Contribution to P802.1Qdt—Suggested text
May 6, 2024

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.

This update includes changes as a consequences of task group comment on an earlier draft.

More significantly this update includes a new set of state machines for the headroom measurement protocol, completely revised from a first draft reviewed in a prior task group meeting.

36. Priority-based Flow Control (PFC)

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

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

a) Network and system considerations and limitations for PFC use (36.2).
b) PFC operation with IEEE 802.3 MAC Control support (36.3).
c) PFC-capable interface stack operation with MACsec (36.4, 36.5), MAC Privacy protection (36.6), and Link Aggregation (36.7).
d) The receive buffering (PFC headroom) required to avoid against frame loss (36.1.1, 36.8).
e) A PFC round-trip delay measurement protocol that supports automatic headroom calculation (36.9).
f) Management of PFC, including parameter exchanges using DCBX/LLDP, the headroom measurement protocol, and MACsec Key Agreement (MKA) (36.11).

The encoding of DCBX/LLDP parameters is specified in Annex D.

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

36.1 PFC overview

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

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

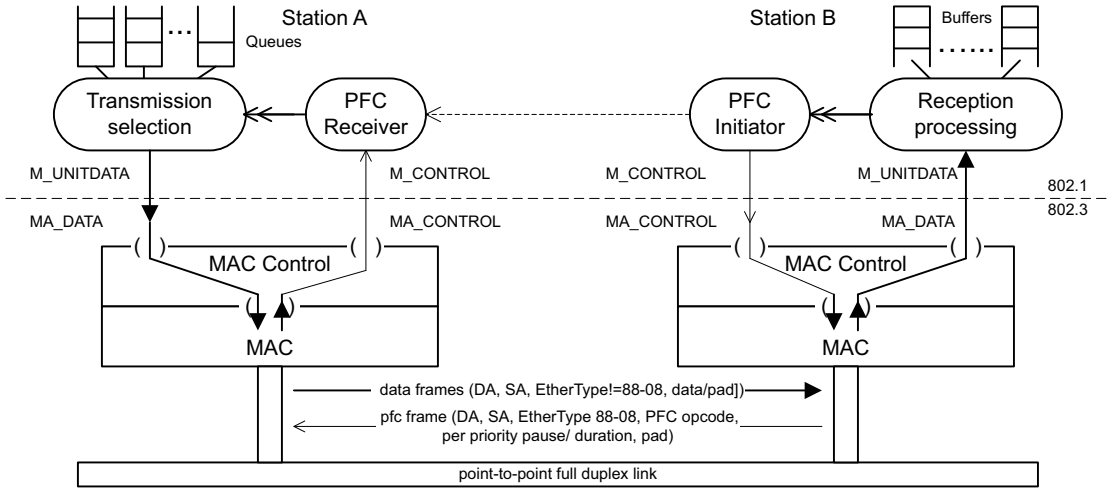

Figure 36-1—PFC example

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

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

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

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

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

36.1.1 PFC headroom

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

a) B's reception processing to calculate the remaining buffering following frame receipt.
b) B's PFC Initiator to initiate PFC following that buffering calculation.
c) Encoding of the PFC frame and any other transmission delays associated with B's interface stack.
d) Any prior in-progress frame transmission by B (possibly of a maximum sized frame that cannot be preempted) to complete.
e) PFC frame transmission on the physical link.
f) The link delay for transmission from B to A.
g) PFC frame reception, including frame validation, by A's interface stack.
h) A's PFC Receiver to decode the PFC frame and halt transmission selection for specified priorities.
i) Any in-progress frame transmission by A (possibly of a maximum sized frame) to complete.
j) The link delay for transmission from A to B.
k) Reception delays associated with B's interface stack, reception processing, and buffering.

The PFC *headroom* is the buffering that needs to remain available to B's reception process before PFC is initiated to ensure that frames are not lost as a result of a shortage of buffers. If, when not PFC paused, data

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

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

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

36.2 Network and system considerations and limitations

36.2.1 Data center network protocol support

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

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

36.2.2 Hop-by-hop operation

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

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

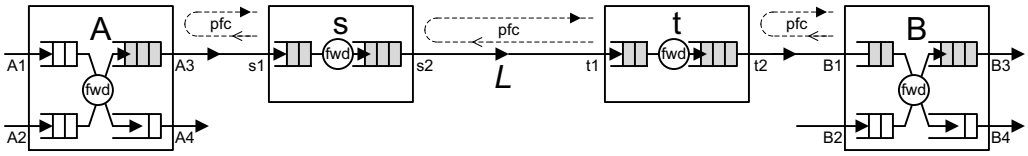

Figure 36-2—PFC hop-by-hop flow control with TPMRs

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

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

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 initiation at s1 back-pressuring A3, as shown in Figure 36-3

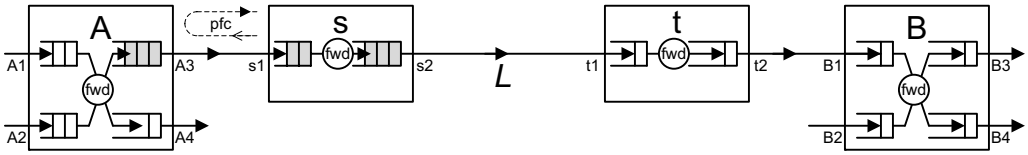

Figure 36-3—PFC hop-by-hop flow control with link rate mismatch

36.2.3 PFC and flow-aware congestion signaling

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

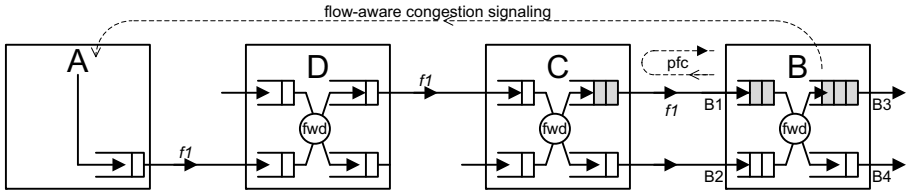

Figure 36-4—Flow-aware congestion signaling with PFC loss prevention

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

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

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

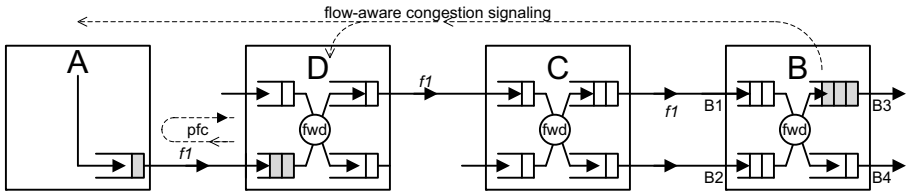

Figure 36-5—Flow-aware congestion signaling with PFC back-pressure

In Figure 36-5, B has sent a message to A requesting a rate reduction for flow *f1*, but A does not implement the congestion signaling protocol. If D intercepts that flow rate reduction message and reduces its own transmission for *f1* or other flows transmitted by A, D's buffers can fill, triggering PFC to pause flows with PFC-enabled priorities. As in Figure 36-4, PFC operation depends only on buffer use and is independent of flow-aware signaling and the details of D's interception of congestion signaling message (not specified by this standard).

36.2.4 Congestion spreading

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

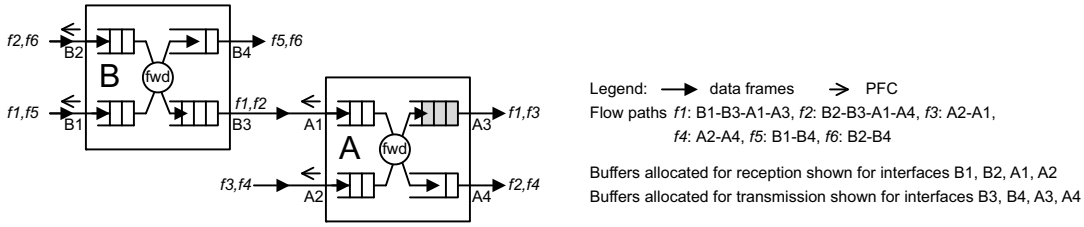

Figure 36-6—PFC congestion spreading

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

36.2.5 Potential for deadlock and delay

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

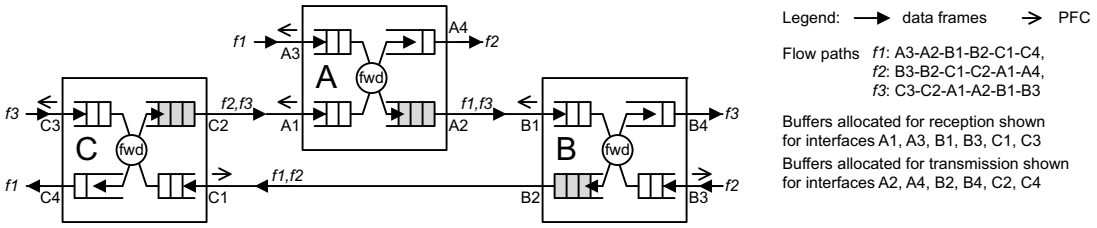

Figure 36-7—PFC deadlock example

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.

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.

36.2.6 PFC and MAC Security

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.

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.

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).

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.

36.2.7 PFC and MAC Privacy protection

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

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

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

36.2.8 Network configuration and management protocols

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

a) Transmission is not subject to PFC, and not excessively delayed by transmission of other frames including high priority forwarded frames.

b) Reception, and delivery to the correct protocol processing and/or forwarding entities does not depend on the processing of frames subject to PFC.

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

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

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

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

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

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

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

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

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

36.2.9 Point-to-point operation

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

36.3 Detailed specification of PFC operation with IEEE 802.3 MAC Control

36.3.1 PFC primitives

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

a) The globally assigned 48-bit multicast address 01-80-C2-00-00-01.

b) The PFC opcode (i.e., 01-01, as specified in Annex 31A of IEEE Std 802.3-2022).

and a request_operand_list with two operands as follows:

c) priority_enable_vector: a 2-octet field, with the most significant octet being reserved (i.e., set to zero on transmission and ignored on receipt). Each bit of the least significant octet indicates if the corresponding field in the time_vector parameter is valid. The bits of the least significant octet are 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 corresponding time[n] value is valid. For each e[n] bit set to zero, the corresponding time[n] value is invalid.

d) time_vector: a list of eight 2-octet fields, named time[0] to time[7]. The eight time[n] values are always present regardless of the value of the corresponding e[n] bit. Each time[n] field is a 2-octet, unsigned integer containing the length of time for which the receiving station is requested to inhibit transmission of data frames associated with priority n. The field is transmitted most significant octet first, and least significant octet second. The time[n] fields are transmitted sequentially, with time[0] transmitted first and time[7] transmitted last. Each time[n] value is measured in units of pause_quanta, equal to the time required to transmit 512 bits of a frame at the data rate of the MAC. Each time[n] field can assume a value in the range of 0 to 65 535 pause_quanta.

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

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

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

36.3.2 Processing PFC M_CONTROL.indications

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

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

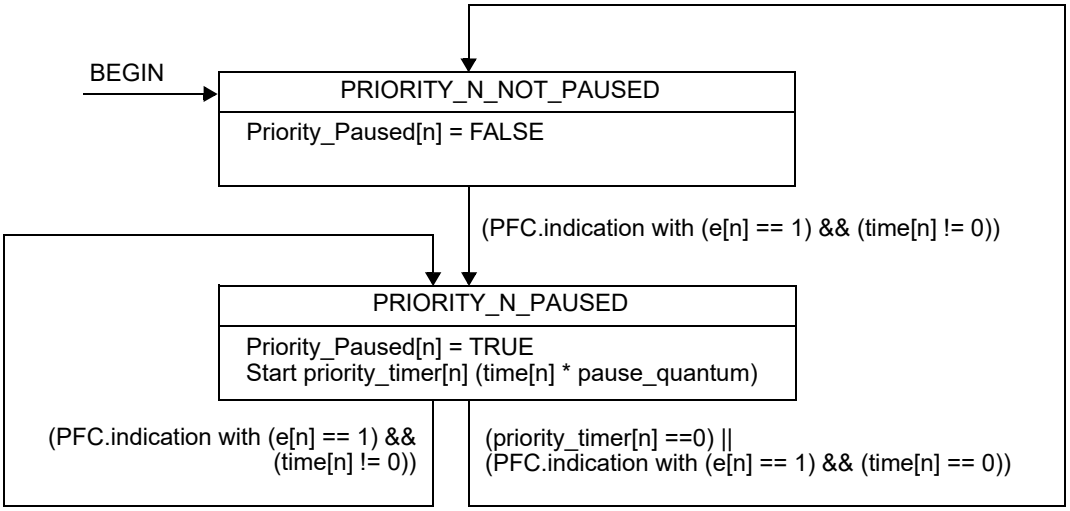

Figure 36-8—PFC Receiver state diagram for priority n

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

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

36.3.3 Timing considerations

A priority flow controlled queue shall go into paused state in no more than 614.4 ns since the reception of a PFC M_CONTROL.indication that paused that priority. This delay is equivalent to 12 pause quanta (i.e., 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 120 pause quanta (i.e., 61 440 bit times) at the speed of 100 Gb/s.

36.4 PFC with MACsec data protection

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

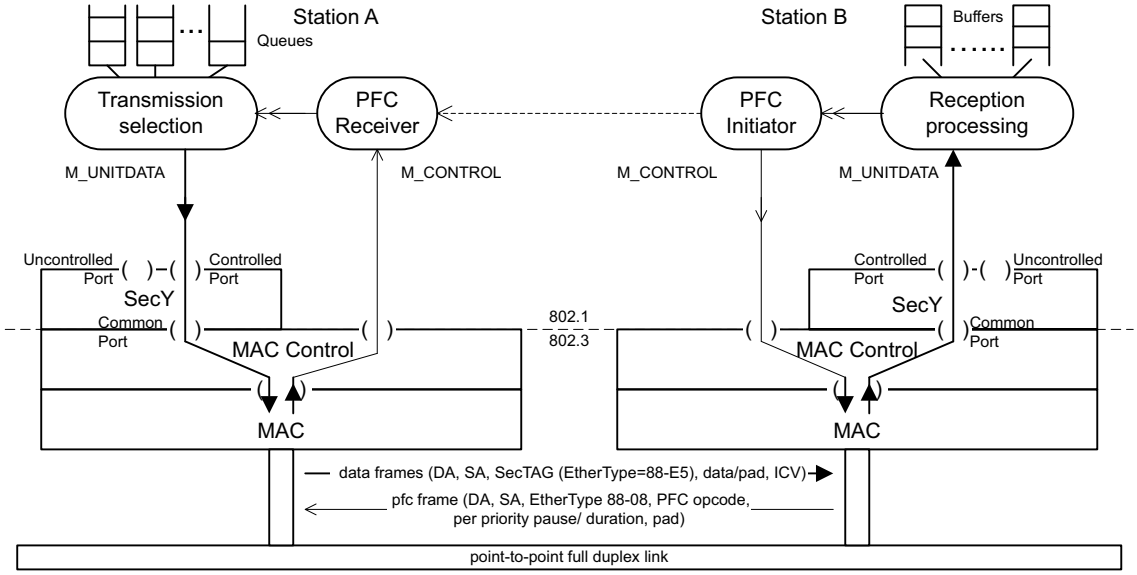

Figure 36-9—PFC with IEEE 802.3 MAC Control and MACsec

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

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

36.4.1 PFC headroom with MACsec data protection

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

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

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

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

36.5 PFC with MACsec protection of user data and PFC frames

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

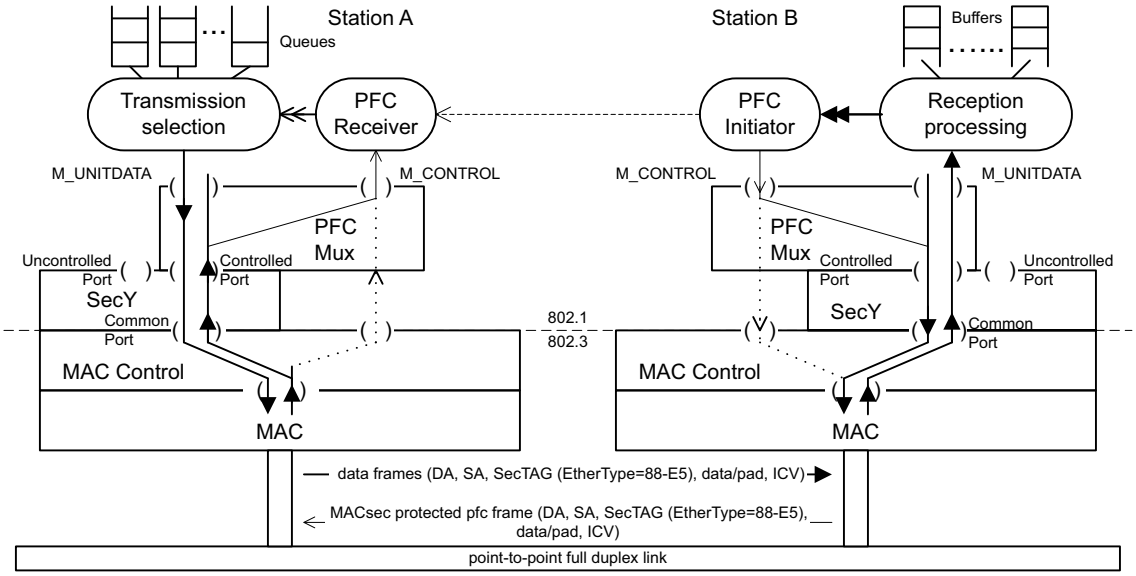

Figure 36-10—MACsec protection of user data and PFC frames

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

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

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

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

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

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

36.5.1 PFC headroom with MACsec protection of PFC and data frames

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

36.6 PFC with MAC Privacy protection

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

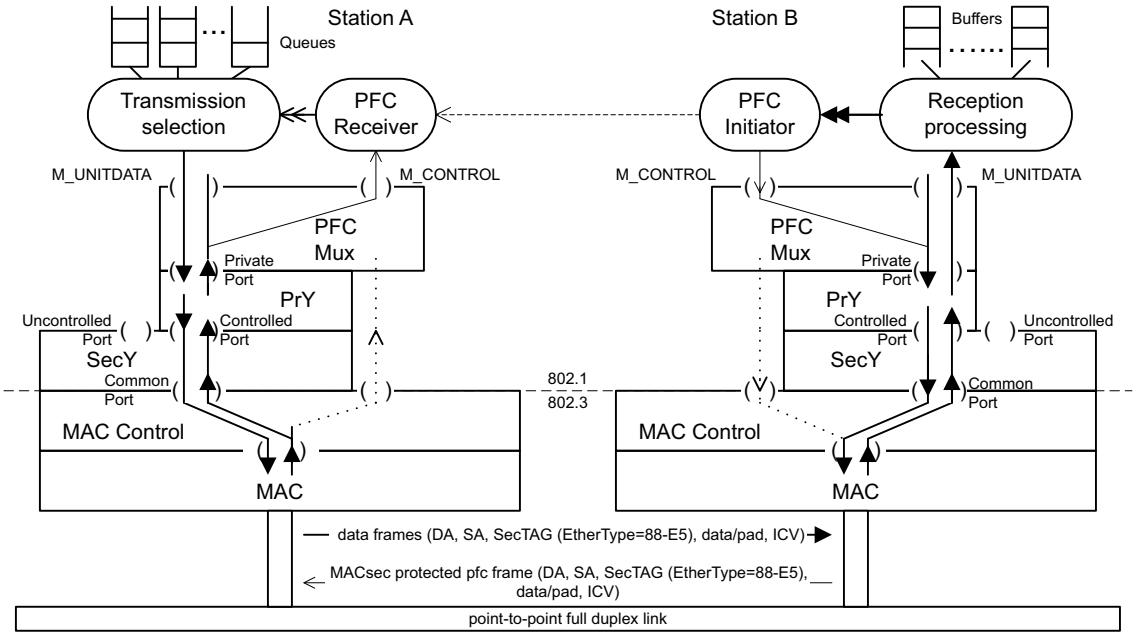

Figure 36-11—MAC Privacy protection and PFC

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

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

NOTE—MAC Privacy protection was first standardized in the IEEE Std 802.1AEdk–2023 amendment to IEEE Std 802.1AE–2018.

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

36.7 PFC with link aggregation

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

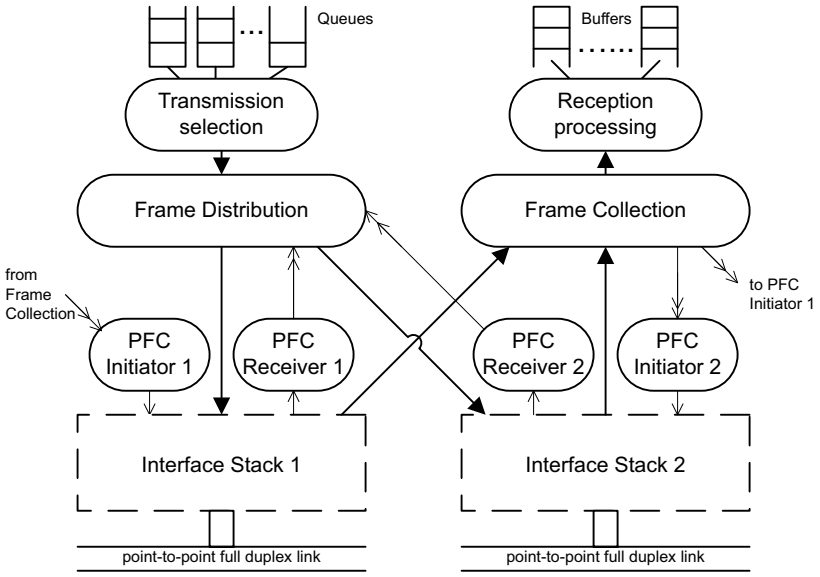

Figure 36-12—PFC operation with link aggregation

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

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

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

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

36.8 PFC headroom calculation

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

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

or

b) The round-trip delays reported by the PFC headroom measurement protocol (36.9), adjusted for:
   1) The system's local knowledge of the maximum delay that would occur between:
      i) buffer consumption by reception processing, and
      ii) the transmission of a PFC

      i.e., [items a) and b) of 36.1.1], further adjusted for any differences between:
      iii) the maximum delay for PFC frame encoding and initiating transmission [item c) of 36.1.1], and
      iv) the delay between selection of a timestamp value to be encoded in a headroom measurement frame and initating transmission of that frame.
   2) The peer system's assessment of the difference between:
      i) the maximum delay from the reception of a PFC to halting transmission selection for the affected priorities [item h) of 36.1.1], and
      ii) the delay between the reception of PFC headroom measurement request, and its processing by the PFC Receiver.

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

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

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

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

The result of PFC headroom calculation is made available to network management (36.11). Automated headroom calculation can take place even when its result is to be overriden by manual configuration, which

can specify an initial value (as the link is typically operational while measurement and calculation proceeds), and maximum and minimum values (36.11).

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

36.8.1 Headroom calculation using link delays

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

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

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

Management parameters for link delay based calculation are specified in 36.11.

36.9 PFC headroom measurement protocol

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

36.9.1 Protocol purpose, goals, and non-goals

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

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

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

— Accuracy. Averaged results of headroom measurement are expected to estimate PFC headroom to within 8 pause quanta (512 octets). Headroom measurement addresses the requirement for buffer allocation, and is not intended as a substitute for clock synchronization. Measurement requests and responses traverse the peer interface stacks in the same way

— Timeliness. Headroom measurements are available shortly after connectivity is established between the peers, even if the peer interfaces become MAC_Operational (6.8.2) at different times. Periodic measurement can be used if the link delays can change, e.g. through optical switching, without explicit interface signaling.

— Efficiency. Timeliness is not achieved by rapid repetitive transmission when the interface becomes MAC_Operational, in competition with other startup protocols.

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

The measurement protocol does not specify:

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

36.9.2 Addressing, protocol identification, and protocol versions

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

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

36.9.3 Protocol parameter values, representation, and encoding

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

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

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

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

36.9.4 Measurement requests and responses

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

A measurement request comprises the following parameters:

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

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

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

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

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

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

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

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

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

(ResponseDelay) + Request Adjustment + Response Adjustment

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

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

[1] Cable delay approximately 5 microseconds per kilometer (5 nanoscconds 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.

36.9.5 Measurement PDU formats

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

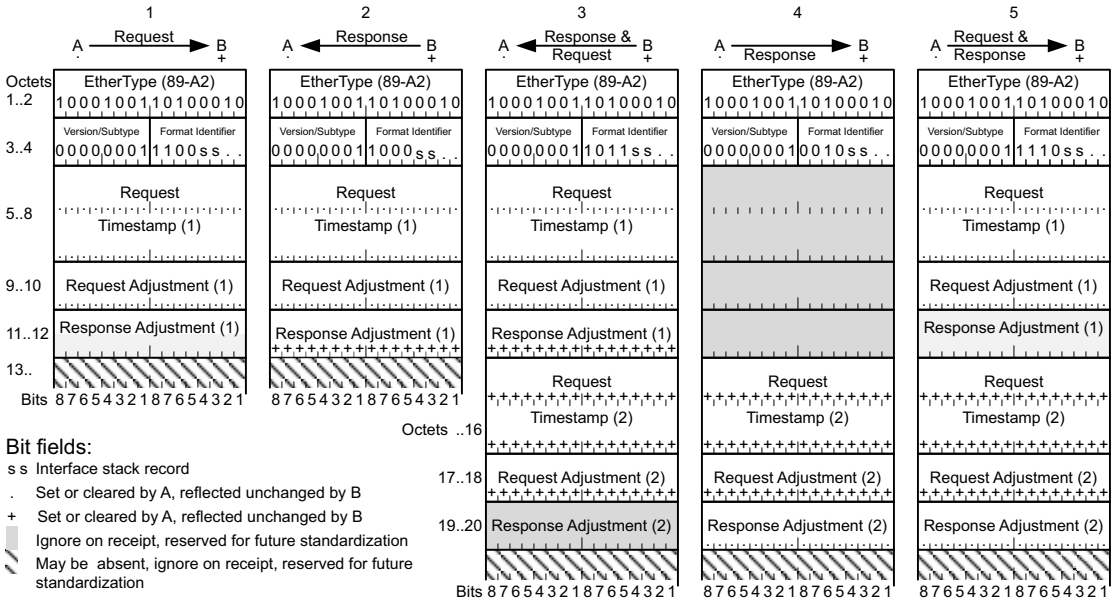

Figure 36-13—HMPDU format (examples)

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.

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:

— 0x03 : The tuple is a measurement request.
— 0x02 : The tuple is a measurement response with a non-zero Response Adjustment.
— 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.
— 0x00 : The tuple is unused.

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.

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.

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.

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

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

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

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

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

36.9.6 Measurement protocol exchanges

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

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

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

A measurement response shall be transmitted whenever a measurement request has been received. Each transmission of a measurement response provides an opportunity for the transmission of a measurement request in the same HMPDU, a measurement request can also be transmitted when a measurement response has been received. Otherwise measurement requests should not be repeated at intervals of less than the system dependent maximum acceptable round-trip delay (36.11). As a consequence of this restriction on request transmission, a protocol participant does not have to buffer more than two HMPDUs provided that it does not delay request or response transmission for longer than its peer’s round-trip delay maximum.

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

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

The measured round-trip delay is calculated (36.9.4) for each measurement response received. If the calculated value is less than a system dependent minimum (36.11), the latter value is substituted. If the value is greater than a system dependent, manageable, maximum (36.11), that maximum is substituted.

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

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

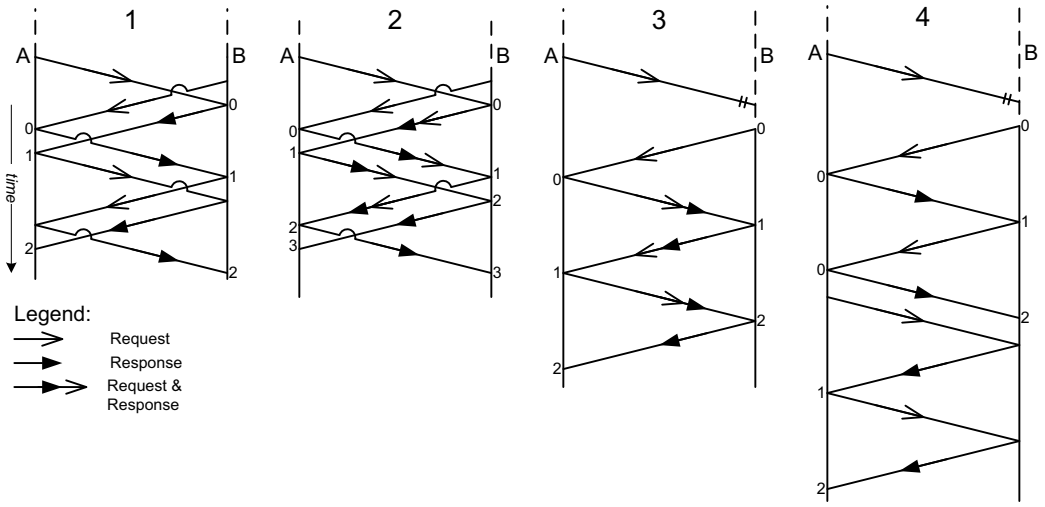

Figure 36-14—Measurement (examples)

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

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

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

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

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

The headroom measurement protocol's ability to determine headroom rapidly, even in the event of initial HMPDU loss, is dependant on the participation of both stations attached to the link, even if one of the stations will not use measurement results. LLDP (IEEE Std 802.1AB) should also be used to exchange information about the each station's use of, and response to, PFC (36.11, <D.2.10>). Use of the headroom measurement protocol can be terminated if neither station needs measurement results.

36.9.7 PFC headroom measurement protocol entities and measurement paths

A protocol entity that transmits and receives HMPDUs to and from a link is associated with its station's PFC Initiator and Receiver for that link. Figure 36-1 and Figures 36-9, 36-10, and 36-11, illustrate the transmission of a PFC from one station (Station B in each figure) to control the transmission of user data frames from the other station (Station A). User data frames and PFCs can be transmitted in both directions, and each station includes both a PFC Initiator and a PFC Receiver for the link, although the use of PFC to control either direction of user data transmission can be independently controlled (36.11).

The round-trip path for measurement requests and the resulting responses follows, as closely as possible, that traversed by PFCs and the user data frames whose transmission they control. When PFCs are received via the MA_CONTROL interface, there is a potential implementation dependent difference between the time taken for their reception and that taken for measurement requests. The variance in the time taken by the reception processing that demultiplexes measurement requests to the headroom measurement protocol entity needs to be within the bounds necessary to support accuracy desired for headroom measurement (36.9.1), as does the difference between the times taken to respond to PFCs and measurement requests respectively unless included as a Response Adjustment (36.9.5).

If PFCs are not MACsec protected (see Figure 36-1 and Figure 36-9) they will be received via the MA_CONTROL interface. Measurement requests will be received via the M_UNITDATA interface (and subsequently via the SecY's Uncontrolled Port if user data frames are MACsec protected).

If PFCs and user data frames are MACsec protected or MAC Privacy protected (see Figure 36-10 and Figure 36-11) PFCs, measurement requests, and measurements responses are all received via the M_UNITDATA interface. The PFC Mux serves to multiplex transmitted PFCs, measurement requests and responses, and to demultiplex received PFCs, measurement requests and responses.

The measurement round-trip path and delay can change while a MAC interface remain MAC_Operational, as MACsec or MAC Privacy protection becomes operational. Measurement requests identify the round-trip path to be measured (36.9.5). Requests and responses that specify a path that is not operational, are discarded. It is possible for different round-trip paths, for different priorities, to be simultaneously active though this standard does not identify any requirement for simultaneous round-trip paths except in times of transition from one to another. The measurement protocol design requires an implementation to be able to buffer, pending transmission or reception processing, a maximum of two HMPDUs at any given time for a round-trip path. A conformant implementation is required to support any configurable round-trip path, but not more than one round-trip path at a time.

Rapid determination of headroom, accomodating several measurements to minimize the effect of link start up effects and other unrepresentational results, without excessive resource competition with other start up protocols is facilitated by using the reception of responses to time the transmission of following requests. The point-to-point link is expected to preserve the order of transmitted HMPDUs, so a participant's reception of two successive requests with no intervening or accompanying response can be taken as an indication that the participant's last request has been lost (see example 4 in Figure 36-14). That participant can then initiate a new request if further measurment is desirable. If the interface stack paths traversed by

request and responses differ (as in 36.4), it might be possible for a participant to process a measurement response and initiate a subsequent request that is submitted to the MAC for transmission prior to response to a previously received request. The possibility of succesive doubling of requests is avoided by limiting each participant to handling at most two HMPDUs at a time, discarding any others received.

36.10 Headroom measurement protocol state machines

The operation of the headroom measurement protocol specified in this clause (Clause 36) is specified by the the state machines shown in Figure 36-15. Each of the state machines is specified in Figure 36-16 through Figure 36-19 using the notation specified in Annex E.

Figure 36-15 is not itself a state machine, but provides an overview of the state machines and the variables used to communicate between the machines. Each of the machines, and the reception process, is initialized by the global variables BEGIN (see Annex E) and hmOperUp. The boolean variable hmOperUp is TRUE if and only if the headroom measurement protocol entity (and its associated PFC Initiator) can both transmit and receive HMPDUs using the current request and response paths, and transitions FALSE if either changes.

The boolean variable measure is controlled externally to the state machine shown. It is set TRUE whenever hmOperUp transitions TRUE, and will remain TRUE until at least two measurement responses have been processed by the Response Processing machine and until the system of which the headroom measurement entity is part has determined that the measurement has provided a sufficient guide for buffer allocation.

The implementation dependent variables requestPath and responsePath identify the interfaces used to receive and transmit measurement requests and to receive and transmit measurement responses respectively. If equal, requests and responses can be conveyed in the same HMPDU (36.3, 36.5, 36.6) and are transmitted by the Common Path Transmission state machine. Otherwise, i.e. PFCs are not MACsec protected but user data is (36.4), the Request Path and Response Path Transmission machines transmit HMPDUs of the same format, but with some fields unused. Unused Path Transmission machines remain in their initialization state.

The implementation dependent state machine variables rcvdRequest, rcvdResponse, and txResponse, point to data structures that identify measurement request and response parameters and the PDUs that convey them. Each is NULL (cleared, with a value of FALSE in boolean expressions) if not currently used. The boolean prompt serves to stimulate request transmission.

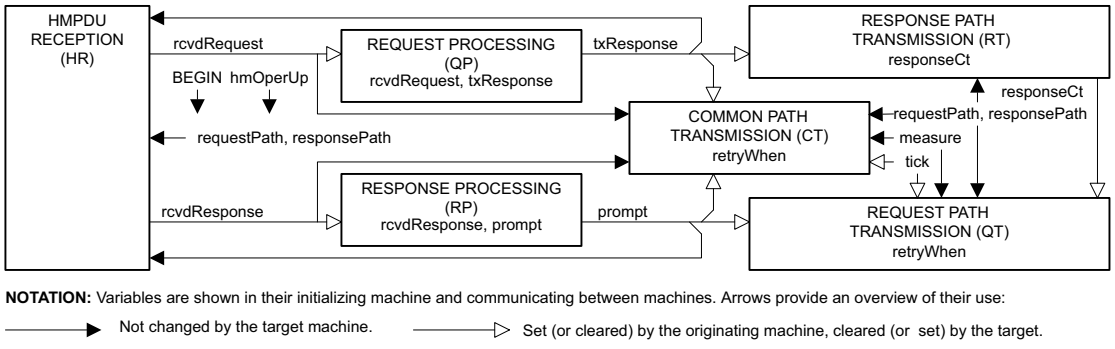

Figure 36-15—Headroom measurement state machines—overview

The HMPDU Reception (HR) is receives and validates HMPDUs. If a received HMPDU contains a request and a response, rcvdRequest and rcvdResponse are set as an atomic state machine action.

Request Processing (QP) processes each rcvdRequest, determining the appropriate txResponse before clearing rcvdRequest. The response is transmitted either by Common Path Transmission (CT) or Response Path Transmission (RT) which will clear txResponse. Response Processing (RP) processes each rcvdResponse, and sets prompt to stimulate transmission of further request. CT or QT will clear prompt, and transmit a further request if measure is set. The counter responseCt is incremented by RT for each response transmitted and cleared by QT on request transmission, a value greater than 1 indicates that two requests have been received without an intervening reponse, and QT transmits a further request.

CT does not transmit if rcvdRequest is set (not NULL) and txResponse is not set, or rcvdResponse is set and prompt is not, but waits until both QP and RP have processed the received HMPDU. Either or both rcvdRequest and rcvdResponse can remain set when txResponse and prompt are set, as a result of another HMPDU reception before they are cleared as by CT. In that case HR will discard further received HMPDUs: no more than a total of two HMPDUs need to be buffered, processed, or awaiting transmission at any instant.

If requestPath and responsePath differ, requests and responses are received (and transmitted) in separate HMPDUs. No more than one received request HMPDU need be buffered, processed, or responded to at any instant, and no more one received response need be buffered and processed at any instance.

NOTE 1—This state machine specification (36.10) models the operation of a participant. Protocol conformance is only in respect of the externally observable behavior. Modeling details have been chosen and/or left unspecified to facilitate mapping to and discussion of a wide range of implementations. Requests and responses can be received in limited dedicated or general buffering, responses can be generated with or without copying unchanged request information, and a request transmission prompted by reception of a response can use that response's buffering.

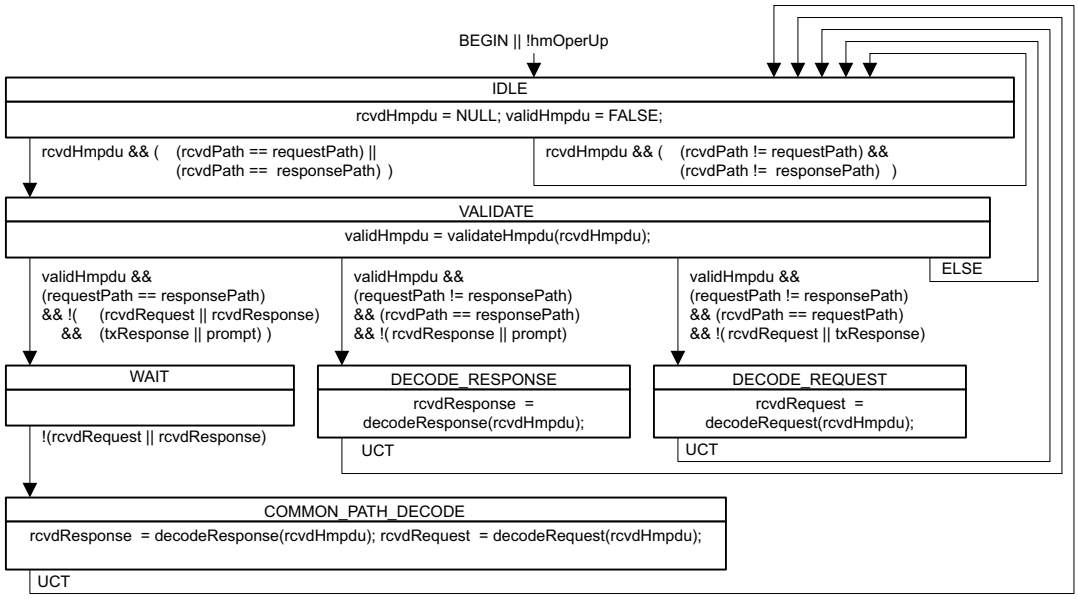

Figure 36-16—HMPDU Reception state machines

In Figure 36-16 (HMPDU Reception), rcvdHmpdu is an implementation dependent reference, allocated on reception, to a received HMPDU from the interface referenced by rcvdPath. A state machines assignment of NULL to rcvdHmpdu, rcvdRequest, rcvdResponse, or txResponse, deallocates the associated resources if not otherwise referenced by one of those variables.

Received HMPDUs are only accepted from the interfaces for the current request or response path, and are discarded otherwise. If those paths are distinct [(requestPath != responsePath)], a response is only decoded from a correctly formatted HMPDU from the response path interface if no prior response is either awaiting processing by the Response Processing state machine (see Figure 36-17) or has generated a prompt for a

further request which is yet to be transmitted. Similarly, if the paths are distinct, a request is only decoded from a correctly formatted HMPDU from the request path interface if no prior request is either awaiting processing by the Request Processing state machine (see Figure 36-17) or has generated a response which is yet to be transmitted.

If request and responses are expected from the same interface, a correctly formatted HMPDU is decoded once any prior requests and responses have been processed by both the Request and Response Processing machines and responses and requests stimulated by their reception transmitted. While a HMPDU is delayed, any other HMPDUs received will be discarded.

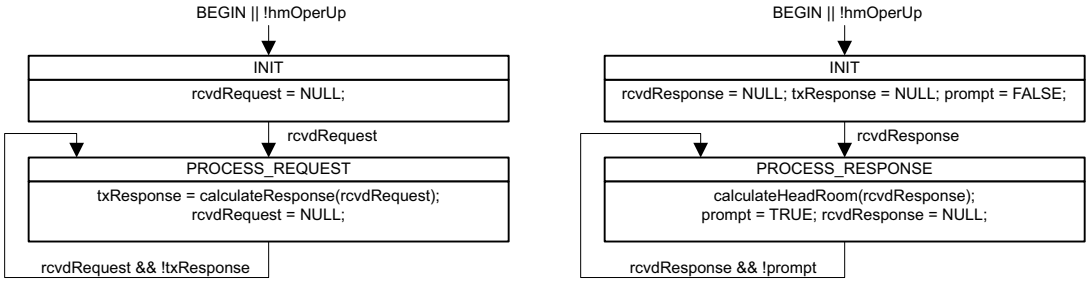

Figure 36-17—Request and Response processing state machines

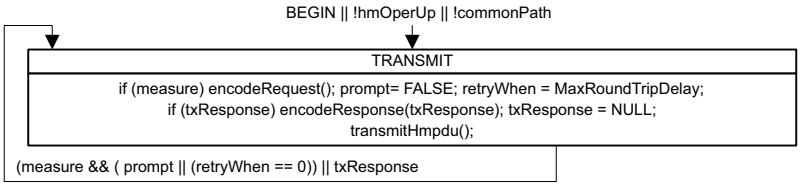

Figure 36-18—Common Path Transmission state machine

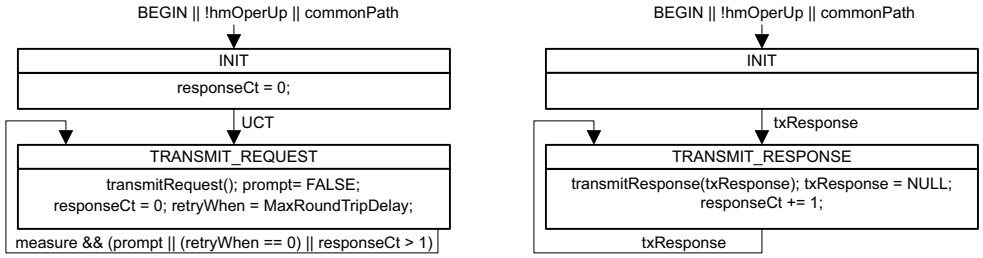

Figure 36-19—Request and Response Path Transmission state machines

36.11 PFC management

Existing 802.1Q Clause 36 issues

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

36.1.1, “data center environment” is undefined:

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

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

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

Open Questions:

PFC request frequency:

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

Additional 802.1Q issues

8.6.8 Transmission selection, NOTE 1 and NOTE 2:

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

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

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

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

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

D.2.10.6, PFC Enable:

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

802.1Q-2022 Annex M, status, adequacy

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

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

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

a) The destination and source MAC addresses. It is vital that a PFC frame not be forwarded, by a bridge or any similar frame forwarding device, from one point-to-point link to another. IEEE Std 802.3 mandates the use of the 01-80-C2-00-00-01 "IEEE MAC-specific Control Protocols group address" for this purpose. Frames with this destination address are not forwarded by any type of bridge (MAC Bridge, VLAN Bridge, TPMR, Provider Bridge, Provider Backbone Bridge) specified by IEEE Std 802.1Q.

b) Protocol discrimination. IEEE Std 802.3 assigned the EtherType 88-08 to identify MAC Control frames (including, but not limited to PFC). Any 802.3 station that implements MAC Control recognizes any and all received frames with this EtherType as MAC Control frames without regard to its destination or source MAC Address, provided that the station is configured to receive frames with the destination MAC Address. While other link layers could use other ways to distinguish PFC frames, it is vital that they be distinguished using a method commom to all

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

IEEE Std 802.1AX-2020 Issues

LACPDU priority

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

IEEE Std 802 issues

802f EtherType 88-08 description:

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

Additional 802.1Q references

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

"Mapping between M_CONTROL.requests/indications and IEEE802.3 MA_CONTROL.requests/indications is performed as specified in IEEE Std 802.1AC. If the MAC supports the MAC Merge sublayer specified in IEEE Std 802.3, then PFC M_CONTROL.requests are mapped onto the MAC control interface associated with the express MAC (eMAC)."

12.23 Priority-based Flow Control objects

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

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

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

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

17.2.17 Structure of the IEEE8021-PFC-MIB

Table 17-23 describes the relationship between the SMIv2 objects defined in the PFC-MIB module (17.7.13) and the variables and managed objects defined in Clause 12 and Clause 36.

17.3.17 Relationship of the IEEE8021-PFC-MIB to other MIB modules

17.4.17 Security considerations of the IEEE8021-PFC-MIB

17.7.17 Definitions for the IEEE8021-PFC-MIB module

```
...
ieee8021PfcLinkDelayAllowance OBJECT-TYPE
SYNTAX Unsigned32
MAX-ACCESS read-write
STATUS current
DESCRIPTION
"The allowance made for round-trip propagation delay of the link in bits."
...
ieee8021PfcRequests OBJECT-TYPE
SYNTAX Counter32
UNITS "Requests"
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"A count of the invoked PFC M_CONTROL.request primitives
ieee8021PfcIndications OBJECT-TYPE
SYNTAX Counter32
UNITS "Indications"
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"A count of the received PFC M_CONTROL.indication primitives.
```

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

References to PFC in items d) and e).

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

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

49. Congestion Isolation

Clause 49 clause begins:

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

In the fourth paragraph:

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

49.1 Congestion isolation objectives

"d) Reduce the frequency of invoking PFC in a lossless environment."

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

49.2.7 System topology and port orientation

Fifth paragraph:

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

D.2.10.6 PFC Enable

"Table D-6 shows the layout of the PFC Enable bit vector."

"A bit vector of 8 bits, one per priority:

a) A one indicates PFC is enabled on the priority.

b) A zero indicates that PFC is disabled on the priority.

c) Local policy in each end of the link decides whether to use the priority if the configuration does not match."

D.5.5 IEEE 802.1 LLDP extension MIB module version 2

Contains a number of PFC related items.

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

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

Second paragraph (extract):

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

IEEE Std 802.3-2022 references

Figure 1-1:

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

2.3.2 MA_DATA.indication, 2.3.2.2 Semantics of the service primitive:

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

2.3.2 MA_DATA.indication, 2.3.2.3 When generated:

"The MA_DATA.indication is passed from the MAC sublayer entity (through the optional MAC Control sublayer, if implemented) to the MAC client entity or entities to indicate the arrival of a frame to the local MAC sublayer entity that is destined for the MAC client. Such frames are reported only if ... and their destination address designates the local MAC entity. Frames destined for the optional MAC Control sublayer are not passed to the MAC client if the MAC Control sublayer is implemented."

2.3.2 MA_DATA.indication, 2.3.2.5 Additional comments:

"If the local MAC sublayer entity is designated by the destination_address parameter of an MA_DATA.request, the indication primitive will also be invoked by the MAC entity to the MAC client entity. This characteristic of the MAC sublayer may be due to unique functionality within the MAC sublayer or characteristics of the lower layers (for example, all frames transmitted to the broadcast address will invoke MA_DATA.indication at all stations in the network including the station that generated the request)."

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

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

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

"A SEQUENCE that meets the requirements of the description below: PAUSE PAUSE command implemented MPCP MPCP implemented PFC PFC implemented EXTENSION EXTENSION MAC Control frame supported"

30.3.3.6 aPFCEnableStatus (enabled or disabled)

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

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

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

31. MAC Control, 31.2 Layer architecture:

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

31.3 Support by interlayer interfaces:

See Figure 31-2.

"All MAC frames validly received by the CSMA/CD MAC are passed to the MAC Control sublayer for interpretation. If the MAC frame is destined for the MAC client, the MAC Control sublayer generates an MCF:MA_DATA.indication primitive, providing complete transparency for normal data exchange between MAC clients. If the MAC frame is destined for the MAC Control sublayer entity, it is interpreted and acted on internal to the MAC Control sublayer. This may result in state changes within the MAC Control sublayer, the generation of MA_CONTROL.indication primitives, or other actions as necessary to support the MAC Control sublayer function. MAC PFC time limtControl sublayer functions shall always sink MAC Control frames."

"In the MAC:MA_DATA.indication primitive, MAC frames destined for the MAC Control sublayer (MAC Control frames) are distinguished from MAC frames destined for MAC clients by a unique Length/Type field identifier."

31.4 MAC Control frames

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

31.4.1.1 Destination Address field

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

31.4.1.3 Length/Type field

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

31.5 Opcode-independent MAC Control sublayer operation

"The MAC passes to the MAC Control sublayer all valid MAC frames via the MA_DATA.indication primitive. Invalid MAC frames are not passed to the MAC Control sublayer (see 3.4)."

31.5.1 Frame parsing and data frame reception

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

"A MAC frame that does not contain the unique Length/Type field specified in 31.4.1.3 is a data frame. The receipt of a data frame results in the generation of a MCF:MA_DATA.indication primitive by the MAC Control sublayer, with its parameters identical to the MAC:MA_DATA.indication primitive."

31.5.2 Control frame reception

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

Annex 31A (normative) MAC Control opcode assignments:

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

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

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

a) The globally assigned 48-bit multicast address 01-80-C2-00-00-01.
b) The PFC opcode.
c) A request_operand list with two operands: priority_enable_vector and time_vector. (See 31D.2.)

Unlike the MAC Control PAUSE operation, the inhibition of frames for the PFC operation occurs in the MAC Control client. Upon receiving a PFC frame, the only action in MAC Control is to generate a MA_CONTROL.indication primitive with the indication_operand list specified in Table 31A9.

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

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

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

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

"The PFC opcode takes the following request_operand_list:

priority_enable_vector:

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

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

time_vector:

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

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

"Upon receipt of a MA_CONTROL.request primitive containing the PFC opcode from a MAC client, the MAC Control sublayer calls the MAC sublayer MAC:MA_DATA.request service primitive with the following parameters:

a) The destination_address is set equal to the destination_address parameter of the MA_CONTROL.request primitive. This parameter is currently restricted to the value specified in 31D.1.
b) The source_address is set equal to the 48-bit individual address of the station.
c) The length/type field (i.e., the first two octets) within the mac_service_data_unit parameter is set to the IEEE 802.3 MAC Control EtherType value assigned in 31.4.1.3.
d) The remainder of the mac_service_data_unit is set equal to the concatenation of the PFC opcode encoding (see Annex 31A), the priority_enable_vector and the time_vector specified in the MA_CONTROL.request primitive, and a field containing zeros of the length specified in 31.4.1.6.
e) The frame_check_sequence is omitted."

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

"Upon receipt of a valid MAC Control frame with the opcode indicating PFC and the destination address indicating the globally assigned multicast address specified in 31D.1, the MAC Control sublayer generates the MA_CONTROL.indication to the MAC Control Client."

IEEE Std 802.3.1-2013 references

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

```
dot3ControlFunctionsSupported OBJECT-TYPE
       SYNTAX BITS {
                     pause(0), -- 802.3 pause flow control
                     mpcp(1), -- 802.3 multi-point control protocol
                     pfc(2) -- 802.3 priority-based flow control
              }
MAX-ACCESS    read-only
STATUS        current
DESCRIPTION   "A list of the possible MAC Control functions
              implemented for this interface."
REFERENCE "IEEE Std 802.3, 30.3.3.2,
              aMACControlFunctionsSupported."
::= { dot3ControlEntry 1 }


dot3PFCTable OBJECT-TYPE
       SYNTAX       SEQUENCE OF Dot3PFCEntry
       MAX-ACCESS   not-accessible
       STATUS       current
```

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

IEEE Std 802.1AC-2016 references

11.4 Control primitives and parameters:

The ISS provides two control primitives, an M_CONTROL.request and an M_CONTROL.indication, andtheir associated parameters.

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

The M_CONTROL.request primitive has the following form:

```
M_CONTROL.request     (
                      destination_address,
                      opcode,
                      request_operand_list
                      )
```

The M_CONTROL.indication primitive has the following form:

```
M_CONTROL.indication  (
                      opcode,
                      indication_operand_list
                      )
```

IEEE Std 802.1AX-2020 references

6.5 Marker protocol, 6.5.1 Introduction:

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

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