IEEE P802.1Qbb/D2.3

Draft Standard for
Local and Metropolitan Area Networks—

Virtual Bridged Local Area Networks — Amendment: Priority-based Flow Control

Sponsor

**LAN MAN Standards Committee
of the
IEEE Computer Society**

Prepared by the Data Center Bridging Task Group of IEEE 802.1

Abstract: This amendment 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 the IEEE 802.3 Annex 31B PAUSE, but operating on individual priorities.
Keywords: local area networks, LANs, transparent bridging, MAC Bridges, VLANs, priority, flow control.

IEEE Standards Department
Copyright and Permissions
445 Hoes Lane, P.O. Box 1331
Piscataway, NJ 08855-1331, USA

Editors’ Foreword

<<Notes>>

<<Throughout this document, all notes such as this one, presented between angle braces, are temporary notes inserted by the Editors for a variety of purposes; these notes and the Editors’ Foreword will all be removed prior to publication and are not part of the normative text.>>

<<Comments and participation in 802.1 standards development

Comments on this draft are encouraged. **PLEASE 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 development of this draft requires individual 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 had a policy of considering ballot 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. Non-members are advised that the email lists exist primarily to allow the members of the working group to develop standards, and are not a general forum.

Comments on this document may be sent to the 802.1 email exploder, to the editors, or to the Chairs of the 802.1 Working Group and Interworking Task Group.

Claudio DeSanti
Editor, P802.1Qbb
Email: cds@cisco.com

Pat Thaler
Chair, 802.1 Data Center Bridging Task Group
Email: pthaler@broadcom.com

Tony Jeffree
Chair, 802.1 Working Group
11A Poplar Grove
Sale
Cheshire
M33 3AX
UK
+44 161 973 4278 (Tel)
+44 161 973 6534 (Fax)
Email: tony@jeffree.co.uk

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

<<The draft text and accompanying information

This document currently comprises:

- — A temporary cover page, preceding the Editors' Forewords. This cover page will be removed following working group approval of this draft, i.e. prior to sponsor ballot.
- — IEEE boilerplate text.
- — The editors' forewords, including this text. These include an unofficial and informal appraisal of history and status, introductory notes to each draft that summarize the progress and focus of each successive draft, and requests for comments and contributions on major issues.
- — A title page for the proposed amendment including an Abstract and Keywords. This title page will be retained following approval.
- — IEEE boilerplate text (identical to the above).
- — The introduction to this standard.
- — A record of participants (not included in early drafts but added prior to publication).
- — The proposed amendment proper.
- — An Annex Z comprising the editors' discussion of issues. This annex will be deleted from the document prior to sponsor ballot.

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 the other discussions that take place in the working group meetings. Preparation of drafts often exposes inconsistencies in editors 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' introductory notes to the current draft, at appropriate points in the draft, and in Annex Z. Significant discussion of more difficult topics will be found in the last of these.

The ballot comments received on each draft, and the editors' proposed and final disposition of comments, 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 web site (for address see above).
>>

<<Introductory notes to Draft 0.1

This document, Draft 0.1, was prepared by Claudio DeSanti as a consequence of the IEEE 802 plenary meeting held in Denver, Colorado, USA, July 14-17, 2008. This draft is not presented for balloting. This is an initial draft.

Introductory notes to Draft 0.2

This document, Draft 0.2, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 interim meeting held in Seoul, Korea, September 15-18, 2008. This draft is intended for an initial ballot.

Introductory notes to Draft 1.0

This document, Draft 1.0, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 interim meeting held in New Orleans, LA, January 12-16, 2008. This draft is presented for task group balloting.

Introductory notes to Draft 1.1

This document, Draft 1.1, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 interim meeting held in Volterra, Tuscany, September 8-11, 2009. This draft is presented for working group balloting.

Introductory notes to Draft 1.2

This document, Draft 1.2, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 plenary meeting held in Atlanta, GA, November 16-19, 2009. This draft is presented for working group balloting.

Introductory notes to Draft 1.3

This document, Draft 1.3, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 interim meeting held in Austin, TX, January 19-22, 2010. This draft is presented for working group balloting.

Introductory notes to Draft 2.0

This document, Draft 2.0, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 plenary meeting held in Orlando, FL, March 15-19, 2010. This draft incorporates the resolution of all comments.

Introductory notes to Draft 2.1

This document, Draft 2.1, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 plenary meeting held in Orlando, FL, March 15-19, 2010. This draft incorporates the resolution of all comments and is presented for sponsor balloting.

Introductory notes to Draft 2.2

This document, Draft 2.2, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 plenary meeting held in Orlando, FL, March 15-19, 2010. This draft incorporates the opcodes and MIB arcs and is presented for sponsor balloting.

Introductory notes to Draft 2.3

This document, Draft 2.3, was prepared by Claudio DeSanti as a consequence of the IEEE 802.1 interim meeting held in Geneva, Switzerland, May 24-28, 2010. This draft incorporates the resolution of all sponsor ballot comments and is presented for sponsor recirculation balloting.

>>

<< Project Authorization Request, Scope, Purpose, and Five Criteria

A PAR (Project Authorization Request) for this project was first discussed in the January 2008 802.1 interim meeting, and forwarded for SEC consideration by vote of the 802.1 Working Group at its closing plenary during the March 2008 meeting of P802.

Scope of Proposed Project:

> This standard specifies protocols, procedures and managed objects that enable flow control per traffic class on IEEE 802 full duplex links. Data Center Bridging networks (bridges and end nodes) are characterized by limited bandwidth-delay product and limited hop-count. Traffic class is identified by the VLAN tag priority values. Priority-based Flow Control (PFC) is intended to eliminate frame loss due to congestion. This is achieved by a mechanism similar to the IEEE 802.3 Annex 31B PAUSE, but operating on individual priorities. This mechanism, in conjunction with other Data Center Bridging technologies, enables support for higher layer protocols that are highly loss sensitive while not affecting the operation of traditional LAN protocols utilizing other priorities. In addition, PFC complements Congestion Notification in Data Center Bridging networks.

Operation of Priority-based Flow Control is limited to a domain controlled by a Data Center Bridging control protocol that controls the application of Priority-based Flow Control, Enhanced Transmission Selection, and Congestion Notification.

Purpose of Proposed Project:

Data Center Bridging networks employ higher layer protocols that depend on the delivery of data frames without frame loss due to congestion. These protocols were designed for an underlying transport that approaches lossless behavior and therefore do not include appropriate response to frame loss due to congestion (e.g. back-off, slow restart, etc.). This amendment enables multiple data center networks, including those serving loss sensitive protocols (e.g. inter-processor communitcation, storage, etc.), to be converged onto an IEEE 802 network.

Need for the Proposed Project:

There is significant customer interest and market opportunity for 802 LANs as a converged Layer 2 solution in high-speed short-range networks such as data centers, backplane fabrics, single and multi-chassis interconnects, computing clusters, and storage networks. These environments currently use Layer 2 networks that do not discard packets due to congestion (e.g., Fibre Channel, InfiniBand). This project will bring comparable frame loss characteristics to 802 LANs in Data Center Bridging environments. This in conjunction with the other Data Center Bridging technologies will enable converged networks. Use of a converged network will realize operational and equipment cost benefits.

1. Broad Market Potential

A standards project authorized by IEEE 802 shall have a broad market potential. Specifically, it shall have the potential for:

a) Broad sets of applicability.

Mechanisms to avoid frame loss due to congestion are essential to support the highly loss sensitive higher layer protocols used in Data Center Bridging networks for data storage, clustering, and backplane fabrics. Back-end data storage networks, clustering networks and backplane fabrics with limited number of hops are amenable to a flow control mechanism that operates hop-by-hop.

The data traffic to be controlled by the proposed flow control mechanism will be segregated using priority values in the VLAN tag, ensuring that traffic types that are not amenable to hop-by-hop flow control may co-exist with those that are.

b) Multiple vendors and numerous users

Multiple equipment vendors, as well as INCITS T11 Technical Committee, have expressed interest in the proposed project. In addition, multiple vendors have announced product supporting similar technologies in a proprietary way. There is strong and continued user interest in combining separate existing networks into a converged infrastructure, based on international standards, resulting in the realization of operational and equipment cost savings.

c) Balanced costs (LAN versus attached stations)

The introduction of this flow control mechanism is not expected to materially alter the balance of costs between end stations and bridges. Significant equipment and operational costs savings are expected as compared to the use of separate networks for traditional LAN connectivity and for loss sensitive applications.

2. Compatibility

IEEE 802 defines a family of standards. All standards shall be in conformance with the IEEE 802.1 Architecture, Management and Interworking standards as follows: 802 Overview and Architecture, 802.1D, 802.1Q and parts of 802.1f. If any variances in conformance emerge, they shall be thoroughly disclosed and reviewed with 802.

Each standard in the IEEE 802 family of standards shall include a definition of managed objects which are compatible with systems management standards.

> The proposed standard will be an amendment to 802.1Q, and will interoperate and coexist with all prior revisions and amendmentsof the 802.1Q standard.
>
> The data traffic to be controlled by the proposed flow control mechanism will be segregated using priority values in the VLAN tag, thus ensuring that traffic types already supported by VLAN Bridges are not affected.
>
> The proposed amendment will contain MIB modules, or additions to existing MIB modules, to provide management operations for configuration and performance monitoring for both end stations and bridges.
>
> The proposed standard will contain managed objects that will enable its use in conjunction with P802.1Qau and P802.1Qaz.

3. Distinct Identity

Each IEEE 802 standard shall have a distinct identity. To achieve this, each authorized project shall be:

a) Substantially different from other IEEE 802 standards.

IEEE Std 802.1Q is the authoritative specification for priority aware Bridges and their participation in LAN protocols. No other IEEE 802 standard addresses priority-based flow control by bridges.

b) One unique solution per problem (not two solutions to a problem).

IEEE 802.3x defines a link flow control that pauses traffic on the whole link. The need to subject certain classes of traffic to flow control mechanisms, while allowing others to operate without flow control, has not been anticipated by any other IEEE 802 specification. Consequently, this proposal is the only solution to the problem of allowing a coexistence of such traffic types.

c) Easy for the document reader to select the relevant specification.

IEEE Std 802.1Q is the natural reference for priority based handling of traffic flows, which will make the capabilities added by this amendment easy to locate. The amendment will clearly state where its use is appropriate.

4. Technical Feasibility

For a project to be authorized, it shall be able to show its technical feasibility. At a minimum, the proposed project shall show:

a) Demonstrated system feasibility.

Similar techniques are widely deployed in other networking technologies of similar extent, such as Fibre Channel and InfiniBand, as well as in proprietary enhancements to 802.1Q bridging. These deployments have demonstrated that the proposed techniques are preferable to discarding packets during congestion for certain traffic types in networks of limited extent.

b) Proven technology, reasonable testing.

These and similar techniques have been proven in real world deployments of Fibre Channel, InfiniBand, in proprietary enhancements to 802.1Q bridging, and other networking technologies of similar extent. These techniques have been shown to be reasonably testable.

c) Confidence in reliability.

These and similar techniques have been proven reliable in real-world deployments of Fibre Channel, InfiniBand, and other networking technologies of similar extent.

d) Coexistence of 802 wireless standards specifying devices for unlicensed operation.

Not applicable.

5. Economic Feasibility

For a project to be authorized, it shall be able to show economic feasibility (so far as can reasonably be estimated), for its intended applications. At a minimum, the proposed project shall show:

a) Known cost factors, reliable data.

The proposed amendment will retain existing cost characteristics of bridges including simplicity of queue structures and will not require maintenance of additional queues beyond the existing per traffic class (priority) queues for conformance to either its mandatory or optional provisions. In particular per flow queuing will not be required.

b) Reasonable cost for performance.

The proposed technology will reduce overall costs where separate networks are currently required by enabling the use of a converged network. The proposed solution allows a network to avoid frame loss due to congestion without significant throughput reduction.

c) Consideration of installation costs.

Installation costs of VLAN Bridges or end stations are not expected to be significantly affected; any increase in network costs is expected to be more than offset by a reduction in the number of separate networks required to be installed and managed.

>>

IEEE P802.1Qbb/D2.3

DRAFT Amendment to IEEE Std 802.1Q -2005
as amended by IEEE Std 802.1ad-2005

May 25, 2010

IEEE P802.1Qbb/D2.3

Draft Standard for
Local and Metropolitan Area Networks —

Virtual Bridged Local Area Networks — Amendment: Priority-based Flow Control

Sponsor

LAN MAN Standards Committee
of the
IEEE Computer Society

Prepared by the Data Center Bridging Task Group of IEEE 802.1

Abstract: This amendment 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: local area networks, LANs, transparent bridging, MAC Bridges, VLANs, priority, flow control.

IEEE Standards Department
Copyright and Permissions
445 Hoes Lane, P.O. Box 1331
Piscataway, NJ 08855-1331, USA

Introduction to IEEE Std 802.1Qbb™

This Standard provides Priority-based Flow Control capabilities useful to 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 can be obtained from

Secretary, IEEE-SA Standards Board
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
USA

Participants

The following is a list of participants in the Interworking activities of the IEEE 802.1 Working Group during the development of 802.1Qbb. Voting members at the time of publication are marked with an asterisk (*).

When the IEEE 802.1 Working Group approved IEEE Std 802.1Qbb, it had the following membership:

Tony Jeffree, Chair
Paul Congdon, Vice Chair
Path Thaler, Chair, Data Center Bridging Task Group
Claudio DeSanti, Editor, 802.1Qbb

<<TBA>>

The following members of the balloting committee voted on 802.1Qbb. Balloters may have voted for approval, disapproval, or abstention.

<<TBA>>

When the IEEE-SA Standards Board approved this standard on <<TBA>>, it had the following membership:

???, Chair ???, Vice Chair ???, Secretary

<<TBA>>

Contents

IEEE P802.1Qbb/D2.3

Draft Standard for
Local and Metropolitan Area Networks—
Amendment to 802.1Q
Virtual Bridged Local Area Networks:
Priority-based Flow Control

Editorial Note

This amendment specifies changes to IEEE Std 802.1Q that provide capabilities for enabling flow control per traffic class on IEEE 802 point-to-point full duplex links.

Changes are applied to the base text generated by applying the amendment IEEE Std 802.1ad-2005 to IEEE Std 802.1Q-2005. 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.

<< Editor's note: Before finalizing this amendment it will be necessary to produce a 'rolled up' base document comprising this amendment and 802.1Q-2005. This rolling up process usually reveals some unintended effects of what would otherwise appear a reasonable amendment. >>

1. Overview

Insert the following after the initial paragraphs 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).

1.1 Scope

Insert the following at end of subclause 1.1, relettering the bullet points so that they follow in order from those in the existing text.

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

a) 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 at the appropriate point:

IEEE Std 802.1AE: *Media Access Control (MAC) Security*

IEEE Std 802.1Qaz: *Enhanced Transmission Selection for Bandwidth Sharing Between Traffic Classes*

IEEE Std 802.3bd: *MAC Control Frame for Priority-based Flow Control*

3. Definitions

Insert the following definitions, renumbering to place them in the appropriate collating order.

3.1 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.

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

3.3 Data center environment: A domain controlled by the Data Center Bridging eXchange protocol (DCBX, see IEEE Std 802.1Qaz clause 38).

4. Abbreviations

Insert the following definitions, placing them in the appropriate collating order (alphabetical).

DCBX	Data Center Bridging eXchange protocol
PFC	Priority-based Flow Control
TLV	Type, Length, Value

5. Conformance

In subclause 5.4.1 VLAN-aware Bridge component options, insert the following additional bullet after current bullet (d):

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

Insert the following subclause after subclause 5.10:

5.11 System requirements for Priority-based Flow Control

A system that conforms to the provisions of this standard for Priority-based Flow Control (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 IEEE Std 802.1Qaz clause 38).

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

Insert the following subclause after subclause 6.6.3:

6.6.4 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.

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

Insert the following paragraph at the end of subclause 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

8.6.8 Transmission selection

Insert the following text after item b) of subclause 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 text at the end of subclause 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 IEEE Std 802.1Qaz subclause 38.5.4.6).

12. Bridge management

Insert a new subclause at the end of Clause 12, as follows, renumbering as necessary.

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

Table 12-1—Priority-based Flow Control objects

Name	Data type	Operations supported[*]	Conformance[†]
PFCLinkDelayAllowance	unsigned integer	RW	BE
PFCRequests	unsigned integer	R	BE
PFCIndications	unsigned integer	R	BE

[*] R = Read only access; RW = Read/Write access
[†] 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 (informative) for an example of how to compute this parameter.

17. MIB Modules

17.1 The Internet Standard Management Framework

17.2 Structure of the MIB

Insert the following line at the appropriate place in 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

Insert the following subclause after Clause 17.2.16:

17.2.17 Structure of the Priority-based Flow Control MIB

Subclause 12.23 of this document defines the information model associated with this standard in a protocol independent manner. Table 17-2 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 17-2—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

Insert the following subclause after Clause 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 subclause 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

Insert the following subclause after Clause 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 non-secure 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

Insert the following subclause after Clause 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 a new Clause 36 as follows.

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.

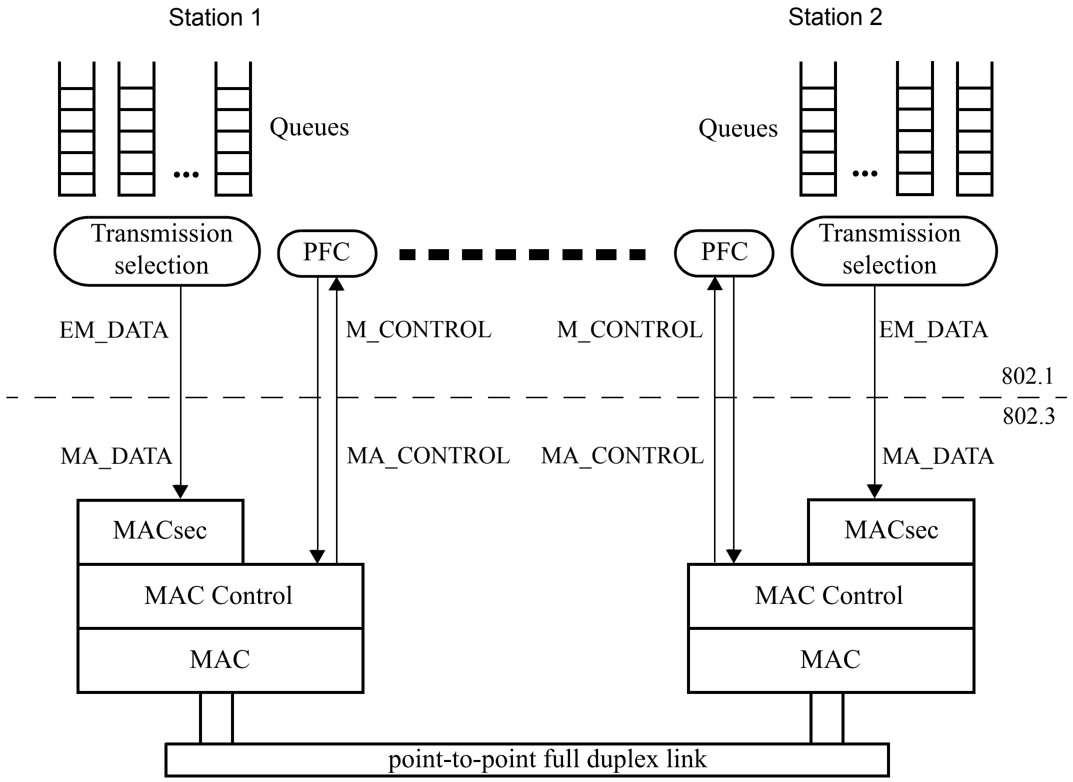

Figure 36-1—PFC Peering

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 6.6.4). 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[1]).

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,

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

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.

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.

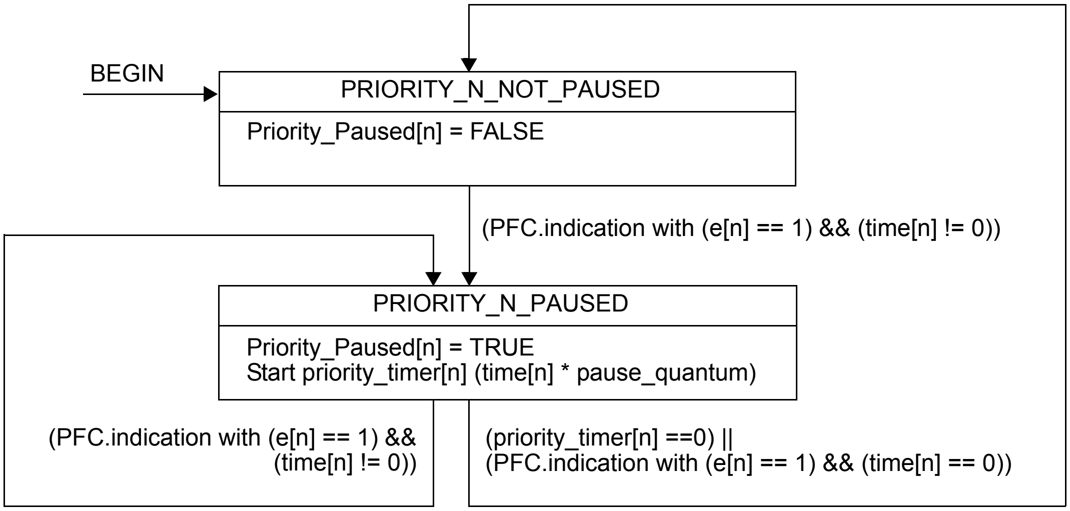
BEGIN
PRIORITY_N_NOT_PAUSED
Priority_Paused[n] = FALSE
(PFC.indication with (e[n] == 1) && (time[n] != 0))
PRIORITY_N_PAUSED
Priority_Paused[n] = TRUE
Start priority_timer[n] (time[n] * pause_quantum)
(PFC.indication with (e[n] == 1) &&
(time[n] != 0))
(priority_timer[n] ==0) ||
(PFC.indication with (e[n] == 1) && (time[n] == 0))

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 non-zero. 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 non-zero 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.

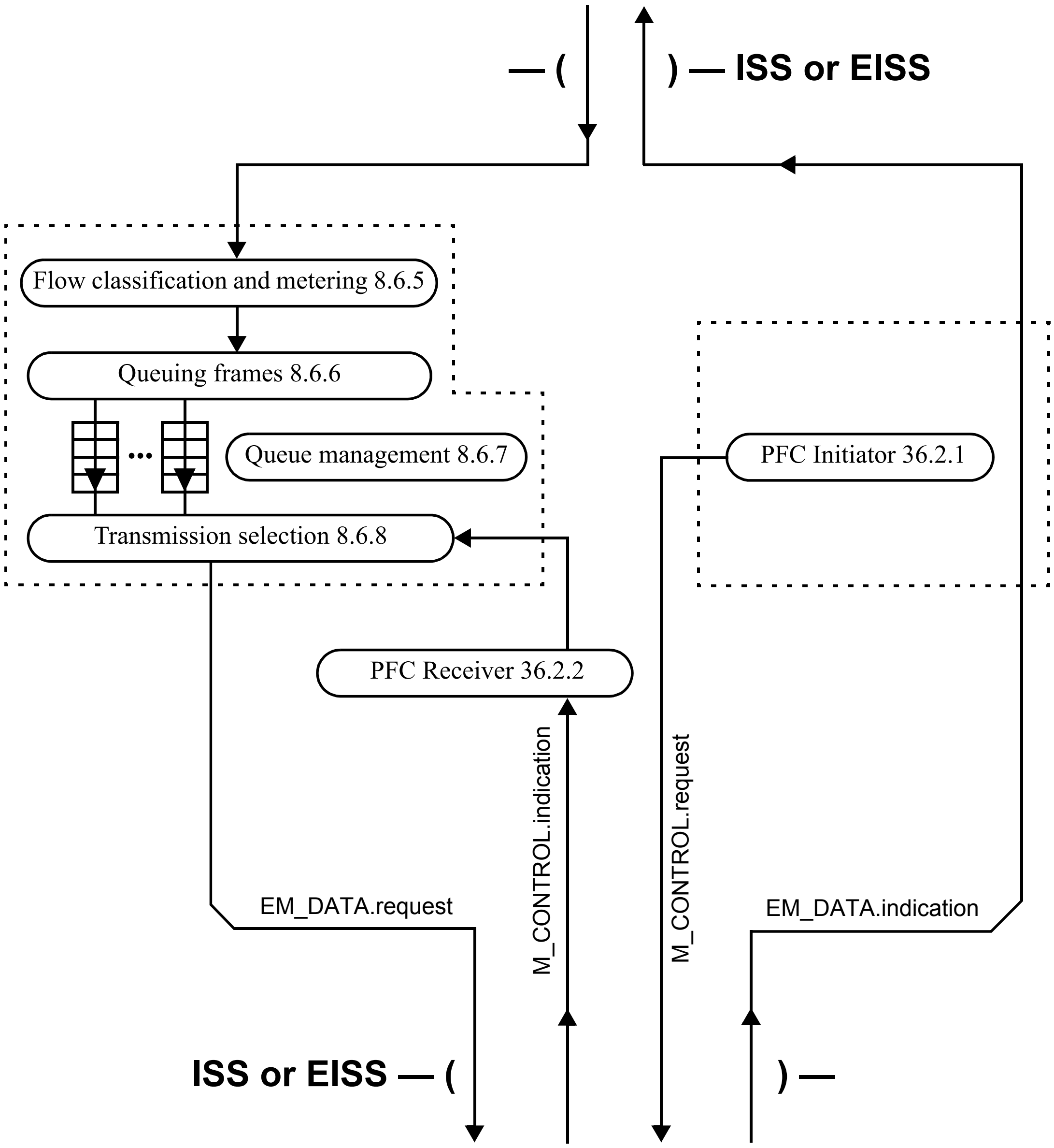

Figure 36-3—PFC aware system queue functions

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.

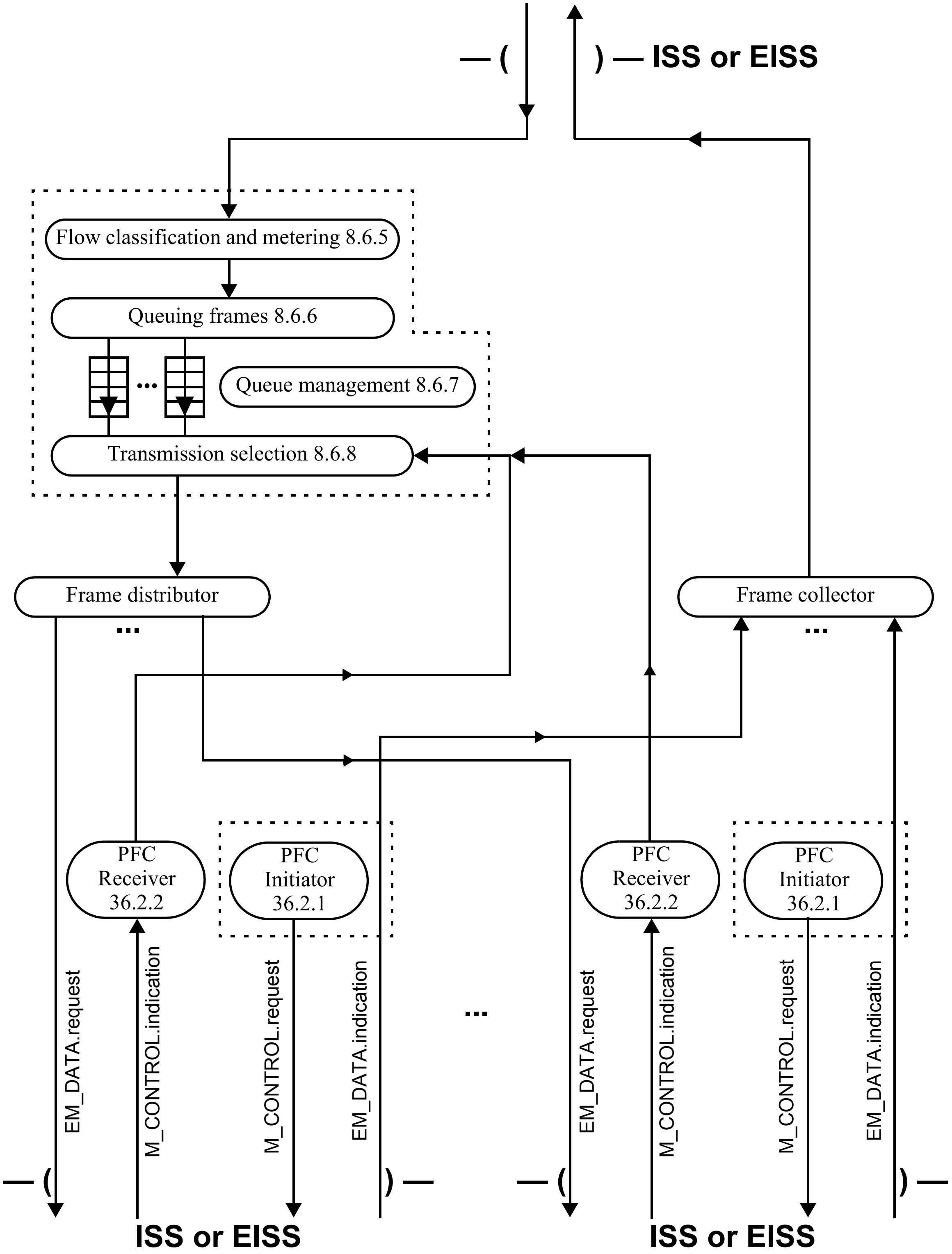

Figure 36-4—PFC aware system queue functions with link aggregation

Annex A (normative)

PICS Proforma[1]

A.1 Major Capabilities (continued)

Insert the following items 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 points at the end of A.14, renumbering items if necessary:

Item	Feature	Status	References	Support
MGT-208	Priority-based Flow Control entities	PFC: O	12.23	Yes [ ] No [ ]

A.24 Management Information Base (MIB)

Insert the following points at the end of A.24, renumbering items if necessary:

Item	Feature	Status	References	Support
MIB-25	Is the IEEE8021-PFC-MIB module fully supported (per its MODULE-COMPLIANCE)?	PFC AND MIB: O	17.7.17	Yes [ ] No [ ]

A.30 Priority-based Flow Control

Insert the following annex subclause:

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 [ ]

[1] *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.

Item	Feature	Status	References	Support
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 a new Annex N:

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 6.6.4). 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[1]. 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.

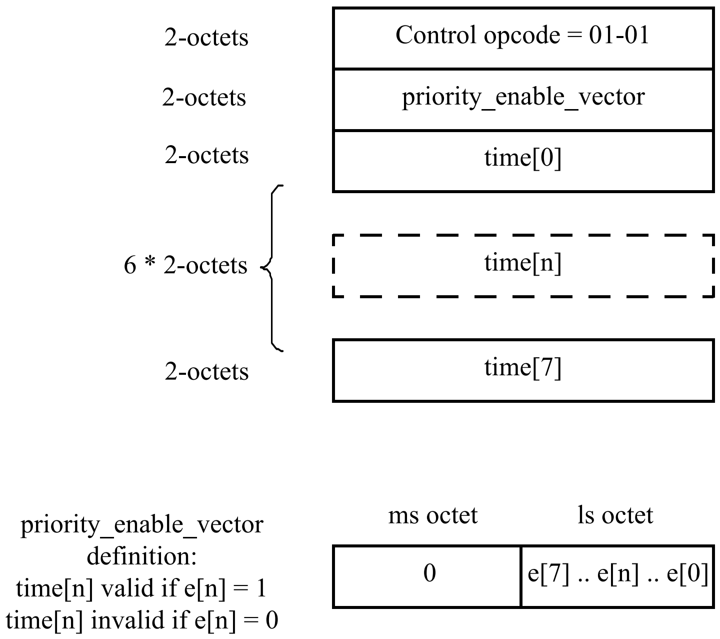

Figure N-1—PFC PDU Format

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.

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

Insert a new Annex O:

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.

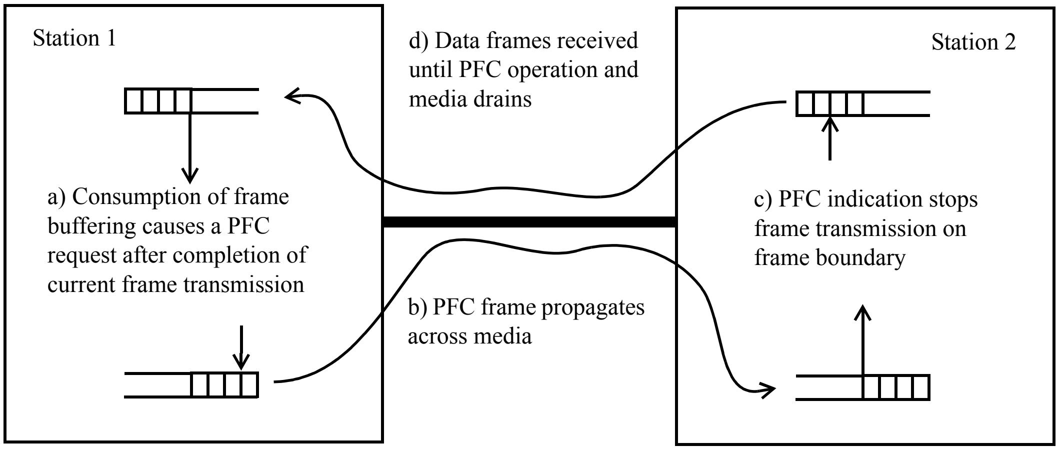
Station 1
d) Data frames received until PFC operation and media drains
Station 2
a) Consumption of frame buffering causes a PFC request after completion of current frame transmission
c) PFC indication stops frame transmission on frame boundary
b) PFC frame propagates across media

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.

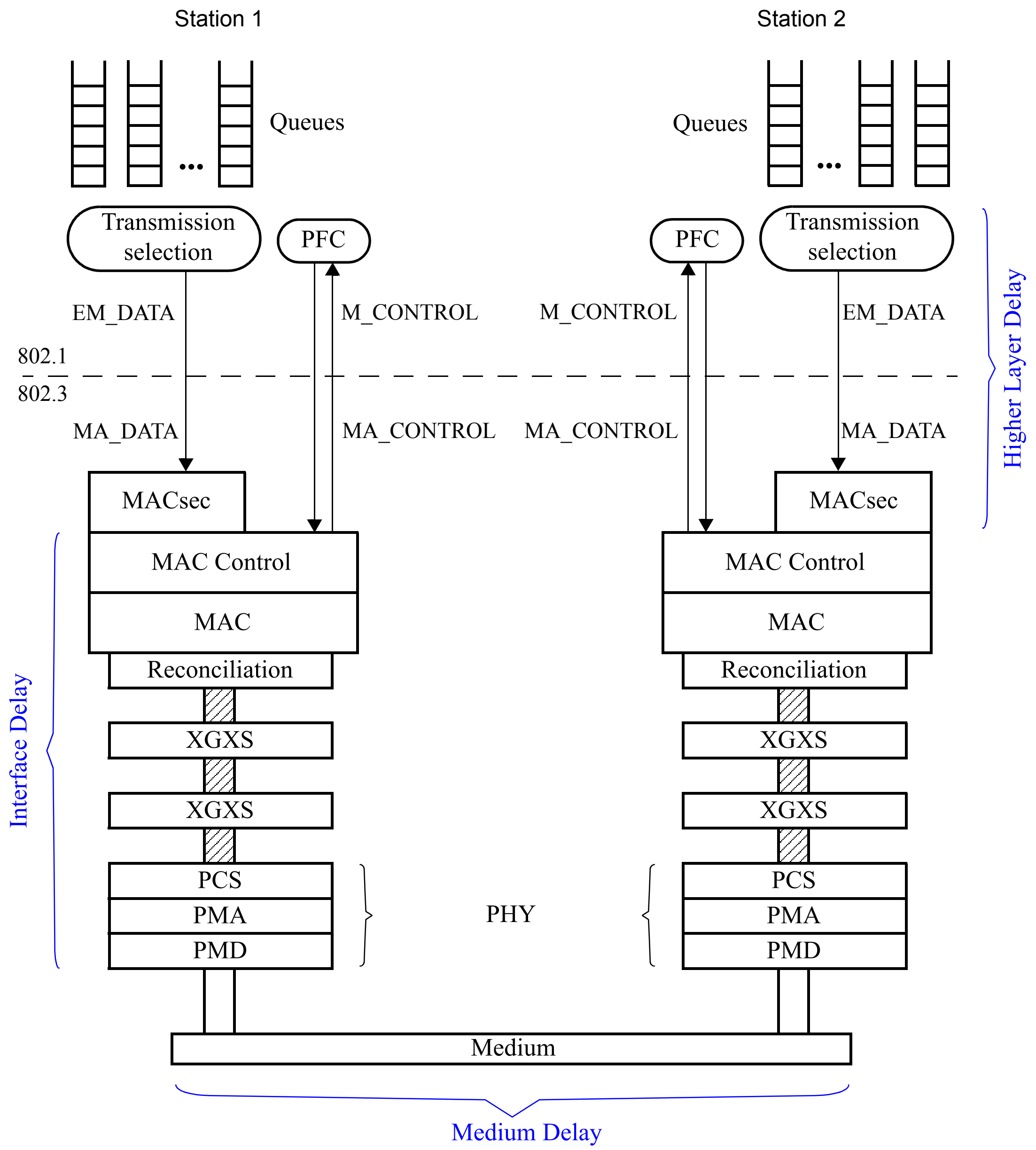

Figure O-2—Delay Model

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.

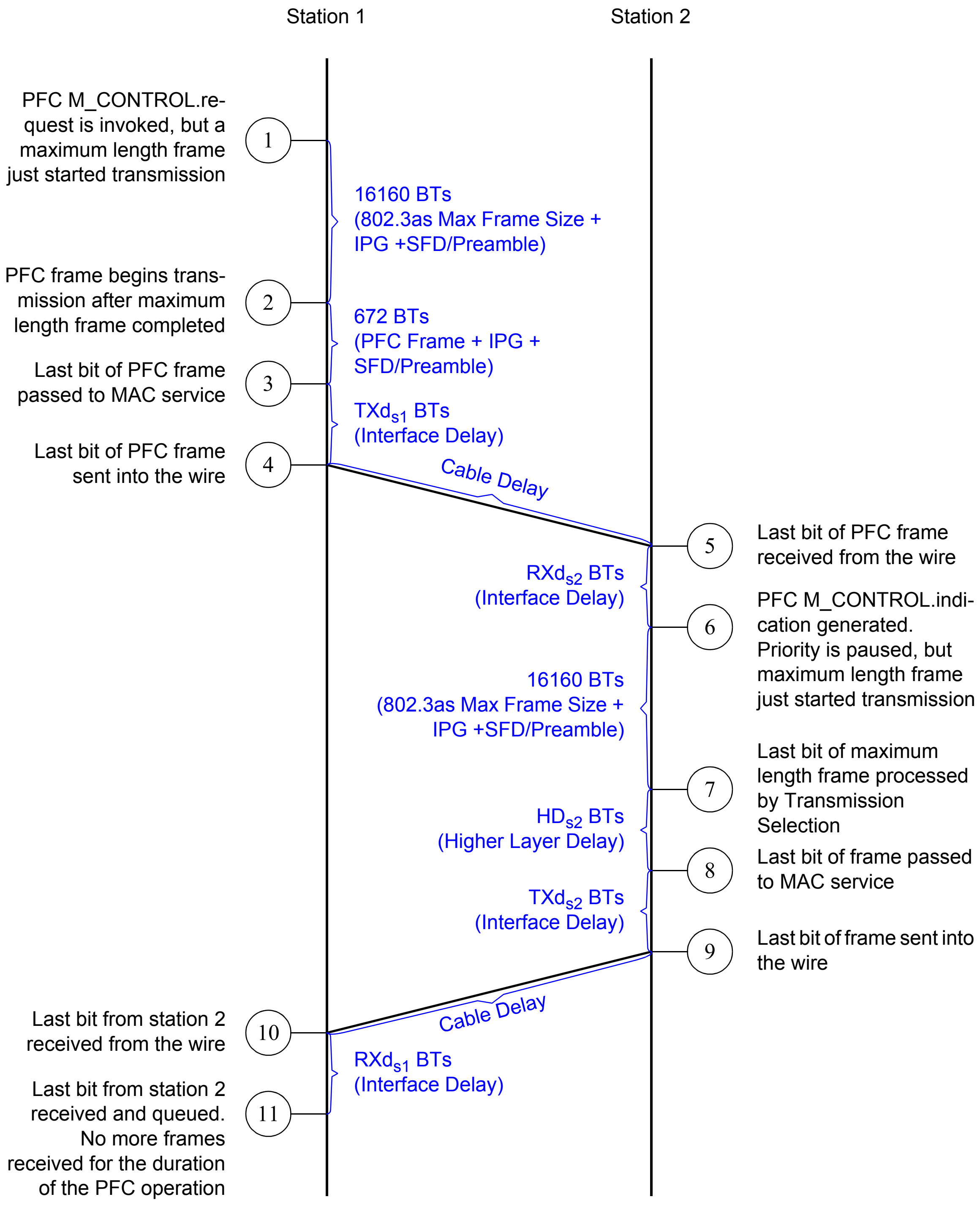
Station 1
Station 2
PFC M_CONTROL.request is invoked, but a maximum length frame just started transmission
1
16160 BTs (802.3as Max Frame Size + IPG +SFD/Preamble)
PFC frame begins transmission after maximum length frame completed
2
672 BTs (PFC Frame + IPG + SFD/Preamble)
Last bit of PFC frame passed to MAC service
3
TXd$_{s1}$ BTs (Interface Delay)
Last bit of PFC frame sent into the wire
4
Cable Delay
5
Last bit of PFC frame received from the wire
RXd$_{s2}$ BTs (Interface Delay)
6
PFC M_CONTROL.indication generated. Priority is paused, but maximum length frame just started transmission
16160 BTs (802.3as Max Frame Size + IPG +SFD/Preamble)
Last bit of maximum length frame processed by Transmission Selection
7
HD$_{s2}$ BTs (Higher Layer Delay)
8
Last bit of frame passed to MAC service
TXd$_{s2}$ BTs (Interface Delay)
9
Last bit of frame sent into the wire
Cable Delay
Last bit from station 2 received from the wire
10
RXd$_{s1}$ BTs (Interface Delay)
Last bit from station 2 received and queued. No more frames received for the duration of the PFC operation
11

Figure O-3—Worst Case Delay

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

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

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

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

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

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

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:

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

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.