

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 (36.7).
- 10 e) A PFC round-trip delay measurement protocol that supports automatic headroom calculation (36.7).
- 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.

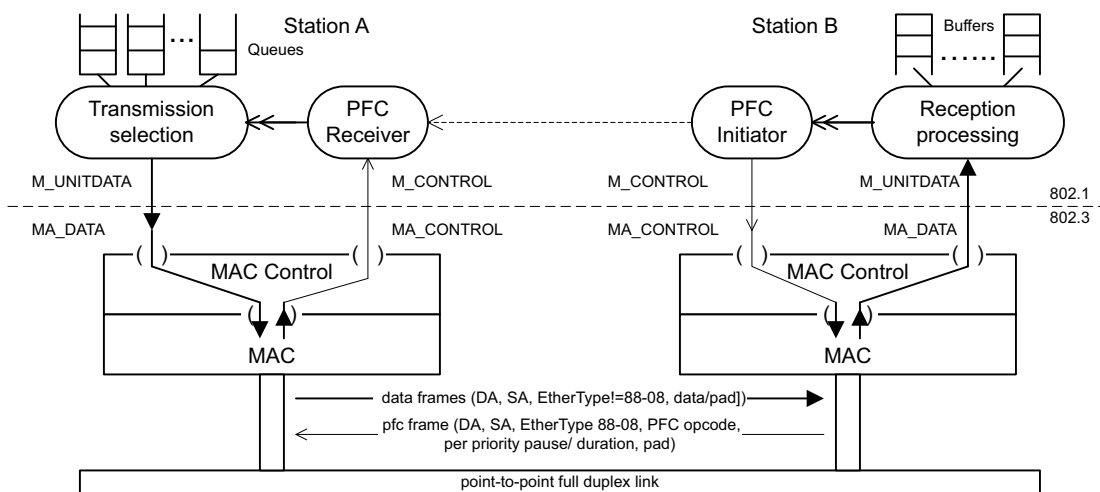


Figure 36-1—PFC example

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 B to A.
- 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—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 36.2 Network and system considerations and limitations

11 36.2.1 Data center network protocol support

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

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

25 36.2.2 Hop-by-hop flow control

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

33 Distributed data centers can use data center protocols over links are significantly longer than those typically
 34 found in an individual data center (e.g. 60 km as opposed to 100 metres) and introduce corresponding PFC
 35 headroom buffering requirements as consequence of the increased transmission delays. When those links are
 36 supported by Two-Port MAC Relays (TPMRs), the headroom requirements for a data center system
 37 interface directly attached to one of those TPMRs is determined by the round-trip delay to that TPMR, and is
 38 unaffected by the length of the link. The headroom requirements for PFC operation over the long link are
 39 met by the TPMR interfaces directly attached to that link, as shown in Figure 36-2.

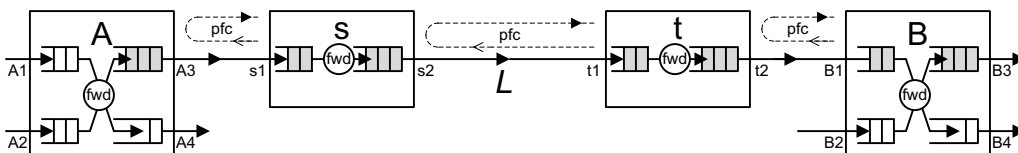


Figure 36-2—PFC hop-by-hop flow control with TPMRs

1 Figure 36-2 shows the buffering of user data frames, as they flow from data center switch (bridge or router)
 2 A 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 will
 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

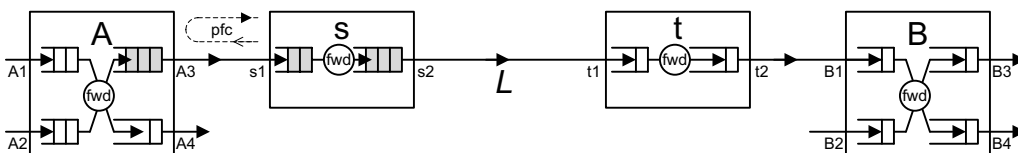


Figure 36-3—PFC hop-by-hop flow control with link rate mismatch

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.

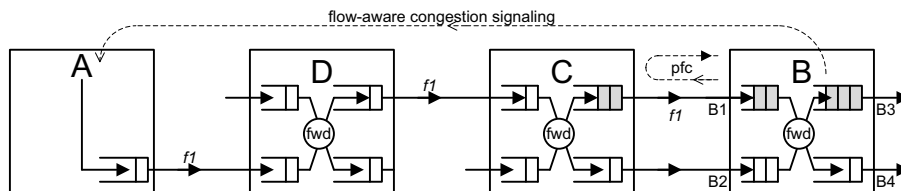


Figure 36-4—Flow-aware congestion signaling with PFC loss prevention

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.

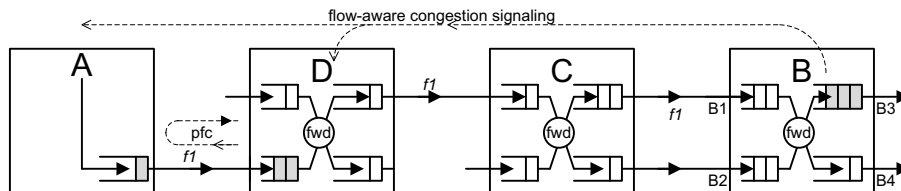


Figure 36-5—Flow-aware congestion signaling with PFC back-pressure

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

7 36.2.4 Congestion spreading

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

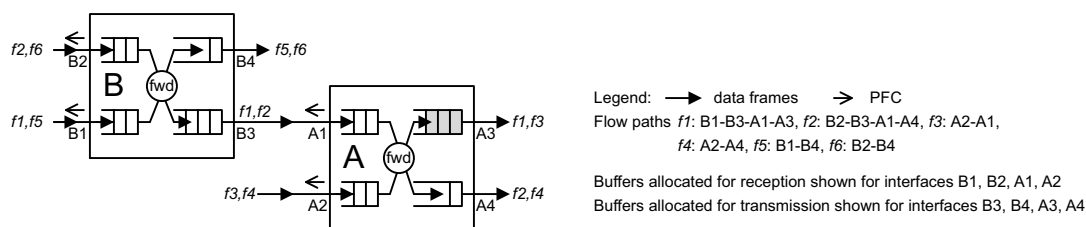


Figure 36-6—PFC congestion spreading

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

20 36.2.5 Potential for deadlock and delay

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

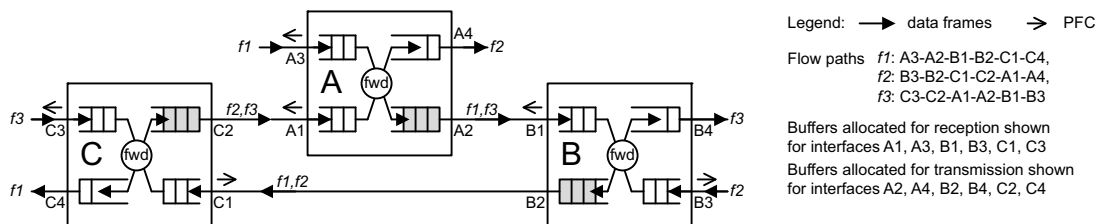


Figure 36-7—PFC deadlock example

22 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
 23 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
 24 A3 and transmitted on C4), there is a circular buffer dependency as PFC operates per-priority and is
 25 independent of any particular flow. If flows f_1 and f_3 cause congestion at A2, A can initiate PFC for the link

1 A1-C2, causing C (after received frames fill buffers for C2) to initiate PFC for C1-B2, and B in turn to
2 initiate PFC for B1-A2. As A2's transmission is now blocked, A cannot let the PFC for A1 lapse without
3 losing frames.

4 Circular buffer dependency is a necessary condition for PFC deadlock, and does not occur in some network
5 topologies (a simple case is where all flows follow the same tree). However, even in networks whose
6 intended topology is circular buffer dependency free, there remains the possibility of such a dependency
7 during network reconfiguration as a consequence of link loss or addition. The operation of network
8 configuration and management protocols should be independent of PFC operation (36.2.7). Each Bridge
9 enforces a maximum Bridge transit delay (6.5.6), discarding frames queued for longer. That discard can
10 suffice to remove a deadlock, if the network converges on a circular buffer dependency free topology.

11 **36.2.6 PFC and MAC Security**

12

13 **36.2.7 Network configuration and management protocols**

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

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

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

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

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

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

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

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

4 NOTE 3—Priority is a parameter both of the EISS, that adds VLAN tags to frames, and of the ISS (6.6,
5 IEEE Std 802.1AC). The priority to be associated with a received frame that is to be forwarded by a Bridge can be
6 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
7 tag, and can be further modified by flow classification and metering (8.6.5).

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

10 **36.2.8 Point-to-point operation**

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

13 **36.3 Detailed specification of PFC operation with IEEE 802.3 MAC Control**

14 **36.3.1 PFC primitives**

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

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

35 As a result of the processing of the PFC M_CONTROL.request, the peering PFC station receives a PFC
36 M_CONTROL.indication with the same multicast address and PFC opcode, and an indication_operand_list
37 with the operands specified for the M_CONTROL.request.

38 NOTE—IEEE Std 802.1AC maps M_CONTROL.requests and M_CONTROL.indications to and from the
39 MA_CONTROL.requests and MA_CONTROL.indications specified by IEEE Std 802.3 respectively.

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

1 36.3.2 Processing PFC M_CONTROL.indications

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

6 Figure 36-8 shows the PFC state diagram for priority n. If PFC is not enabled for priority n, then the PFC
 7 state diagram does not apply to priority n and Priority_Paused[n] is FALSE.

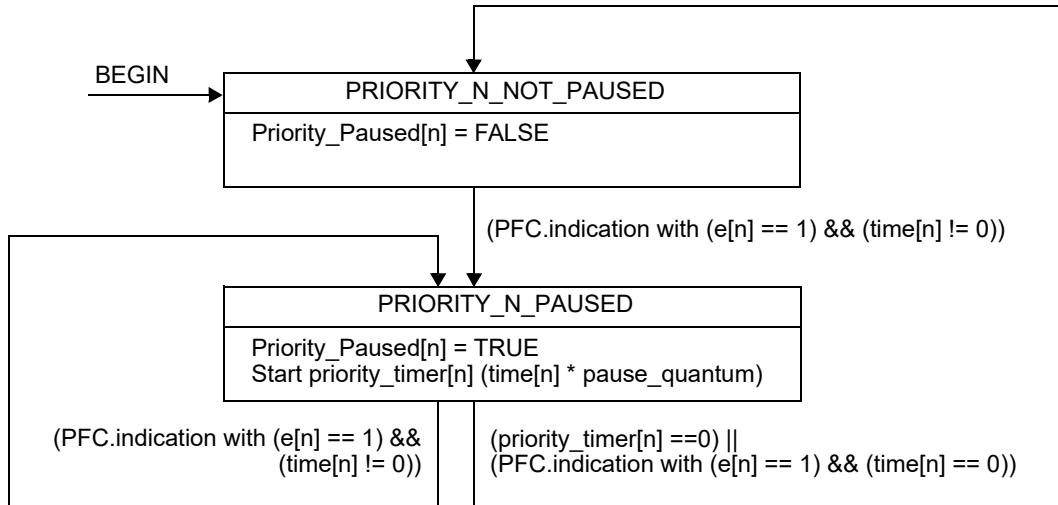


Figure 36-8—PFC Receiver state diagram for priority n

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

17 NOTE—A priority_enable_vector with all bits set to zero is legal and equivalent to a no-op.

18 36.3.3 Timing considerations

19 A priority flow controlled queue shall go into paused state in no more than 614.4 ns since the reception of a
 20 PFC M_CONTROL.indication that paused that priority. This delay is equivalent to 12 pause quanta (i.e.,
 21 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
 22 120 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.

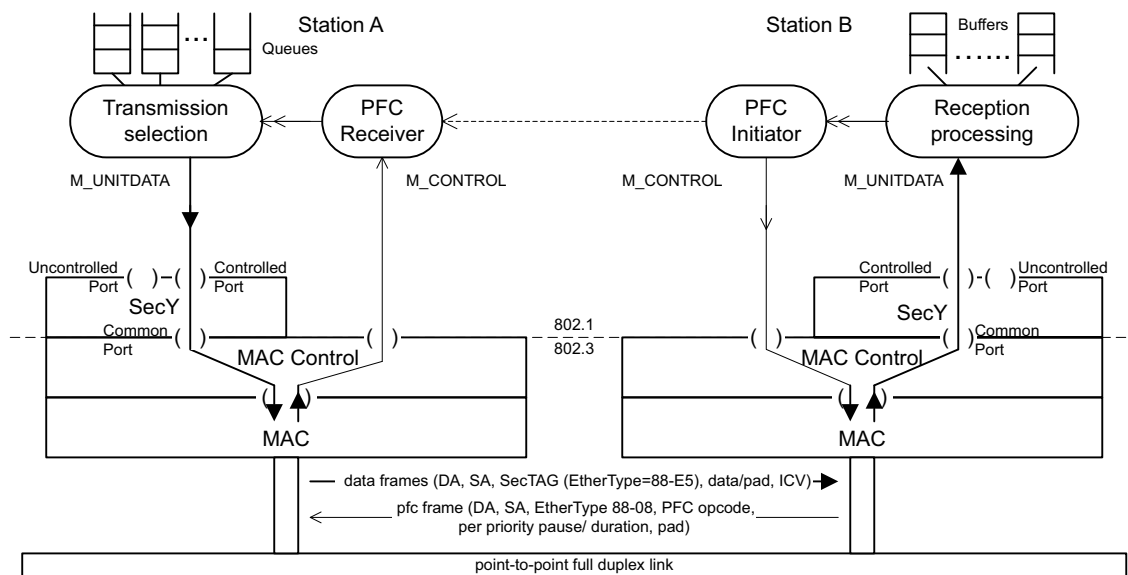


Figure 36-9—PFC with IEEE 802.3 MAC Control and MACsec

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.

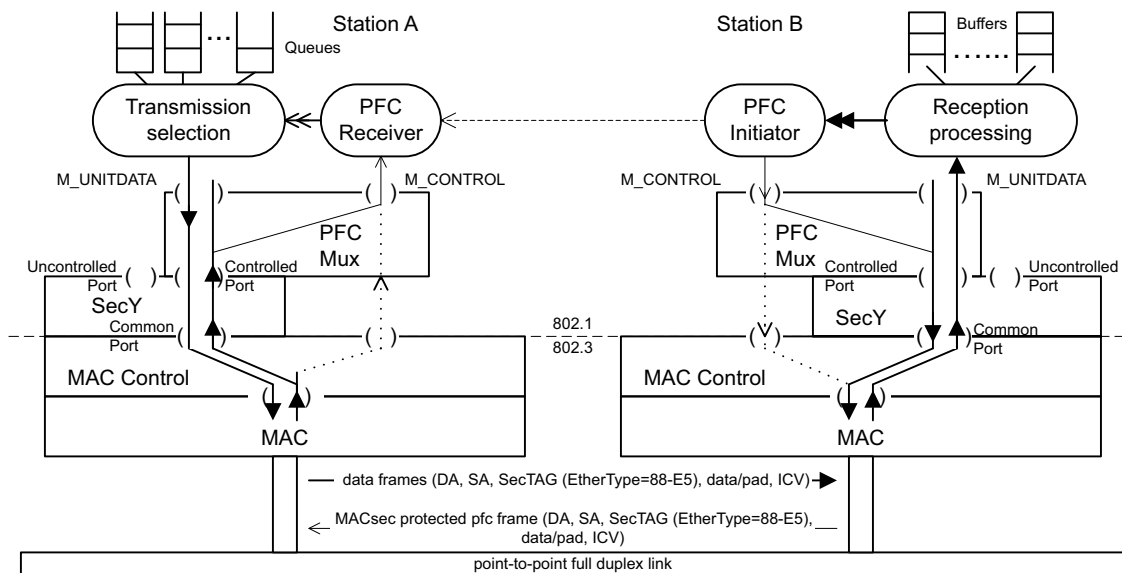


Figure 36-10—MACsec protection of user data and PFC frames

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.

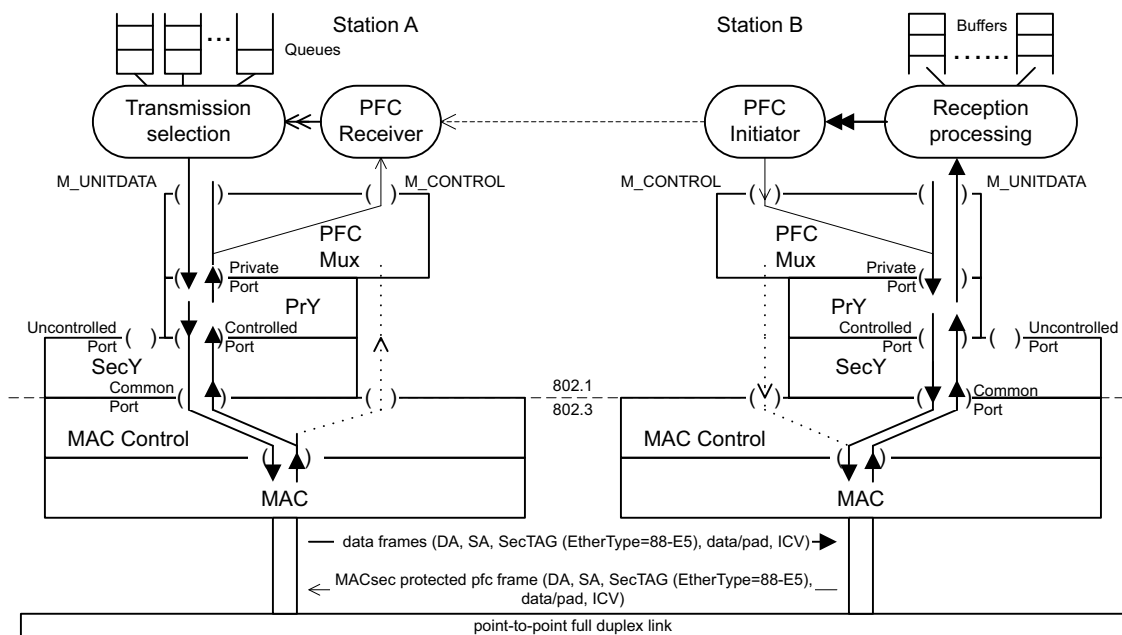


Figure 36-11—MAC Privacy protection and PFC

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^{dk}–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.

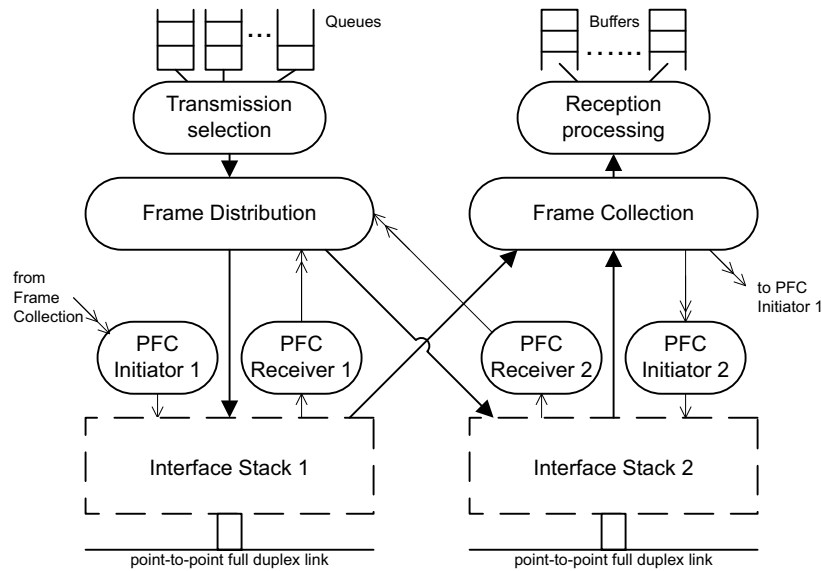


Figure 36-12—PFC operation with link aggregation

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 (36.11), including:
- 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], 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 ns/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 36.11 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 headroom requirement. A PFC-enabled system can inform its peer of the maximum user data frame bit
40 transmission rate that it can sustain over a one-way link delay period (Δ).

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

44 The result of PFC headroom calculation is made available to network management (Δ). Automated
45 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 (<>).

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.¹ 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 **Headroom measurement frames are protected or not protected just as for PFC frames.**

19

20 **36.10 PFC management**

21 **36.11 PFC headroom measurement protocol**

22 **Headroom measurement frames are protected or not protected just as for PFC frames.**

23

¹ Q. What units should be used? Is it likely that the interface speed will be negotiated at link up time? In which case the delay should be in nanoseconds, which conflicts with the use of pause quanta elsewhere. Or should we use pause quanta at the highest nominal speed, allowing adjustment if the speed is different.

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 2×10^5
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.