P802.1Qdt/D0.3
June 18, 2024
(Amendment to IEEE Std 802.1Q-2022 as amended by IEEE P802.1Qdt/D0.3)

Draft Standard for Local and Metropolitan Area Networks—

Bridges and Bridged Networks

Amendment: Priority-based Flow Control Enhancements

Prepared by the
Time-Sensitive Networking (TSN) Task Group of IEEE 802.1

Sponsor

LAN/MAN Standards Committee of the IEEE Computer Society

This and the following cover pages are not part of the draft. They provide revision and other information for IEEE 802.1 Working Group members and will be updated as convenient. **New participants: Please read these cover pages**, they contain information that should help you contribute effectively to this standards development project. The Introduction to the current draft should be useful to all readers.

The text proper of this draft begins with the Title page.

This document is a draft amendment to IEEE Std 802.1Q-2022 as updated by published and draft amendments (if, and as, noted on the Title page), and may include (in addition to the main subject of the amendment, as per the PAR) the agreed or proposed resolution of Maintenance items and technical and editorial corrections, to the description of existing functionality(see below).

These cover pages provide an Introduction to the current draft, an introduction to Participation in 802.1 standards development, a summary of the PAR (Project Authorization Request) and CSD,for this project, and a general discussion of Draft development.

These cover pages will be replaced for SA Ballot by a briefer version providing information for that ballot, with space for commentary on, and hyperlinks to, changes that occur in SA Ballot.

Introduction to the current draft[1]

This draft, P802.1Qdt/D0.3, has been prepared for a first Task Group Ballot.

Maintenance items and technical and editorial corrections

This draft does not include proposed or agreed resolutions of maintenance items for the base standard.

This draft does not include any technical corrections to the base standard beyond the project subject matter.

This draft does not include any editorial corrections to the base standard beyond the project subject matter.

YANG modules

The YANG modules specified by this standard are not ready yet.

Sources

This draft, P802.1Qdt/D0.3, has been prepared from a set of Framemaker files with conditional text that supports the production of an amendment draft and a preliminary rollup of that amendment draft into the text of the base standard, IEEE Std 802.1Q-2022 as amended by prior amendments.

These sources are those used for P802.1Q-2022-Rev/D1.1, which include the text of the published and in-process amendments (at the time of preparation of this draft).

This particular amendment does not depend on the in-process amendments (P802.1Qdj and P802.1Qdx) and should be unaffected by any changes made to those amendments as part of SA Ballot, with the minor exception of possible (though unlikely) changes to clause numbering.

For a description of the use of conditional text and other FrameMaker and IEEE Std 802.1Q Style considerations applicable to this draft see the EDITOR-PLEASE-READ-ME file in the FrameMaker books used to generate this draft.

[1]The whole or parts of the introduction, possibly updated, to past drafts may be retained at the Editor's discretion, with the most recent introduction first. The introduction to each draft may solicit input on specific subjects.

Participation in 802.1 standards development

All participants in IEEE 802.1 activities should be aware of the Working Group Policies and Procedures, and their obligations under the IEEE Patent Policy, the IEEE Standards Association (SA) Copyright Policy, and the IEEE SA Participation Policy. For information on these policies see 1.ieee802.org/rules/ and the slides presented at the beginning of each of our Working Group and Task Group meeting.

The IEEE SA PAR (Project Authorization Request) and CSD (Criteria for Standards Development established by IEEE 802) are summarized in these cover pages and links are provided to the full text of both PAR and CSD. As part of the IEEE 802® process, the text of the PAR and CSD of each project is reviewed regularly to ensure their continued validity. A vote of "Approve" on this draft is also an affirmation by the voter that the PAR and CSD for this project are still valid.

Comments on this draft are encouraged. NOTE: All issues related to IEEE standards presentation style, formatting, spelling, etc. are routinely handled between the 802.1 Editor and the IEEE Staff Editors prior to publication, after balloting and the process of achieving agreement on the technical content of the standard is complete. Readers are urged to devote their valuable time and energy only to comments that materially affect either the technical content of the document or the clarity of that technical content. Comments should not simply state what is wrong, but also what might be done to fix the problem.

Full participation in the work of IEEE 802.1 requires attendance at IEEE 802 meetings. Information on 802.1 activities, working papers, and email distribution lists etc. can be found on the 802.1 Website:

http://ieee802.org/1/

Use of the email distribution list is not presently restricted to 802.1 members, and the working group has a policy of considering comments from all who are interested and willing to contribute to the development of the draft. Individuals not attending meetings have helped to identify sources of misunderstanding and ambiguity in past projects. The email lists exist primarily to allow the members of the working group to develop standards, and are not a general forum. All contributors to the work of 802.1 should familiarize themselves with the IEEE patent policy and anyone using the email distribution list will be assumed to have done so. Information can be found at http://standards.ieee.org/db/patents/

Comments on this draft may be sent to the 802.1 email exploder, to the Editors, or to the Chairs of the 802.1 Working Group and Time-Sensitive Networking (TSN) Task Group.

Lily Yunping Lyu
Editor, P802.1Qdt
Email:lvyunping@huawei.com

Mick Seaman
Editor, IEEE Std 802.1Q
Email:mickseaman@gmail.com

Janos Farkas
Chair, 802.1 TSN Task Group
Email:Janos.Farkas@ericsson.com

Glenn Parsons
Chair, 802.1 Working Group
+1 514-379-9037
Email: glenn.parsons@ericsson.com

NOTE: Comments whose distribution is restricted in any way cannot be considered, and may not be acknowledged.

All participants in IEEE standards development have responsibilities under the IEEE patent policy and should familiarize themselves with that policy, see http://standards.ieee.org/about/sasb/patcom/materials.html

PAR (Project Authorization Request) and CSD

Extracts from the PAR, as approved by IEEE NesCom June 5th, 2023:

https://development.standards.ieee.org/myproject-web/public/view.html#pardetail/10473

and the CSD (Criteria for Standards Development):

https://mentor.ieee.org/802-ec/dcn/22/ec-22-0083-01-ACSD-p802-1qdt.pdf

follow. The Scope and Purpose of the base standard remains unchanged from IEEE Std 802.1Q-2022.

PAR Scope of the Project:

This amendment specifies procedures and managed objects for automated Priority-based Flow Control (PFC) headroom calculation and Media Access Control Security (MACsec) protection of PFC frames, using point-to-point roundtrip measurement and enhancements to the Data Center Bridging Capability Exchange protocol (DCBX). This amendment places emphasis on the requirements for low latency and lossless transmission in large-scale and geographically dispersed data centers. This amendment also addresses errors of the existing IEEE Std 802.1Q functionality.

PAR Need for the Project:

PFC is used to avoid packet loss in low latency, high reliability Ethernet data centers and data center interconnects. For PFC to function properly and without wasting memory, the amount of headroom buffer must be calculated. Deployment in large scale data center networks and long distance interconnects is currently problematic and requires manual configuration. There are customer requirements for the integrity and confidentiality protection of all frames transmitted between geographically distributed data centers. The current specification is inconsistent and incomplete regarding the operation of PFC and MACsec together.

PAR Possible registration activity related to this project:

No.

CSD Broad market potential [extract]:

The data center market continues to grow very fast. Remote Direct Memory Access over Converged Ethernet (RoCEv2) is widely deployed, both within data centers and across data center interconnects. RoCEv2 requires lossless operation on Ethernet to avoid wasteful retransmissions. Priority-based Flow Control (PFC, specified in IEEE Std 802.1Q) enhancements make Ethernet technology more applicable and appealing for data center environments. There is a wide interest in the industry to enhance priority-based Flow Control (PFC, specified in IEEE Std 802.1Q) to make Ethernet technology more applicable and appealing for data center environment, such as cloud vendor, large enterprises, financial institutions, and other high-performance computing environments.

CSD Economic feasibility [extract]:

a) The proposed project can reduce cost of data center bridges by avoiding wasting memory.
b) The proposed project does not change the cost characteristics of bridges and end stations.
c) A modest reduction in installation cost of new equipment is expected. No incremental installation costs are expected from introducing round-trip delay measurement and associated DCBX enhancements.
d) The proposed project can reduce operational cost by configuration automation.

Draft development

During the early stages of draft development, 802.1 editors have a responsibility to attempt to craft technically coherent drafts from the resolutions of ballot comments and from the other discussions that take place in the working group meetings. Preparation of drafts often exposes inconsistencies in editor's instructions or exposes the need to make choices between approaches that were not fully apparent in the meeting. Choices and requests by the editors' for contributions on specific issues will be found in the editors' Introduction to the current draft and at appropriate points in the draft.

Any text with a Cyan background (as in this sentence) is temporary, with conditional tag 'Editor comment', inserted by the Editors to solicit comment, suggest a future change, or act simply as an aide memoire. Text can also highlighted to be draw it to the readers' attention, using conditional tag 'Editor highlight'. In both these case conditional tagging helps location, and eventual removal, of text or highlighting and can control whether or not it is displayed.

The ballot comments received on each draft, and the editors' proposed and final disposition of comments on working group drafts, are part of the audit trail of the development of the standard and are available, along with all the revisions of the draft on the 802.1 website (for address see above).

During the early stages of draft development the proposed text can be moved around a great deal, and even minor rearrangement can lead to a lot of 'change', not all of which is noteworthy from the point of the reviewer, so the use of automatic change bars is not very effective. In early drafts change bars may be omitted or applied manually, with a view to drawing the readers attention to the most significant areas of change. Readers interested in viewing every change are encouraged to use Adobe Acrobat to compare the document with their selected prior draft. Note that the FrameMaker change bar feature is useless when it comes to indicating changes to Figures.

This draft has been prepared from a set of Framemaker files with conditional text that supports the production of an amendment draft and a preliminary roll up of that amendment draft into the text of the base standard, i.e. IEEE Std 802.1Q as of the last Revision as amended by prior amendments (usually as of the close of their successful SA ballots) as noted on the Title Page and the first Cover Page. The editor may make preliminary roll ups available to check consistency with the base standard and cross-references to text that does not appear in this amendment. Roll ups may also be recorded as part of the approved P802.1Q Revision project.

For a description of the use of conditional text and other FrameMaker and IEEE Std 802.1Q Style considerations applicable to this draft see the EDITOR-PLEASE-READ-ME file in the FrameMaker books used to generate these drafts.

There are generally multiple amendments under development at any time, and while they will add or amend different clauses in the base standard, there are some clauses (notably Clauses 12, 48, and the PICS Annexes that all are likely to change). They need to be fully integrated before or during SA Ballot, and complete that ballot in serial order to avoid future problems.

Records of participants in the development of the standard are added after SA Ballot, as part of pre-publication editing by IEEE Staff.

P802.1Qdt/D0.3
June 18, 2024
(Amendment to IEEE Std 802.1Q-2022 as amended by IEEE P802.1Qdt/D0.3)

Draft Standard for Local and Metropolitan Area Networks—

Bridges and Bridged Networks

Amendment: Priority-based Flow Control Enhancements

Prepared by the

Time-Sensitive Networking (TSN) Task Group of IEEE 802.1

Sponsor

LAN/MAN Standards Committee
of the
IEEE Computer Society

IEEE Standards Department
445 Hoes Lane
Piscataway, NJ 08854, USA

Amendment: Priority-based Flow Control Enhancements

Abstract: This amendment to IEEE Std 802.1Q-2022 as amended by IEEE Std 802.1Qcz-2023, IEEE Std 802.1Qcw-2023, IEEE Std 802.1Qcj-2023, IEEE Std 802.1Qdj-2024, and IEEE Std 802.1Qdx-2024 addresses Multiple Spanning Tree Protocol (MSTP) requirements arising from industrial automation networks. It specifies YANG for bridge and bridge component RSTP and MSTP configuration and status reporting.

Keywords: Bridged Network, IEEE 802.1Q™, IEEE 802.1Qdy™, LAN, local area network, MAC Bridge, metropolitan area network, MSTP, Multiple Spanning Tree Protocol, MIB, Rapid Spanning Tree Protocol, RSTP, Virtual Bridged Network, virtual LAN, VLAN Bridge,YANG.

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

PDF: ISBN 978-X-XXX-XXX-X STDXXXXX
Print: ISBN 978-X-XXX-XXX-X STDPDXXXXX

Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (https://standards.ieee.org/ipr/disclaimers.html), appear in all standards and may be found under the heading "Important Notices and Disclaimers Concerning IEEE Standards Documents."

Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within IEEE Societies and subcommittees of IEEE Standards Association (IEEE SA) Board of Governors. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE Standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers are not necessarily members of IEEE or IEEE SA and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE makes no warranties or representations concerning its standards, and expressly disclaims all warranties, express or implied, concerning this standard, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE standards documents are supplied "AS IS" and "WITH ALL FAULTS."

Use of an IEEE standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document, should rely upon his or her own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Translations

The IEEE consensus development process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English version published by IEEE is the approved IEEE standard.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual shall not be considered or inferred to be the official position of IEEE or any of its committees and shall not be considered to be, nor be relied upon as, a formal position of IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter's views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group. Statements made by volunteers may not represent the formal position of their employer(s) or affiliation(s).

Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents**.

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and subcommittees of the IEEE SA Board of Governors are not able to provide an instant response to comments, or questions except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or in revisions to an IEEE standard is welcome to join the relevant IEEE working group. You can indicate interest in a working group using the Interests tab in the Manage Profile & Interests area of the IEEE SA myProject system.[1] An IEEE Account is needed to access the application.

Comments on standards should be submitted using the Contact Us form.[2]

Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

Copyrights

IEEE draft and approved standards are copyrighted by IEEE under US and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These include both use, by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, neither IEEE nor its licensors waive any rights in copyright to the documents.

[1] Available at: https://development.standards.ieee.org/myproject-web/public/view.html#landing.
[2] Available at: https://standards.ieee.org/content/ieee-standards/en/about/contact/index.html.

Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; https://www.copyright.com/. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit IEEE Xplore or contact IEEE.[3] For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

Errata

Errata, if any, for all IEEE standards can be accessed on the IEEE SA Website.[4] Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in IEEE Xplore. Users are encouraged to periodically check for errata.

Patents

IEEE Standards are developed in compliance with the IEEE SA Patent Policy.[5]

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at https://standards.ieee.org/about/sasb/patcom/patents.html. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

[3] Available at: https://ieeexplore.ieee.org/browse/standards/collection/ieee.
[4] Available at: https://standards.ieee.org/standard/index.html.
[5] Available at: https://standards.ieee.org/about/sasb/patcom/materials.html.

IMPORTANT NOTICE

IEEE Standards do not guarantee or ensure safety, security, health, or environmental protection, or ensure against interference with or from other devices or networks. IEEE Standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, and interference protection practices and all applicable laws and regulations.

Amendment: Priority-based Flow Control Enhancements

Participants

<<The following lists will be updated in the usual way prior to publication>>

At the time this standard was submitted to the IEEE-SA Standards Board for approval, the IEEE 802.1 Working Group had the following membership:

Glenn Parsons, *Chair*
Jessy V. Rouyer, *Vice Chair*
János Farkas, *Chair, Time-Sensitive Networking Task Group*
Craig Gunther, *Vice Chair, Time-Sensitive Networking Task Group*
Martin Mittelberger, *Editor*

<<TBA>>

The following members of the individual balloting committee voted on this standard. Balloters may have voted for approval, disapproval, or abstention.

<<TBA>>

When the IEEE-SA Standards Board approved this standard on XX Month 20xx, it had the following membership:

<<TBA>>

<<TBA>>

*Member Emeritus

Introduction

This introduction is not part of IEEE Std 802.1Qdy™-2024, IEEE Standard for Local and metropolitan area networks— Bridges and Bridged Networks—Amendment 40: YANG for Multiple Spanning Trees.

IEEE Std 802.1Qdy™-2024: YANG for Multiple Spanning Trees addresses requirements arising from industrial automation networks, specifying YANG for bridge and bridge component MSTP configuration and status reporting.

This standard contains state-of-the-art material. The area covered by this standard is undergoing evolution. Revisions are anticipated within the next few years to clarify existing material, to correct possible errors, and to incorporate new related material. Information on the current revision state of this and other IEEE 802 standards may be obtained from

Secretary, IEEE-SA Standards Board
445 Hoes Lane
Piscataway, NJ 08854-4141
USA

1. Overview

1.3 Introduction

a)

Add a paragraph to introduce DCBX function as follows:

<<Editor notes: DCBX function is only simply mentioned in ETS introduction paragraph. There should be a dedicate paragraph introducing DCBX.>>

This standard defines the Data Center Bridging eXchange protocol (DCBX), which is used by Data Center Bridging (DCB) devices to exchange configuration information with directly connected peers.

Change the paragraph beginning "This standard specifies protocols, procedures, and managed objects to support Priority-based Flow Control (PFC)" as follows:

This standard specifies protocols, procedures, and managed objects to support Priority-based Flow Control (PFC). These allow a Virtual Bridged Network, or a portion thereof, to enable flow control per traffic class on IEEE 802 point-to-point full-duplex links. To this end, it:

bh) Defines a means for a system to inhibit transmission of data frames on certain priorities from the remote system on the link.

bi) Defines PFC-capable interface stack operation with MACsec, MAC Privacy protection, and Link Aggregation.

bj) Defines a means for two participating systems to automatically calculate the minimum buffer requirements to assure lossless operation.

Change the paragraph beginning "This standard specifies protocols, procedures, and managed objects for enhancement of transmission selection to support allocation of bandwidth among traffic classes" as follows:

<<Editor notes: remove DCBX to a separate paragraph.>>

bk) This standard specifies protocols, procedures, and managed objects for Enhanced Transmission Selection (ETS) ~~enhancement of transmission selection~~ to support allocation of bandwidth among traffic classes. When the offered load in a traffic class does not use its allocated bandwidth, ~~Enhanced Transmission Selection (~~ETS~~) will~~ can allow other traffic classes to use the available bandwidth. Bandwidth is used by traffic classes subject to ETS when there are no frames to be transmitted for traffic classes subject to strict priority or credit-based shaper algorithms. ~~It defines the Data Center Bridging eXchange protocol (DCBX), which controls the application of ETS and PFC.~~

2. Normative references

The following referenced documents are indispensable for the application of this document (i.e., they must be understood and used, so each referenced document is cited in text and its relationship to this document is explained). For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments or corrigenda) applies.

Insert the following items into the list of Normative References:

ANSI X3.159, American National Standards for Information Systems—Programming Language—C.[2]

IEEE Std 802®, IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture.[3, 4]

IEEE Std 802d™-2017, IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture—Amendment 1: Allocation of Uniform Resource Name (URN) Values in IEEE 802® Standards.

IEEE Std 802.1AB™, IEEE Standard for Local and metropolitan area networks—Station and Media Access Control Connectivity Discovery.

IEEE Std 802.1AC™, IEEE Standard for Local and metropolitan area networks—Media Access Control (MAC) Service Definition.

IEEE Std 802.1AE™, IEEE Standard for Local and metropolitan area networks—Media Access Control (MAC) Security.

IEEE Std 802.1AS™, IEEE Standard for Local and metropolitan area networks—Timing and Synchronization for Time-Sensitive Applications in Bridged Local Area Networks.

IEEE Std 802.1AX™, IEEE Standard for Local and metropolitan area networks—Link Aggregation.

IEEE Std 802.1BR™, IEEE Standard for Local and metropolitan area networks—Virtual Bridged Local Area Networks—Bridge Port Extension.

IEEE Std 802.1CB™, IEEE Standard for Local and metropolitan area networks—Frame Replication and Elimination for Reliability.

IEEE Std 802.1CS™, IEEE Standard for Local and Metropolitan Area Networks—Link-local Registration Protocol.

IEEE Std 802.1X™, IEEE Standard for Local and Metropolitan Area Networks—Port-Based Network Access Control.

IEEE Std 802.3™, IEEE Standard for Ethernet.

IEEE Std 802.11™, Standard for Information Technology—Telecommunications and Information Exchange between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.

IEEE Std 802.20™, IEEE Standard for Local and metropolitan area networks—Part 20: Air Interface for Mobile Broadband Wireless Access Systems Supporting Vehicular Mobility—Physical and Media Access Control Layer Specification.

IEEE Std 1588™, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems.

IETF RFC 768 (STD0006), User Datagram Protocol, August 1980.

---

[2] ANSI publications are available from the IHS Standards Store (https://global.ihs.com/).

[3] The IEEE standards or products referred to in Clause 2 are trademarks owned by The Institute of Electrical and Electronics Engineers, Incorporated.

[4] IEEE publications are available from The Institute of Electrical and Electronics Engineers (https://standards.ieee.org/).

IETF RFC 791 (STD0005), Internet Protocol—DARPA Internet Program Protocol Specification, September 1981.[5]

IETF RFC 1035 (STD 13), Domain Names—Implementation and Specification, November 1987.

IETF RFC 1042, A Standard for the Transmission of IP Datagrams over IEEE 802 Networks, February 1988.

IETF RFC 1390 (STD 36), Transmission of IP and ARP over FDDI Networks, January 1993.

IETF RFC 2104, HMAC: Keyed-Hashing for Message Authentication, February 1997.

IETF RFC 2119 (BCP 14), Key words for use in RFCs to Indicate Requirement Levels, March 1997.

IETF RFC 2205, Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification, September 1997.

IETF RFC 2271, An Architecture for Describing SNMP Management Frameworks, January 1998.

IETF RFC 2474, Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, December 1998.

IETF RFC 2578 (STD 58), Structure of Management Information Version 2 (SMIv2), April 1999.

IETF RFC 2579 (STD 58), Textual Conventions for SMIv2, April 1999.

IETF RFC 2580 (STD 58), Conformance Statements for SMIv2, April 1999.

IETF RFC 2685, Virtual Private Networks Identifier, September 1999.

IETF RFC 2737, Entity MIB (Version 2), December 1999.

IETF RFC 2750, RSVP Extensions for Policy Control, January 2001.

IETF RFC 2863, The Interfaces Group MIB, June 2000.

IETF RFC 3046, DHCP Relay Agent Information Option, January 2000.

IETF RFC 3168, The Addition of Explicit Congestion Notification (ECN) to IP, September 2001.

IETF RFC 3232, Assigned Numbers: RFC 1700 is Replaced by an On-line Database, January 2002.

IETF RFC 3410, Introduction and Applicability Statements for Internet Standard Management Framework, December 2002.

IETF RFC 3411, An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks, December 2002.

IETF RFC 3413 (STD 62), Simple Network Management Protocol (SNMP) Applications, December 2002.

IETF RFC 3414 (STD 62), User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3), December 2002.

IETF RFC 3415 (STD 62), View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP), December 2002.

IETF RFC 3417 (STD 62), Transport Mappings for the Simple Network Management Protocol (SNMP), December 2002.

IETF RFC 3418 (STD 62), Management Information Base (MIB) for the Simple Network Management Protocol (SNMP), December 2002.

IETF RFC 3419, Textual Conventions for Transport Addresses, December 2002.

IETF RFC 4122, A Universally Unique IDentifier (UUID) URN Namespace, July 2005.

[5] IETF RFCs are available from the Internet Engineering Task Force (https://www.ietf.org/).

IETF RFC 4188, Definitions of Managed Objects for Bridges, September 2005.

IETF RFC 4291, IP Version 6 Addressing Architecture, February 2006.

IETF RFC 4318, Definitions of Managed Objects for Bridges with Rapid Spanning Tree Protocol, December 2005.

IETF RFC 4363, Definitions of Managed Objects for Bridges with Traffic Classes, Multicast Filtering, and Virtual LAN Extensions, January 2006.

IETF RFC 4789, Simple Network Management Protocol (SNMP) over IEEE 802 Networks, November 2006.

IETF RFC 5120, M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs), February 2008.

IETF RFC 5303, Three-Way Handshake for IS-IS Point-to-Point Adjacencies, October 2008.

IETF RFC 5305, IS-IS Extensions for Traffic Engineering, October 2008.

IETF RFC 5307, IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS), October 2008.

IETF RFC 6165, Extensions to IS-IS for Layer-2 Systems, April 2011.

IETF RFC 6335, Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry, August 2011.

IETF RFC 7365, Framework for Data Center (DC) Network Virtualization, October 2014.

IETF RFC 7810, IS-IS Traffic Engineering (TE) Metric Extensions, May 2016.

IETF RFC 7811, An Algorithm for Computing IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR) , June 2016.

IETF RFC 7950, The YANG 1.1 Data Modeling Language, August 2016.

IETF RFC 8200 (STD0086), Internet Protocol, Version 6 (IPv6) Specification, July 2017.

IETF RFC 8343, A YANG Data Model for Interface Management, March 2018.

IETF RFC 8394, Split Network Virtualization Edge (Split-NVE) Control-Plane Requirements, May 2018.

ISO/IEC 7498-1, Information technology—Open Systems Interconnection—Basic Reference Model: The Basic Model.[6]

ISO/IEC 8802-2, Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 2: Logical link control.

ISO/IEC 8802-11, Telecommunications and information exchange between systems—Specific requirements for local and metropolitan area networks—Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications.

ISO/IEC TR 9577:1999, Information technology—Protocol identification in the network layer.

ISO/IEC 10589:2002, Information technology—Telecommunications and information exchange between systems—Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473).

[6] ISO/IEC publications are available from the International Organization for Standardization (https://www.iso.org/) and the International Electrotechnical Commission (https://www.iec.ch/). ISO/IEC publications are also available from the American National Standards Institute (https://www.ansi.org/).

ISO/IEC TR 11802-5:1997, Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Technical reports and guidelines—Part 5: Media Access Control (MAC) Bridging of Ethernet V2.0 in Local Area Networks.

ITU-T Recommendation X.690 (2002), Information technology—ASN.1 encoding rules: Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER).[7]

ITU-T Recommendation G.8013/Y.1731, Operation, administration and maintenance (OAM) functions and mechanisms for Ethernet-based networks.

MEF Technical Specification 10.3 (MEF 10.3), Ethernet Services Attributes Phase 3, October 2013.[8]

[7] ITU-T publications are available from the International Telecommunications Union (https://www.itu.int/).
[8] MEF publications are available from the MEF Forum (https://www.mef.net/).

5. Conformance

5.11 System requirements for Priority-based Flow Control (PFC)

A system that conforms to the provisions of this standard for PFC (see Clause 36) shall:

Change below items as follows:

a) Support, on one or more ports, enabling PFC on at least one priority (~~36.1.2~~36.3.1).
b) Support, for each PFC Priority, processing PFC M_CONTROL.requests (~~36.1.3.1~~36.3.1).
c) Support, for each PFC Priority, processing PFC M_CONTROL.indications (~~36.1.3.3~~36.3.2).
d) Abide by the PFC delay constraints (~~36.1.3.3~~36.3.3).
e) Provide PFC-aware system queue functions (~~36.2~~36.3.4).
f) Enable use of PFC only in a domain controlled by DCBX (Clause 38).

A system that conforms to the provisions of this standard for PFC may:

g) Support enabling PFC on up to eight priorities per port.
h) Support the IEEE8021-PFC-MIB (17.7.17).

Insert new list items after item h) in 5.11 as follows:

i) Support PFC-capable interface stack operation with MACsec (36.5).
j) Support PFC-capable interface stack operation with MAC Privacy protection (36.6).
k) Support PFC-capable interface stack operation with Link Aggregation (36.7).
l) Support automatic calculation of PFC minimum buffer requirements for lossless operation (36.8)

12. Bridge management

12.23 Priority-based Flow Control objects

<< Editor notes: This sub-clause defines PFC managed objects.

1. PFCLinkDelayAllowance is defined as PFC headroom, but the value is set manually. We need a new managed object for automatic calculated headroom.

2. How is manual setting and automatic setting compatible?

1) Add a new managed object 'PFCHeadroomAllowance' for automatic calculated headroom.

2) If automatic way is defined in DCBX, use PFCHeadroomAllowance. Otherwise, use PFCLinkDelayAllowance as before.

3) The default value of PFCHeadroomAllowance is recommended to be PFCLinkDelayAllowance.

>>

Add a new object into the sub-clause and change the content as follows:

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

a) **PFCLinkDelayAllowance:** the default 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
d) **PFCHeadroomAllowance:** the automatic calculated round-trip propagation delay of the link as PFC headroom in bits

Table 12-1 shows the format and applicability of these objects.

Table 12-1—Priority-based Flow Control objects

Name	Data type	Operations supported[a]	Conformance[b]
PFCLinkDelayAllowance	unsigned integer	RW	BE
PFCRequests	unsigned integer	R	BE
PFCIndications	unsigned integer	R	BE
PFCHeadroomAllowance	unsigned integer	RW	BE

[a] R = Read only access; RW = Read/Write access.
[b] B = Required for Bridge or Bridge Component support of PFC; E = Required for end station support of PFC.

NOTE—The PFC Initiator (see 36.2.1) can use the PFCLinkDelayAllowance or PFCHeadroomAllowance parameter as one of the factors to determine when to issue a PFC M_CONTROL.request in order to not discard frames. The PFCLinkDelayAllowance parameter can be ~~written~~ set manually 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. When PFC headroom calculation (36.8) function is enabled, the PFCLinkDelayAllowance parameter takes effect.

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.

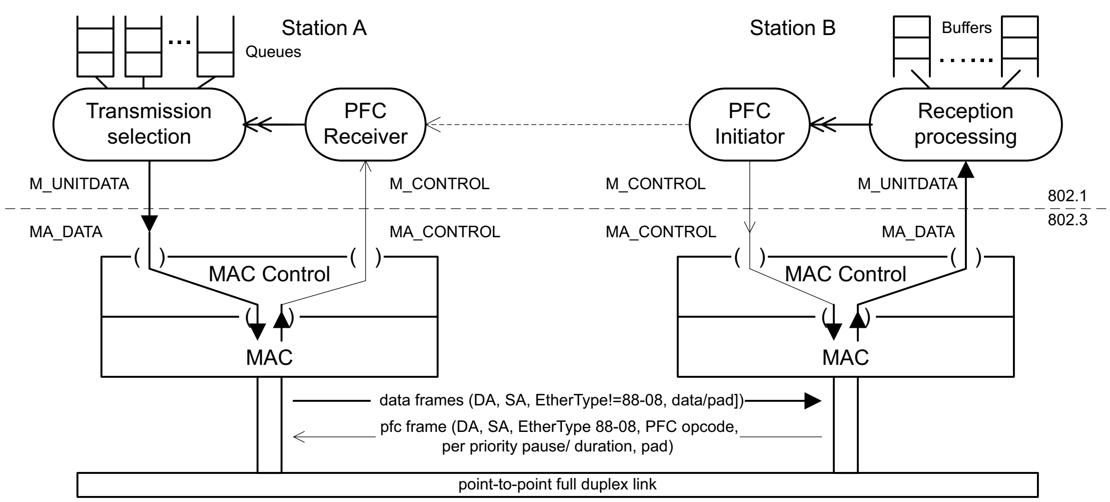

Figure 36-1—PFC example

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.

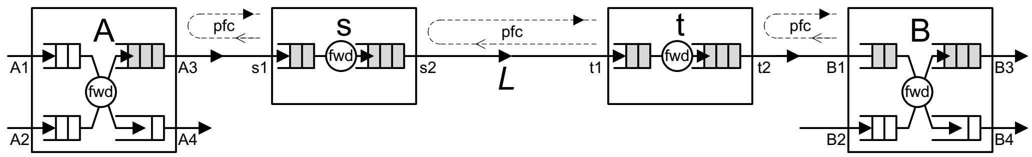

Figure 36-2—PFC hop-by-hop flow control with TPMRs

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

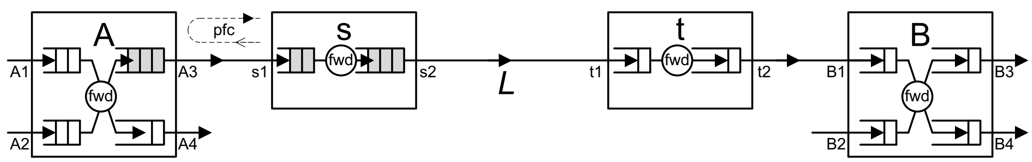

Figure 36-3—PFC hop-by-hop flow control with link rate mismatch

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.

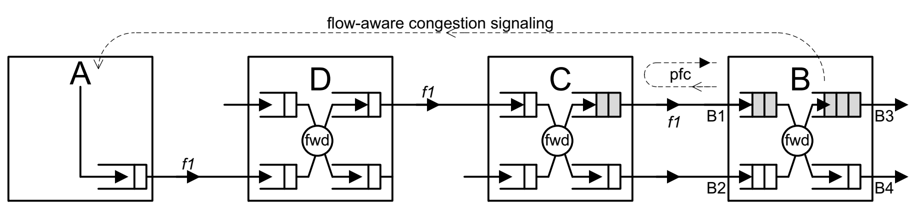

Figure 36-4—Flow-aware congestion signaling with PFC loss prevention

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.

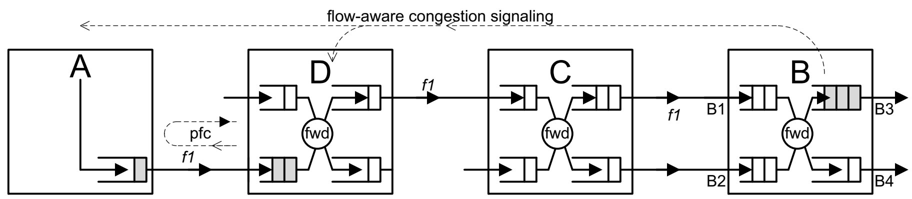

Figure 36-5—Flow-aware congestion signaling with PFC back-pressure

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.

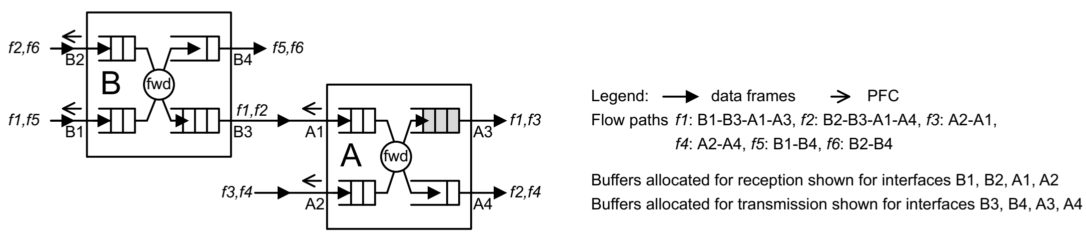

Figure 36-6—PFC congestion spreading

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.

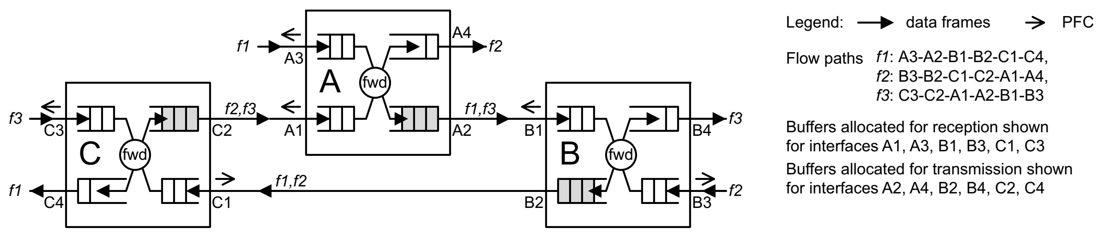

Figure 36-7—PFC deadlock example

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.

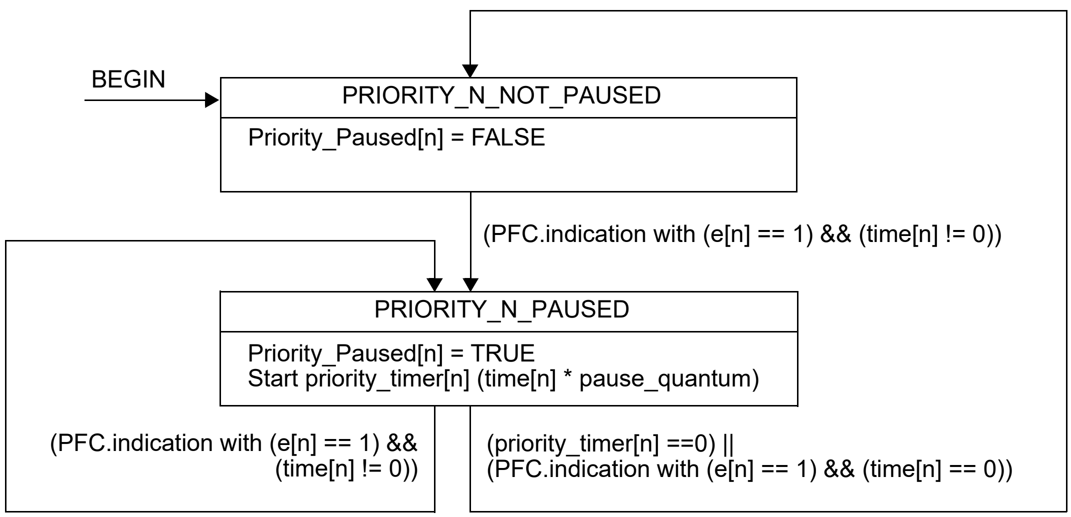

Figure 36-8—PFC Receiver state diagram for priority n

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.

Add new section to describe PFC-aware system queue functions. (Reuse text from original 36.2)

36.3.4 PFC-aware system queue functions

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.

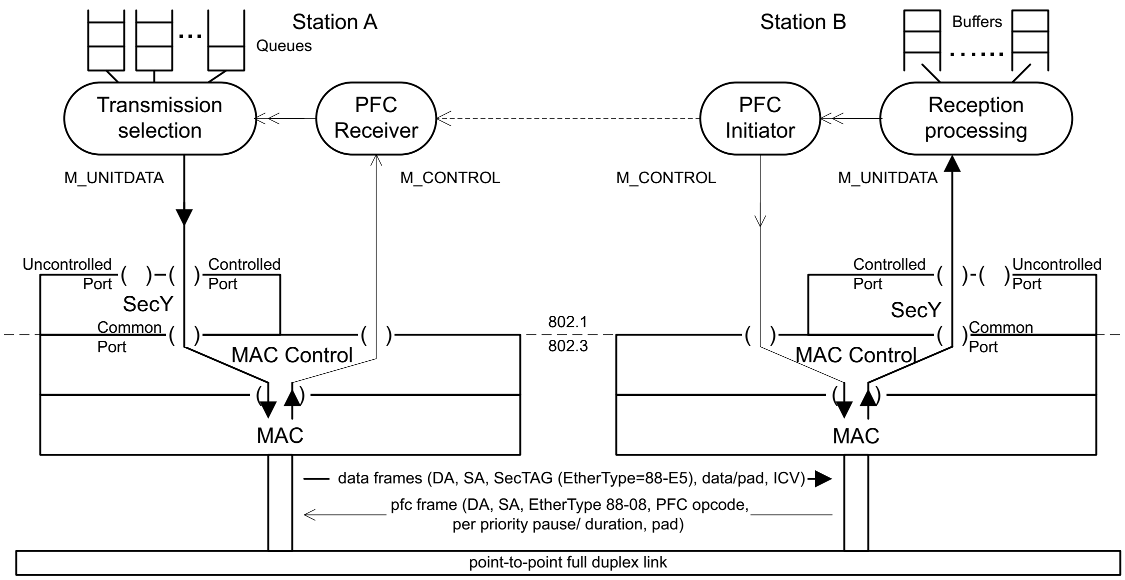

Figure 36-9—PFC with IEEE 802.3 MAC Control and MACsec

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.

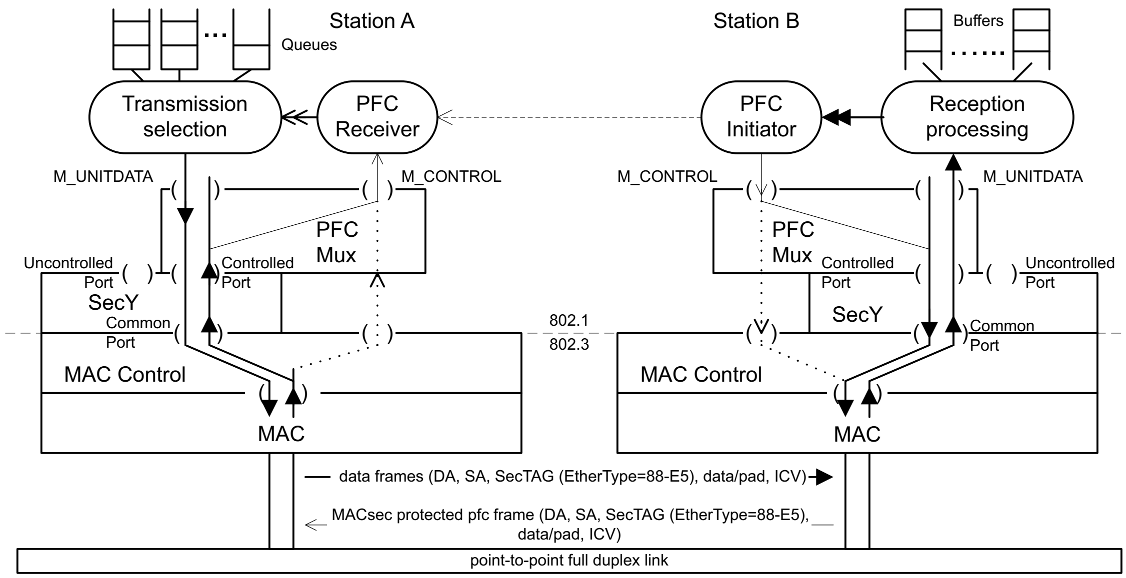

Figure 36-10—MACsec protection of user data and PFC frames

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.

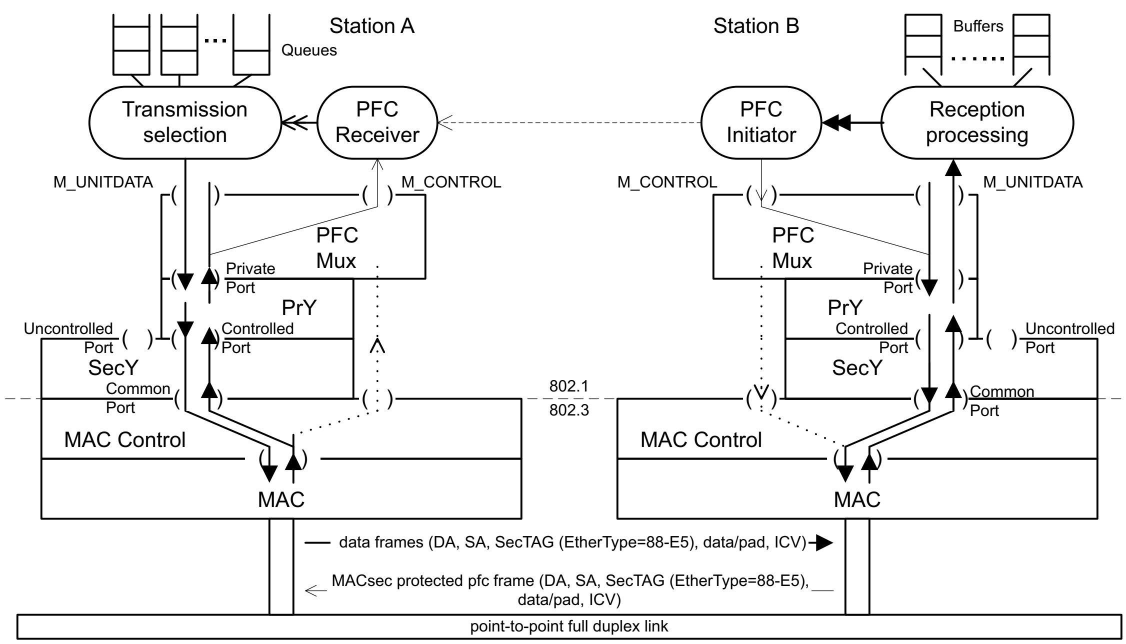

Figure 36-11—MAC Privacy protection and PFC

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.

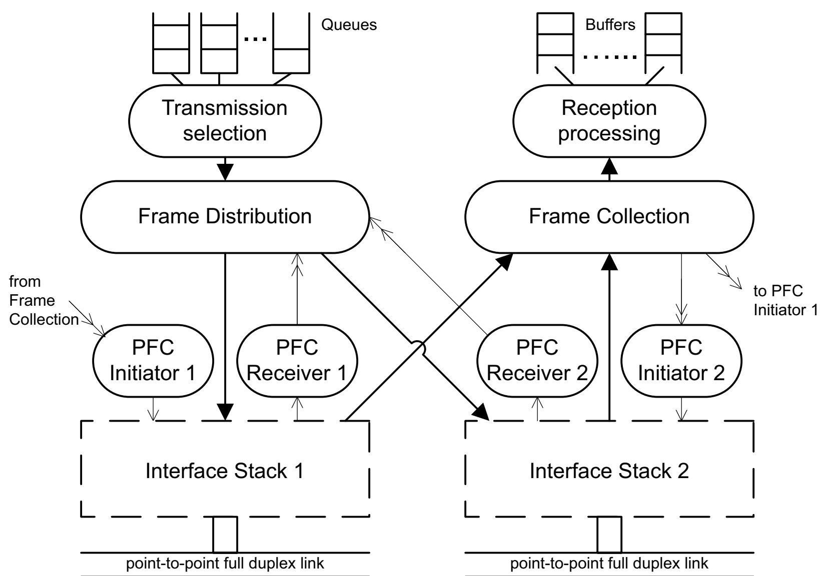

Figure 36-12—PFC operation with link aggregation

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 $0\ldots2^{n}-1$ 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 $-2^{n-1}\ldots2^{n-1}-1$ 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.[9]

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.

[9] Cable delay approximately 5 microseconds per kilometer (5 nanosconds 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.

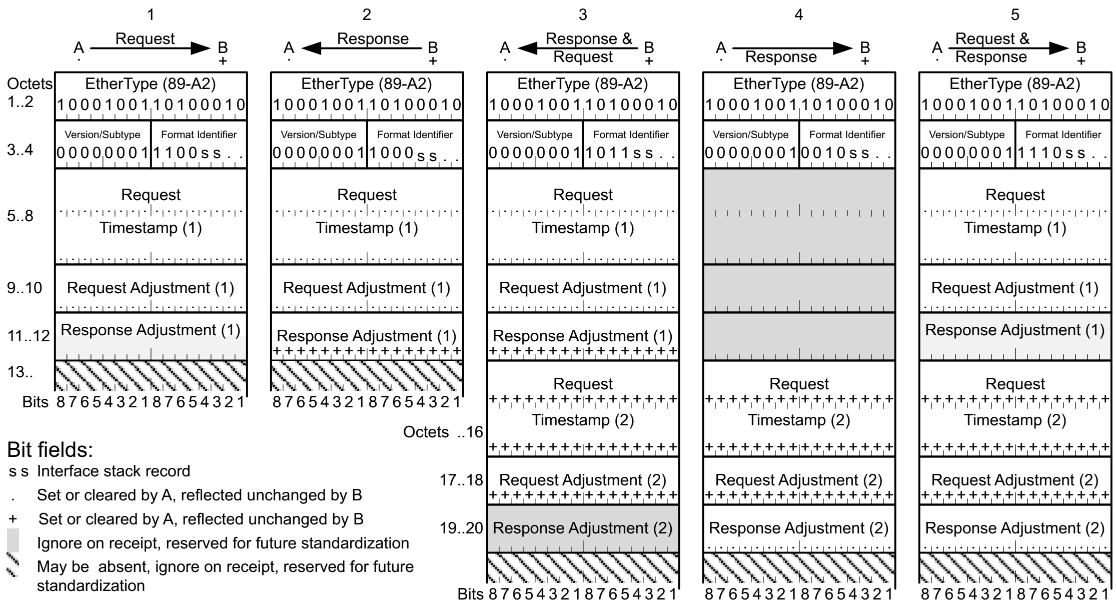

Figure 36-13—HMPDU format (examples)

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.

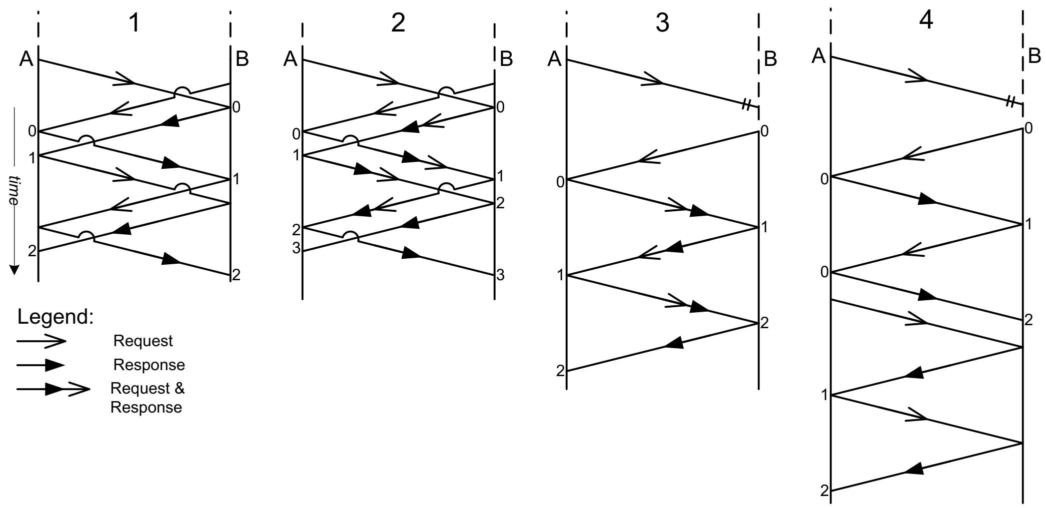

Figure 36-14—Measurement (examples)

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.

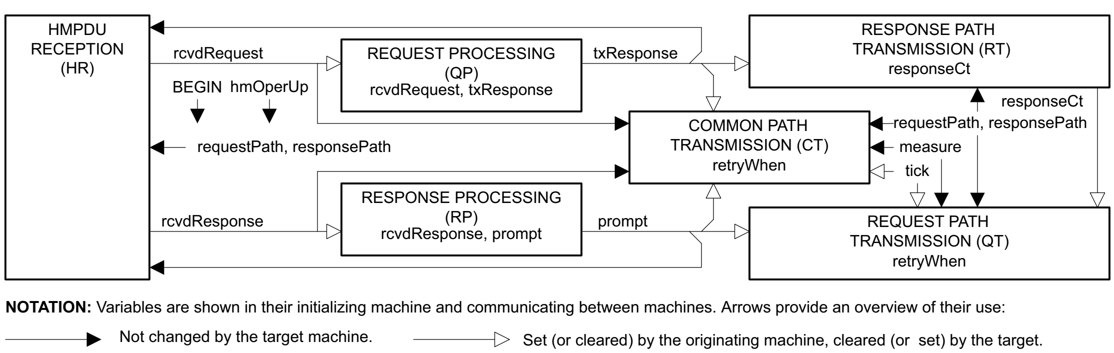

Figure 36-15—Headroom measurement state machines—overview

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.

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.

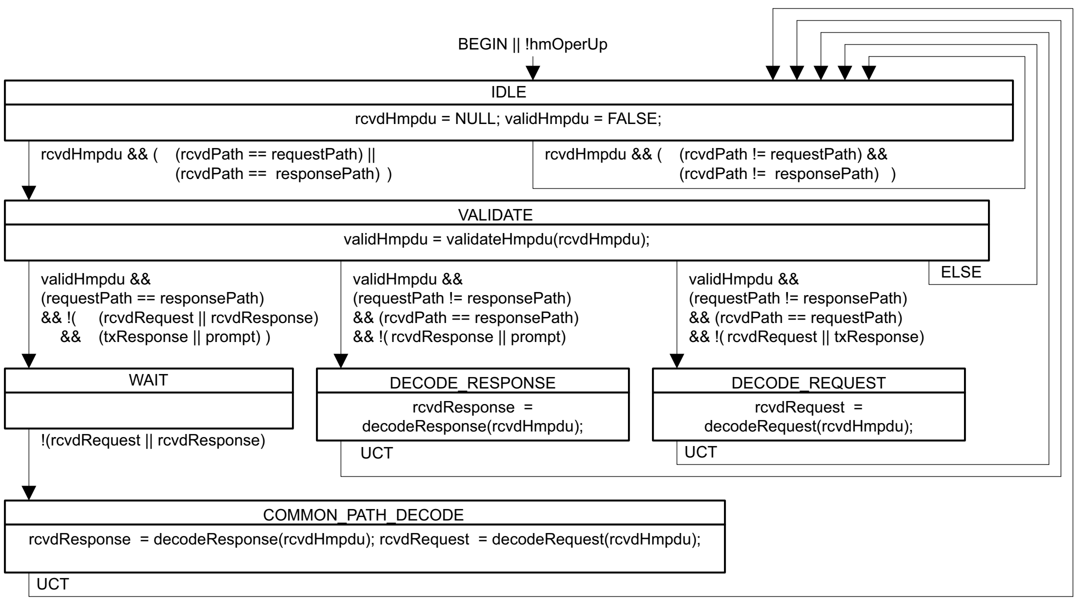

Figure 36-16—HMPDU Reception state machines

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.

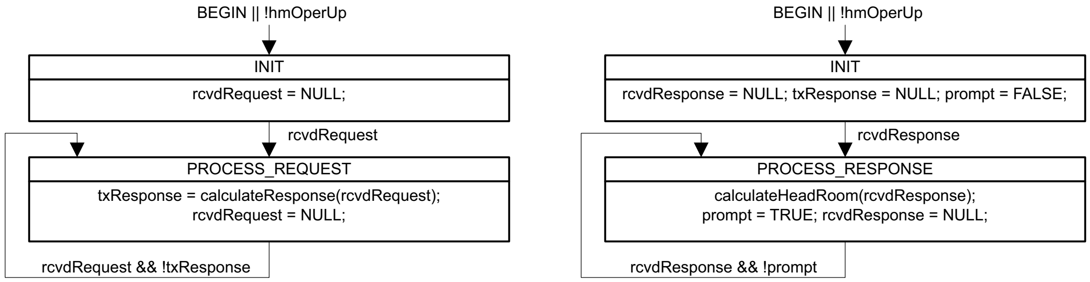

Figure 36-17—Request and Response processing state machines

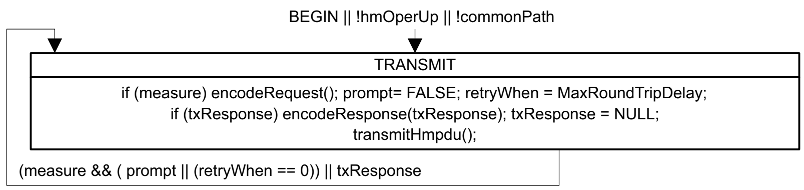

Figure 36-18—Common Path Transmission state machine

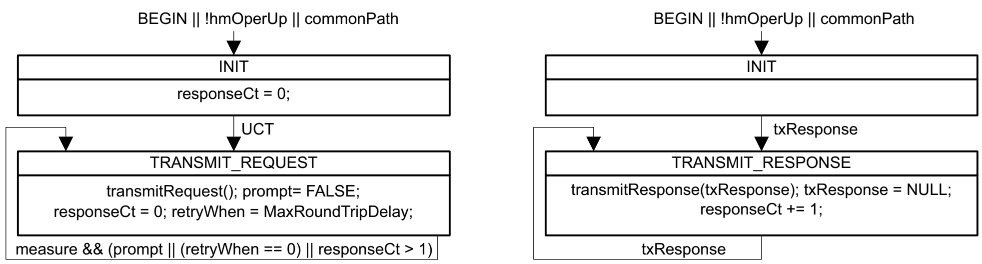

Figure 36-19—Request and Response Path Transmission state machines

36.11 PFC management

38. Data Center Bridging eXchange protocol (DCBX)

38.1 Overview

This clause details the DCBX, which is used by DCB devices to exchange configuration information with directly connected peers. The protocol may also be used for misconfiguration detection and for configuration of the peer.

This standard describes the base protocol, which comprises state machines and TLVs for capability exchange. For each feature that is supported by DCBX, the attributes that are to be exchanged specify the following:

a) The attributes to be exchanged
b) How the attributes are used for detecting misconfiguration
c) What action needs to be taken when a misconfiguration is detected

<<Editor notes: The description of why and how DCBX should be used for the features below is missing. In Clause 36 (PFC), this description should be added, and item e) could reference this description.>>

The information listed above is specified for the following:

d) ETS
e) PFC
f) Application Priority TLV
g) Application VLAN TLV

38.2 Goals

<<Editor notes: This is a high-level description of the goals of DCBX. The detailed goals for specific features (e.g., PFC) should be described in the main clause corresponding to those features.>>

The goals of DCBX are as follows:

a) Discovery of DCB capability in a peer port; for example, it can be used to determine if two link peer ports support PFC.
b) DCB feature misconfiguration detection: DCBX can be used to detect misconfiguration of a feature between the peers on a link. Misconfiguration detection is feature-specific because some features allow asymmetric configuration.
c) Peer configuration of DCB features: DCBX can be used by a device to perform configuration of DCB features in its peer port if the peer port is willing to accept configuration.

48. YANG Data Models

<<Editor notes: YANG model for PFC is missing. Sub-clauses in Clause 48 need to be checked, and the YANG model for PFC should be added. >>

Add one item in the paragraph beginning with "The YANG data models specified in this clause include the following:" as follows:

The YANG data models specified in this clause include the following:

— A VLAN Bridge components data model (48.2.1) that allows control and status monitoring of one or more C-VLAN or S-VLAN Bridge components (8.2) that compose all or part of a system's functionality, and the Bridge Port interfaces that support those components.

— A Two-Port MAC Relay data model (48.2.2) that both subsets and augments the VLAN Bridge components model to model a VLAN-unaware TPMR (3.292)

— A Customer VLAN Bridge model (48.2.3) that comprises a single VLAN Bridge component from the VLAN Bridge components model.

— A Provider Bridges model that uses one or multiple components from the VLAN Bridge components model to compose an S-VLAN component Provider Bridge or a Provider Edge Bridge.

— Connectivity Fault Management (CFM) models (48.2.3) for use with the VLAN Bridge components and related models in systems that provide CFM functionality.

— A Stream filters and stream gates model (48.2.6) that augments the VLAN Bridge components model.

— An Asynchronous Traffic Shaping (ATS) model that augments the VLAN Bridge components model and the Stream filters and stream gates model.

— An Priority-Based Flow Control (PFC) model that augments the VLAN Bridge components model.

Annex D

(normative)

IEEE 802.1 Organizationally Specific TLVs

D.1 Requirements of the IEEE 802.1 Organizationally Specific TLV sets

Change Table D-1 as follows:

Table D-1—IEEE 802.1 Organizationally Specific TLVs

IEEE 802.1 subtype	TLV name	TLV set name	TLV reference	Feature clause reference
0x01	Port VLAN ID	basicSet	D.2.1	6.9
0x02	Port And Protocol VLAN ID	basicSet	D.2.2	6.12
0x03	VLAN Name	basicSet	D.2.3	12.10.2.1.3
0x04	Protocol Identity	basicSet	D.2.4	D.2.4
0x05	VID Usage Digest	basicSet	D.2.5	D.2.5
0x06	Management VID	basicSet	D.2.6	D.2.6
0x07	Link Aggregation TLV	basicSet	IEEE Std 802.1AX	IEEE Std 802.1AX
0x08	Congestion Notification	cnSet	D.2.7	Clause 33
0x09	ETS Configuration TLV	dcbxSet	D.2.8	Clause 38
0x0A	ETS Recommendation TLV	dcbxSet	D.2.9	Clause 38
0x0B	Priority-based Flow Control Configuration TLV	dcbxSet	D.2.10	Clause 38
0x0C	Application Priority TLV	dcbxSet	D.2.11	Clause 38
0x0D	EVB TLV	evbSet	D.2.12	D.2.12
0x0E	CDCP TLV	evbSet	D.2.13	D.2.13
0x10	Application VLAN TLV	dcbxSet	D.2.14	Clause 38
0x11	LRP ECP Discovery TLV	lrpSet	IEEE Std 802.1CS	IEEE Std 802.1CS
0x12	LRP TCP Discovery TLV	lrpSet	IEEE Std 802.1CS	IEEE Std 802.1CS
0x13	Congestion Isolation TLV	ciSet	D.2.15	49.4.4
0x14	Topology Recognition TLV	trSet	D.2.16	49.5
0x15	PBBN Auto Attach System TLV	aaSet	D.2.17	Clause 50
0x16	PBBN Auto Attach Assignment TLV	aaSet	D.2.18	Clause 50
0x17	Priority-based Flow Control Local Delay TLV	dcbxSet	D.2.19	Clause 38

D.2 Organizationally Specific TLV definitions

D.2.10 Priority-based Flow Control Configuration TLV

The TLV illustrated in Figure D-10 is encoded into each LLDP message and may be transmitted by a system in order to indicate how PFC should be configured. Shall be sent using Symmetric attribute passing.

<<Editor note: Can both 'RTM HDRM' and 'PTP HDRM' be true at the same time? Can both 'MACsec cap' and 'Privacy cap' be true at the same time?>>

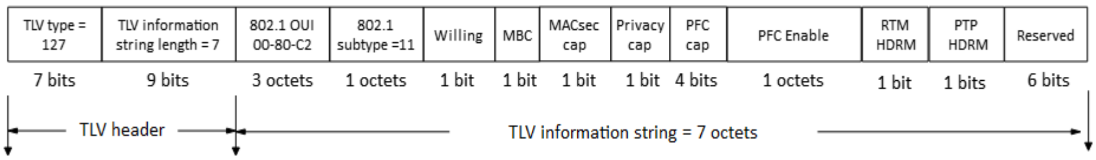

Figure D-10—Priority-based Flow Control Configuration TLV format

D.2.10.1 TLV type

D.2.10.2 TLV information string length

Change the paragraph as follows:

A 9-bit unsigned integer, occupying the LSB of the first octet of the TLV (the MSB of the TLV information string length) and the entire second octet of the TLV, containing the total number of octets in the TLV information string of the Priority-based Flow Control Configuration TLV. This does not count the TLV type and TLV information string length fields. It is equal to ~~6~~ 7.

D.2.10.3 Willing

D.2.10.4 MBC

<<Editor note: MBC definition is ambiguous, need to clarify. >>

The MACsec Bypass Capability Bit. If set to zero, the sending station is capable of bypassing MACsec processing when MACsec is disabled. If set to one, the sending station is not capable of bypassing MACsec processing when MACsec is disabled (see Clause 36).

Insert new paragraph for MACsec capability as follows:

D.2.10.5 MACsec cap

A 1-bit unsigned integer, MACsec cap (MACsec protection capability) indicates the device support for MACsec protection on PFC frames (see 36.2.6). If the MACsec cap bit is 1, and PFC is enabled on at least one traffic class, the MACsec protection is enabled.

Insert new paragraph for Privacy capability as follows:

D.2.10.6 Privacy cap

A 1-bit unsigned integer, Privacy cap (Privacy protection capability) indicates the device support for privacy protection on PFC frames (see 36.2.7). If the Privacy cap bit is 1, and PFC is enabled on at least one traffic class, the privacy protection is enabled.

D.2.10.7 PFC cap

D.2.10.8 PFC Enable

Insert new paragraph for headroom round-trip measurement capability as follows:

D.2.10.9 RTM HDRM

A 1-bit unsigned integer, RTM HDRM (round-trip measurement) indicates the device support for PFC headroom calculation using round-trip measurement (see 36.9). If the RTM HDRM bit is 1, and PFC is enabled on at least one traffic class, the round-trip measurement is enabled.

Insert new paragraph for headroom PTP based measurement capability as follows:

D.2.10.10 PTP HDRM

A 1-bit unsigned integer, PTP HDRM (PTP based measurement) indicates the device support for PFC headroom calculation using link delay (see 36.8.1). If the PTP HDRM bit is 1, and PFC is enabled on at least one traffic class, the PTP based measurement is enabled.

Insert new PFC informational TLV as follows:

D.2.19 Priority-based Flow Control Local Delay TLV

The TLV illustrated in Figure D-19— is encoded into each LLDP message and may be transmitted by a system in order to indicate local delays which facilitate PFC headroom calculation using link delays. This TLV is informational and used to indicate a peer station the local value.

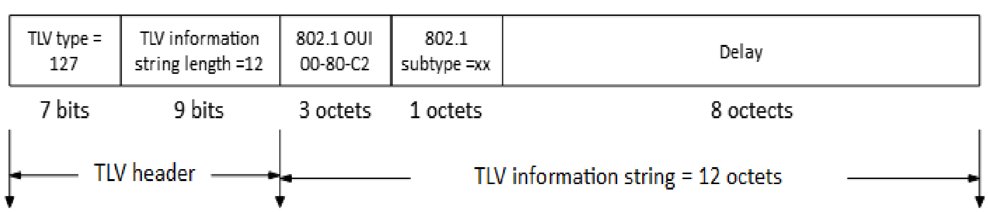

Figure D-19—Priority-based Flow Control Internal Delay TLV Format

D.2.19.1 TLV type

A 7-bit integer value occupying the most significant bits of the first octet of the TLV. Always contains the value 127.

D.2.19.2 TLV information string length

A 9-bit unsigned integer, occupying the LSB of the first octet of the TLV (the MSB of the TLV information string length) and the entire second octet of the TLV, containing the total number of octets in the TLV information string of the Priority-based Flow Control Internal Delay TLV. This does not count the TLV type and TLV information string length fields. It is equal to 12.

D.2.19.3 Subtype

The subtype used to identify the TLV format is as shown in Table D-1.

D.2.19.4 Delay

A 8-octet signed integer, indicating local system delays (see item a) 3) of 36.8) to PFC headroom calculation.

D.3 IEEE 802.1 Organizationally Specific TLV management

D.3.1 IEEE 802.1 Organizationally Specific TLV selection management

D.3.2 IEEE 802.1 managed objects—TLV variables

Insert a new paragraph as follows:

D.3.2.14 Priority-base Flow Control TLV managed objects

<<Editor notes: Add PFC DCBX managed objects>>

<<Editor notes: Need to add MIB/YANG for PFC configuration TLV and PFC local delay TLV in D.5 and D.6>>

Annex N

(informative)

Buffer requirements for PFC

Change the clause text as below.

N.1 Overview

To ensure that data frames are not lost due to lack of receive buffer space, receivers must ensure that a PFC M_CONTROL.request primitive is invoked while there is sufficient receive buffer to absorb the data that can continue to be received during the time needed by the remote system to react to the PFC operation. The PFC headroom (see 36.1.1) is the minimum buffer size that needs to be available when PFC frame is transmitted. It can be consumed before reception stops. It helps implementation to allocate buffer for PFC-enabled priorities. But the ~~The~~ precise calculation of this buffer requirement and buffer allocation are ~~is~~ highly implementation dependent. ~~This annex provides an example of how it can be calculated based on a hypothetical delay model.~~ This annex explains delay model of PFC headroom, and provides an example of buffer allocation based on the PFC headroom calculation. Setting the PFCLinkDelayAllowance or PFCHeadroomAllowance (see 12.23) to less than ~~the round-trip delay~~ the headroom value can result in frames loss.

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

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

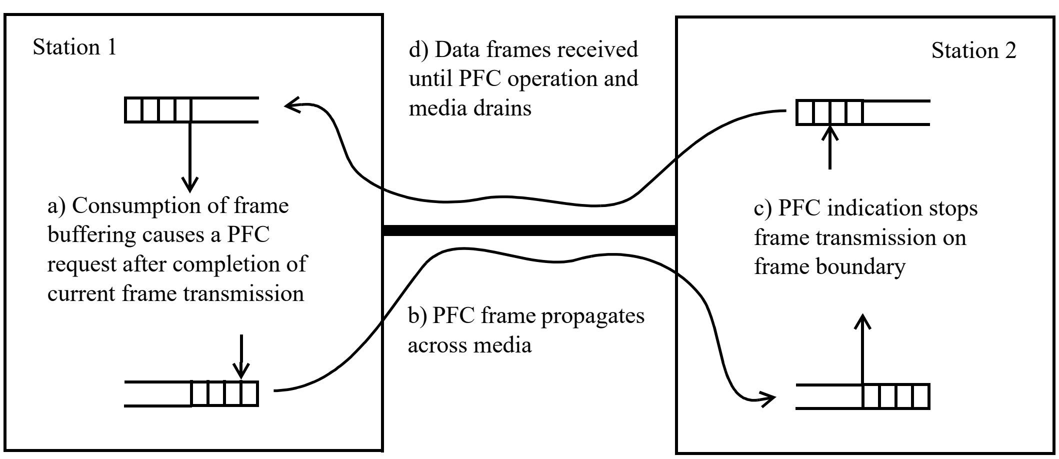

Figure N-1—~~PFC delays~~

N.2 Delay model of PFC headroom

PFC headroom calculation considers various delays accumulated. Figure N-1 provides a high-level view of the various delays to consider, in which station B is the station with PFC initiator and station A is the station with PFC receiver.

a) Stage 1: PFC frame transmission in station B
b) Stage 2: PFC frame transmission across link from station B to station A
c) Stage 3: PFC frame reception in station A (including PFC taking action)
d) Stage 4: User data transmission in station A
e) Stage 5: User data transmission across link from station A to station B
f) Stage 6: User data reception in PFC initiator station B

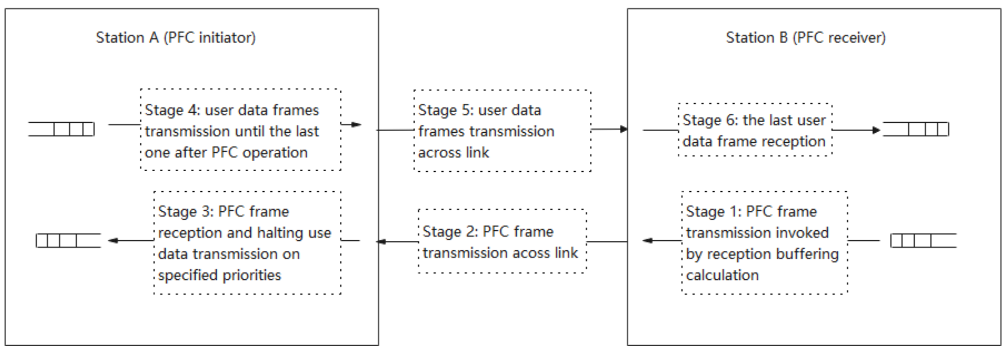

Figure N-1—Various delays

Each stage is further divided into several actions, from item a) to n).

Stage 1: PFC frame transmission in station B.

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

Stage 2: PFC frame transmission across link from station B to station A.

f) Last bit of PFC frame sent from B on the physical link.
g) Last bit of PFC frame arriving at A on the physical link.

Stage 3: PFC frame reception in station A (including PFC taking action).

h) PFC frame reception by A's interface stack and decoded by PFC receiver.
i) Transmission selection for specified priorities halted.

Stage 4: User data transmission in station A

j) Any in-progress frame transmission (possibly of a maximum sized frame).
k) Last bit of last frame sent on the physical link.

Stage 5: User data transmission across link from station A to station B:

l) Last bit of last frame sent from A on the physical link.
m) Last bit of last frame arriving at B on the physical link.

Stage 6: User data reception in PFC initiator station B:

n) Last frame reception by B’s interface stack as well as buffering.

There is a worst case scenario considered in stage 1 and stage 4. In stage 1, item c) considers a maximum sized frame transmission just before the PFC frame transmission at PFC initiator. In stage 4, item j) considers a maximum sized frame transmission just after the PFC frame taking effect at PFC receiver. The delay introduced by such worst case scenario is worst case delay (WD).

The delays in stage 1,3,4 and 6 but without WD are internal processing delays (ID). These delays represent the time spent on frame processing within the stations. Examples of such delays include interface stack delay, buffering delay, queue status change delay, assuming no prior in-progress frame transmission. Stage 1 and 6 occur at the PFC initiator station, with stage 1 in the transmitting direction and stage 6 in the receiving direction.

The delays in stage 2 and 5 are link delays (LD). These delays represent the time spent on physical link between two stations. Stage 2 is from PFC initiator to PFC receiver, while stage 5 is in the reverse direction.

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

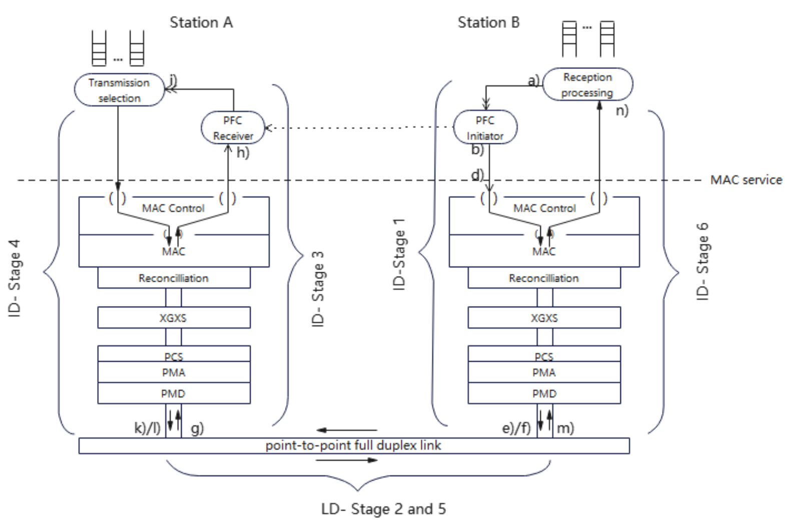

Figure N-2—Delay model

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

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

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

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

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

Figure N-3 shows a possible worst-case delay example where MACsec is disabled for PFC frame and user data.

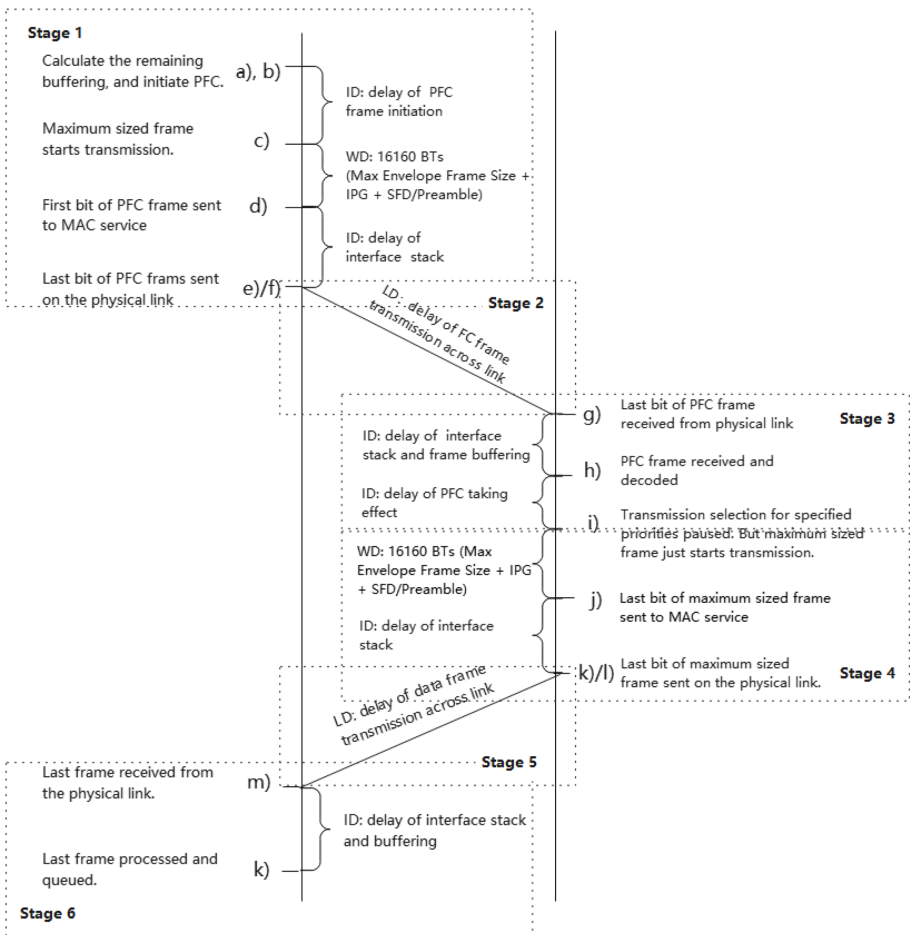

Figure N-3—Worst case delay

The total Delay Value (DV) is the sum of all delays shown in Figure N-3. It is the round-trip delay from PFC initiation to last frame reception and buffered, plus 2 maximum sized frames represented by bit times.

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

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

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

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

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

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

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

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

For link speeds of up to 10Gb/s, MACsec constrains each of the transmit delay and the receive delay to a maximum of 19 360 bit times (see 36.3.3).

This standard defines a queue shall go into paused state in no more than 614.4 ns (see 36.3.3). This delay is equivalent to 6144 bit times at the speed of 10Gb/s.

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

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

Table N-1—IEEE 802.3 Interface Delays

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

N.4 ~~Cable Delay~~ Link Delay

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

$$\cancel{\text{Cable}}\ \text{Link Delay} = \text{Medium Length} \times \frac{1}{\text{BT} \times \upsilon}$$

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

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

The Worst-case Delay comprises 2 parts.

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

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

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

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

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

N.6 Buffer allocation ~~Computation~~ example

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

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

The total Delay Value in this scenario results as follows:

~~DV = 2 × (Max Frame) + (PFC Frame) + 2 × (Cable Delay) + 2 × ID + HDs2~~

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

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

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

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

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

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

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