IEEE STANDARDS ASSOCIATION

IEEE Standard for

Local and metropolitan area networks—

Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—

Amendment 17: Priority-based Flow Control

IEEE Computer Society

Sponsored by the
LAN/MAN Standards Committee

IEEE
3 Park Avenue
New York, NY 10016-5997
USA

30 September 2011

IEEE Std 802.1Qbb™-2011
(Amendment to
IEEE Std 802.1Q™-2011
as amended by IEEE 802.1Qbe™-2011
and IEEE Std 802.1Qbc™-2011)

IEEE Std 802.1Qbb™-2011
(Amendment to
IEEE Std 802.1Q™-2011
as amended by IEEE Std 802.1Qbe™-2011
and IEEE Std 802.1Qbc™-2011)

IEEE Standard for
Local and metropolitan area networks—

Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—

Amendment 17: Priority-based Flow Control

Sponsor

LAN/MAN Standards Committee
of the
IEEE Computer Society

Approved 16 June 2011

IEEE-SA Standards Board

Approved 26 April 2012

American National Standards Institute

Abstract: This amendment to IEEE Std 802.1Q-2011 specifies protocols, procedures, and managed objects that enable flow control per traffic class on IEEE 802® point-to-point full duplex links. This is achieved by a mechanism similar to IEEE 802.3 Annex 31B PAUSE, but operating on individual priorities.
Keywords: flow control, IEEE 802.1Qbb, LANs, local area networks, MAC Bridges, priority, transparent bridging, VLANs

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

PDF: **ISBN 978-0-7381-6657-5** **STD97113**
Print: **ISBN 978-0-7381-6658-2** **STDPD97113**

IEEE Standards documents are developed within the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Association (IEEE-SA) Standards Board. The IEEE develops its standards through a consensus development process, approved by the American National Standards Institute, which brings together volunteers representing varied viewpoints and interests to achieve the final product. Volunteers are not necessarily members of the Institute and serve without compensation. While the IEEE administers the process and establishes rules to promote fairness in the consensus development process, the IEEE does not independently evaluate, test, or verify the accuracy of any of the information contained in its standards.

Use of an IEEE Standard is wholly voluntary. The IEEE disclaims liability for any personal injury, property or other damage, of any nature whatsoever, whether special, indirect, consequential, or compensatory, directly or indirectly resulting from the publication, use of, or reliance upon this, or any other IEEE Standard document.

The IEEE does not warrant or represent the accuracy or content of the material contained herein, and expressly disclaims any express or implied warranty, including any implied warranty of merchantability or fitness for a specific purpose, or that the use of the material contained herein is free from patent infringement. IEEE Standards documents are supplied "**AS IS**."

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. Every IEEE Standard is subjected to review at least every five years for revision or reaffirmation, or every ten years for stabilization. When a document is more than five years old and has not been reaffirmed, or more than ten years old and has not been stabilized, 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 publishing and making this document available, the IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity. Nor is the IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing this, and any other IEEE Standards document, should rely upon his or her 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.

Interpretations: Occasionally questions may arise regarding the meaning of portions of standards as they relate to specific applications. When the need for interpretations is brought to the attention of IEEE, the Institute will initiate action to prepare appropriate responses. Since IEEE Standards represent a consensus of concerned interests, it is important to ensure that any interpretation has also received the concurrence of a balance of interests. For this reason, IEEE and the members of its societies and Standards Coordinating Committees are not able to provide an instant response to interpretation requests except in those cases where the matter has previously received formal consideration. A statement, written or oral, that is not processed in accordance with the IEEE-SA Standards Board Operations Manual shall not be considered the official position of IEEE or any of its committees and shall not be considered to be, nor be relied upon as, a formal interpretation of the IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position, explanation, or interpretation of the IEEE.

Comments for revision of IEEE Standards are welcome from any interested party, regardless of membership affiliation with IEEE. Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Recommendations to change the status of a stabilized standard should include a rationale as to why a revision or withdrawal is required. Comments on standards and requests for interpretations should be addressed to:

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

Introduction

This introduction is not part of IEEE Std 802.1Qbb-2011, IEEE Standard for Local and metropolitan area networks—Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—Amendment 17: Priority-based Flow Control.

This amendment to IEEE Std 802.1Q-2011 provides Priority-based Flow Control capabilities useful to Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks to enable flow control per traffic class on IEEE 802® point-to-point full duplex links.

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

Notice to users

Laws and regulations

Users of these documents should consult all applicable laws and regulations. Compliance with the provisions of this standard does not imply 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.

Copyrights

Updating of IEEE documents

Users of IEEE standards 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. 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 the IEEE Standards Association website at

http://ieeexplore.ieee.org/xpl/standards.jsp, or contact the IEEE at the address listed previously. For more information about the IEEE Standards Association or the IEEE standards development process, visit the IEEE-SA website at http://standards.ieee.org.

Errata

Errata, if any, for this and all other standards can be accessed at the following URL: http://standards.ieee.org/findstds/errata/index.html. Users are encouraged to check this URL for errata periodically.

Interpretations

Current interpretations can be accessed at the following URL: http://standards.ieee.org/findstds/interps/index.html.

Patents

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 with respect to the existence or validity of any patent rights in connection therewith. The IEEE shall not be responsible for identifying patents or patent applications for which a license may be required to implement an IEEE standard or for conducting inquiries into the legal validity or scope of those patents that are brought to its attention. A patent holder or patent applicant has filed a statement of assurance that it will grant licenses under these rights without compensation or under reasonable rates and nondiscriminatory, reasonable terms and conditions to applicants desiring to obtain such licenses. The IEEE makes no representation as to the reasonableness of rates, terms, and conditions of the license agreements offered by patent holders or patent applicants. Further information may be obtained from the IEEE Standards Department.

Participants

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

Anthony Jeffree, *Chair*
Paul Congdon, *Vice Chair*
Claudio DeSanti, *Editor*
Patricia Thaler, *Chair, Data Center Bridging Task Group*

Zehavit Alon
Yafan An
Ting Ao
Peter Ashwood-Smith
Christian Boiger
Paul Bottorff
Rudolf Brandner
Craig Carlson
Claudio DeSanti
Zhemin Ding
Donald Eastlake
Janos Farkas
Donald Fedyk
Norman Finn
Ilango Ganga
Geoffrey Garner
Anoop Ghanwani
Eric Gray
Yingjie Gu
Craig Gunther
Hitoshi Hayakawa
Hal Keen
Yongbum Kim
Philippe Klein
Oliver Kleineberg
Michael Krause
Li Li
Jeff Lynch
John Messenger
John Morris
Eric Multanen
David Olsen
Donald Pannell
Glenn Parsons
Joseph Pelissier
Rene Raeber
Karen Randall
Dan Romascanu
Jessy Rouyer
Panagiotis Saltsidis
Michael Seaman
Rakesh Sharma
Kevin B. Stanton
Robert Sultan
Michael Teener
Patricia Thaler
Chait Tumuluri
Maarten Vissers

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

Thomas Alexander
Danilo Antonelli
Madhusudan Banavara
Hugh Barrass
Tomo Bogataj
Nancy Bravin
Edward Carley
James Carlo
Juan Carreon
Keith Chow
Charles Cook
Claudio DeSanti
Wael Diab
Thomas Dineen
Carlo Donati
Sourav Dutta
Frank Effenberger
Jose Dominic Espejo
C. Fitzgerald
Yukihiro Fujimoto
Ilango Ganga
Ignacio Marin Garcia
Devon Gayle
Anoop Ghanwani
Randall Groves
C. Guy
Marek Hajduczenia
John Hawkins
Ian Hilliard
Ryan Hirth
Akio Iso
Atsushi Ito
Raj Jain
Anthony Jeffree
Vincent Jones
Shinkyo Kaku
Piotr Karocki
Stuart J. Kerry
Max Kicherer
Yongbum Kim
Bruce Kraemer
Glen Kramer
Bruce Kwan
Juan L. Lazaro
Brian L'Ecuyer
Li Li
Greg Luri
Elvis Maculuba
Mark Maloney
Arthur Marris
Peter Martini
Sean Maschue
Jonathon McLendon
Jose Morales
Joseph Moran
Shimon Muller
Michael S. Newman
Glenn Parsons
Mark Pilip
Subburajan Ponnuswamy
Maximilian Riegel
Robert Robinson
Jessy Rouyer
Randall Safier
Peter Saunderson
Bartien Sayogo
Rich Seifert
Gil Shultz
Kapil Sood
Manikantan Srinivasan
Thomas Starai
Walter Struppler
Joseph Tardo
Michael Johas Teener
Patricia Thaler
David Thompson
Mark-Rene Uchida
Prabodh Varshney
Peter Yan
Oren Yuen
Wenhao Zhu

When the IEEE-SA Standards Board approved this amendment on 16 June 2011, it had the following membership:

Richard H. Hulett, *Chair*
John Kulick, *Vice Chair*
Robert M. Grow, *Past President*
Judith Gorman, *Secretary*

Masayuki Ariyoshi
William Bartley
Ted Burse
Clint Chaplin
Wael Diab
Jean-Philippe Faure
Alexander Gelman
Paul Houzé
Jim Hughes
Joseph L. Koepfinger*
David J. Law
Thomas Lee
Hung Ling
Oleg Logvinov
Ted Olsen
Gary Robinson
Jon Walter Rosdahl
Sam Sciacca
Mike Seavey
Curtis Siller
Phil Winston
Howard Wolfman
Don Wright

*Member Emeritus

Also included are the following nonvoting IEEE-SA Standards Board liaisons:

Satish K. Aggarwal, *NRC Representative*
Richard DeBlasio, *DOE Representative*
Michael Janezic, *NIST Representative*

Michelle Turner
IEEE Standards Program Manager, Document Development

Kathryn Bennett
IEEE Standards Program Manager, Technical Program Development

Contents

1. Overview 2

1.3 Introduction 2

2. References 3

3. Definitions 4

4. Abbreviations 5

5. Conformance 6

5.4 VLAN-aware Bridge component requirements 6
5.4.1 VLAN-aware Bridge component options 6
5.10 Provider Bridge conformance 6
5.11 System requirements for Priority-based Flow Control 6

6. Support of the MAC Service 7

6.6 Internal Sublayer Service 7
6.6.4 Stream Reservation Protocol (SRP) Domain status parameters 7
6.6.5 Control primitives and parameters 7
6.7 Support of the Internal Sublayer Service by specific MAC procedures 7
6.7.1 Support of the Internal Sublayer Service by IEEE Std 802.3 (CSMA/CD) 7

8. Principles of bridge operation 8

8.6 The Forwarding Process 8
8.6.8 Transmission selection 8

12. Bridge management 9

12.22 SRP entities 9
12.23 Priority-based Flow Control objects 9

17. Management Information Base (MIB) 10

17.2 Structure of the MIB 10
17.2.16 Structure of the MIRP MIB 10
17.2.17 Structure of the Priority-based Flow Control MIB 10
17.3 Relationship to other MIB modules 10
17.3.16 Relationship of the IEEE8021-MIRP-MIB to other MIB modules 10
17.3.17 Relationship of the Priority-based Flow Control MIB to other MIB modules 10
17.4 Security considerations 11
17.4.16 Security considerations of the IEEE8021-MIRP-MIB 11
17.4.17 Security considerations for the Priority-based Flow Control MIB 11
17.7 MIB modules 11
17.7.16 MIRP MIB module 11
17.7.17 Priority-based Flow Control MIB module 11

36. Priority-based Flow Control 15

36.1 Priority-based Flow Control operation 15
36.1.1 Overview 15
36.1.2 PFC Primitives 16

36.1.3 Detailed specification of PFC operation 17
36.2 PFC aware system queue functions 18
36.2.1 PFC Initiator 19
36.2.2 PFC Receiver 19

Annex A (normative) PICS Proforma 21

Annex N (normative) Support for PFC in link layers without MAC Control 23

Annex O (informative) Buffer requirements for Priority-based Flow Control 24

Figures

Figure 36-1— PFC peering 15
Figure 36-2— PFC Receiver state diagram for priority n 17
Figure 36-3— PFC aware system queue functions 19
Figure 36-4— PFC aware system queue functions with link aggregation 20
Figure N-1— PFC PDU format 23
Figure O-1— PFC delays 24
Figure O-2— Delay model 25
Figure O-3— Worst-case delay 26

Tables

Table 12-15— Priority-based Flow Control objects..... 9
Table 17-1— Structure of the MIB Modules..... 10
Table 17-23— Variables, managed object tables, and MIB objects..... 10
Table O-1— IEEE 802.3 Interface Delays 27

IEEE Standard for
Local and metropolitan area networks—

Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—

Amendment 17: Priority-based Flow Control

This amendment to IEEE Std 802.1Q™-2011 provides capabilities for enabling flow control per traffic class on IEEE 802® point-to-point full duplex links. Changes are applied to the base text of IEEE Std 802.1Q-2011 as amended by IEEE Std 802.1Qbe™-2011 and IEEE Std 802.1Qbc™-2011.

NOTE—The editing instructions contained in this amendment define how to merge the material contained therein into the existing base standard and its amendments to form the comprehensive standard. Text shown in bold italics in this amendment defines the editing instructions necessary to changes to this base text. Three editing instructions are used: ***change***, ***delete***, and ***insert***. ***Change*** is used to make a change to existing material. The editing instruction specifies the location of the change and describes what is being changed. Changes to existing text may be clarified using ~~strikeout~~ markings to indicate removal of old material and underscore markings to indicate addition of new material. ***Delete*** removes existing material. ***Insert*** adds new material without changing the existing material. Insertions may require renumbering. If so, renumbering instructions are given in the editing instruction. Editorial notes will not be carried over into future editions of IEEE Std 802.1Q.[1]

[1]Notes in text, tables, and figures are given for information only and do not contain requirements needed to implement the standard.

1. Overview

Insert the following paragraph at the end of Clause 1:

This standard specifies protocols, procedures and managed objects that enable Priority-based Flow Control (PFC) on IEEE 802 point-to-point full duplex links in Data Center Bridging (DCB) networks (bridges and end stations) that are characterized by limited bandwidth delay product and limited hop count. PFC is intended to eliminate frame loss due to congestion on a link; this is achieved by a mechanism similar to IEEE 802.3 Annex 31B PAUSE, but operating on individual priorities. This mechanism, in conjunction with other DCB technologies, enables support for higher layer protocols that are highly loss sensitive while not affecting the operation of traditional LAN protocols utilizing other priorities. Operation of Priority-based Flow Control is limited to a data center environment (i.e., a domain controlled by the Data Center Bridging eXchange protocol, DCBX).

1.3 Introduction

Insert the following text at end of 1.3:

This standard specifies protocols, procedures, and managed objects to support Priority-based Flow Control. These allow a Virtual Bridged Local Area 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

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

2. References

Insert the following references into Clause 2 in alphanumeric order:

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

IEEE Std 802.1Qaz™-2011, IEEE Standard for Local and metropolitan area networks—Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—Amendment 18: Enhanced Transmission Selection for Bandwidth Sharing Between Traffic Classes.

IEEE Std 802.3bd™-2011, IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications—Amendment 8: MAC Control Frame for Priority-based Flow Control.

3. Definitions

Insert the following definitions into Clause 3 in alphabetical order, number them appropriately, and renumber the remaining definitions in Clause 3 accordingly:

3.x bit time: The duration of one bit as transferred to and from the Media Access Control (MAC). The bit time is the reciprocal of the bit rate.

3.x Paused state: A state of a queue in which the transmission selection entity does not select frames from the queue.

3.x Data center environment: A domain controlled by the Data Center Bridging eXchange (DCBX) Protocol.

NOTE—See Clause 38 in IEEE Std 802.1Qaz-2011.[2]

[2]Information on references can be found in Clause 2.

4. Abbreviations

Insert the following abbreviations into Clause 4 in alphabetical order:

DCBX Data Center Bridging eXchange protocol

PFC Priority-based Flow Control

TLV Type, Length, Value

5. Conformance

5.4 VLAN-aware Bridge component requirements

5.4.1 VLAN-aware Bridge component options

Insert the following list item after item d) in 5.4.1, and reletter the remaining list items in 5.4.1 accordingly:

e) Support Priority-based Flow Control (5.11);

5.10 Provider Bridge conformance

Insert the following subclause, 5.11, after 5.10.2, and renumber the remaining subclauses in Clause 5 accordingly:

5.11 System requirements for Priority-based Flow Control

A system that conforms to the provisions of this standard for Priority-based Flow Control (PFC) (see Clause 36) shall:

a) Support, on one or more ports, enabling PFC on at least one priority (see 36.1.2);
b) Support, for each PFC Priority, processing PFC M_CONTROL.requests (see 36.1.3.1);
c) Support, for each PFC Priority, processing PFC M_CONTROL.indications (see 36.1.3.2);
d) Abide by the PFC delay constraints (see 36.1.3.3);
e) Provide PFC aware system queue functions (see 36.2); and
f) Enable use of PFC only in a domain controlled by the DCBX protocol (see Clause 38 in IEEE Std 802.1Qaz-2011).

A system that conforms to the provisions of this standard for Priority-based Flow Control may:

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

6. Support of the MAC Service

6.6 Internal Sublayer Service

6.6.4 Stream Reservation Protocol (SRP) Domain status parameters

Insert the following subclause, 6.6.5, after 6.6.4:

6.6.5 Control primitives and parameters

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

The M_CONTROL.request primitive has the form:

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

The M_CONTROL.indication primitive has the form:

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

See 36.1.2 for a description of the M_CONTROL parameters used for Priority-based Flow Control.

6.7 Support of the Internal Sublayer Service by specific MAC procedures

6.7.1 Support of the Internal Sublayer Service by IEEE Std 802.3 (CSMA/CD)

Insert the following paragraph at the end of 6.7.1:

An M_CONTROL.request primitive is mapped to an IEEE 802.3 MA_CONTROL.request primitive having the same parameters. An IEEE 802.3 MA_CONTROL.indication primitive is mapped to an M_CONTROL.indication primitive having the same parameters.

8. Principles of bridge operation

8.6 The Forwarding Process

8.6.8 Transmission selection

Insert the following text after item b) of 8.6.8:

In a port of a Bridge or station that supports PFC, a frame of priority n is not available for transmission if that priority is paused (i.e., if Priority_Paused[n] is TRUE (see 36.1.3.2)) on that port. When Transmission Selection is running above Link Aggregation, a frame of priority n is not available for transmission if that priority is paused on the physical port to which the frame is to be distributed.

NOTE 1—Two or more priorities can be combined in a single queue. In this case if one or more of the priorities in the queue are paused, it is possible for frames in that queue not belonging to the paused priority to not be scheduled for transmission.

NOTE 2—Mixing PFC and non-PFC priorities in the same queue results in non-PFC traffic being paused causing congestion spreading, and therefore is not recommended.

8.6.8.2 Credit-based shaper algorithm

Insert the following paragraph at the end of 8.6.8.2:

Traffic classes using the credit-based shaper algorithm shall not use PFC and shall ignore the setting of the bits related to such classes in the PFC Enable bit vector (see 38.5.4.6 in IEEE Std 802.1Qaz-2011).

12. Bridge management

12.22 SRP entities

Insert the following subclause, 12.23, after 12.22.5:

12.23 Priority-based Flow Control objects

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

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

Table 12-15 shows the format and applicability of these objects.

Table 12-15—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

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

17. Management Information Base (MIB)

17.2 Structure of the MIB

Insert the following row at the end of Table 17-1:

Table 17-1—Structure of the MIB Modules

Module	Subclause	Defining standard	Reference	Notes
IEEE8021-PFC-MIB	17.2.17	802.1Qbb	36	Initial version in 802.1Qbb

17.2.16 Structure of the MIRP MIB

Insert the following subclause, 17.2.17 (including Table 17-23), after 17.2.16, and renumber the subsequent tables in Clause 17 accordingly:

17.2.17 Structure of the Priority-based Flow Control MIB

Subclause 12.23 defines the information model associated with this standard in a protocol independent manner. Table 17-23 describes the relationship between the SMIv2 objects defined in the MIB module in 17.7.17 and the variables and managed objects defined in Clause 12 and Clause 36.

Table 17-23—Variables, managed object tables, and MIB objects

Variable	Reference	MIB object (17.7.17)
PFC Interface Table	17.7.17	**ieee8021PfcIfTable**
(AUGMENTS ifEntry)	—	—
PFCLinkDelayAllowance	12.23	ieee8021PfcLinkDelayAllowance
PFCRequests	12.23	ieee8021PfcRequests
PFCIndications	12.23	ieee8021PfcIndications

17.3 Relationship to other MIB modules

17.3.16 Relationship of the IEEE8021-MIRP-MIB to other MIB modules

Insert the following subclause, 17.3.17, after 17.3.16:

17.3.17 Relationship of the Priority-based Flow Control MIB to other MIB modules

Subclause 17.7.17 defines a Priority-based Flow Control MIB (PFC MIB) module. A system implementing the PFC MIB module in 17.7.17 shall also implement at least the System Group of the SNMPv2-MIB defined in IETF RFC 3418 and the Interfaces Group (the Interfaces MIB module, or IF-MIB) defined in IETF RFC 2863. The Interfaces Group has one conceptual row in a table for every interface in a system. Section 3.3 of IETF RFC 2863, the Interface MIB Evolution, defines hierarchical relationships among interfaces. IETF RFC 2863 also requires that any MIB module that is an adjunct of the Interface Group

clarify specific areas within the Interface MIB module. These areas were intentionally left vague in IETF RFC 2863 to avoid over constraining the MIB, thereby precluding management of certain media types. These areas are clarified in other clauses which define the MIB modules in this standard. Even if a system supports none of these, if it supports the PFC MIB module, and hence, the Interfaces Group, the clarifications from the other clauses shall be applied to the Interfaces Group. The relationship between IETF RFC 2863 and IETF RFC 3418 interfaces and ports is also described in previous subclauses of 17.3.

17.4 Security considerations

17.4.16 Security considerations of the IEEE8021-MIRP-MIB

Insert the following subclause, 17.4.17, after 17.4.16:

17.4.17 Security considerations for the Priority-based Flow Control MIB

One management object defined in the IEEE8021-PFC-MIB module has a MAXACCESS clause of read-write. Such object can be considered sensitive or vulnerable in some network environments. The support for SET operations in a nonsecure environment without proper protection can have a negative effect on network operations. The management object is:

PFCLinkDelayAllowance

Improper setting of this management object can result in improper network operations. If the value of this management object is too high, then PFC can be invoked excessively, negatively impacting the link bandwidth. If the value of this management object is too low, then PFC can be invoked too late and frame loss can occur.

17.7 MIB modules

17.7.16 MIRP MIB module

Insert the following subclause, 17.7.17, after 17.7.16:

17.7.17 Priority-based Flow Control MIB module

In the MIB definition below, if any discrepancy between the DESCRIPTION text and the corresponding definition in Clause 12 occur, the definition in Clause 12 takes precedence.

```
IEEE8021-PFC-MIB DEFINITIONS ::= BEGIN

-- ************************************************************
-- IEEE P802.1Qbb(TM) Priority-based Flow Control MIB
-- ************************************************************

IMPORTS
    MODULE-IDENTITY,
    OBJECT-TYPE,
    Counter32,
    Unsigned32                FROM SNMPv2-SMI    -- [RFC2578]
    MODULE-COMPLIANCE,
    OBJECT-GROUP              FROM SNMPv2-CONF   -- [RFC2580]
    ifEntry,
```

```
    ifGeneralInformationGroup
                            FROM IF-MIB        -- [RFC2863]
    systemGroup             FROM SNMPv2-MIB    -- [RFC3418]
   ;

ieee8021PFCMib MODULE-IDENTITY
    LAST-UPDATED "201002080000Z"    -- 02/08/2010 00:00GMT
    ORGANIZATION "IEEE 802.1 Working Group"
    CONTACT-INFO
       "WG-URL:   http://grouper.ieee.org/groups/802/1/index.html
        WG-EMail: stds-802-1@ieee.org

        Contact:  Claudio DeSanti

                  Cisco Systems
                  170 W. Tasman Drive
                  San Jose, CA 95134, USA

        E-mail:   cds@cisco.com"
   DESCRIPTION
     "Priority-based Flow Control module for managing IEEE 802.1Qbb"
   REVISION       "201002080000Z"    -- 02/08/2010 00:00GMT
   DESCRIPTION
      "Included in IEEE P802.1Qbb

        Copyright (C) IEEE."
    ::= { iso(1) org(3) ieee(111)
         standards-association-numbers-series-standards (2)
         lan-man-stds (802) ieee802dot1 (1) ieee802dot1mibs (1) 21 }


ieee8021PfcMIBObjects      OBJECT IDENTIFIER ::= { ieee8021PFCMib 1 }
ieee8021PfcConformance     OBJECT IDENTIFIER ::= { ieee8021PFCMib 2 }


ieee8021PfcIfTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF Ieee8021PfcIfEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
       "A table of PFC information for all interfaces of a system."
    REFERENCE
       "802.1Qbb clause 12.18"
    ::= { ieee8021PfcMIBObjects 1 }

ieee8021PfcIfEntry OBJECT-TYPE
    SYNTAX      Ieee8021PfcIfEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
       "Each entry contains information about
        the PFC function on a single interface."
    REFERENCE
       "802.1Qbb clause 12.18"
```



```
    AUGMENTS { ifEntry }
    ::= { ieee8021PfcIfTable 1 }

Ieee8021PfcIfEntry ::= SEQUENCE {
        ieee8021PfcLinkDelayAllowance       Unsigned32,
        ieee8021PfcRequests                 Counter32,
        ieee8021PfcIndications              Counter32
    }

ieee8021PfcLinkDelayAllowance    OBJECT-TYPE
    SYNTAX      Unsigned32
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "The allowance made for round-trip propagation delay
        of the link in bits.

        The value of this object MUST be retained across
        reinitializations of the management system."
    ::= { ieee8021PfcIfEntry 1 }

ieee8021PfcRequests    OBJECT-TYPE
    SYNTAX      Counter32
    UNITS       "Requests"
    MAX-ACCESS  read-only
    STATUS      current
    DESCRIPTION
        "A count of the invoked PFC M_CONTROL.request primitives.

         Discontinuities in the value of this counter can occur at
         re-initialization of the management system, and at other
         times as indicated by the value of
         ifCounterDiscontinuityTime."
    ::= { ieee8021PfcIfEntry 2 }

ieee8021PfcIndications    OBJECT-TYPE
    SYNTAX      Counter32
    UNITS       "Indications"
    MAX-ACCESS  read-only
    STATUS      current
    DESCRIPTION
        "A count of the received PFC M_CONTROL.indication primitives.

         Discontinuities in the value of this counter can occur at
         re-initialization of the management system, and at other
         times as indicated by the value of
         ifCounterDiscontinuityTime."
    ::= { ieee8021PfcIfEntry 3 }
```

```
-- ****************************************************************
-- IEEE 802.1Qbb MIB Module - Conformance Information
-- ****************************************************************

   ieee8021PfcCompliances
       OBJECT IDENTIFIER ::= { ieee8021PfcConformance 1 }
   ieee8021PfcGroups
       OBJECT IDENTIFIER ::= { ieee8021PfcConformance 2 }


-- ****************************************************************
-- Units of conformance
-- ****************************************************************

ieee8021PfcGlobalReqdGroup OBJECT-GROUP
    OBJECTS {
      ieee8021PfcLinkDelayAllowance,
      ieee8021PfcRequests,
      ieee8021PfcIndications
    }
    STATUS      current
    DESCRIPTION
       "Objects in the global required group."
    ::= { ieee8021PfcGroups 1 }


-- ****************************************************************
-- MIB Module Compliance statements
-- ****************************************************************

ieee8021PfcCompliance MODULE-COMPLIANCE
    STATUS      current
    DESCRIPTION
       "The compliance statement for support by a system of
        the IEEE8021-PFC-MIB module."

    MODULE SNMPv2-MIB -- The SNMPv2-MIB, RFC 3418
        MANDATORY-GROUPS {
            systemGroup
        }

    MODULE IF-MIB -- The interfaces MIB, RFC 2863
        MANDATORY-GROUPS {
            ifGeneralInformationGroup
        }

    MODULE
        MANDATORY-GROUPS {
            ieee8021PfcGlobalReqdGroup
        }
    ::= { ieee8021PfcCompliances 1 }


END
```


Insert the following text, Clause 36, after 35.2.7:

36. Priority-based Flow Control

This clause specifies the operation of Priority-based Flow Control (PFC) (see 36.1) and the architecture of Priority-based Flow Control in a PFC aware system (see 36.2).

The models of operation in this clause provide a basis for specifying the externally observable behavior of Priority-based Flow Control, 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 Priority-based Flow Control operation

36.1.1 Overview

Operation of Priority-based Flow Control is limited to a data center environment. PFC enables to not discard frames due to congestion for protocols that require this property. However, PFC can cause congestion spreading behavior therefore it is intended for use on networks of limited extent. When PFC is used, deployment of Congestion Notification (see clause 30) can reduce the frequency with which PFC is invoked.

PFC is a function defined only for a pair of full duplex MACs (e.g., 802.3 MACs operating in point-to-point full duplex mode) connected by one point-to-point link. Use of PFC on shared media such as EPON is out of the scope of this standard. Figure 36-1 shows an example of PFC peering when 802.3 point-to-point full duplex MACs are used.

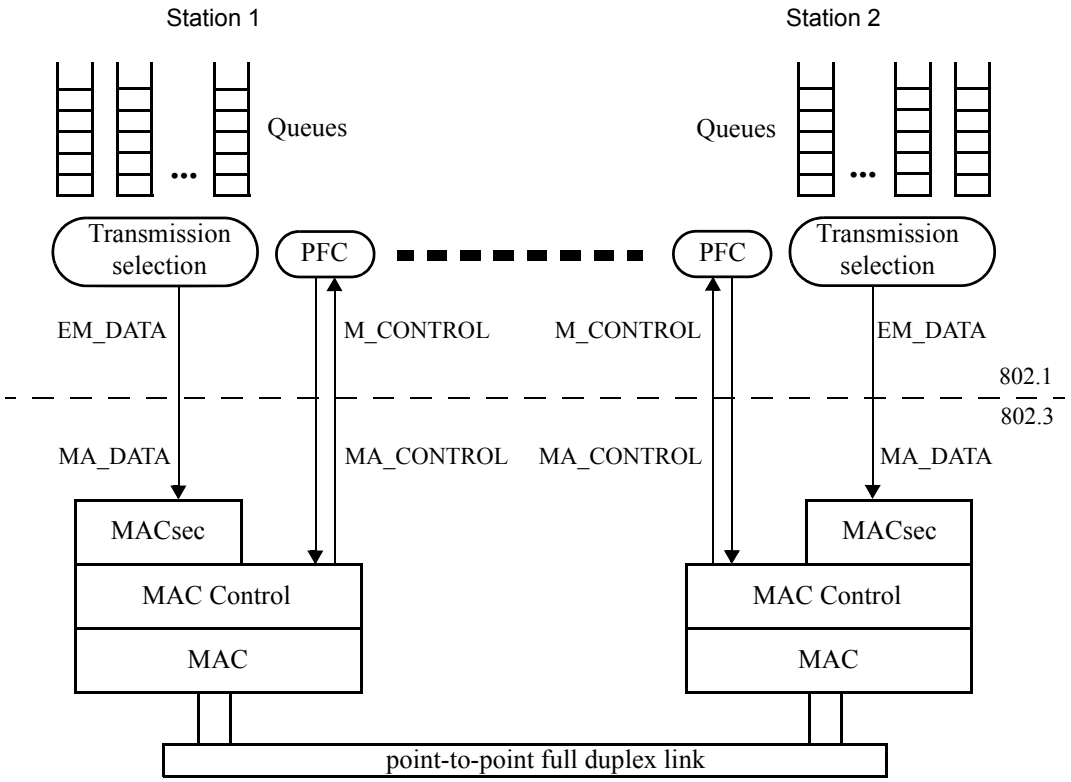

Figure 36-1—PFC peering

PFC allows link flow control to be performed on a per-priority basis. In particular, PFC is used to inhibit transmission of data frames associated with one or more priorities for a specified period of time. PFC can be enabled for some priorities on the link and disabled for others.

A VLAN unaware end station can use PFC by sending traffic as priority-tagged and by ignoring the VLAN ID in received frames. Given that BPDUs, for example, are sent untagged and can bypass the output queues, it is strongly recommended for the default priority of a port to not have PFC enabled.

NOTE—The LLC-SAP of a bridge port can host a management protocol stack that uses PFC-enabled priorities, and these management frames can bypass the output queues. In this situation PFC can fail to provide insurance against these frames overflowing the buffer in the remote station of the link.

36.1.2 PFC primitives

PFC is invoked through the M_CONTROL PFC primitives (see 6.6.5). A system client wishing to inhibit transmission of data frames on certain priorities from the remote system on the link generates an M_CONTROL.request primitive specifying:

a) The globally assigned 48-bit multicast address 01-80-C2-00-00-01;
b) The PFC opcode (i.e., 01-01); and
c) A request_operand_list with two operands indicating respectively the set of priorities addressed and the lengths of time for which it wishes to inhibit data frame transmission of the corresponding priorities.

NOTE—By definition, a point-to-point full duplex link comprises exactly two stations, thus there is no ambiguity regarding the destination station's identity. The use of a well-known multicast address does not require a station to know, and maintain knowledge of, the individual 48-bit address of the other station.

Over an IEEE 802.3 link layer, when PFC is enabled on a port for at least one priority, the IEEE 802.3 Annex 31B PAUSE mechanism is not used for that port (see IEEE Std 802.3 Annex 31D[3]).

As a result of the processing of the PFC M_CONTROL.request, the peering PFC station receives a PFC M_CONTROL.indication primitive.

The parameters of the PFC M_CONTROL.indication are:

d) The PFC opcode (i.e., 01-01); and
e) A indication_operand_list with two operands indicating respectively the set of priorities addressed and the lengths of time for which data frame transmission of the corresponding priorities has to be inhibited.

The request_operand_list of a PFC M_CONTROL.request and the indication_operand_list of a PFC M_CONTROL.indication are composed of the following operands:

f) 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 least significant bit) to e[7] (the most significant bit). Bit e[n] refers to priority n. For each e[n] bit set to one, the corresponding time[n] value is valid. For each e[n] bit set to zero, the corresponding time[n] value is invalid.
g) 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

[3]At the time of publication of this standard, IEEE Std 802.3 Annex 31D was contained in IEEE Std 802.3bd-2011.

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.

36.1.3 Detailed specification of PFC operation

36.1.3.1 Processing PFC M_CONTROL.requests

Invoking the PFC M_CONTROL.request results in the invocation of the appropriate link layer service request. For IEEE 802.3 link layers the PFC M_CONTROL.request is mapped to a PFC MA_CONTROL.request (see 6.7.1). If PFC is not enabled for priority n, then PFC requests with e[n] set to one and time[n] different than zero (see 36.1.2) should not be generated.

NOTE—In the 802.1Q architecture frames coming from the LLC, including BPDUs, bypass the priority queues and therefore are not subject to PFC. However, in some implementations frames coming from the LLC can pass through the priority queues. In this case, it is not recommended to enable PFC for the priority to which BPDUs are assigned (usually priority 7).

36.1.3.2 Processing PFC M_CONTROL.indications

The PFC Receiver entity (see 36.2.2) 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 36-2 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.

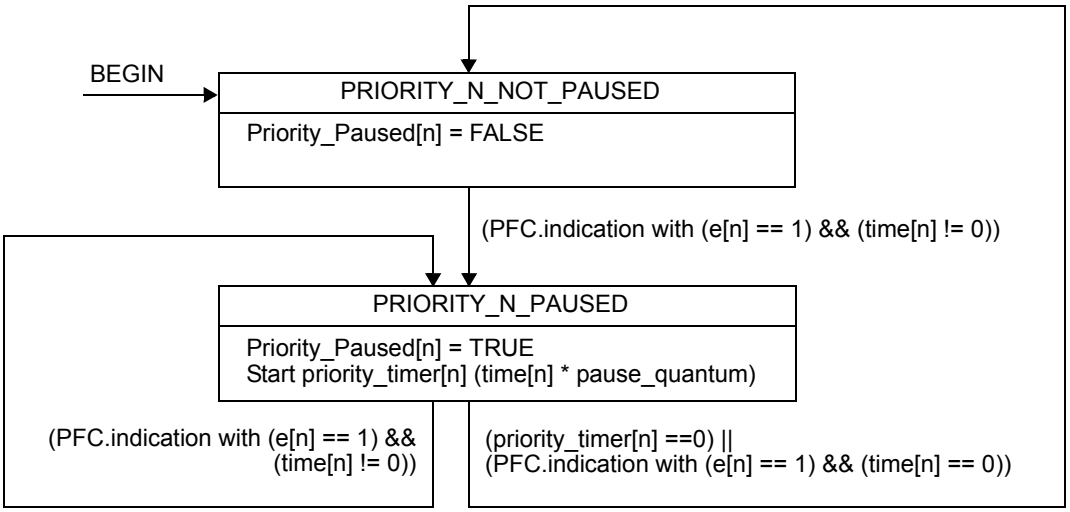

Figure 36-2—PFC Receiver state diagram for priority n

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.

36.1.3.3 Timing considerations

For effective flow control on a point-to-point full duplex link, it is necessary to place an upper bound on the length of time that a device can transmit data frames after receiving a PFC M_CONTROL.indication with e[n] set to one in the priority_enable_vector and a nonzero time[n] in the time_vector operands.

If MACsec is not supported, a 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., 6 144 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.

If MACsec is used, a queue shall go into paused state in no more than 614.4 ns + ‘SecY transmit delay’ (see Table 10-1 of IEEE Std 802.1AE) since the reception of a PFC M_CONTROL.indication that paused that priority. The ‘SecY transmit delay’ is defined as the wire transmit time for a maximum sized MPDU + 4 times the wire transmit time for 64 octet MPDUs. For a 2 000 bytes frame the ‘SecY transmit delay’ is 8*(2 000+20) + 8*4*(64+12+4+20) = 19 360 bit times.

NOTE—19 360 bit times is an appropriate value for ‘SecY transmit delay’ for speeds up to 10 Gb/s. Support for the speeds of 40 Gb/s and 100 Gb/s can require a higher value.

If MACsec is supported but not used, the delay computation has to take into account the MACsec Bypass Capability (MBC) bit in the PFC configuration TLV of DCBX (see IEEE Std 802.1Qaz subclause 38.5.4), that indicates if the link peer needs the extra time for MACsec. If the MBC bit is set to zero, the maximum PFC delay is 614.4 ns. If the MBC bit is set to one, the maximum PFC delay is 614.4 ns + ‘SecY transmit delay’.

NOTE—In addition to the above delays, system designers should take into account the delay of the PHY and of the link segment when designing devices that implement the PFC operation to ensure frames are not lost due to congestion (see Annex O (informative) for additional discussion on this topic).

36.2 PFC aware system queue functions

Figure 36-3 illustrates the architecture of the queue functions of a PFC aware system when link aggregation is not used. These functions offer a service to higher layers that utilizes a single instance of the ISS or EISS to connect to the lower layers. In Figure 36-3, two major blocks are outlined with dotted boundaries:

a) The PFC Initiator block, in the right of Figure 36-3 (see 36.2.1); and
b) The outbound queue block, in the left of Figure 36-3 (see Figure 22-2).

The remaining entities illustrated in Figure 36-3, other than the PFC Receiver entity, are part of the 802.1 architecture and are not further discussed here.

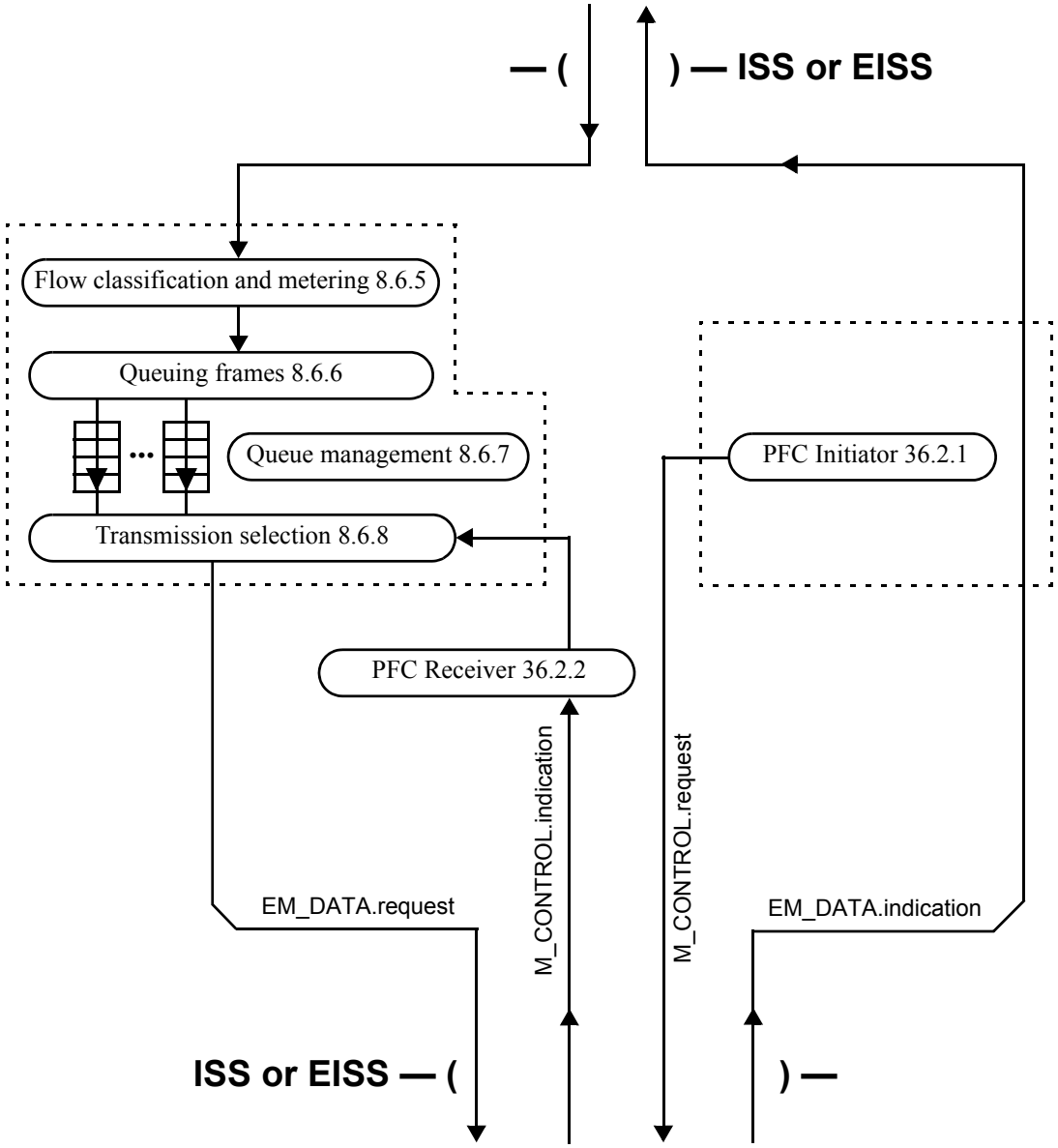

Figure 36-3—PFC aware system queue functions

36.2.1 PFC Initiator

The PFC Initiator entity generates M_CONTROL PFC requests using the M_CONTROL.request primitive (see 36.1.3.1) when appropriate (e.g., when an input buffer reaches a certain threshold).

36.2.2 PFC Receiver

The PFC Receiver entity processes the M_CONTROL.indication primitives as specified in 36.1.3.2. In addition, 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.

The PFC Receiver entity acts per physical port. When Transmission Selection is running above Link Aggregation, each PFC Receiver entity processes the M_CONTROL.indication primitives as specified in 36.1.3.2, and maintains and makes available to Transmission Selection the vector of the Priority_Paused[n] variables, indicating the state of each of the eight priorities of that physical link, as shown in Figure 36-4.

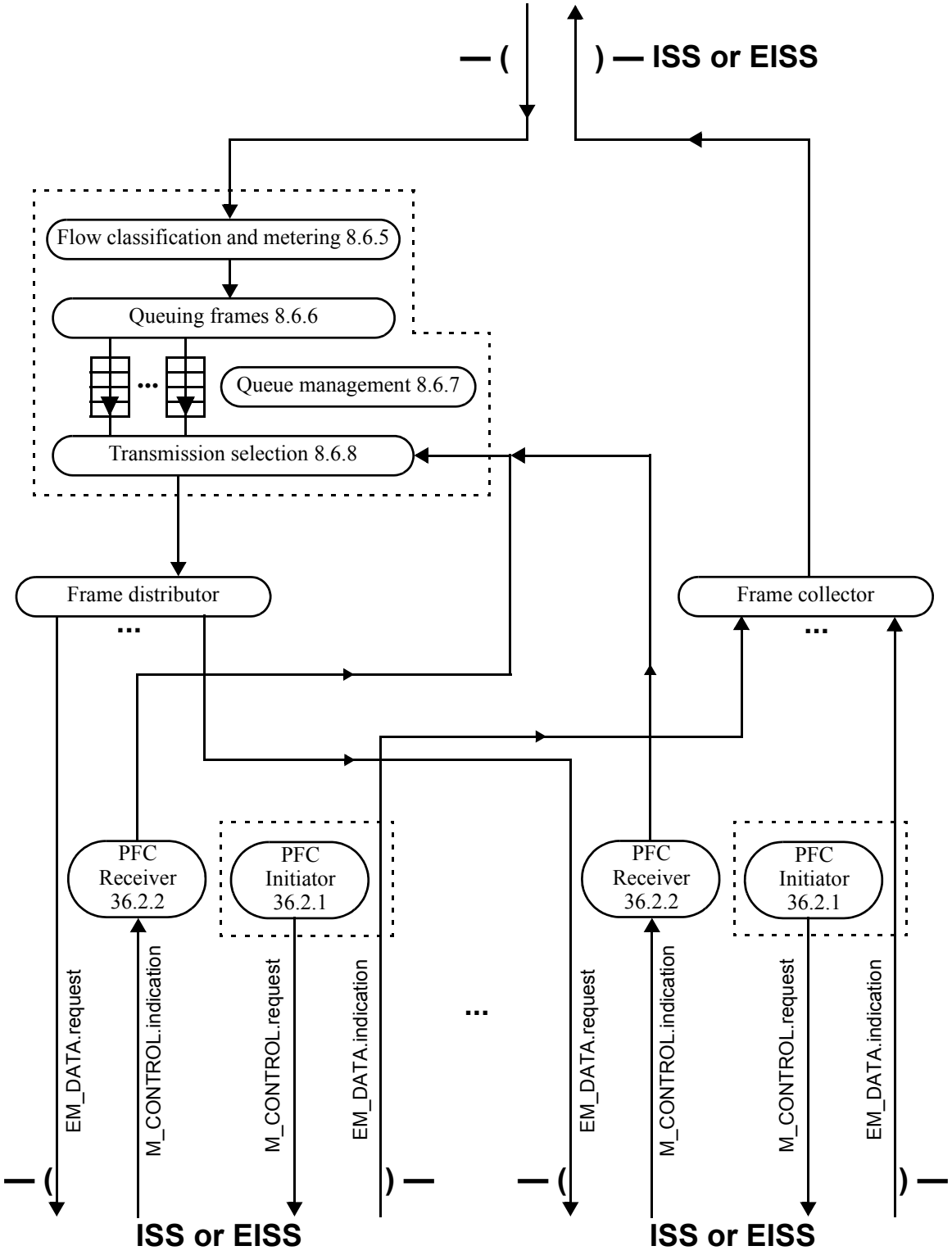

Figure 36-4—PFC aware system queue functions with link aggregation

Annex A

(normative)

PICS proforma—Bridge implementations[4]

A.5 Major capabilities

Insert the following row at the end of A.5:

Item	Feature	Status	References	Support
PFC	Is Priority-based Flow Control implemented?	O	5.11, 36	Yes [ ] No [ ]

A.14 Bridge Management

Insert the following row at the end of A.14:

Item	Feature	Status	References	Support
MGT-213	Priority-based Flow Control entities	PFC: O	12.23	Yes [ ] No [ ]

A.24 Management Information Base (MIB)

Insert the following row at the end of A.24:

Item	Feature	Status	References	Support
MIB-36	Is the IEEE8021-PFC-MIB module fully supported (per its MODULE-COMPLIANCE)?	PFC AND MIB: O	17.7.17	Yes [ ] No [ ]

[4]*Copyright release for PICS proformas:* Users of this standard may freely reproduce the PICS proforma in this annex so that it can be used for its intended purpose and may further publish the completed PICS.

A.32 MIRP

Insert the following subclause, A.33, after A.32:

A.33 Priority-based Flow Control

Item	Feature	Status	References	Support
PFC-1	Enabling PFC on at least one priority	PFC: M	36.1.2	Yes [ ]
PFC-2	Processing PFC Requests	PFC: M	36.1.3.1	Yes [ ]
PFC-3	Processing PFC Indications	PFC: M	36.1.3.2	Yes [ ]
PFC-4	PFC delay constraints	PFC: M	36.1.3.3	Yes [ ]
PFC-5	PFC aware system queue functions	PFC: M	36.2	Yes [ ]
PFC-6	DCBX	PFC: M	5.11	Yes [ ]
PFC-7	Enabling PFC on up to eight priorities	PFC: O	36.1.2	Yes [ ] No [ ]
PFC-8	PFC not enabled for traffic classes using the credit-based shaper algorithm	PFC: M	8.6.8.2	Yes [ ]

Insert the following annexes, Annex N and Annex O, after Annex M:

Annex N

(normative)

Support for PFC in link layers without MAC Control

N.1 Overview

Priority-based Flow Control is a function defined for only point-to-point full duplex links in terms of the M_Control primitives (see 6.6.5). For IEEE 802.3 link layers the M_CONTROL primitives are mapped into the MAC Control MA_CONTROL primitives (see 6.7.1), that use the PDU format defined in IEEE Std 802.3 Annex 31D[5]. Other link layers supporting point-to-point full duplex operations need to define their mapping of the M_CONTROL primitives. This annex describes a PDU format suitable to support PFC.

N.2 PFC PDU format

Figure N-1 shows a PDU format suitable to support PFC.

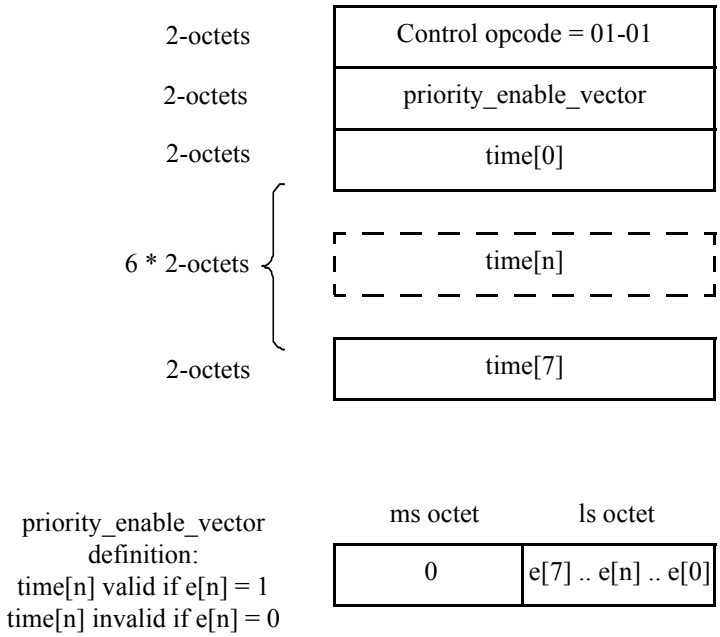

Figure N-1—PFC PDU format

The Control opcode field contains a 2-octet operation code indicating the Control function.

The remaining fields contain the parameters defined in 36.1.2.

[5]At the time of publication of this standard, IEEE Std 802.3 Annex 31D was contained in IEEE Std 802.3bd-2011.

Annex O

(informative)

Buffer requirements for Priority-based Flow Control

O.1 Overview

To assure 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 precise calculation of this buffer requirement is highly implementation dependent. This annex provides an example of how it can be calculated based on a hypothetical delay model. Setting the PFCLinkDelayAllowance (see 12.23) to less than the round-trip delay value can result in frames loss.

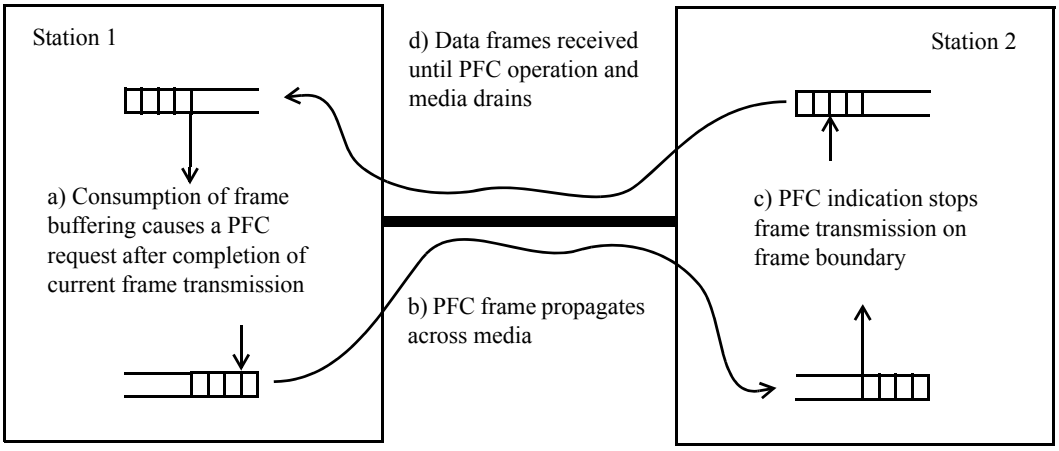

Figure O-1—PFC delays

Figure O-1 provides an high level view of the various delays to consider, that include:

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; and
d) Propagation delay across the media on the return path.

O.2 Delay model

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

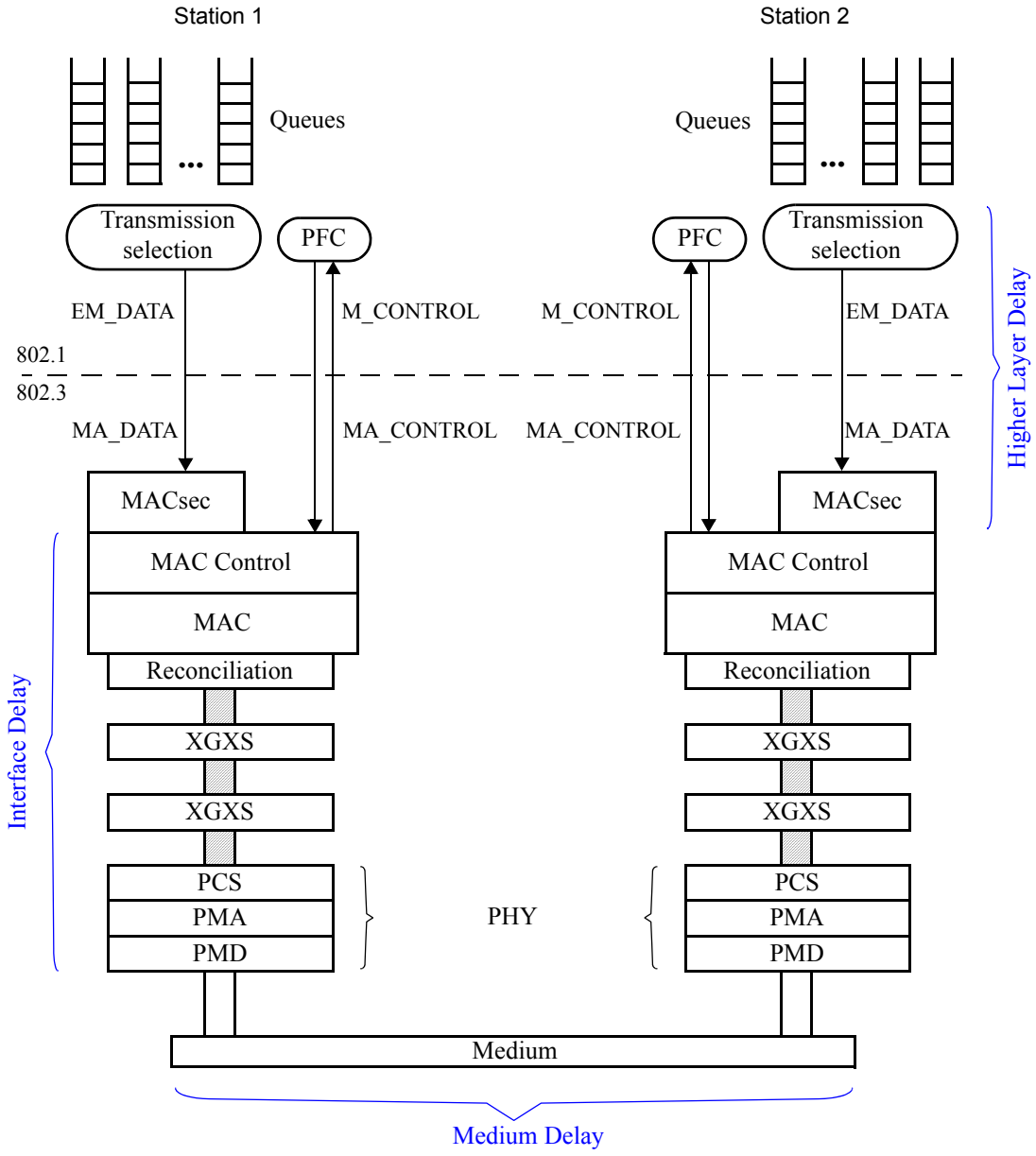

Figure O-2—Delay model

The main delay components shown in Figure O-2 are:

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. Interface Delay is 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.

Figure O-3 shows a possible worst case delay example.

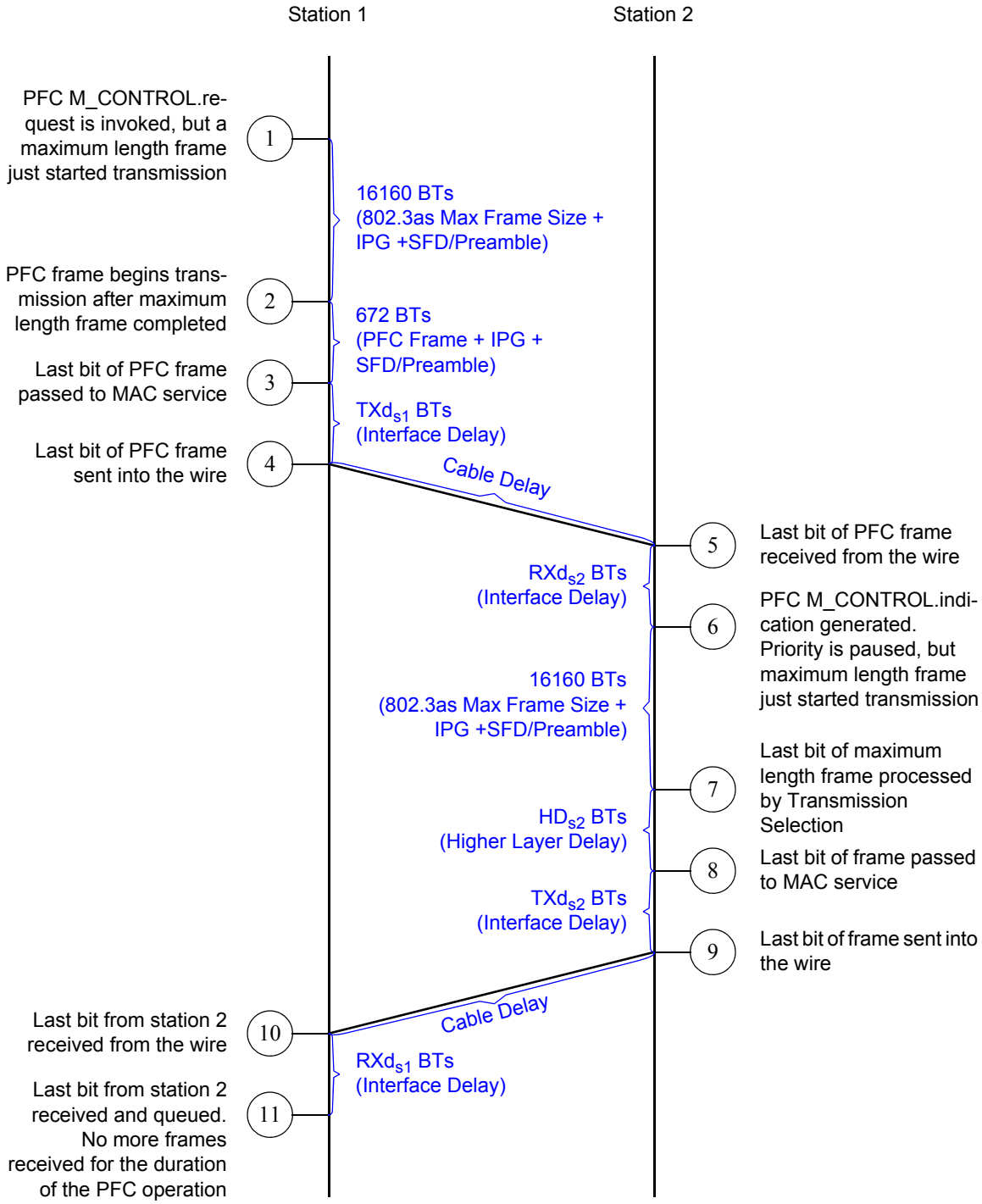

Figure O-3—Worst-case delay

The total Delay Value (DV) is the sum of all delays shown in Figure O-3:

DV = 2*(Max Frame) + (PFC Frame) + 2*(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*(Max Frame) + (PFC Frame) + 2*(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*(Max Frame) + (PFC Frame) + 2*(Cable Delay) + 2*ID + HD_{s2}

O.3 Interface Delay

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

Table O-1—IEEE 802.3 Interface Delays

Sublayer	Maximum RTT (bit times)	Maximum RTT (pause quanta)	Reference (subclause of 802.3)
10G MAC Control, MAC, and RS	8 192	16	46.1.4
XGXS and XAUI	2 048	4	48.5
10GBASE-X PCS	2 048	4	49.2.15
10GBASE-R PCS	3 584	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

O.4 Cable Delay

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

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

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

O.5 Higher Layer 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 6 144 bit times at the speed of 10Gb/s.

O.6 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:

— 802.3as Maximum frame size: 2 000 octets, 16 160 bit times;
— PFC frame size: 64 octets, 672 bit times;
— XGMII MAC/RS and XAUI interface: 8 192 + 2 * 2 048 = 12 288 bit times;
— 10GBASE-T Delay: 25 600 bit times;
— 100 meters Cat6 cable: 5 556 bit times (computed assuming $\upsilon = 0.6 * c$, where c is the speed of the light in meters per second);
— HD = 6 144.

The total Delay Value in this scenario results to be:

DV = 2*(Max Frame) + (PFC Frame) + 2*(Cable Delay) + 2*ID + HD_{s2}

DV = 2 * (16 160) + (672) + 2 * (5 556) + 2 * (25 600) + 2 * (12 288) + 6 144= 126 024 bit times

For this case, the amount of buffering needed to ensure no frame loss due to congestion results to be 126 024 bit times, roughly equivalent to 15.5 KBytes.

If MACsec is used, the High Layer Delay is incremented by 19 360 bit times, therefore the total Delay Value results to be:

DV = 2 * (16 160) + (672) + 2 * (5 556) + 2 * (25 600) + 2 * (12 288) + 25 504 = 145 384 bit times

For this case, the amount of buffering needed to ensure no frame loss due to congestion results to be 145 384 bit times, roughly equivalent to 18 KBytes.