IEEE P802.1Qcz/D0.0

Draft Standard for Local and Metropolitan Area Networks—

Bridges and Bridged Networks—Amendment: Congestion Isolation

Sponsor
LAN/MAN Standards Committee of the IEEE Computer Society

Prepared by the Time Sensitive Networking Task Group of IEEE 802.1

Abstract: This amendment to IEEE Std 802.1Q-2018 specifies protocols, procedures and managed objects that support the isolation of congested data flows within data center environments.
Keywords: Bridged Local Area Networks, local area networks (LANs), MAC Bridges, data center networks, Virtual Bridged Local Area Networks (virtual LANs), Data Center Bridging (DCB), Congestion Isolation (CI).

DRAFT STATUS:

Individual contribution draft, issued for TG discussion.

Introduction to IEEE P802.1Qcz/D0.0™

This introduction is not part of IEEE P802.1Qcz/D0.0, Draft Standard for Local and Metropolitan Area Networks—Bridges and Bridged Networks—Amendment: Congestion Isolation

This amendment specifies protocols, procedures and managed objects that support the isolation of congested data flows within data center environments. This is achieved by enabling systems to individually identify flows creating congestion, adjust transmission selection for packets of those flows, and signal to neighbors. This mechanism reduces head-of-line blocking for uncongested flows sharing a traffic class in lossless networks. Congestion Isolation is intended to be used with higher layer protocols that utilize end-to-end congestion control in order to reduce packet loss and latency. This amendment also addresses errors and omissions in the description of existing functionality.

There is significant customer interest and market opportunity for large scale, low-latency, lossless Ethernet data centers to support high-performance computing and distributed storage applications. Congestion is the primary cause of loss and delay. These environments currently use higher layer end-to-end congestion control coupled with priority-based flow control at Layer 2 to avoid performance degradation from packet loss due to congestion. As the Ethernet data center network scales in size, speed and number of concurrent flows, the current environment creates head-of-line blocking for flows sharing the same traffic class. Isolating flows that cause congestion reduces latency for flows not causing congestion and improves the scale and performance of the Ethernet data center network. This amendment will support the identification and isolation of the higher layer protocol flows that are creating congestion. The amendment will interoperate with existing congestion management. Use of a consolidated Ethernet data center network will realize operational and equipment cost benefits.

<<Editor's Note: The text of the PAR for this project can be found here:

https://development.standards.ieee.org/get-file/P802.1Qcz.pdf?t=98411700003

The CSD including the "5 Criteria" that were approved by 802.1 and the 802 EC at PAR submission can be found here:

http://www.ieee802.org/1/files/public/docs2018/cz-CSD-0718-v01.pdf

As part of our IEEE 802 process, the text of the PAR and the CSD should be reviewed on a regular basis in order to ensure their continued validity. A vote of "Approve" on this draft is assumed also to be an affirmation by the balloter that the text of the PAR and CSD are still valid.>>

Contents

Figures

Tables

IEEE IEEE P802.1Qcz/D0.0

Draft Standard for Local and Metropolitan Area Networks—

Bridges and Bridged Networks—Amendment: Congestion Isolation

(This amendment is based on IEEE Std 802.1Q™-2018.)

1. Overview

1.3 Introduction

This amendment specifies protocols, procedures and managed objects that support the isolation of congested data flows within data center environments. This is achieved by enabling systems to individually identify flows creating congestion, adjust transmission selection for packets of those flows, and signal to neighbors. This mechanism reduces head-of-line blocking for uncongested flows sharing a traffic class in lossless networks. Congestion Isolation is intended to be used with higher layer protocols that utilize end-to-end congestion control in order to reduce packet loss and latency. This amendment also addresses errors and omissions in the description of existing functionality. To this purpose it:

a) Defines a means for VLAN-aware Bridges that support congestion isolation for identifying flows that are creating congestion.
b) Defines a means for adjusting transmission selection for packets of congested flows
c) Provides for a means for discovering peer VLAN-aware Bridges and stations that support congestion isolation
d) Defines a means for signaling congestion isolation to supporting peer Bridges and stations.

2. Normative references

Insert the following reference in the appropriate collating sequence:

3. Definitions

Insert the following definitions in the appropriate collating sequence, re-numbering as appropriate:

3.1 Congestion Isolation Aware System: A bridge component conforming to the congestion isolation provisions of this standard.

3.2 Congested Flow: A sequence of frames the end-to-end congestion controlled higher-layer protocol treats as belonging to a single flow that is experiencing congestion within a Congestion Isolation Aware System.

3.3 Congestion Isolation Message (CIM): A message transmitted by a Congestion Isolation Aware System, conveying congestion Congested Flow information used by the upstream peer Congestion Isolation Aware System.

3.4 Congestion Isolation Point (CIP): A Congestion Isolation Aware System that monitors a set of queues for Congested Flows and can generate Congestion Isolation Messages.

4. Abbreviations

Insert the following abbreviations in the appropriate sequence, re-ordering as appropriate:

CF Congested Flow
CI Congestion Isolation
CIM Congestion Isolation Message
CIP Congestion Isolation Point

5. Conformance

Insert the following subclause after Clause 5.4.6:

5.4.7 VLAN Bridge requirements for congestion isolation

A VLAN-aware Bridge implementation that conforms to the provisions of this standard for congestion isolation (XX) shall:

a) Support, on one or more Ports, the creation of at least one Congestion Isolation Point (xx.x.x);
b) Support, at each Congestion Isolation Point, the generation of Congestion Isolation Messages (xx.x);
c) Support the ability to configure the variables controlling the operation of each Congestion Isolation Point (xx.x.x);
d) Conform to the required capabilities of the LLDP of 5.2 of IEEE Std 802.1AB-2009;
e) Support the use of the Congestion Isolation TLV in LLDP (xx.x.x)

A VLAN Bridge implementation that conforms to the provisions of this standard for congestion isolation may:

a) Support the creation of up to four CIPs on a Bridge Port (xx.x.x)
b) Support Congested Traffic Enhanced Traffic Selection (yy.y.y)
c) Support the Congestion Isolation YANG model (xx.x.x))

6. Support of the MAC Service

6.10.1 Data Indication

Insert the following sentence after the first sentence:

If the PIP is congestion isolation aware (5.4.7) and the initial octets of the mac_service_data_unit contain a valid CIM encapsulation, the received frame is processed according to XX.XX.

8. Principles of bridge operation

8.6.6 Queuing Frames

Replace the last paragraph with to following paragraph:

In a congestion aware Bridge (Clause 30) or a congestion isolation aware Bridge (Clause XX), the act of queuing a frame for transmission on a Bridge Port can result in the Forwarding Process generating a CNM or a CIM. The CNM is and CIM are injected back into the Forwarding Process (8.6.1) as if it had been received on that Bridge Port.

8.6.8 Transmission Selection

Replace the fourth paragraph with to following paragraph:

The strict priority transmission selection algorithm defined in 8.6.8.1 shall be supported by all Bridges as the default algorithm for selecting frames for transmission. The credit-based shaper transmission selection algorithm defined in 8.6.8.2, the ETS algorithm defined in 8.6.8.3, and the Congestion Isolation algorithm defined in 8.6.8.4 may be supported in addition to the strict priority algorithm. Further transmission selection algorithms, selectable by management means, may be supported as an implementation option so long as the requirements of 8.6.6 are met.

Replace Table 8-5 with the following table:

Transmission selection algorithm	Identifier
Strict priority (8.6.8.1)	0
Credit-based shaper (8.6.8.2)	1
Enhanced Transmission Selection (ETS) (8.6.8.3)	2
Congestion isolation (8.6.8.4)	3
Reserved for future standardization	4-254
Vendor-specific Transmission Selection algorithm value for use with DCBX (D.2.8.8)	255
Vendor-specific	A four-octet integer, where the most significant 3 octets hold an OUI or CID value, and the least significant octet holds an integer value in the range 0–255 assigned by the owner of the OUI or CID.

Insert the following clause after 8.6.8.3 and renumber subsequent clauses accordingly:

8.6.8.4 The Congestion Isolation Algorithm

Traffic selection for the congestion isolation algorithm requires that two participating traffic classes are enabled for congestion isolation. The higher priority class is designated the non-congested traffic class and the lower priority class is the congested traffic class. Either traffic class may be blocked from transmission selection by the controlling variables defined in clause XX.XX. When the traffic classes are permitted to participate in traffic selection, bandwidth may be distributed among the classes such that each class is allocated available bandwidth in proportion to its TCBandwidth (see Clause 37).

When a queue that supports Congestion Isolation is permitted to participate in transmission selection the algorithm determines a frame is available for transmission if the following conditions are true:

a) The queue contains one or more frames
b) The congestion isolation algorithm (xx.xx) determines that a frame should be transmitted from the queue; and
c) There are no frames available for transmission for any queues running strict priority, credit-base shaper or ETS algorithms.

12. Bridge management

Insert Congestion Isolation objects into the existing list of managed objects in 12.1.1

m) The ability to create and delete the functional elements of congestion isolation and to control their operation.

Insert Congestion Isolation in the list of VLAN Bridge Objects of 12.2

s) The congestion isolation entities (12.xx).

Insert a new clause 12.xx Congestion Isolation managed objects

12.xx Congestion Isolation managed objects

Several variables control the operation of Congestion Isolation in a congestion isolation aware bridge. The managed objects are as follows:

a) CI component managed object (12.xx.1)
b) Congestion Isolation Point (CIP) component managed object (12.xx.2)

12.xx.1 CI component managed object

A single instance of the CI component managed object shall be implemented by a Bridge component or end station that is congestion isolation aware. It comprises all the variables included in the CI component variables (XX.X) as illustrated in Table 12-X.

Table 12-X – Congestion Isolation component managed object

Name	Data type	Operations supported	Conformance	References
ciMasterEnable	Boolean	RW		XX.X.X
ciCimTransmitPriority	Unsigned integer [0..7]	R		XX.X.X

12.xx.2 CIP component managed object

There is one congestion isolation point (CIP) managed object for each CIP in a Bridge component or end station that is congestion isolation aware. The CIP managed object comprises the some of the variables included in the CI variables (XX.X) as illustrated in table 12-Y.

Table 12-Y – Congestion Isolation Point component managed object

Name	Data type	Operations supported	Conformance	References
cipMonitoredQueues	unsigned integer [0..255]	RW		XX.X.X

cipCongestedQueue	unsigned integer [0..7]	RW		XX.X.X
cipMinHeaderOctets	unsigned integer	RW		XX.X.X
cipTransmittedCims	counter	R		XX.X.X
cipReceivedCims	counter	R		XX.X.X

99. Congestion Isolation

<< Editor's Note: This section includes the following introduction material:

- Reference diagram
- Problems being solved; preventing head-of-line blocking and reducing latency of short-lived flows while providing time for the end-to-end congestion control loop to react.
- Requirements and objectives for the solution
- Methods for identifying congested flows
- Operation of adjusting traffic selection for congested flows
- Allocating and deallocating congested flows in a congested flow table
- Role and operation of signaling to peers
- Relationship and comparison with Congestion Notification

.>>

In current data center networks, traffic can be a mix of various multi-tenant TCP and UDP flows across both the physical underlay and virtual overlay network. Intermittent congestion within the network can be caused by the unfortunate mix of flows across the fabric. A small number of long duration elephant flows can align in such a way to create queuing delays for the larger number of short but critical mice flows. The queuing delays deter the end-to-end congestion control loop and cannot prevent PFC flow control from being invoked. When buffers fill and eventual flow-control kicks in, mice flows can be blocked by the unfortunate burst alignment of elephant flows. If PFC flow control is not being used, packet loss on short mice flows can result in full retransmission timeouts, significantly penalizing the latency of mice flows used for control and synchronization within the parallel application.

Congestion Isolation (CI) avoids head-of-line blocking caused by the frequent-use of PFC while supporting lossless behavior. CI identifies the flows that are causing congestion, isolates them to a separate traffic class and signals to the upstream neighbor to do the same. CI effectively moves the congested flows out of the way, temporarily, while the end-to-end control loop has time to take effect.

Figure 99-1—Congestion Isolation Model

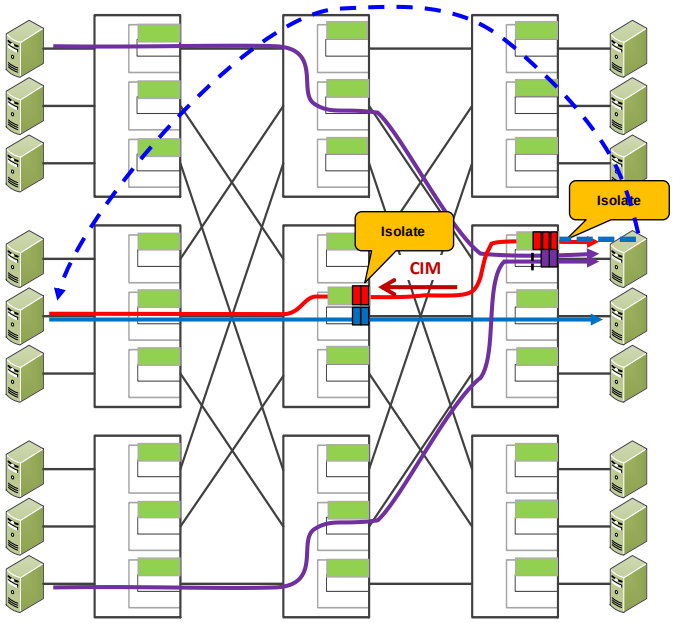


Figure 99-1 shows the operation of CI. When flows unfortunately collide at the egress port of a bridge, congestion is detected, and the offending flows are identified. Subsequent packets from the offending flows are routed through a dedicated congested flow queue (i.e. they are effectively moved out of the way). Once the congested flow queue reaches a threshold, the CI functionality signals to the upstream bridge using a Congestion Isolation Message (CIM) that contains enough information for the upstream bridge to identify the same congested flow. The upstream bridge also isolates the same flow and begins to monitor the depth of the congested flow queue. The packets in the congested flow queue are drained at a lower priority than other non-congested queues, so when congestion persists, the congested flow queue may fill. When the congested flow queue fills, the ingress port feeding that queue can issue PFC to avoid packet loss. Flow control is only blocking the congested flow queues and other well-behaved mice and elephant flows are free to traverse the fabric via non-congested queues.

99.1 Congestion Isolation Entity Operation

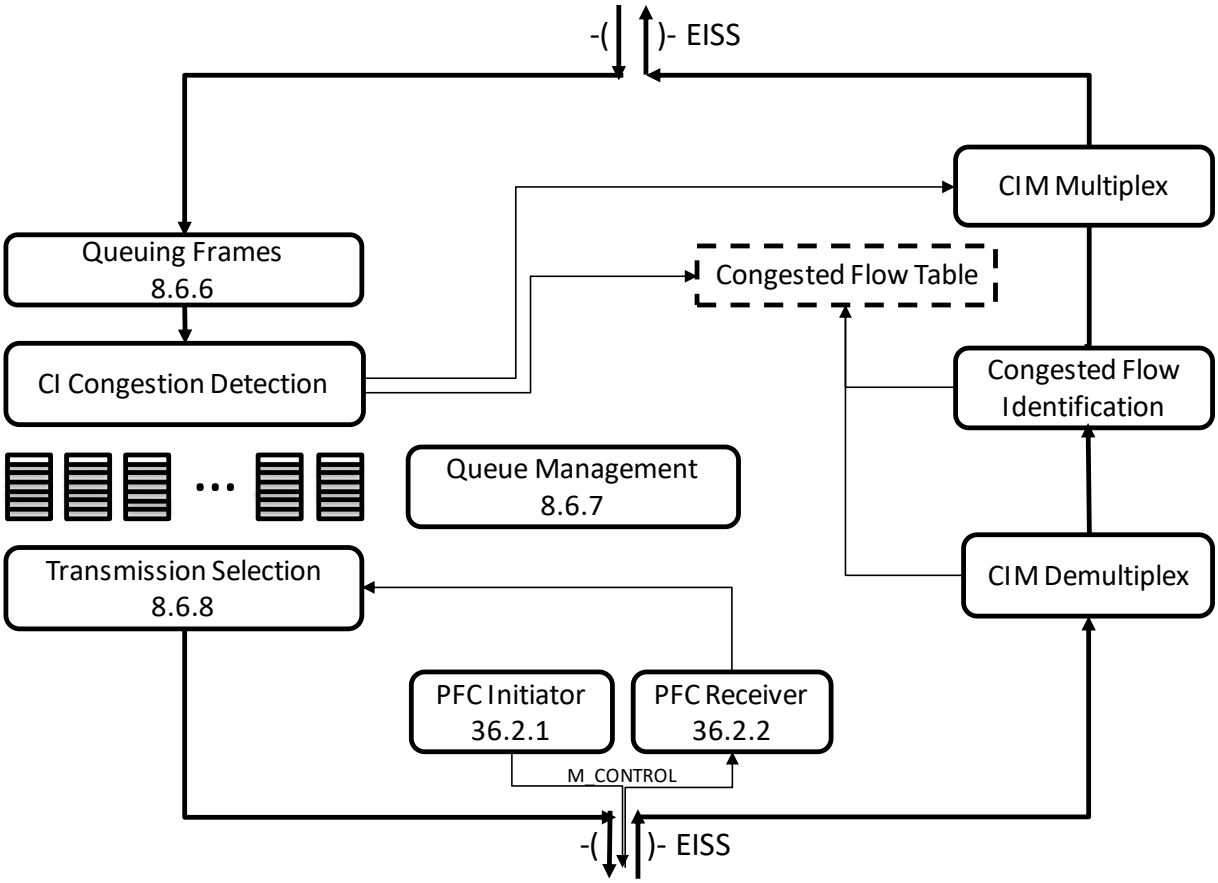
The text of this section.
<< Editor's Note: The text should cover the following:

- Congestion Isolation aware Bridge Forwarding Process diagram (see next slide)
- Congestion Isolation Point (CIP)
- Congestion Isolation Input Multiplexor – how CIMs are decoded and received.
- Congested Flow Identification and Table

>>

Figure 99-2—Congestion Isolation Reference Diagram

`



99.2 Congestion Isolation Protocol

More text for this normative section.
<< Editor's Note – This section needs to include:

- Variables controlling operation
- State machine and associated variables and procedures that control the CIM and congested flow table
- The Congestion Isolation Protocol
  - Generating CIMs – State Machines
  - Creating and Deleting entries in the Congested Flow Table
  - Queuing frames via EM_UNITDATA.request
  - Processing received CIMs
- Encoding of CIM PDUs
- Congestion Isolation LLDP TLV definition
- UML and YANG model

>>

99.2.1 Congestion Isolation LLLDP TLV

Text here.

Annex A (normative)PICS proforma—Bridge implementations

Annex D

(normative)

IEEE 802.1 Organizationally Specific TLVs

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

Add the following row to Table D-1

Table D-1—IEEE 802.1 Organizationally Specific TLVs

IEEE 802.1 subtype	TLV name	TLV set name	TLV reference	Feature clause reference
xx	Congestion Isolation	ciSet	D.2.16	Clause 99

<< Editor's Note: the value xx will be assigned at sponsor ballot >>

D.2 Organizationally Specific TLV definitions

Add the following clauses to the end of D.2

D.2.16 Congestion Isolation TLV

The Congestion Isolation TLV is an optional TLV that allows an IEEE 802.1Q-compliant bridge and an IEEE 802.1Q-compatible IEEE 802 LAN station to discover each other and exchange configuration information.

TLV type = 127	TLV information string length = X	802.1 OUI 00-80-C2	802.1 subtype = X	Monitored Queues	Congested Queue	CIM encap header length
7 bits	9 bits	3 octects	1 octect	1 octect	1 octect	1 octect

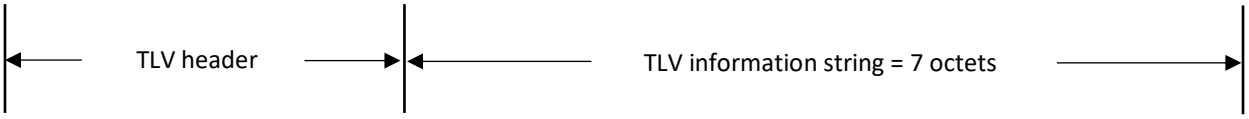


Figure D-16—Congestion Isolation TLV Format

D.2.16.1 TLV type

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

D.2.16.2 TLV information string length

A 9-bit unsigned integer, occupying the least-significant bit of the first octet of the TLV (the most significant bit of the TLV information string length) and the entire second octet of the TLV, containing the total number of octets in the TLV information string of the Element TLV. This does not count the TLV type and TLV information string length fields. The length for the Congestion Isolation TLV is fixed and equal to 7 octets.

D.2.16.3 Monitored Queues

A bit vector, one per priority value, containing all eight of the traffic classes support by the Bridge or end station. The LSB of the octet carries priority 0, and the MSB is that of priority 7.

D.2.16.4 Congested Queue

The numerical value in the range of 0 to 7 that represents the traffic class that will act as the congested queue. The congested queue is lower priority than the monitored queues.

D.2.1.5 CIM encap header length

A single octet unsigned integer representing the requested length in bytes of the header encapsulated into a CIM message by a peer. The default value is 64.

D.3 IEEE 802.1 Organizationally Specific TLV management

Add the following clauses to the end of D.3

D.3.11 Congestion Notification TLV managed objects

a) **monitored queues**: see D.2.16.3
b) **congested queue**: see D.2.16.4
c) **encaps header length**: see D.2.16.5

D.4 PICS proforma for IEEE 802.1 Organizationally Specific TLV extensions[4,5]

[4]Instructions for completing the PICS Proforma are given in A.3.

[5]*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.

D.5 IEEE 802.1/LLDP extension MIB

D.5.1 Internet Standard Management Framework

D.5.2 Structure of the IEEE 802.1/LLDP extension MIB

D.5.3 Relationship to other MIBs

D.5.4 Security considerations for IEEE 802.1 LLDP extension MIB module

Annex Z (informative)Outstanding issues

<<Editor's Note: This annex documents issues that have yet to be resolved during the balloting process, along with their resolution status where appropriate. It will be removed prior to Sponsor ballot.>>