P802.1Qcz
March 11, 2019
(Individual contribution)

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

Unapproved draft

Prepared by Paul Congdon

Editors' Foreword

This draft standard is the subject of an amendment project. The scope of changes to the initial standard as modified by its standardized amendments is thus strictly limited, as detailed in the PAR.

Information on participation in this project, and in the IEEE 802.1 Working Group can be found here.

This second draft is an individual contribution issued for Task Group discussion.

Participation in 802.1 standards development

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

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

http://ieee802.org/1/

Use of the email distribution list is not presently restricted to 802.1 members, and the working group has a policy of considering 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. The email lists exist primarily to allow the members of the working group to develop standards, and are not a general forum. All contributors to the work of 802.1 should familiarize themselves with the IEEE patent policy and anyone using the mail distribution will be assumed to have done so. Information can be found at http://standards.ieee.org/db/patents/

Comments on this document may be sent to the 802.1 email exploder, to the Editor of this individual contribution, or to the Chairs of the 802.1 Working Group and Time Sensitive Networking Task Group.

Paul Congdon	János Farkas	Glenn Parsons
Editor	Chair, 802.1 TSN Task Group	Chair, 802.1 Working Group
paul.congdon@tallac.com	janos.farkas@ericsson.com	glenn.parsons@ericsson.com

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

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

As part of our IEEE 802 process, the text of the PAR and CSD (Criteria for Standards Development, formerly referred to as the 5 Criteria or 5C's) is reviewed on a regular basis in order to ensure their continued validity. A vote of "Approve" on this draft is also an affirmation by the balloter that the PAR is still valid.

Project Authorization Request, Scope, Purpose, and Criteria for Standards Development (CSD)

The complete PAR, as approved by IEEE NesCom 28th September 2017, can be found at:

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

Scope:

The scope of this standard specifies Bridges that interconnect individual LANs, each supporting the IEEE 802 MAC Service using a different or identical media access control method, to provide Bridged Networks and VLANs.

Purpose:

Bridges, as specified by this standard, allow the compatible interconnection of information technology equipment attached to separate individual LANs.

Need for the Project:

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.

Contents

1. Overview ..... h
   1.3 Introduction ..... h
2. Definitions i
4. Abbreviations and acronyms ..... j
5. Conformance ..... k
6. Support of the MAC Service l
8. Principals of bridge operation ..... m
12. Bridge management ..... p
   12.1 Congestion Isolation managed objects ..... p
98. Congestion Isolation ..... r
   98.1 Congestion isolation requirements and objectives ..... s
   98.2 Identifying congested flows ..... t
   98.3 Registering congested flows with the 802.1CB stream identification function ..... t
   98.4 Congestion Isolation Entity Operation ..... t
   98.5 Congestion Isolation Protocol ..... t

Annex D v
   D.1 Requirements of the IEEE 802.1 Organizationally Specific TLV sets v
   D.2 Organizationally Specific TLV definitions v
   D.3 IEEE 802.1 Organizationally Specific TLV management ..... w
   D.4 PICS proforma for IEEE 802.1 Organizationally Specific TLV extensions x

Annex X ..... y
   X.1 Queue Markers for Order Preservation ..... z

e

Figures

Figure 98-1 Congestion Isolation Model ..... r
Figure 98-2 Congestion Isolation reference diagram ..... u
Figure X-1 Out-of-order packet example ..... y
Figure X-2 Using queue markers and counters to preserve order ..... z

Tables

Table 12.1 Congestion Isolation component managed object ..... p
Table 12.2 Congestion Isolation Point component managed object ..... q
Table D-1 IEEE 802.1 Organizationally Specific TLVs v

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]

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.

The editing instructions are shown in ***bold italic***. Four editing instructions are used: change, delete, insert, and replace. ***Change*** is used to make corrections in existing text or tables. The editing instruction specifies the location of the change and describes what is being changed by using ~~strikethrough~~ (to remove old material) and underscore (to add new material). ***Delete*** removes existing material. ***Insert*** adds new material without disturbing the existing material. Deletions and insertions may require renumbering. If so, renumbering instructions are given in the editing instruction. ***Replace*** is used to make changes in figures or equations by removing the existing figure or equation and replacing it with a new one. Editing instructions, change markings, and this NOTE will not be carried over into future editions because the changes will be incorporated into the base standard.

1. Overview

1.3 Introduction

Insert the following text after the eleventh paragraph of 1.3 and renumber accordingly

This standard 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:

bd) Defines a means for VLAN-aware Bridges that support congestion isolation for identifying flows that are creating congestion.

be) Defines a means for adjusting transmission selection for packets of congested flows

bf) Provides for a means for discovering peer VLAN-aware Bridges and stations that support congestion isolation

bg) Defines a means for signaling congestion isolation to supporting peer Bridges and stations.

2. Definitions

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

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

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.

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

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 and acronyms

Insert the following acronym(s) and abbreviation(s), in the appropriate collating sequence:

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)
a) Support Congested Traffic Enhanced Traffic Selection (yy.y.y)
a) 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. Principals of bridge operation

8.6.5 Flow classification and metering

Replace the last sentence of the third paragraph with to following:

Item e), specifying a connection_identifier, is only applicable to bridges that support PSFP and/or congestion isolation.

Insert the following clause after 8.6.5.1:

8.6.5.2 Congestion Isolation flow classification

A Bridge or an end station may support Congestion Isolation that allows traffic class modification of congested flows and subsequent frame queuing decisions (8.6.6.2) to be made on a per-stream basis for received frames.

Support of Congestion Isolation requires implementation of the Stream identification function specified in Clause 6 of IEEE Std 802.1CB-2017 [B14], as the stream_handle provided by this function is used to identify frames from congested flows and queuing decisions taken by congestion isolation.

8.6.5.2.1 Congestion isolation use of the stream filter instance table

Congestion isolation uses a single stream filter instance from the PSFP stream filter instance table (8.6.5.1.1) to select a stream gate instance that will modify the priority of congested flows. The stream filter instance for congestion isolation contains the following elements:

a) A *stream filter instance identifier*. An integer value that uniquely identifies the filter instance. Since there is a single filter instance for each monitored non-congested traffic class, the ordinal value of this identifier is insignificant.
b) A *stream_handle specification*. A common stream_handle value, as specified in IEEE Std 802.1CB.
c) A priority specification. A single priority value that specifies the monitored non-congested traffic class.
d) A *stream gate instance identifier*. Identifies a stream gate instance that is configured specifically for congestion isolation. The stream gate allows congestion isolation to specify the internal priority value of the congested traffic class for the matching stream filter instance. The state of the stream gate will always be open for congestion isolation.
e) There will be no *filter specifications* used by congestion isolation.
f) A count of frames matching both the stream_handle and the priority specification. The other frame counters specified by the stream filter instance table (8.6.5.1.1) are not used by congestion isolation.
g) The S*treamBlockedDueToOversizeFrameEnable* parameter that is set FALSE to disable this function not used by congestion isolation.

The stream_handle and priority parameters associated with a received frame select the stream filter instance of congestion isolation for a particular monitored non-congested traffic class. The purpose of the stream filter is to select the stream gate instance that will modify the priority of congested flow frames.

8.6.5.2.2 Congestion isolation use of the stream gate instance table

Congestion isolation uses a single stream gate instance from the PSFP stream gate instance table (8.6.5.1.2) to modify the priority of congested flows. The single stream gate instance for congestion isolation contains the following elements:

a) A *stream gate instance identifier*. An integer value identifying the stream gate instance
b) An operational and an administrative *stream gate state* (8.6.10.4, 8.6.10.5, 12.31.3) that is set to Open for congestion isolation.
c) An operational and an administrative *internal priority value* specification (IPV, 8.6.10.6, 8.6.10.7, 12.31.3). The IPV is used in place of the priority value associated with the frame to determine the frame's traffic class, using the Traffic Class Table as specified in 8.6.6.
d) The *GateClosedDueToInvalidRxEnable* parameter that is set to FALSE to disable this function not used by congestion isolation.
e) The *GateClosedDueToOctetsExceededEnable* parameter that is set to FALSE to disable this function not used by congestion isolation.
f) A null *stream gate control list* since this feature is not used by congestion isolation.

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.

Rename clause 8.6.6.1 as follows:

8.6.6.1 PSFP and Congestion Isolation queuing

Replace the first paragraph of 8.6.6.1 with to following paragraph:

If PSFP (8.6.5.1) or Congestion Isolation (8.6.5.2) are supported, and the IPV associated with the stream filter that passed the frame is anything other than the null value, then that IPV is used to determine the traffic class of the frame, in place of the frame's priority, via the Traffic Class Table specified in 8.6.6. In all other respects, the queuing actions specified in 8.6.6 are unchanged.

Add the following new clause after 8.6.8.4:

8.6.8.5 Enhancements for congestion isolation

A Bridge or an end station may support enhancements to isolate the frames of congested flows to a designated congested traffic class. In order to meet the ordering requirements of 8.6.6, when a frame's priority is modified by the congestion isolation stream gate instance (8.6.5.2.2), a transmission gate is associated with each monitored non-congested queue and each congested queue. The state of the transmission gate determines whether or not queued frames can be selected for transmission. For a given queue, the transmission gate can be either *Open* or *Closed* as described in 8.6.8.4.

Congestion isolation specifies that congested and non-congested queues use the same transmission selection algorithm. Additionally, the congested queues have lower priority than the monitored non-congested queues. When the transmission selection algorithm is strict priority, the state of the transmission gate is permanently *open*. When the transmission selection algorithm is anything other than strict priority the state of the transmission gate is controlled by the state machines specified in (8.6.11)

Add the following new clause after 8.6.10:

8.6.11 Congestion Isolation transmission gate state machines

<<Editor's note: lots of new text, or simply punt on this detail and refer to the example in Annex X? The preference would be to simply describe the transmission gate control variables, but not provide a detailed state-machine that sets them.>>

8.6.11.1 Isolate state machine

8.6.11.2 De-isolate state machine

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

12.1 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 (12.1.1)
b) Congestion Isolation Point (CIP) component managed object (12.1.2)

12.1.1 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 12-1.

Table 12.1—Congestion Isolation component managed object

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

12.1.1 Congestion Isolation Point 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 12-2.

Table 12.2—Congestion Isolation Point component managed object

Name	Data type	Operations supported	Conformance	References
cipMonitoredQueues	Unsigned integer [0..255]	RW		XX.XX
cipCongestedQueue	Unsigned integer [0..7]	RW		XX.XX
cipMinHeaderOctets	Unsigned integer	RW		XX.XX
cipTransmittedCims	counter	R		XX.XX
cipReceivedCims	counter	R		XX.XX

98. Congestion Isolation

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, in a lossless environment, cannot prevent priority based flow control (PFC) 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 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 application control and synchronization.

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 lower priority traffic class and signals to the upstream neighbor to do the same. CI effectively moves the congested flows out of the way, temporarily delaying the delivery of congested frames, while the higher-layer congestion feedback control loop has time to take effect.

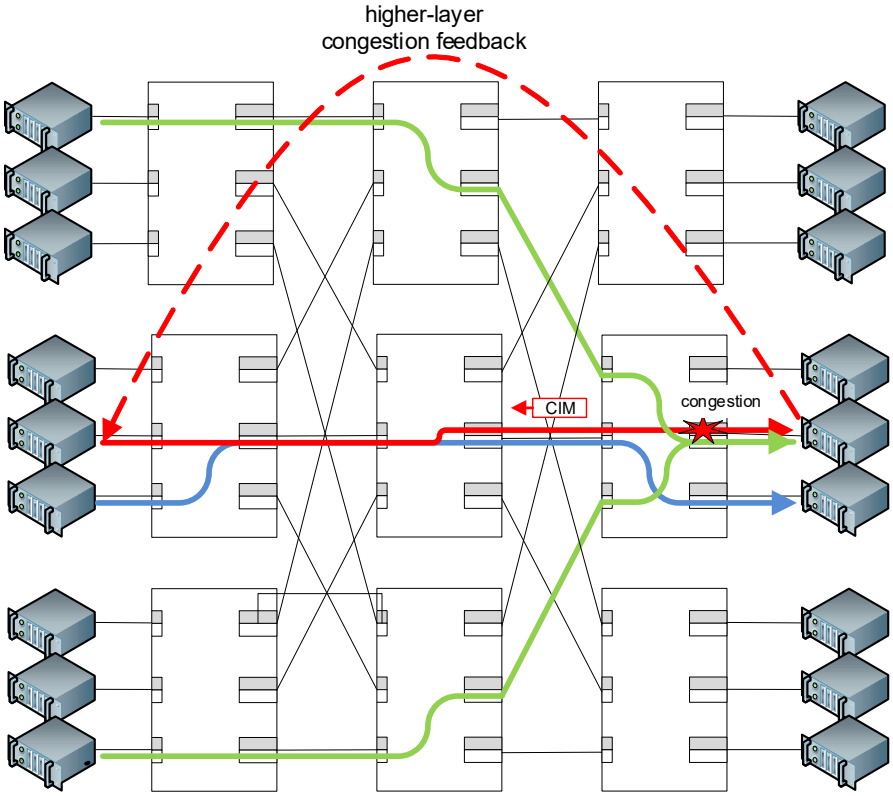

Figure 98-1—Congestion Isolation Model

Figure 98-1 shows the operation of CI. In the figure, server to server traffic is flowing from left to right across the CLOS fabric. When flows unfortunately collide at the egress port of a bridge, congestion is detected, and the offending flows are identified. Higher-layer congestion feedback is provided by the destination system, but takes time to impact the rate of injection by the source system. 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) containing flow description information

necessary for the upstream bridge to identify the same congested flow. The upstream bridge also isolates the same flow by placing subsequent packets into the lower priority congested queue 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

This clause introduces the concepts and protocols essential to congestion isolation as follows:

a) The requirements and objectives for congestion isolation (98.1)
b) Methods for identifying congested flows
c) Registering congested flows with the 802.1CB stream identification function
d) Process for modifying the priority of congested flows
e) Signaling congestion isolation messages to peers
f) Relationship and comparison with Congestion Notification

98.1 Congestion isolation requirements and objectives

The operation, procedures and protocols of congestion isolation are designed to meet the following objectives by category:

- Functionality
  1) With high probability, identify the flows that are causing congestion
  2) Quickly adjust the traffic class of congested flows
  3) Avoid head-of-line blocking of non-congested flows by signaling to upstream peers the information needed to isolate the same congested flows
  4) Reduce the frequency of invoking PFC in a lossless environment
- Compatibility
  1) Work in legacy environments by automatically detecting legacy peers and automatically disabling functionality
  2) Work in existing lossless environments using Priority-based flow control without requiring additional traffic classes
  3) Work in conjunction with higher-layer end-to-end congestion control protocols (e.g Explicit Congestion Notification, RDMA over Converged Ethernet, QCN)
  4) Coexist with existing traffic scheduling paradigms on other traffic classes
- Performance
  1) Reduce average flow completion times across the data center network
  2) Reduce the amount of pause time when PFC is enabled
  3) Reduce overall frame loss when PFC is not enabled
  4) Reduce head-of-line blocking of victim flows at upstream peers from PFC
  5) Reduce overall congestion control signaling
  6) Increase link utilization
- Scale
  1) Work in arbitrary data center network topologies with a mix of link speeds
  2) Limit messaging overhead by restricting message propagation to hop-by-hop
  3) Reduce stream identification table requirements by only requiring the registration of congested flows and providing facilities to rapidly remove flows that are no longer congested.

— Implementation complexity
  1) Limit the impact of existing traffic selection algorithms
  2) Achieve the benefits of congestion isolation without additional buffer requirements
  3) Support implementations of existing traffic classes
  4) Leverage existing standard functionality for congested flow identification and stream identification
— Manageability
  1) Only require a small set of configuration parameters which are consistent across multiple bridge deployments
  2) Eliminate the ability to configure an inoperable environment
  3) Provide auto discovery of peer capabilities using existing LLDP messages and without creating additional hello and auto-configuration protocols

98.2 Identifying congested flows

An essential step in the process of congestion isolation is the identification of congested flows by an Active Queue Management (AQM) scheme. There are many potential methods of identifying congested flows and interoperable implementations can exist using different approaches. IEEE Std 802.1Q defines the CP algorithm (30.2.1) for detecting congested controlled flows in congestion aware bridges. This approach may be used to detect congested flows in a congestion isolation aware system. A number of other possible approaches, including those that support the end-to-end ECN congestion control, are discussed in IETF RFC 7567.

Once a congested flow has been identified, it is necessary for the implementation to assert the ciCongestedFlow variable and provide the initial 64 bytes of the frame immediately following the MAC header. The ciCongestedFlow variable along with the contents of the frame are used to invoke the CIM send statemachine (x.x.x) to cause the transmission of a CIM message to the upstream congestion isolation peer.

98.3 Registering congested flows with the 802.1CB stream identification function

98.4 Congestion Isolation Entity Operation

<< Editor's Note: The text should cover the following:

a) Congestion Isolation aware Bridge Forwarding Process diagram (see next slide)
b) Congestion Isolation Point (CIP)
c) Congestion Isolation Input Multiplexer - how CIMs are decoded and received.
d) Congested Flow Identification and Table

>>

98.5 Congestion Isolation Protocol

<< Editor's Note - This section needs to include:

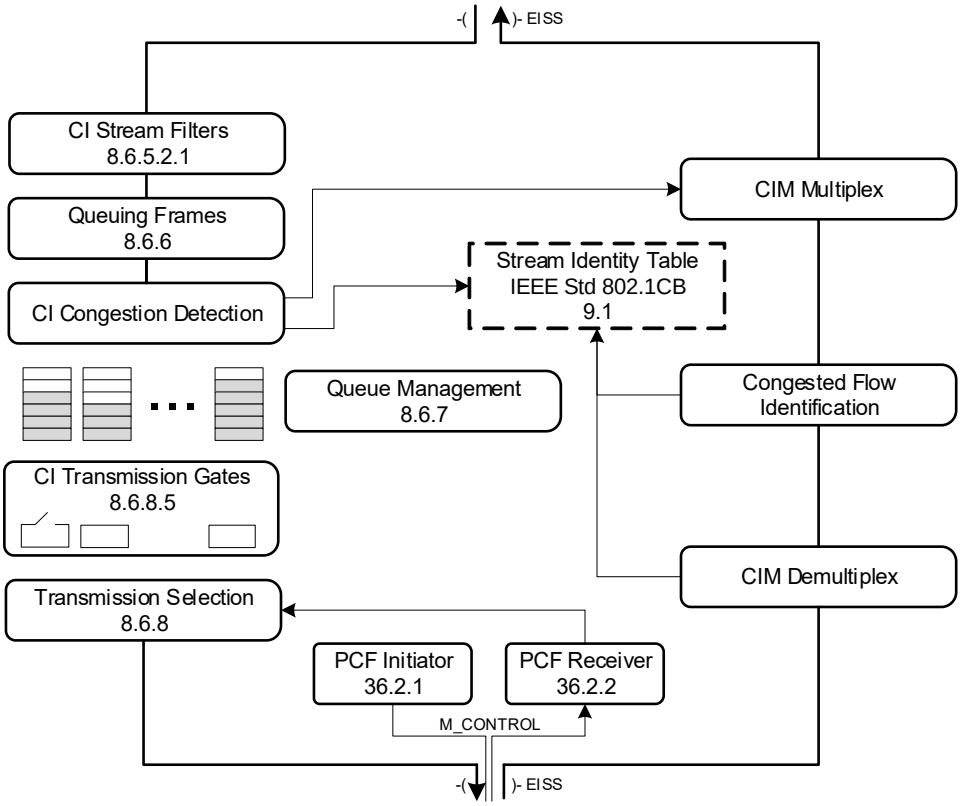

Figure 98-2—Congestion Isolation reference diagram

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

>>

98.5.1 Congestion Isolation LLLDP TLV

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 98

<< 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.

Figure D-16 shows the VLAN Name TLV format.

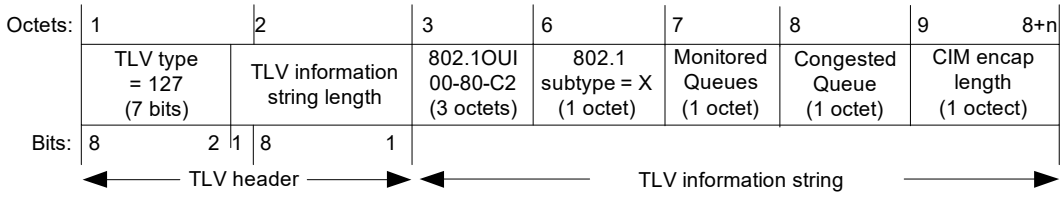

Figure D-16—VLAN Name TLV format

<< Editor's Note: the subtype value of X will be assigned at sponsor ballot >>

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

The TLV information string length field of the Congestion Isolation TLV is fixed and shall contain the value 7.

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.2.16.5 CIM encap length

A single octet unsigned integer representing the requested length in octets of the data from the frame of a congested flow to be encapsulated into a CIM message by a peer. The default value is 64.

D.2.16.6 Congestion Isolation TLV usage rules

The priority of the congested queue shall be lower than the priority of all monitored queues.

D.3 IEEE 802.1 Organizationally Specific TLV management

D.3.2 IEEE 802.1 managed objects—TLV variables

Add the following clauses to the end of D.3.2

D.3.2.11 Congestion Isolation TLV managed objects

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

D.4 PICS proforma for IEEE 802.1 Organizationally Specific TLV extensions

D.4.3 Major capabilities and options

Add the following to the end of D.4.3

Item	Feature	Status	References	Support
ciSet	Is the IEEE 802.1 Organizationally Specific TLV ciSet implemented?	O.1	D.1, Table D-1	Yes [ ] No [ ]
ciQueuePri	Are the monitored queues higher priority than the congested queue?	ciSet:M	D.2.16.6	Yes [ ]

Annex X

(informative)

Maintaining Packet Order with Congestion Isolation

The process of congestion isolation involves identifying the packets of a congested flow and subsequently modifying the egress traffic class of those packets based on the level of congestion in the monitored and congested queues. During this process, it is possible that packets for the same flow can exist in multiple queues at the same time, resulting in the possibility of an out-of-order packet delivery. As an illustration, consider the following example depicted in Figure X-1.

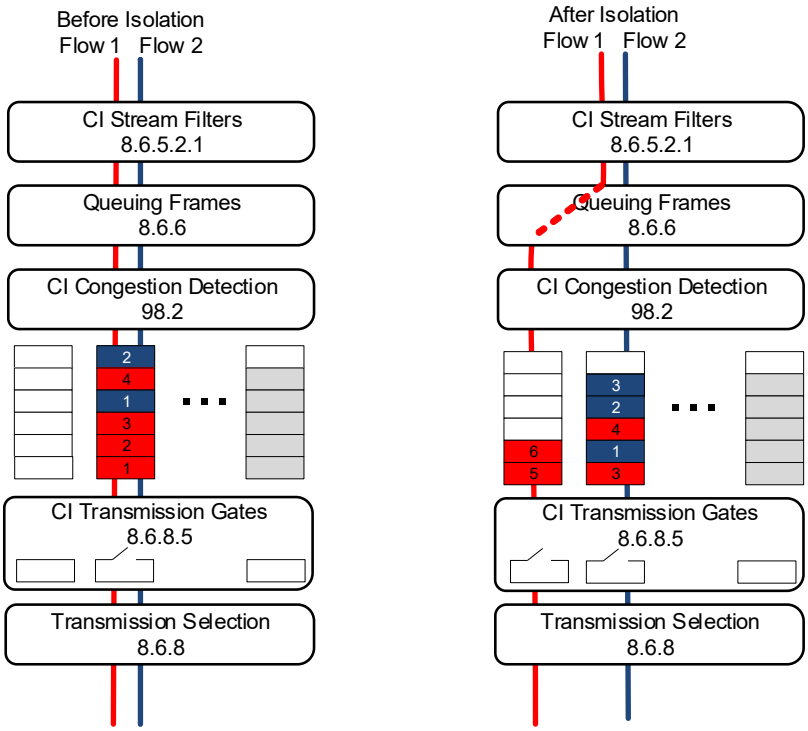

Figure X-1—Out-of-order packet example

In the example in Figure X-1, the packets of two flows, red and blue, are intermixed while traversing a common monitored queue. As the monitored queue fills, the red flow is determined to be a congested flow and subsequent packets of that flow will be reclassified and queued in the congested queue. Previously received packets for that flow, numbered 1 through 6, may reside in the monitored queue. Since the congested queue is empty, the subsequent packets, numbered 7 and 8, are placed at the head of the congested queue. Depending upon the traffic selection algorithm, it may be possible for packets 7 and 8 to be selected for transmission before some of packets 1 through 6.

The priority of the congested queue is intended to be lower than the priority of the monitored queue. The strict priority traffic selection algorithm (8.6.8.1) will assure that no out-of-order packet delivery occurs, however there is a risk of starvation for congested flows and alternative traffic selection algorithms may be desired. The enhanced transmission selection algorithm (8.6.8.3) or other vendor specific algorithms may not assure out-of-order delivery on their own.

Congestion isolation defines a transmission gate for the monitored and congested queues that make those queues available to the transmission selection algorithm. When the transmission selection algorithm is strict priority, the transmission gate is permanently open. The transmission gate is controlled by the ciGateControl variable for other transmission selection algorithms that can not assure out-of-order delivery. The management of ciGateControl is implementation dependent, but must be asserted in a way to assure the externally visible behavior of the bridge supporting congestion isolation is to maintain packet order.

The following informative description provides an example mechanism to preserve packet order for transmission selection algorithms other than strict priority.

X.1 Queue Markers for Order Preservation

The mechanism below provides control of the ciGateControl variable for a congested queue in order to preserve the order of packets for a contested flow. It involves a queue position marker and marker counter for both the monitored and congested queues. The mechanism is described using the example shown in Figure X-2.

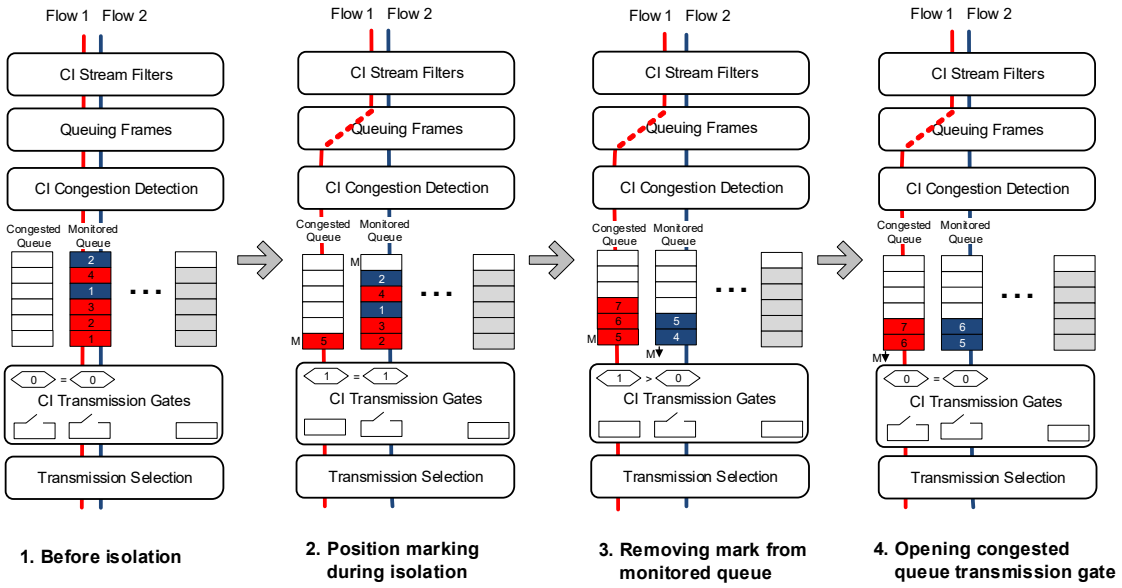

Figure X-2—Using queue markers and counters to preserve order

The example in Figure X-2 shows the state of the position markers and marker counters during four different phases of congestion isolation operation; before a flow is isolated, position marking during the isolation of a flow, the closing of the congested queue transmission gate and the opening of the congested queue transmission gate. When a queue transmission gate is open, that queue is available to the traffic selection algorithm. When it is closed, the queue is not available to the traffic selection algorithm.

In the example, the packets of two flows are intermixed in the monitored queue as they traverse the switch. Since no flows have been isolated yet, the position marker counters of both the congested queue and the monitored queue are set to 0. Once a flow has been isolated and subsequent packets of that flow are placed in the congested queue, a marker is placed in both the congested queue and the monitored queue. The marker counter is incremented for both queues once the position is marked. The congested queue is empty at the time the first flow is isolated, so the marker will be at the head of the queue. When a marker is at the head of the congested queue and the marker counters are equal, the ciGateControl variable for the congested queue is set to closed. The monitored queue continues to drain and eventually the position marker will reach the

head of the monitored queue. The marker counter for the monitored queue will be decremented when the packet associated with the position marker is scheduled for transmission and exits the queue. When the value of the congested queue marker counter is greater than the monitored queue position counter, it is possible to set the ciGateControl variable to open and begin to schedule the congested queue. When ever a packet that aligns with a position marker is scheduled for transmission, the associated marker counter is decremented.