
 17 received with e[n] set to one, then the time[n] parameter is ignored (i.e., the primitive is processed as if e[n]  
 18 was set to zero).

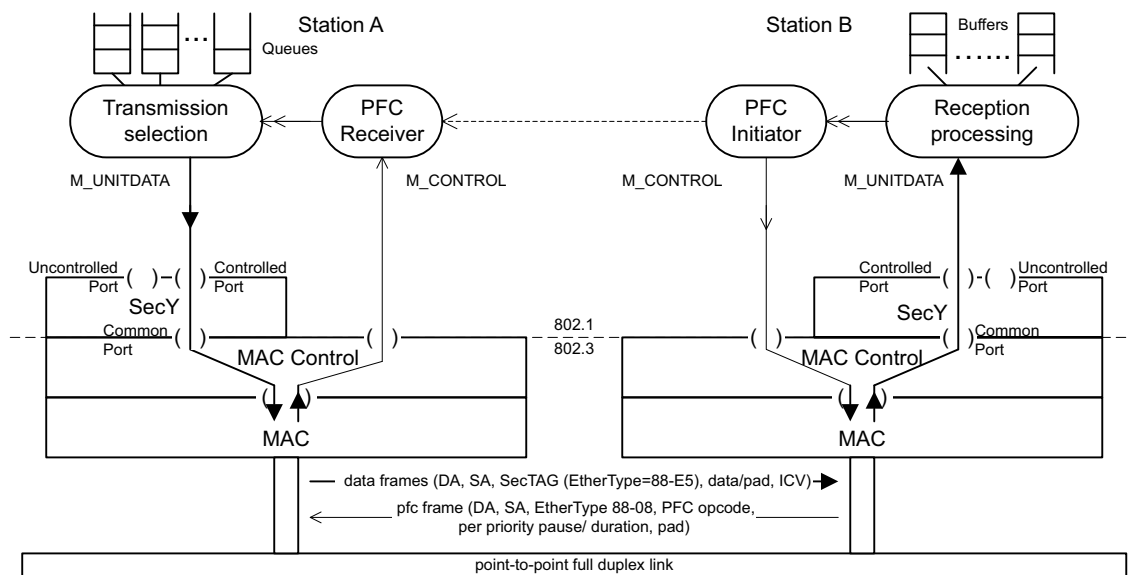
19 NOTE—A priority\_enable\_vector with all bits set to zero is legal and equivalent to a no-op.

20 **36.3.3 Timing considerations**

21 A priority flow controlled queue shall go into paused state in no more than 614.4 ns since the reception of a  
 22 PFC M\_CONTROL.indication that paused that priority. This delay is equivalent to 12 pause quanta (i.e.,  
 23 6144 bit times) at the speed of 10 Gb/s, 48 pause quanta (i.e., 24 576 bit times) at the speed of 40 Gb/s, and  
 24 120 pause quanta (i.e., 61 440 bit times) at the speed of 100 Gb/s.

### 1 36.4 PFC with MACsec data protection

2 Figure 36-9 illustrates IEEE 802.3 MAC Control support of PFC primitives together with the use of the  
 3 MAC Security protocol (MACsec, IEEE Std 802.1AE) to provide data integrity, data origin authenticity, and  
 4 (optionally) confidentiality protection for data frames.



**Figure 36-9—PFC with IEEE 802.3 MAC Control and MACsec**

5 In Figure 36-9, the MAC Security Entity (SecY) in Station A applies MACsec protection to data frames  
 6 transmitted through its Controlled Port. The SecY in Station B validates, and if necessary decrypts, those  
 7 protected frames before passing them to the user(s) of its Controlled Port. The operation of MACsec and its  
 8 supporting key agreement protocol is as specified in IEEE Std 802.1AE and IEEE Std 802.1X. PFC  
 9 communication from the PFC Initiator in Station B to the PFC Receiver is not MACsec protected, and  
 10 operates as specified in 36.3.

11 A SecY can map (10.5, 10.7.17 of IEEE Std 802.1AE-2018) the frame’s user priority (the priority for the  
 12 M\_UNITDATA.request made at its Controlled Port) to an access priority (the priority for the corresponding  
 13 M\_UNITDATA.request that the SecY makes of the supporting interface stack at its Common Port). Each  
 14 PFC’s per-priority parameters apply to the user priority (used by transmission selection in the figure).

#### 15 36.4.1 PFC headroom with MACsec data protection

16 IEEE Std 802.1AE places requirements on the performance of the MAC Security Entity (SecY), limiting the  
 17 transmit and receive delays attributable to MACsec (10.10 of IEEE Std 802.1AE-2018).

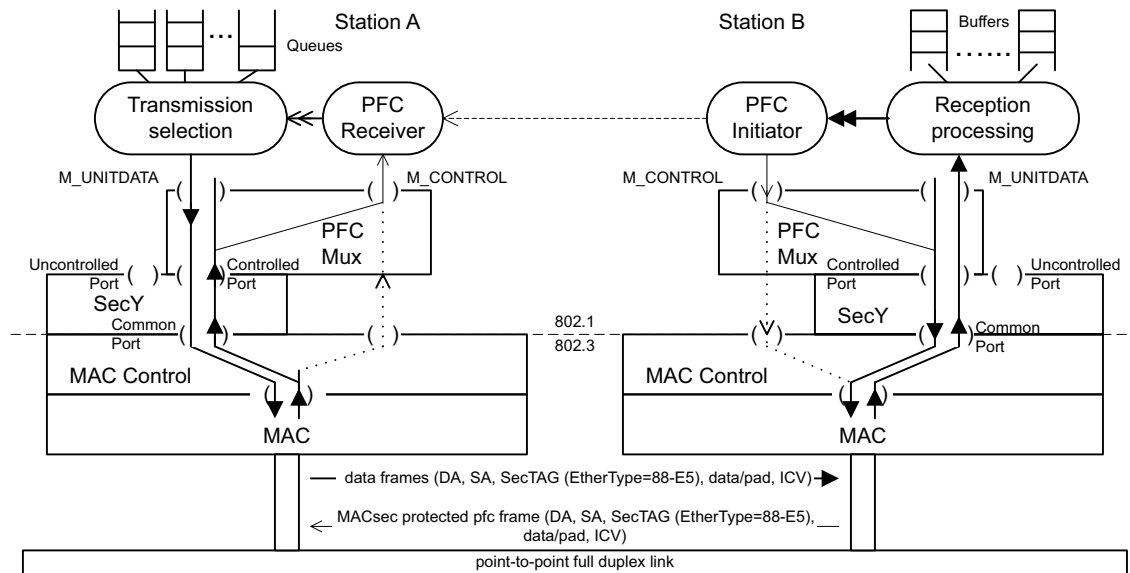
18 NOTE 1—IEEE Std 802.1AC-2018 specifies a maximum SecY transmit delay as the physical transmission time, at wire  
 19 speed, for a maximum sized MPDU and four 64-octet MPDUs, with an equal maximum SecY receive delay. If the  
 20 maximum sized MPDUs comprises 2000 octets, each of these delays is  $19\ 360$  bit times  $[8 \times (2000 + 20) + 8 \times 4 \times (64 +$   
 21  $12 + 4 + 20)$  bit times]. These maximums are appropriate for speeds up to 10 Gb/s.

22 Protection and validation at LAN speeds with the specified delay limits is facilitated by the parallelism  
 23 supported by the standardized MACsec Cipher Suites, and can be pipelined with frame transmission and  
 24 reception. IEEE Std 802.1AE-2018 did not separately limit delays for data frames passing through the SecY  
 25 when MACsec protection and validation are not applied, and some pipelined implementations can introduce  
 26 the same delay. The PFC configuration TLV of DCBX (D.2.10) includes a MACsec Bypass Capability

1 (MBC) bit. If MBC is set to one, the TLV’s recipient needs to take its peer SecY’s transmit and receive  
 2 delays into account when calculating PFC headroom (36.1.1), even when MACsec is not being used.

### 3 36.5 PFC with MACsec protection of user data and PFC frames

4 Figure 36-10 illustrates communication with MACsec protection of both PFC and data frames.



**Figure 36-10—MACsec protection of user data and PFC frames**

5 In Figure 36-10, Station B’s PFC Initiator makes an M\_CONTROL.request to a PFC Multiplexer, which  
 6 makes an ISS M\_UNITDATA.request to the SecY to initiate PFC. The parameters of the request comprise  
 7 the MAC destination address, the MAC source address of the station, priority, and a MAC Service Data Unit  
 8 (MSDU) comprising the EtherType 88-08 followed by the PFC opcode and the operand list as specified for  
 9 IEEE 802.3 MAC Control [item c) and d) in 36.3.1]. The effect of this request will be the transmission of a  
 10 MACsec protected (by B’s SecY) PFC frame. Its transmission is not subject to PFC control by the  
 11 transmitting station’s immediate peer (Station A in the figure). Since the MACsec EtherType (88-E5), rather  
 12 than the EtherType for MAC Control frames (88-08), immediately follows the frame’s source MAC  
 13 Address, the MAC Control sublayers treat this protected PFC frame as a data frame (31.3, 31.4  
 14 IEEE Std 802.3-2022). In Station A it is passed directly to the SecY, which validates (and, if necessary,  
 15 decrypts) the frame, removing the SecTAG with the MACsec EtherType and the ICV, before passing it to the  
 16 PFC multiplexer. The PFC Multiplexer recognizes the 88-08 EtherType and the PFC opcode, and invokes an  
 17 M\_CONTROL.indication to pass the MAC DA, opcode, and operand list to the PFC Receiver which  
 18 processes that indication as specified in 36.3.2. The PFC Multiplexer passes received frames with initial  
 19 protocol identifiers other than the 88-08 EtherType to the other user(s) of the SecY’s Controlled Port, and  
 20 discards received frames with the 88-08 EtherType that do not include the PFC opcode.

21 NOTE 1—When MACsec protected, the PFC frame and data frames are always Length/Type encoded. If media access  
 22 control method is not as specified in IEEE Std 802.3 and uses the SNAP SAP (see IEEE Std 802 ) to convey EtherTypes,  
 23 frames submitted to, and delivered by, the SecY can use the protocol identifier encoding specified for that method. In  
 24 that case their initial protocol identifier will be translated to and from Length/Type encoding as the SecTAG is added and  
 25 removed. See G.3.

26 If Station B’s MAC is configured to support preemption (6.7.2), PFC frames are transmitted as express  
 27 frames. A PFC Receiver communicates the need to pause transmission to system determined entities (such  
 28 as a Bridge’s Forwarding Process’s Transmission Selection function) and is thus capable of pausing  
 29 transmission for forwarded frames while still permitting PFC, network control, and management

1 transmission of frames of the same priority. However, a SecY's choice of preemption and Secure Channel  
 2 (SC) is based on the user priority accompanying each ISS M\_UNITDATA.request at its Controlled Port  
 3 (10.5, 10.7.17 of IEEE Std 802.1AE-2018), and is not a separate parameter of the ISS. To avoid delays to  
 4 PFC frames when both they and user data frames are protected by MACsec, PFC frames should be  
 5 transmitted with a priority that is assigned to an SC not used by preemptable frames (see Annex R). Other  
 6 frames not subject to PFC can be transmitted using the same SC.

7 Figure 36-10 also shows an alternate path for PFC frames, which is used if data frames are not protected by  
 8 MACsec. This is possible (see IEEE Std 802.1X) even if both stations implement MACsec. In that case the  
 9 PFC Multiplexer makes and accepts M\_CONTROL requests and indications directly to and from the MAC  
 10 Control sublayer.

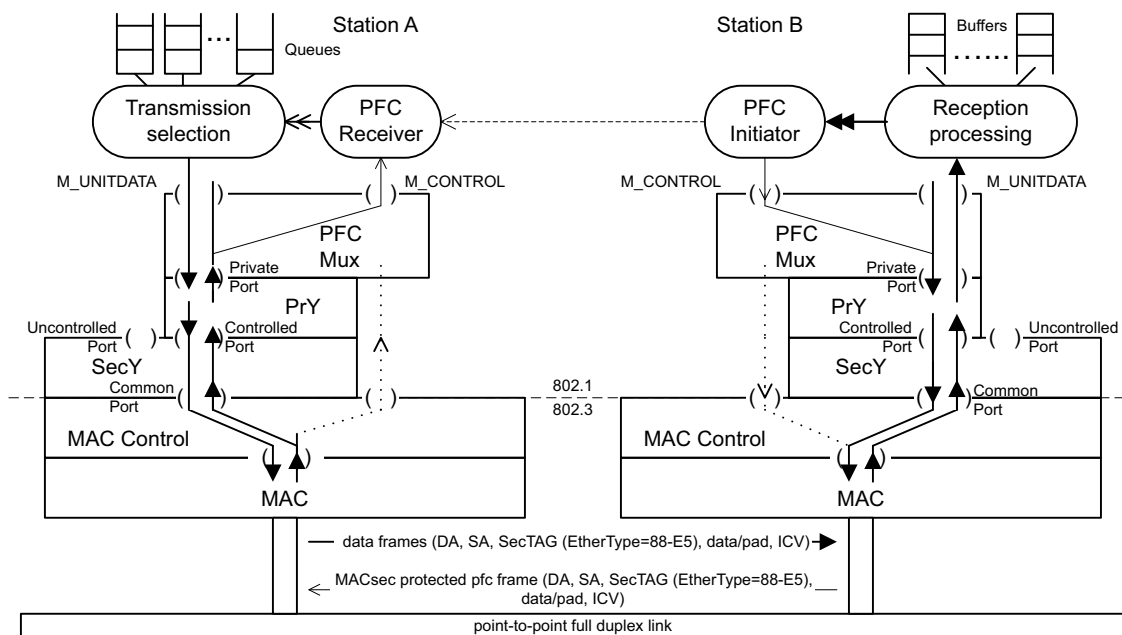
11 NOTE 2—If one of the peer stations does not implement the MAC Control sublayer it can transmit and receive PFC  
 12 frames which are not subsequently protected through the SecY's Controlled Port. If that station's peer implements MAC  
 13 Control, received PFC frames will give rise to M\_CONTROL indications.

### 14 36.5.1 PFC headroom with MACsec protection of PFC and data frames

15 When both PFC frames and data frames are MACsec protected, the headroom criteria in 36.4.1 are  
 16 applicable, with the additional consideration of delays introduced by PFC frame protection and validation.

### 17 36.6 PFC with MAC Privacy protection

18 Figure 36-11 illustrates communication with MAC Privacy protection of user data and PFC frames.



**Figure 36-11—MAC Privacy protection and PFC**

19 In Figure 36-11, user data and PFC frames are submitted to the MAC Privacy protection Entity (PrY). If  
 20 (and only if) the SecY is providing confidentiality protection, the PrY can add padding to obscure its  
 21 original length or can encapsulate the frame (possibly with other frames) to obscure its length, MAC

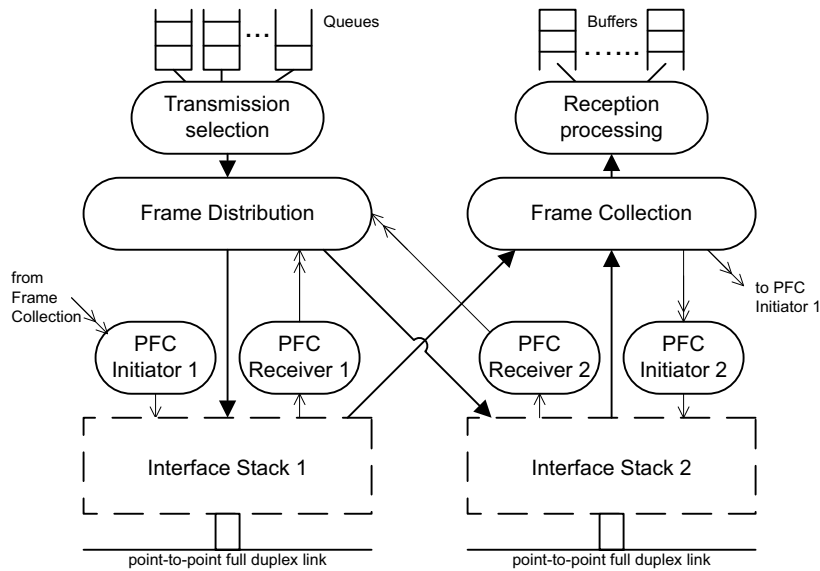
1 addresses, and the fact of its transmission (i.e., transmission unprotected, as an individual Privacy Frame, or  
 2 in a Privacy Channel as specified in Clauses 17 through 20 of IEEE Std 802.1AE).

3 NOTE—MAC Privacy protection was first standardized in the IEEE Std 802.1AE<sub>dk</sub>–2023 amendment to  
 4 IEEE Std 802.1AE–2018.

5 In addition to the possible mapping of priority by the SecY (36.5), the PrY can map the priority of Privacy  
 6 Frames and encapsulate multiple user data frames of different original user priority in a single Privacy  
 7 Channel frame. Where the MAC service data unit of the user data transmit request made to the PFC  
 8 Multiplexer (and passed unmodified to the PrY’s Private Port) includes a VLAN tag, that tag is both  
 9 integrity and confidentiality protected by the SecY, and can be used (in the figure, by the Reception  
 10 processing in Station B) to recover user priority (6.9.3, 6.9.4). Each PFC’s per-priority parameters apply to  
 11 that original user priority.

### 12 36.7 PFC with link aggregation

13 Figure 36-12 illustrates PFC operation for a port (a system interface, possibly a Bridge Port) that aggregates  
 14 two or more links.



**Figure 36-12—PFC operation with link aggregation**

15 The system includes a PFC Initiator and a PFC Receiver for each PFC-capable link, as shown in  
 16 Figure 36-11. The interface stacks shown can be any of those specified in 36.3 through 36.6. Each PFC  
 17 Receiver maintains Priority\_Paused variables for its link, for each priority, as specified in 36.3.2. If a system  
 18 wishes to pause reception on all the links in an aggregate it initiates PFC requests on each of those links.  
 19 Neither this standard nor IEEE Std 802.1AX constrains the organization and allocation of the buffering used  
 20 by reception processing, so an imminent buffer shortage can result in PFC initiation on a single, several, or  
 21 all, of the links in an aggregate.

22 Link Aggregation Control Protocol PDUs (LACP, IEEE Std 802.1AX), which support automated  
 23 configuration and reconfiguration of aggregates as link availability changes, are not be subject to PFC.

24 Repetitive pausing of transmission on a link can lead to redistribution of flows to other links. If a flow is  
 25 subject to PFC, so are the PDUs transmitted by the Marker protocol (6.5 of IEEE Std 802.1AX-2020) that

1 can be used to ensure in-order delivery of frames that are redistributed, potentially slowing redistribution.  
2 Conversation-Sensitive Collection and Distribution (6.6 of IEEE Std 802.1AX-2020) can also be used to  
3 redistribute flows, and uses LACPDUs.

#### 4 **36.8 PFC headroom calculation**

5 A system may determine the round-trip delay for PFC operation (36.1.1) for a given interface using either:

- 6 a) The sum of:
- 7 1) The system's local knowledge of its own implementation delays for PFC initiation and  
8 transmission [items a) through e) of 36.1.1].
  - 9 2) The link delay for transmission to and from the peer interface [items f) and j) of 36.1.1].
  - 10 3) System provided or configured values for the peer station's PFC reception, transmission  
11 selection pausing, and transmission completion delays [items g), h), and i) of 36.1.1].
  - 12 4) The system's local knowledge of its own implementation delays for user data frame reception  
13 [item k) of 36.1.1].

14 or

- 15 b) The round-trip delays reported by the PFC headroom measurement protocol (36.9), adjusted for:
- 16 1) The system's local knowledge of the maximum delay that would occur between:
    - 17 i) buffer consumption by reception processing, and
    - 18 ii) the transmission of a PFC  
19 i.e., [items a) and b) of 36.1.1], further adjusted for any differences between:
      - 20 iii) the maximum delay for PFC frame encoding and initiating transmission  
21 [item c) of 36.1.1], and
      - 22 iv) the delay between selection of a timestamp value to be encoded in a headroom  
23 measurement frame and initiating transmission of that frame.
  - 24 2) The peer system's assessment of the difference between:
    - 25 i) the maximum delay from the reception of a PFC to halting transmission selection for the  
26 affected priorities [item h) of 36.1.1], and
    - 27 ii) the delay between the reception of PFC headroom measurement request, and its  
28 processing by the PFC Receiver.

29 NOTE 1—The link delay or cable delay, i.e. the time required for frame propagation between stations is approximately 5  
30 microseconds per kilometer for optical fiber. At a notional data rate of 100 Gb/s, this adds approximately 125 kB/km of  
31 link length to PFC headroom (accounting for delays in both directions). For 10 Gb/s transmission cable delay becomes  
32 the dominant headroom factor for stations more than 1.2 km apart (120 meters for 100 Gb/s). Transmitted frames can  
33 include fields (e.g., SFD/Preamble for the IEEE 802.3 MAC) that do not require buffering following receipt, differences  
34 in the headroom required depend on frame length (a reduction of between 24% and 1% for the IEEE 802.3 MAC).

35 Further details of headroom calculation using link delay information [item a) above] and the PFC headroom  
36 measurement protocol [item b) above] are specified in 36.8.1 and 36.9.4 respectively.

37 At data rates of 100 Gb/s and above, a given PFC implementation's maximum sustained user data frame  
38 transmission rate can be less than implied by the nominal interface bit rate, thus reducing its peer's PFC  
39 buffering requirement.

40 NOTE 2—The sustainable user data frame bit rate for PFC-enabled priorities can also be reduced by the configuration of  
41 other system parameters that allocate bandwidth for different priorities or identified flows. Maximum rate reduction  
42 considerations are only significant for links with delays equivalent to many frame transmission times.

43 The result of PFC headroom calculation is made available to network management (36.10). Automated  
44 headroom calculation can take place even when its result is to be overridden by manual configuration, which

1 can specify an initial value (as the link is typically operational while measurement and calculation  
2 proceeds), and maximum and minimum values (36.10).

3 NOTE 3—The actual allocation of system memory as a consequence of headroom calculation is system dependent,  
4 reflects the structure of system buffering, and can be more or less efficient depending on frame size.

### 5 **36.8.1 Headroom calculation using link delays**

6 The PFC round-trip delay can be calculated by summing link, local, and remote delays [item a) of 36.8].

7 If the communicating PFC-capable stations participate in IEEE 1588, the sum of the link delays  
8 [item a) 2) of 36.8] should be as reported by IEEE 1588. Otherwise a locally configured value is used. The  
9 contribution of local system delays to the headroom calculation [items a) 1), a) 3), and a) 4) of 36.8] reflect  
10 delays with respect to the times that the frame's last bit passes each station's timing reference plane.

11 NOTE 1—While IEEE 1588 reports timing (for an IEEE 802.3 MAC, see IEEE Std 802.3cx–2023) with respect to  
12 transmission or reception of the first octet following the start of frame delimiter (SFD), the link delay from first octet  
13 transmitted to first octet received is the same (to the accuracy required for headroom calculation) as that from the  
14 transmission of the last frame bit to its reception. This standard references last bit transmission and reception times for  
15 consistency with the original specification and description in Annex O of IEEE Std 802.1Qbb–2011.

16 Management parameters for link delay based calculation are specified in 36.10.

## 17 **36.9 PFC headroom measurement protocol**

18 The headroom measurement protocol comprises transmission and reception of PFC measurement requests  
19 and PFC measurement responses in Headroom Measurement Protocol Data Units (HMPDUs, 36.9.5), and  
20 the recalculation of PFC headroom following reception of a PFC measurement response.

### 21 **36.9.1 Protocol purpose, goals, and non-goals**

22 Technological limitations on the location of buffering capable of supporting high data rates constrain the  
23 amount of buffering that is economically viable for some interfaces. In the absence of per interface  
24 configuration or determination of PFC headroom, buffering and bandwidth can be under-utilized (if a high  
25 'safe' default value is assumed, PFCs can be sent unnecessarily) and some otherwise viable network  
26 configurations can be unsupported (interfaces attached to long links are deprived of an appropriate share of  
27 buffering as a consequence of unnecessary allocations to those attached to short links).

28 The PFC headroom measurement protocol removes or reduces the need for administrative buffer allocation  
29 for lossless operation with PFC-enabled priorities for a station connected to a point-to-point link. It  
30 determines the maximum number of octets that the station could receive, assuming the peer station transmits  
31 at the full line rate, following a potentially imminent receive buffering exhaustion condition that results in  
32 PFC transmission before a pause in reception resulting from the peer's receipt of the PFC.

33 The measurement protocol design and implementations meet requirements for the following:

- 34 — Accuracy. Averaged results of headroom measurement are expected to estimate PFC headroom to  
35 within 8 pause quanta (512 octets). Headroom measurement addresses the requirement for buffer  
36 allocation, and is not intended as a substitute for clock synchronization. Measurement requests and  
37 responses traverse the peer interface stacks in the same way
- 38 — Timeliness. Headroom measurements are available shortly after connectivity is established between  
39 the peers, even if the peer interfaces become MAC\_Operational (6.8.2) at different times. Periodic  
40 measurement can be used if the link delays can change, e.g. through optical switching, without  
41 explicit interface signaling.
- 42 — Efficiency. Timeliness is not achieved by rapid repetitive transmission when the interface becomes  
43 MAC\_Operational, in competition with other startup protocols.



- 1 — Link length independence. The protocol operates with links of any length, irrespective of the  
2 number or frequency of measurement attempts, and without the requester or the responder having to  
3 maintain a record of those attempts.
- 4 — Coexistence. The measurement protocol can still be used if PFCs or PFC measurement protocol  
5 frame transmission is restricted, e.g., by stream gate configuration.
- 6 — Implementation independence. Peer communicating systems can use different transmission  
7 strategies and frequencies without compromising interoperability.

8 The measurement protocol does not specify:

- 9 — Buffer allocation. The buffering required to support PFC-enabled priorities depends on a number of  
10 implementation and situationally dependent factors. These include the PFC headroom, the degree to  
11 which buffering should exceed that loss-preventing minimum in order to avoid degrading bandwidth  
12 utilization and excessive PFC use, the organization of buffering within the system, the efficiency  
13 with which frames are expected to be stored in those buffers, and the possible utilization of the link  
14 by PFC-enabled priority traffic over the timescale corresponding to the PFC headroom.

### 15 **36.9.2 Addressing, protocol identification, and protocol versions**

16 The destination MAC address of each headroom measurement PDU (HMPDU) is the IEEE MAC-specific  
17 Control Protocols group address 01-80-C2-00-00-01, and the source MAC address is the individual MAC  
18 address of the transmitting station. The headroom measurement protocol is identified by the  
19 IEEE 802.1Q Congestion Isolation Message EtherType 89-A2 (Table 49-1) and the Subtype 01 (49.4.3.1.2).  
20 This standard specifies Version 0 (49.4.3.1.2) of the protocol. A conformant implementation shall process  
21 received HMPDUs of any received version as specified by this standard.

22 NOTE—As of this revision of this standard, future headroom measurement protocol versions are expected to support  
23 extensibility and interoperability using the following rules which are consistent with other IEEE 802.1 protocol  
24 specifications. HMPDUs with a Version field value lower than the protocol version implemented by the receiving station  
25 are processed according to the specification for the received Version field value. HMPDUs with a Version field value  
26 that is equal to or greater than that of the implemented version are processed as specified for the implemented version.  
27 The value communicated in the Version field of transmitted HMPDUs identifies the implemented version, and is not  
28 change by management or as a result of protocol exchanges with peer protocol participants. Each version specification  
29 identifies fields that are to be ignored, and are thus available for protocol extensions, and those that are reserved for  
30 future standardization by revision or amendment of this standard.

### 31 **36.9.3 Protocol parameter values, representation, and encoding**

32 Protocol parameters are specified as unsigned integers, signed integers, or flags. All HMPDUs comprise an  
33 integral number of octets. When shown in a figure these octets are numbered starting from 1, the first octet  
34 of the assigned EtherType, and bits within an octet are numbered from 8 (the most significant bit) to 1 (the  
35 least significant bit) and the most significant bit is shown to the left, with the remaining bits shown in  
36 decreasing order of bit significance.

37 When a parameter is specified as an unsigned integer, a meaning is attributed to all values in the range  
38  $0 \dots 2^n - 1$  for some specified integer  $n$ , and the value is encoded as a binary numeral in  $n$  bits in contiguous  
39 octets and contiguous bits within those octets with the most significant bit in the lowest numbered octet.  
40 Values can be represented in hexadecimal, with the most-significant nibble to the left preceded by '0x'. A  
41 decimal representation, without prefix or suffix, can also be used.

42 When a parameter is specified as a signed integer, a meaning is attributed to all values in the range  
43  $-2^{n-1} \dots 2^{n-1} - 1$  for some specified integer  $n$ , and the value is encoded as a two's complement binary numeral  
44 in  $n$  bits in contiguous octets and contiguous bits within those octets with the most significant bit in the  
45 lowest numbered octet. The values of unsigned integer parameters can be represented in hexadecimal, with  
46 the most-significant nibble to the left preceded by '0x'. A decimal representation, without prefix or suffix,  
47 can also be used with negative numbers preceded by '-'.  
48

1 Where a parameter is specified as a flag, it takes the value 0 or the value 1, and is encoded as binary numeral  
2 in a single bit. A value of 1 can also be represented as ‘set’ or ‘true’, and the value 0 as ‘clear’ or ‘reset’. The  
3 operations of ‘setting’ or ‘is set’ applied to the flag makes its value 1, independently of its prior value, and  
4 those of ‘clearing’ or ‘is cleared’ makes its value 0. The value of a sequence of flags encoded in contiguous  
5 bits can be represented by the hexadecimal representation of the identically encoded unsigned integer.

#### 6 **36.9.4 Measurement requests and responses**

7 An HMPDU can contain a measurement request or a response, or both a request and a response (36.9.5).

8 A measurement request comprises the following parameters:

- 9 — Request Timestamp. An implementation specific parameter, encoded in 32 bits.
- 10 — Request Adjustment. A number of pause quanta (36.3.1), a 16-bit signed integer.

11 A measurement response comprises the unchanged (reflected) parameters of the request, and the following:

- 12 — Response Adjustment. A number of pause quanta, a 16-bit signed integer.

13 The Request Timestamp does not have to be interpreted by the responder. The implementation specific  
14 content has to be sufficient to allow the requestor to calculate the elapsed time between acquiring the  
15 timestamp value encoded in the request and receiving the response with that reflected value.

16 NOTE 1—The Request Timestamp 32-bit field is sufficient to accommodate a wrapping unsigned integer that is  
17 continually updated at pause quanta (512 bit) intervals, without wrapping more than once during the round-trip time for  
18 1 Tb/s terrestrial transmission between data centers. However the initiator of the measurement request is not restricted to  
19 encoding a clock value in this field, but can encode any value that can be conveniently used to ascertain the elapsed time  
20 when the field is returned unchanged in a measurement response.

21 The Request Adjustment accounts for requesting system delays [b)1) of 36.8].

22 NOTE 2—The Request Adjustment parameter is included in HMPDUs to accommodate possible request by request  
23 variations in transmission timing, as might occur, e.g., as a result of transmission gate operation. Including the parameter  
24 removes any need for the requestor to reconcile a response with specific request, and allows multiple requests to be  
25 outstanding at any time. Implementations that do not need to account for transmission timing variation can make encode  
26 a zero or other fixed value and make any adjustment locally.

27 The Response Adjustment accounts for responding system delays [item b)2) of 36.8].

28 The round-trip delay for PFC operation is calculated, in pause quanta, as:

$$29 \quad (\text{ResponseDelay}) + \text{Request Adjustment} + \text{Response Adjustment}$$

30 where ResponseDelay is the value of the interval (in pause quanta) obtained on receipt of the response by  
31 comparing the Request Timestamp with the current timestamp, and deducting locally known fixed delays for  
32 request transmission and response processing. If the transmission of the measurement request is less timely  
33 (takes longer) after this adjustment than allowed for PFC transmission, the Request Adjustment will be  
34 negative (and encoded as a negative integer in the HMPDU). Similarly, if the peer system knows that its  
35 measurement response is less timely than the worst case for halting transmission the Response Adjustment  
36 will be negative (and encoded as a negative integer in the HMPDU).

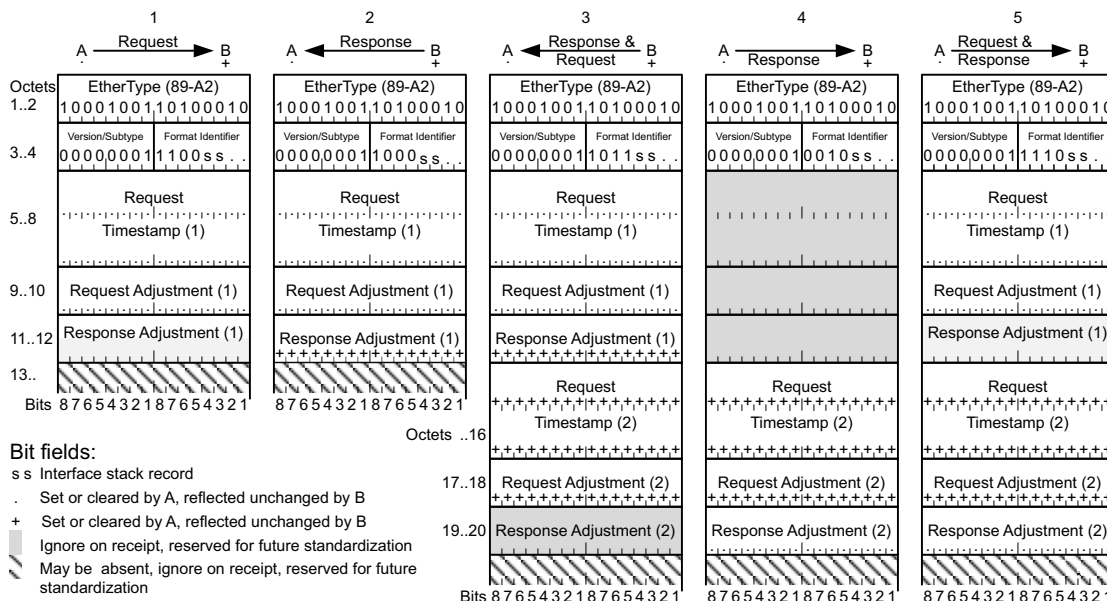
37 NOTE 3—If, e.g., a measurement response is delayed because several other frames are to be transmitted first, a negative  
38 Response Adjustment is appropriate. Contrariwise, if there are no prior frames to be transmitted, but one or more frames  
39 could already be selected for transmission when a PFC is received, a positive adjustment can be appropriate.

---

<sup>1</sup> Cable delay approximately 5 microseconds per kilometer (5 nanoseconds per meter) for optical fiber. 1 pause quanta is time to transmit 512 bits (~500 bits), delay at 100 Gb/s is ~1 pause quanta/meter.  $2^{31}$  meters ~ $2^{21}$  km, data center separation  $2^{20}$  km ~1million km. Circumference of earth ~40,000 km, round-trip through geostationary satellite ~160,000 km.

1 **36.9.5 HMPDU format**

2 Each HMPDU comprises a single octet Format Identifier followed by one or two {Timestamp Field,  
 3 Request Adjustment Field, Response Adjustment Field} tuples, as illustrated in Figure 36-13.



**Figure 36-13—HMPDU format (examples)**

4 Each HMPDU comprises the assigned EtherType 89-A2, its Version/Subtype, and a single octet Format Identifier followed by one or two {Timestamp Field, Request Adjustment Field, Response Adjustment Field} tuples.

7 The use of the first field tuple is determined by the values of bits 8 and 7 of the Format Identifier, and that of the second by the values of bits 6 and 5 as follows:

- 9 — 0x03 : The tuple is a measurement request.
- 10 — 0x02 : The tuple is a measurement response with a non-zero Response Adjustment.
- 11 — 0x01 : The tuple is a measurement response with a zero Response Adjustment, i.e. the content of the Response Adjustment field should be ignored on receipt.
- 12 — 0x00 : Reserved for future standardization. Version 0 implementations treat as unused.

14 If the Format Identifier identifies either field tuple as unused, any values encoded in the fields of that tuple are reserve for future standardization and are ignored on receipt. If the Format Identifier identifies the second field tuple as unused, those fields are not necessarily present in the HMPDU.

17 NOTE 1—The use of a full octet for the Format Identifier places the following 4-octet Timestamp Field on a 4-octet boundary with respect to the first octet of the preceding EtherType.

19 To measure the delay from PFC issuance to cessation of data reception, a measurement request traverses (as closely as possible) the interface stack path followed by the PFC, while the measurement response follows that used by the PFC-enabled data frames. Consequently when data frames are MACsec protected, but PFC frames are not (36.4), any given HMPDU will convey a measurement request or a measurement response, but not both, and the second field tuple will not be used. The latter also applies if PFC-enabled data frames are protected by a Privacy Channel but PFC frames are not.

1 Bits 4 and 3 of the Format Identifier convey interface stack information for the round-trip measurement:

- 2 — 0x00 : PFCs and user data frames are not MACsec protected .
- 3 — 0x01 : PFCs are not MACsec protected, user data frames are MACsec protected.
- 4 — 0x02 : PFCs and user data frames are MACsec protected.
- 5 — 0x03 : PFCs are transmitted in the Express Privacy Channel, user data frames are also transmitted in  
6 a Privacy Channel. PFC measurement requests and responses are both transmitted in the Express  
7 Privacy Channel. While the round-trip return in a Preemptable Channel can take longer, that extra  
8 time is not available for the transmission of user data frames and therefore does not result in a PFC  
9 headroom increment.

10 NOTE 2— A SecY can be configured to accept unprotected data frames before protection is operable, and PFCs can be  
11 both unprotected and protected, so more than one PFC measurement path is possible at a time. While the interface stack  
12 information in bits 4 and 3 could be available to a PFC Initiator or Receiver, interface stack sublayers intentionally  
13 remove the responsibility of understanding details of their operation from their clients. Frame by frame information  
14 availability is limited to that specified for the ISS.

15 NOTE 3—The measurement protocol determines a single PFC headroom value for all priorities for which PFCs and the  
16 data frames they pause are transmitted in the same way (protected, MACsec protected, or protected by a Privacy  
17 Channel) and does not account for the possibility of differing maximum length frames for different priorities.

18 Bits 2 and 1 of the Format Identifier are transmitted as zero and ignored on receipt.

### 19 **36.9.6 Measurement protocol operation**

20 HMPDUs are only transmitted when the transmitting station is also capable of receiving HMPDUs, and both  
21 transmitting and receiving user data frames (unprotected or MACsec protected, as configured).

22 A measurement request can be transmitted when a station that is configured to transmit PFCs to pause data  
23 frame reception wishes to improve its current estimate of PFC headroom, e.g., when:

- 24 a) An interface becomes MAC\_Operational (6.8.2).
- 25 b) CFM (Clause 18), or some other connectivity management protocol, has detected an interruption in  
26 connectivity that could indicate a change in link delay.
- 27 c) Frames received with PFC-enabled priorities are being discarded due to buffer shortage.
- 28 d) A change in measured headroom suggests additional measurement is desirable.

29 A measurement response shall be transmitted whenever a measurement request has been received. Each  
30 transmission of a measurement response provides an opportunity for the transmission of a measurement  
31 request in the same HMPDU, a measurement request can also be transmitted when a measurement response  
32 has been received. Otherwise measurement requests should not be repeated at intervals of less than the  
33 system dependent maximum acceptable round-trip delay (36.10). As a consequence of this restriction on  
34 request transmission, a protocol participant does not have to buffer more than two HMPDUs provided that it  
35 does not delay request or response transmission for longer than its peer's round-trip delay maximum.

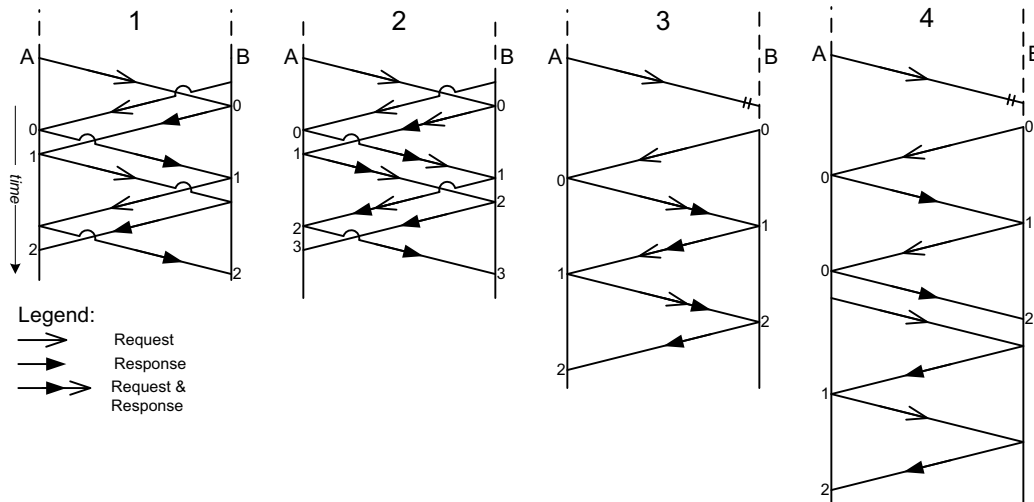
36 A measurement response tuple can be generated from a request tuple by replacing bits 8 and 7 of the Format  
37 Identifier (if the request was encoded in the first tuple) or bits 6 and 5 (if the request was encoded in the  
38 second tuple) with the appropriate code for the response. A Response Adjustment need not be added if its  
39 value would be 4 or less. Bits 4 through 1 are always reflected unchanged—the interface stack path whose  
40 delay is to measured is determined by the initial request.

41 Prior to measurement response reception, the PFC headroom estimate is an implementation dependent  
42 average of prior measurements, which can be persistent across transitions of MAC\_Operational or  
43 temporary interruptions in connectivity. If such a prior estimate is unavailable, an initial value is used ().

1 The measured round-trip delay is calculated (36.9.4) for each measurement response received. If the  
 2 calculated value is less than a system dependent minimum (36.10), the latter value is substituted. If the value  
 3 is greater than a system dependent, manageable, maximum (36.10), that maximum is substituted.

4 NOTE—Bounding round-trip delay times guards against poisoning the average of multiple measurements. While rapid  
 5 determination of round-trip delay after link up is desirable, that is also a time when other configuration protocols attempt  
 6 to achieve rapid results, with an increased likelihood of exceptional response delays.

7 Figure 36-14 provides examples of measurement protocol operation between stations A and B.



**Figure 36-14—Measurement (examples)**

8 In the first example in Figure 36-14, Station B is able to receive and transmit HMPDUS within the one-way  
 9 link delay time of Station A transmitting its first measurement request, and also transmits its first  
 10 measurement request within that time. Each station responds to each measurement request received,  
 11 encoding only (in this example) the measurement response in its next HMPDU transmission. The number of  
 12 responses received by each station as the PDU exchange proceeds is shown, each station being satisfied with  
 13 its PFC headroom estimate when two responses have been received. Each station's use of a measurement  
 14 response to prompt transmission of its next request paces their transmission to the link delay, yielding timely  
 15 results but avoiding having an excessive number of measurements in progress at any one time, and avoiding  
 16 the need for the implementation of brief interval timers.

17 In the second example, both stations encode requests and responses in the same HMPDU, each obtaining the  
 18 results of two measurements in slightly less time than in the first example, and (even though two might be  
 19 enough) the results of three in the time previously taken for two.

20 In the third example, Station B is not ready to receive the first request transmitted by A, but transmission of  
 21 both a request and a response in the HMPDUs that follow B's first request provide each of the stations with  
 22 two measurement results in less than three round-trip times after B transmits that first request.

23 In the fourth example, A's first transmission is also lost, and both stations transmit requests and responses in  
 24 separate HMPDUs. A retransmits a request after receiving two requests from B without an intervening  
 25 response, which indicates that A's initial request has been lost (since B's second response would not have  
 26 been sent until it had received A's later response). A's repeated request enables it to make two round-trip  
 27 measurements in less than four round-trip times after B's first request.

<sup>1</sup> The use of separate HMPDUs to convey requests and responses in the first and fourth example might be a  
<sup>2</sup> consequence of using unprotected PFCs to pause MACsec protected data frames, with an expected  
<sup>3</sup> difference in the one-way transmission delays for those two frame types.

#### <sup>4</sup> **36.10 PFC management**

<sup>5</sup>

1

## 2 **Existing 802.1Q Clause 36 issues**

3 A PDF markup of 802.1Q-2022 Clause 36 is/may be attached to this proposal. The notes immediately  
4 following only cover part of my concerns, some are just reminders for future investigation.

### 5 **36.1.1, “data center environment” is undefined:**

6 36.1.1 said “Operation of PFC is limited to a data center environment”. There is no definition of “data center  
7 environment’ in 802.1Q. The term is used elsewhere in 802.1Q, but in those case use of specified protocols  
8 outside that environment is not prohibited, so a general or loose understanding of the term is adequate. That  
9 is not the case here. Given the use for distributed data centers, connected by 60 km or longer links, the utility  
10 of the prohibition it is unclear. Technically it would seem possible for a PFC Initiator to transmit a stream of  
11 PFCs each specifying a transmission pause of less than the link delay and shorter than the PFC interval in  
12 order to pace reception, so it is also not clear that PFC is inoperable over long links or necessarily inferior to  
13 link window rotation protocols.

### 14 **Elegant variation, invoke/invocation where initiate/initiator/initiation would be consistent:**

15 Use initiate/initiator/initiation consistently for flow control requests rather than mixing in invoke/invocation.

### 16 **Open Questions:**

#### 17 **PFC request frequency:**

18 The existing specification of PFC places no limit on the frequency of PFC requests. Should there be one?  
19 This is not an easy question. The lack of a limit permits considerable flexibility as to PFC use. For example,  
20 a PFC Initiator could rate control transmission on a long link (say 200 Km, with a one way delay of  $2 \times 10^5$   
21  $/2 \times 10^8$  seconds = 1 millisecond) by transmitting PFCs, each specifying a pause of a fraction of the link  
22 delay at intervals that are themselves a fraction of the link delay (say, 100 microsecond pauses at 200  
23 microsecond intervals to halve the link rate). Quite apart from any discussion of whether that is a good way  
24 to rate control a link, is such behavior reasonable? What processing frequency is a PFC receiver expected to  
25 support? This is after all probably less of a burden that adding a sequence number to the front of every frame  
26 and adding sending explicit acks to rotate the reception window.

### 27 **Additional 802.1Q issues**

#### 28 **8.6.8 Transmission selection, NOTE 1 and NOTE 2:**

29 These notes contradict each other. NOTE 1 says pausing of transmission for other priorities assigned to the  
30 same traffic class as a paused priority *can* be paused, NOTE 2 says it *will* be paused. The notes should be  
31 replaced by a single, more carefully constructed, note.

#### 32 **8.6.8.4 Enhancements for scheduled traffic, NOTE-3 (currently misnumbered):**

33 Something needs to be done about this note, which currently reads:

34 “NOTE—3 The use of PFC is likely to interfere with a traffic schedule, because PFC is transmitted by a higher layer  
35 entity (see Clause 36).”

36 Clause 36 does not say that the PFC transmitting entity is a 'higher layer entity'. "likely" is a vague  
37 judgement call, "can" would be appropriate if there is no coordination between PFC transmission and the  
38 transmission selection gates specified in 8.6.8.4, there is nothing in the standard which says such  
39 coordination is forbidden. Text could be added permitting such coordination, noting that further receive  
40 buffering is required to accomodate the delay in PFC transmission.

1 **D.2.10.6, PFC Enable:**

2 The description of PFC Enable in [D.2.10.6 PFC Enable](#) does not say what “enable” actually means in this  
3 context. Does it mean that a PFC can be transmitted, by the station transmitting the TLV, if there is an  
4 imminent risk of overflow for the specified priorities (so the peer receiving the DCBX Priority-based Flow  
5 Control Configuration TLV, should be prepared to act on a PFC specifying one or more of those priorities)?  
6 Alternatively does it mean that a received PFC specifying one those priorities will be acted upon by the TLV  
7 transmitter, so that transmission of a PFC with such a priority is not a futile act. The definition of the  
8 “Willing bit” in D.2.10.3 “A value of one indicates that the station is willing to accept configurations from  
9 the remote station” is no help because it does not say what effect “accept configurations” will have in this  
10 case. Nor does [D.2.10.6 PFC Enable](#) item c): “Local policy in each end of the link decides whether to use the  
11 priority if the configuration does not match.” There may have been some thought that PFC configurations  
12 should be symmetric, but why the fact that one station (perhaps part of an edge switch with minimal total  
13 buffering to forward frames into the network) should have the same PFC requirements as its immediate peer  
14 (perhaps an end station, running a different operating system with a completely different memory  
15 architecture) for flows proceeding *in the opposite direction* is beyond me. The PFC MIBs that I am aware of  
16 do not support direct configuration of “PFC Enable” or an equivalent management variable, nor have I  
17 found any accompanying commentary on how their controls (starting from configuration of traffic classes)  
18 would affect the PFC configuration TLV.

19 **802.1Q-2022 Annex M, status, adequacy**

20 802.1Q Annex M states that it is (a) Normative, and (b) describes a PDU format suitable to support PFC in  
21 link layers that support point-to-point full-duplex operation, other than those specified in IEEE Std 802.3.

22 First, Annex M cannot be normative in the scope of IEEE Std 802.1Q because the ‘other link layers’ are not  
23 within the scope of IEEE Std 802.1Q and a standard cannot define normative provisions outside their scope.  
24 IEEE Std 802.1AC includes the MAC Control primitives that support PFC, but, properly, does not specify  
25 the details of individual MAC Support for those primitives. The ISS mapping provisions for IEEE Std 802.3  
26 for MAC Control are described in 6.7.1 of IEEE Std 802.1Q, but align completely with the IEEE Std 802.3  
27 specification.

28 Second, Annex M is deficient in its description of PDU format because it does not describe the context in  
29 which the suggested PDU is encoded, in particular it does not describe:

- 30 a) The destination and source MAC addresses. It is vital that a PFC frame not be forwarded, by a  
31 bridge or any similar frame forwarding device, from one point-to-point link to another. IEEE Std  
32 802.3 mandates the use of the 01-80-C2-00-00-01 “IEEE MAC-specific Control Protocols group  
33 address” for this purpose. Frames with this destination address are not forwarded by any type of  
34 bridge (MAC Bridge, VLAN Bridge, TPMR, Provider Bridge, Provider Backbone Bridge) specified  
35 by IEEE Std 802.1Q.
- 36 b) Protocol discrimination. IEEE Std 802.3 assigned the EtherType 88-08 to identify MAC Control  
37 frames (including, but not limited to PFC). Any 802.3 station that implements MAC Control  
38 recognizes any and all received frames with this EtherType as MAC Control frames without regard  
39 to its destination or source MAC Address, provided that the station is configured to receive frames  
40 with the destination MAC Address. While other link layers could use other ways to distinguish PFC  
41 frames, it is vital that they be distinguished using a method common to all

42 Finally, with the above omissions, Annex M says nothing other than that other link layers should use a same  
43 packet format as IEEE 802.3. While this might be a sensible choice, it hardly warrants a Normative Annex.  
44 Annex M should be removed.



1 **IEEE Std 802.1AX-2020 Issues**

2 **LACPDU priority**

3 I can find no statement as to the priority to be used to transmit LACPDU.

4 **IEEE Std 802 issues**

5 **802f EtherType 88-08 description:**

6 The ‘Short Description’ ‘Multipoint Control Protocol (MPCP)’ in P802f/D2.4 of the EtherType 88-08 used  
7 to distinguish MAC Control frames is misleading. This EtherType is used for all 802.3 frames processed by  
8 MAC Control, not just for MPCP.

9

1

## 2 **Additional 802.1Q references**

### 3 **6.7.1 Support of the ISS by IEEE Std 802.3 (Ethernet):**

4 “Mapping between M\_CONTROL.requests/indications and IEEE802.3 MA\_CONTROL.requests/  
5 indications is performed as specified in IEEE Std 802.1AC. If the MAC supports the MAC Merge sublayer  
6 specified in IEEE Std 802.3, then PFC M\_CONTROL.requests are mapped onto the MAC control interface  
7 associated with the express MAC (eMAC).”

### 8 **12.23 Priority-based Flow Control objects**

9 The following Priority-based Flow Control objects exist for each port that support PFC:

- 10 a) PFCLinkDelayAllowance: the allowance made for round-trip propagation delay of the link in bits
- 11 b) PFCRequests: a count of the invoked PFC M\_CONTROL.request primitives
- 12 c) PFCIndications: a count of the received PFC M\_CONTROL.indication primitives

13 Table 12-21 shows the format and applicability of these objects.

14 NOTE-The PFC Initiator (see 36.2.1) can use the PFCLinkDelayAllowance parameter as one of the factors to determine  
15 when to issue a PFC M\_CONTROL.request in order to not discard frames. The parameter can be written to adjust to  
16 different link characteristics that affect the link delay (e.g., link length or link technology). See Annex N for an example  
17 of how to compute this parameter.

### 18 **17.2.17 Structure of the IEEE8021-PFC-MIB**

19 Table 17-23 describes the relationship between the SMIV2 objects defined in the PFC-MIB module (17.7.13)  
20 and the variables and managed objects defined in Clause 12 and Clause 36.

### 21 **17.3.17 Relationship of the IEEE8021-PFC-MIB to other MIB modules**

### 22 **17.4.17 Security considerations of the IEEE8021-PFC-MIB**

### 23 **17.7.17 Definitions for the IEEE8021-PFC-MIB module**

24 ...

25 ieee8021PfcLinkDelayAllowance OBJECT-TYPE

26 SYNTAX Unsigned32

27 MAX-ACCESS read-write

28 STATUS current

29 DESCRIPTION

30 "The allowance made for round-trip propagation delay of the link in bits."

31 ...

32 ieee8021PfcRequests OBJECT-TYPE

33 SYNTAX Counter32

34 UNITS "Requests"

35 MAX-ACCESS read-only

36 STATUS current

37 DESCRIPTION

38 "A count of the invoked PFC M\_CONTROL.request primitives

39 ieee8021PfcIndications OBJECT-TYPE

40 SYNTAX Counter32

41 UNITS "Indications"

42 MAX-ACCESS read-only

43 STATUS current

44 DESCRIPTION

45 "A count of the received PFC M\_CONTROL.indication primitives.

1 **37. Enhanced Transmission Selection (ETS), 37.3 ETS algorithm:**

2 References to PFC in items d) and e).

3 **38. Data Center Bridging eXchange protocol (DCBX), 38.2 Goals:**

- 4 a) Discovery of DCB capability in a peer port; for example, it can be used to determine if two link peer  
5 ports support PFC.

6 **49. Congestion Isolation**

7 Clause 49 clause begins:

8 “Congestion Isolation (CI) mitigates head-of-line blocking caused by the frequent use of PFC in lossless  
9 networks and reduces frame loss in lossy networks that are not using PFC.”

10 In the fourth paragraph:

11 “Queuing delays deter the end-to-end congestion control loop, and in a lossless environment, cannot prevent  
12 Priority-based Flow Control (PFC) from being invoked (see Clause 36). When buffers fill and eventual  
13 flow-control kicks in (for lossless networks), non-congesting flows can be blocked by the backlog of frames  
14 from congesting flows. If PFC is not being used, frame loss for non-congesting flows can result in long  
15 retransmission timeouts,...”

16 **49.1 Congestion isolation objectives**

17 “d) Reduce the frequency of invoking PFC in a lossless environment.”

18 “m) Reduce head-of-line blocking of victim flows at upstream peers from PFC.”

19 **49.2.7 System topology and port orientation**

20 Fifth paragraph:

21 “Lossless networks enabled by PFC have been shown, in certain circumstances, to have circular buffer  
22 dependencies that can cause deadlocks when traffic is re-routed due to link failures [B5]. Again, knowing  
23 the position in the topology assists in knowing when traffic has been re-routed and can be used to break  
24 circular buffer dependent deadlocks [B4].”

25 **D.2.10.6 PFC Enable**

26 “Table D-6 shows the layout of the PFC Enable bit vector.”

27 “A bit vector of 8 bits, one per priority:

- 28 a) A one indicates PFC is enabled on the priority.  
29 b) A zero indicates that PFC is disabled on the priority.  
30 c) Local policy in each end of the link decides whether to use the priority if the configuration does not  
31 match.”

32 **D.5.5 IEEE 802.1 LLDP extension MIB module version 2**

33 Contains a number of PFC related items.

## 1 **W.2 Congestion Isolation queuing and Priority-based Flow Control**

2 Discusses the subject in general, with some text particular to PFC even without Congestion Isolation. First  
3 paragraph describes PFC implementation buffering flexibility.

4 Second paragraph (extract):

5 “PFC is known to cause congestion spreading and has recommended use within the data center because of  
6 its limited extent (36.1.1). One of the key objectives for congestion isolation is to reduce the frequency of  
7 PFC requests and avoid head-of-line blocking in lossless data center networks. By reducing the frequency of  
8 PFC requests the impact of congestion spreading can be reduced.”

## 9 **IEEE Std 802.3-2022 references**

### 10 **Figure 1-1:**

11 Figure 1-1 (and many others) shows the relationship of the MAC Control optional sublayer to the MAC  
12 (below) and MAC Clients (above).

### 13 **2.3.2 MA\_DATA.indication, 2.3.2.2 Semantics of the service primitive:**

14 “This primitive defines the transfer of data from the MAC sublayer entity (through the optional MAC  
15 Control sublayer, if implemented) to the MAC client entity or entities in the case of group addresses.”

### 16 **2.3.2 MA\_DATA.indication, 2.3.2.3 When generated:**

17 “The MA\_DATA.indication is passed from the MAC sublayer entity (through the optional MAC Control  
18 sublayer, if implemented) to the MAC client entity or entities to indicate the arrival of a frame to the local  
19 MAC sublayer entity that is destined for the MAC client. Such frames are reported only if ... and their  
20 destination address designates the local MAC entity. Frames destined for the optional MAC Control  
21 sublayer are not passed to the MAC client if the MAC Control sublayer is implemented.”

### 22 **2.3.2 MA\_DATA.indication, 2.3.2.5 Additional comments:**

23 “If the local MAC sublayer entity is designated by the destination\_address parameter of an  
24 MA\_DATA.request, the indication primitive will also be invoked by the MAC entity to the MAC client  
25 entity. This characteristic of the MAC sublayer may be due to unique functionality within the MAC sublayer  
26 or characteristics of the lower layers (for example, all frames transmitted to the broadcast address will  
27 invoke MA\_DATA.indication at all stations in the network including the station that generated the request).”

### 28 **4.1 Functional model of the MAC method, 4.1.1 Overview (fourth paragraph):**

29 “An optional MAC control sublayer, architecturally positioned between LLC (or other MAC client) and the  
30 MAC, is specified in Clause 31. This MAC Control sublayer is transparent to both the underlying MAC and  
31 its client (typically LLC). The MAC sublayer operates independently of its client; i.e., it is unaware whether  
32 the client is LLC or the MAC Control sublayer. This allows the MAC to be specified and implemented in  
33 one manner, whether or not the MAC Control sublayer is implemented. References to LLC as the MAC  
34 client in text and figures apply equally to the MAC Control sublayer, if implemented.”

## 35 **30. Management, 30.3 Layer management for DTEs, 30.3.3 MAC control entity object class,** 36 **30.3.3.2 aMACControlFunctionsSupported**

37 “A SEQUENCE that meets the requirements of the description below: PAUSE PAUSE command  
38 implemented MPCP MPCP implemented PFC PFC implemented EXTENSION EXTENSION MAC  
39 Control frame supported”

### 1 **30.3.3.6 aPFCEnableStatus (enabled or disabled)**

2 “A read-only value that indicates whether PFC MAC Control operation is enabled. The value enabled  
3 indicates that operation of PFC MAC Control is enabled and operation of PAUSE MAC Control is disabled.  
4 The value disabled indicates that transmission and reception of PFC MAC Control is not enabled and  
5 PAUSE MAC Control may operate if it has been enabled through another mechanism.

6 NOTE 1—aPFCEnableStatus is read-only to avoid the risk of it being set to a conflicting value with enablement of PFC  
7 in the MAC Control Client. It is intended that an implementation locally sets the value to enabled when the MAC  
8 Control Client has PFC enabled for any priority and to disabled when the MAC Control Client has PFC disabled for all  
9 priorities.

10 NOTE 2—There is no mechanism in this Clause to enable and disable PAUSE transmit and receive for PHYs without  
11 Auto-Negotiation. IEEE Std 802.3.1 provides dot3PauseAdminMode to enable and disable PAUSE in the absence of  
12 Auto-Negotiation.”

### 13 **31. MAC Control, 31.2 Layer architecture:**

14 “The MAC Control sublayer is a client of the CSMA/CD MAC. Figure 311 depicts the architectural  
15 positioning of the MAC Control sublayer with respect to the CSMA/CD MAC and the MAC Control client.  
16 MAC Control clients may include the Bridge Relay Entity, LLC, or other applications.”

### 17 **31.3 Support by interlayer interfaces:**

18 See Figure 31-2.

19 “All MAC frames validly received by the CSMA/CD MAC are passed to the MAC Control sublayer for  
20 interpretation. If the MAC frame is destined for the MAC client, the MAC Control sublayer generates an  
21 MCF:MA\_DATA.indication primitive, providing complete transparency for normal data exchange between  
22 MAC clients. If the MAC frame is destined for the MAC Control sublayer entity, it is interpreted and acted  
23 on internal to the MAC Control sublayer. This may result in state changes within the MAC Control sublayer,  
24 the generation of MA\_CONTROL.indication primitives, or other actions as necessary to support the MAC  
25 Control sublayer function. MAC PFC time limitControl sublayer functions shall always sink MAC Control  
26 frames.”

27 “In the MAC:MA\_DATA.indication primitive, MAC frames destined for the MAC Control sublayer (MAC  
28 Control frames) are distinguished from MAC frames destined for MAC clients by a unique Length/Type  
29 field identifier.”

### 30 **31.4 MAC Control frames**

31 “MAC Control frames are distinguished from other MAC frames only by their Length/Type field identifier.”

#### 32 **31.4.1.1 Destination Address field**

33 “The Destination Address field of a MAC Control frame contains the 48-bit address of the station(s) for  
34 which the frame is intended. It may be an individual or multicast (including broadcast) address. Permitted  
35 values for the Destination Address field may be specified separately for each MAC Control opcode in the  
36 annexes to Clause 31.”

#### 37 **31.4.1.3 Length/Type field**

38 “The Length/Type field of a MAC Control frame is a 2-octet field that shall contain the hexadecimal value:  
39 88-08. This value carries the EtherType interpretation (see 3.2.6), and has been universally assigned for  
40 MAC Control of CSMA/CD LANs.”

### 41 **31.5 Opcode-independent MAC Control sublayer operation**

42 “The MAC passes to the MAC Control sublayer all valid MAC frames via the MA\_DATA.indication  
43 primitive. Invalid MAC frames are not passed to the MAC Control sublayer (see 3.4).”

### 1 31.5.1 Frame parsing and data frame reception

2 “Upon receipt, the MAC Control sublayer parses the incoming MAC frame to determine whether it is  
3 destined for the MAC client (data frame) or for a specific function within the MAC Control sublayer entity  
4 itself (MAC Control frame).”

5 “A MAC frame that does not contain the unique Length/Type field specified in 31.4.1.3 is a data frame. The  
6 receipt of a data frame results in the generation of a MCF:MA\_DATA.indication primitive by the MAC  
7 Control sublayer, with its parameters identical to the MAC:MA\_DATA.indication primitive.”

### 8 31.5.2 Control frame reception

9 “If the MAC Control sublayer entity does not support the function requested by the specified opcode, it  
10 discards the MAC Control frame. The discard of a frame in this manner may be reported to network  
11 management.”

### 12 Annex 31A (normative) MAC Control opcode assignments:

13 Table 31A1 shows the currently defined opcode values and interpretations: 01-01 is assigned to PFC,  
14 specified in Annex 31D and IEEE Std 802.1Q: “Requests that the recipient stops transmissions in the  
15 priorities indicated in the parameters of the function for a period of time also indicated in the parameters.”

### 16 Annex 31D (normative) MAC Control PFC operation, 31D.1 PFC description:

17 “The Priority-based Flow Control (PFC) operation is used to inhibit transmission of data frames on one or  
18 more priorities for a specified period of time. The behavior of a MAC Control client supporting PFC  
19 operation is specified in IEEE Std 802.1Q. A MAC Control client wishing to inhibit transmission of data  
20 frames from the link partner generates a MA\_CONTROL.request primitive specifying:

- 21 a) The globally assigned 48-bit multicast address 01-80-C2-00-00-01.
- 22 b) The PFC opcode.
- 23 c) A request\_operand list with two operands: priority\_enable\_vector and time\_vector. (See 31D.2.)

24 Unlike the MAC Control PAUSE operation, the inhibition of frames for the PFC operation occurs in the  
25 MAC Control client. Upon receiving a PFC frame, the only action in MAC Control is to generate a  
26 MA\_CONTROL.indication primitive with the indication\_operand list specified in Table 31A9.

27 The PFC operation does not inhibit transmission of MAC Control frames.

28 PFC operation shall not be enabled on DTEs configured to the half-duplex mode of operation. PFC is  
29 intended for use over full-duplex point-to-point links. Use on shared media such as EPON is out of the scope  
30 of this standard.

31 The globally assigned 48-bit multicast address 01-80-C2-00-00-01 has been assigned for use in MAC  
32 Control frames. Bridges conformant to IEEE Std 802.1Q will not forward frames sent to this multicast  
33 destination address, regardless of the state of the bridges ports, and whether or not the bridge implements the  
34 MAC Control sublayer. To allow PFC full duplex flow control, stations implementing the PFC operation  
35 shall instruct the MAC (e.g., through layer management) to enable reception of frames with destination  
36 address equal to this multicast address.”

### 37 Annex 31D (normative) MAC Control PFC operation, 31D.2 Parameter semantics:

38 “The PFC opcode takes the following request\_operand\_list:

39 priority\_enable\_vector:

40 A 2-octet vector. The most significant octet is reserved (i.e., set to zero on transmission and ignored  
41 on receipt). Each bit of the least significant octet indicates if the corresponding field in the  
42 time\_vector parameter is valid. The bits of the least significant octet are named e[0] (the least  
43 significant bit) to e[7] (the most significant bit). Bit e[n] refers to Priority n. For each e[n] bit set to

1 one, the corresponding time[n] value is valid. For each e[n] bit set to zero, the corresponding time[n]  
2 value is invalid.

3 **time\_vector:**

4 A list of eight 2-octet fields named time[0] to time[7]. The eight time[n] values are always present  
5 regardless of the value of the corresponding e[n] bit. Each time[n] field is a 2-octet, unsigned integer  
6 containing the length of time for which the receiving station is requested to inhibit transmission of  
7 data frames associated with Priority n. The field is transmitted most significant octet first, and least  
8 significant octet second. The time[n] fields are transmitted sequentially, with time[0] transmitted  
9 first and time[7] transmitted last. Each time[n] value is measured in units of pause\_quanta, equal to  
10 the time required to transmit 512 bits of a frame at the data rate of the MAC. Each time[n] field can  
11 assume a value in the range of 0 to 65 535 pause\_quanta.”

#### 12 **Annex 31D (normative) MAC Control PFC operation, 31D.3 PFC transmit:**

13 “Upon receipt of a MA\_CONTROL.request primitive containing the PFC opcode from a MAC client, the  
14 MAC Control sublayer calls the MAC sublayer MAC:MA\_DATA.request service primitive with the  
15 following parameters:

- 16 a) The destination\_address is set equal to the destination\_address parameter of the  
17 MA\_CONTROL.request primitive. This parameter is currently restricted to the value specified in  
18 31D.1.
- 19 b) The source\_address is set equal to the 48-bit individual address of the station.
- 20 c) The length/type field (i.e., the first two octets) within the mac\_service\_data\_unit parameter is set to  
21 the IEEE 802.3 MAC Control EtherType value assigned in 31.4.1.3.
- 22 d) The remainder of the mac\_service\_data\_unit is set equal to the concatenation of the PFC opcode  
23 encoding (see Annex 31A), the priority\_enable\_vector and the time\_vector specified in the  
24 MA\_CONTROL.request primitive, and a field containing zeros of the length specified in 31.4.1.6.
- 25 e) The frame\_check\_sequence is omitted.”

#### 26 **Annex 31D (normative) MAC Control PFC operation, 31D.5 PFC receive**

27 “Upon receipt of a valid MAC Control frame with the opcode indicating PFC and the destination address  
28 indicating the globally assigned multicast address specified in 31D.1, the MAC Control sublayer generates  
29 the MA\_CONTROL.indication to the MAC Control Client.”

#### 30 **IEEE Std 802.3.1-2013 references**

##### 31 **10. Ethernet-like interface MIB module, 10.4 MIB module definition, excerpts:**

```
32 dot3ControlFunctionsSupported OBJECT-TYPE
33     SYNTAX BITS {
34         pause(0), -- 802.3 pause flow control
35         mpcp(1), -- 802.3 multi-point control protocol
36         pfc(2) -- 802.3 priority-based flow control
37     }
38 MAX-ACCESS    read-only
39 STATUS        current
40 DESCRIPTION   "A list of the possible MAC Control functions
41               implemented for this interface."
42 REFERENCE    "IEEE Std 802.3, 30.3.3.2,
43               aMACControlFunctionsSupported."
44 ::= { dot3ControlEntry 1 }
45
46
47 dot3PFCTable OBJECT-TYPE
48     SYNTAX      SEQUENCE OF Dot3PFCEntity
49     MAX-ACCESS  not-accessible
50     STATUS      current
```

```
1      DESCRIPTION  "A table of descriptive and status information
2                  about the MAC Control Priority-based Flow Control
3                  function on the Ethernet-like interfaces attached to
4                  a particular system. There will be one row in
5                  this table for each Ethernet-like interface in
6                  the system which supports the MAC Control PFC
7                  function (i.e., the pfc bit in the
8                  corresponding instance of
9                  dot3ControlFunctionsSupported is set). If some,
10                 but not all, of the Ethernet-like interfaces in
11                 the system implement the MAC Control PFC
12                 function (for example, if some interfaces only
13                 support half-duplex), there will be fewer rows
14                 in this table than in the dot3StatsTable."
15      ::= { ieee8023etherMIBObjects 14 }
16 dot3PFCEntry OBJECT-TYPE
17     SYNTAX          Dot3PFCEntry
18     MAX-ACCESS      not-accessible
19     STATUS          current
20     DESCRIPTION    "An entry in the table, containing information
21                   about the MAC Control PFC function on a single
22                   Ethernet-like interface."
23     INDEX { dot3StatsIndex }
24     ::= { dot3PFCTable 1 }
25
```

## 26 IEEE Std 802.1AC-2016 references

### 27 11.4 Control primitives and parameters:

28 The ISS provides two control primitives, an M\_CONTROL.request and an M\_CONTROL.indication,  
29 and their associated parameters.

30 NOTE—These control primitives are used in IEEE Std 802.1Q in order to support Priority-Based Flow Control (5.11  
31 and Clause 36 of IEEE Std 802.1Q-2014).

32 The M\_CONTROL.request primitive has the following form:

```
33 M_CONTROL.request  (
34                   destination_address,
35                   opcode,
36                   request_operand_list
37                   )
```

38 The M\_CONTROL.indication primitive has the following form:

```
39 M_CONTROL.indication (
40                   opcode,
41                   indication_operand_list
42                   )
```

## 43 IEEE Std 802.1AX-2020 references

### 44 6.5 Marker protocol, 6.5.1 Introduction:

45 ...“Marker/Marker Response PDUs are subject to the operation of flow control, where supported on the link.  
46 Hence, if the Frame Distribution function requests transmission of a Marker PDU on a given link and does  
47 not transmit any further frames that relate to a given set of conversations until the corresponding Marker  
48 Response PDU is received from that link, then it can be certain that there are no frames related to those  
49 conversations still to be received by the Partners Frame Collection function.

50 NOTE—The use of the Marker protocol is further discussed in Annex B. An alternative to the Marker protocol is  
51 defined in 6.6.”



1

## 2 **Related P802.1Qdt contributions**

3 Contributions listed in date order. Detail in later contributions may supersede that in earlier ones.

### 4 **Headroom Measurement Protocol Design (2023-07-11, Lily Lv, Fei Chen)**

5 <https://www.ieee802.org/1/files/public/docs2023/dt-lv-design-topics-0723-v1.pdf>

6 Note—Does not completely supersede dt-lv-design-topics-0517-v1.pdf (see below).

### 7 **Headroom Measurement Protocol Design (2023-05-17, Lily Lv, Fei Chen)**

8 <https://www.ieee802.org/1/files/public/docs2023/dt-lv-design-topics-0523-v2.pdf>

9 Detailed discussion of timestamping points.

10 Records decision to use Qcz (CI) EtherType 89-A2, with subtype for measurement protocol.

### 11 **PFC round trip timing protocol (2023-03-14, Seaman)**

12 <https://www.ieee802.org/1/files/public/docs2023/dt-seaman-round-trip-timing-0323-v0.pdf>

### 13 **Headroom Measurement Protocol Design (2023-03-13, Lily Lv, Fei Chen)**

14 <https://www.ieee802.org/1/files/public/docs2023/dt-lv-design-topics-0323-v2.pdf>

15 Records decision to use Qcz (CI) EtherType 89-A2, with subtype for measurement protocol.

### 16 **dt-lv-PFC-management-objects-0123-v1.pdf**

17 <https://www.ieee802.org/1/files/public/docs2023/dt-lv-PFC-management-objects-0123-v1.pdf>

18 Note—includes many objects that need to be included in the proposed Clause 36 text.

### 19 **Qdt Development--- Feedback from TSN presentation...(2022-12-12, Lily Lv)**

20 <https://www.ieee802.org/1/files/public/docs2022/dt-lv-headroom-discussion-feedback-1222-v1.pdf>

21 This presentation records feedback from the TSN WG discussion of the PAR revision that refocused PFC  
22 delay measurement on the round trip delay from detection of potential buffer shortage, through PFC  
23 initiation, receipt, and cessation of data transmission, to the PFC initiators buffering of the last data frame to  
24 be transmitted by the peer station before its transmission is paused. It contains useful information on the  
25 adoption of 802.1AS/IEEE 1588 and the measurement of Pdelay.

26 Note—<https://www.ieee802.org/1/files/public/docs2022/dt-lv-headroom-measurement-discussion-1122-v1.pdf> was part  
27 of the referenced presentation.

### 28 **PFC Enhancement Discussion ---PFC Management (2022-07-11)**

29 <https://www.ieee802.org/1/files/public/docs2022/dt-lv-pfc-management-discussion-0722-v01.pdf>

### 30 **P802.1Qdt Discussion (2022-06-07, Lily Lv, Paul Congdon, Mick Seaman)**

31 <https://www.ieee802.org/1/files/public/docs2022/dt-lv-pfc-enhancement-Discussion-0622-v01.pdf>

### 32 **P802.1Qdt Text Contribution (2022-05-11, Lily Lv)**

33 <https://www.ieee802.org/1/files/public/docs2022/dt-lv-text-contribution-0522-v01.pdf>

34 Note—Supersedes prior contribution with same name 2022-03-24.