Consideration of SFC in 802.1Q

Paul Congdon (Huawei)
JK Lee (Intel)
Lily Lv (Huawei)

January 27, 2022

Outline

- Brief review of proposal
- Proposed scope of work
- Considerations of 802.1Q clauses

Future 802.1 Congestion Management Tools

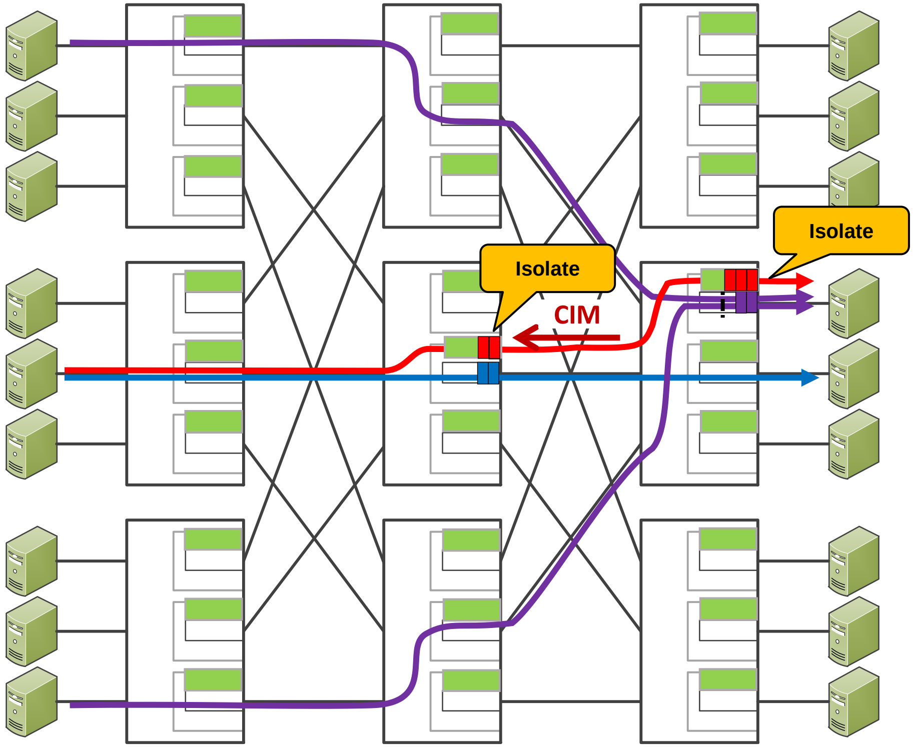
Congested Flow →

Victim Flow →

Congested Queue 🟩

Normal Queue □

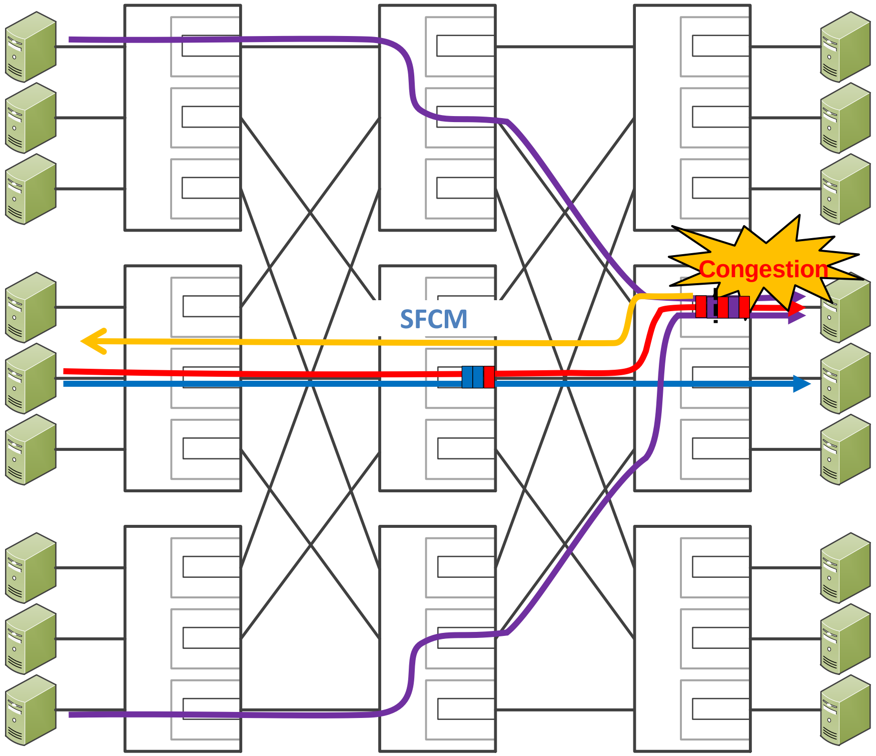
P802.1Qcz - Congestion Isolation



Implementation details

- Congesting flows are isolated locally first
- As queues continue to congest, CIM is generated and sent to upstream bridge/router
- CIM can be L2 or L3 message to support L3 networks (common deployment model).

Source Flow Control



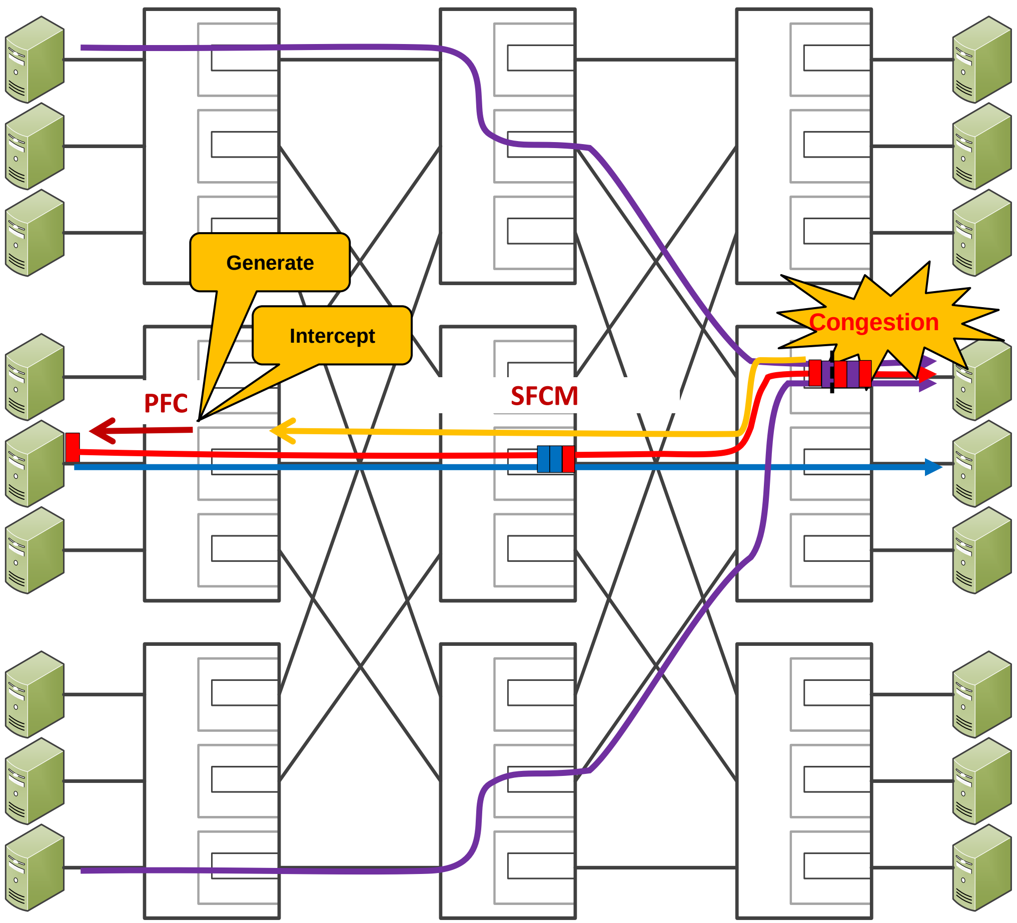
Details

- Can be combined with Congestion Isolation
- If congestion persists, Edge-to-Source signaling using L3 message
- Somewhat like a L3 version of 802.1Qau (L3-QCN), but no Reaction Point (RP) rate controller defined – instead, this is Flow Control
- Optional source Top-of-Rack switch involvement (see next slide)

Top-of-Rack Source Flow Control (proxy)

- Important use case for early deployment.
- ToR intercepts SFCM at egress port connected to non-supporting host using an egress stream_filter matching SFCM UDP port number
- ToR generates traditional PFC frame from SFCM

Source Flow Control (ToR Proxy)



What is needed in the SFCM signaling message?

- Ability to identify the offending ‘flow’ of congestion
    - Source and destination IP addresses of the data pkt
        - SRC IP for reverse forwarding
        - (Optional) DST IP for caching pause time per dst IP at source ToR
        - simply swap src IP <-> dst IP from the data pkt into the signal packet
    - Other L4 tuples (TCP/UDP port number, L3 protocol number)
- DSCP and/or PCP, as needed to identify the PFC priority @ source NIC
- Pause time duration **<=** minimal drain time to reach the target queue level
- A congestion locator such as Topology Recognition level to identify ‘incast’ congestion verses ‘in-network’ congestion.

Levering Qcz Congestion Isolation Message (CIM)

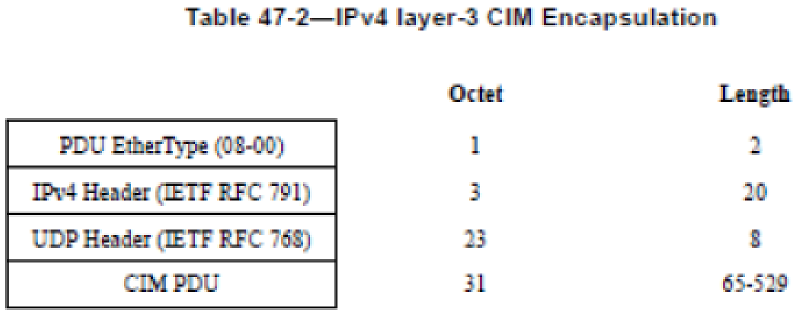


Table 47-2—IPv4 layer-3 CIM Encapsulation

	Octet	Length
PDU EtherType (08-00)	1	2
IPv4 Header (IETF RFC 791)	3	20
UDP Header (IETF RFC 768)	23	8
CIM PDU	31	65-529

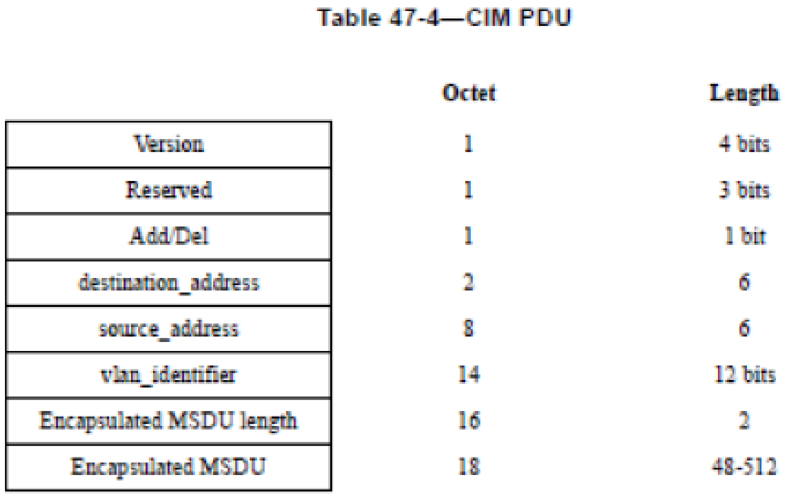


Table 47-4—CIM PDU

	Octet	Length
Version	1	4 bits
Reserved	1	3 bits
Add/Del	1	1 bit
destination_address	2	6
source_address	8	6
vlan_identifier	14	12 bits
Encapsulated MSDU length	16	2
Encapsulated MSDU	18	48-512

- Qcz CIM has Layer-2 and Layer-3 formats
- The CIM PDU contains enough of the payload to identify the offending flow
- Carrying the needed information:
  - Src / Dest IP addresses
  - DSCP
  - Additional tuples of the data pkt to identify the flow endpoint at the source
  - Congestion locator
- What’s missing?
  - Pause time
  - Simplified format of above information (i.e not MSDU)
  - Selection of SFCM Destination IP (NOT previous hop)

Scope of work summary

- Consider an amendment to 802.1Q that does the following:
  - << proposed PAR scope text >>
- Specific functionality includes:
  - Configuration elements enabling/disabling the feature (system wide)
  - Specification of SFC messages and how to generate them
  - Specification of monitoring queues for congesting flows?
  - Specification of SFCM suppression (timeout) mechanism
  - Mechanism and configuration of SFCM ToR proxy capability
  - YANG support

List of impacted 802.1Q clauses

Effort Estimation	
Small	• 1.3 Introduction
Small	• 3 Definitions
Small	• 4 Abbreviations
Small	• 5.11 System requirements for Priority-based Flow Control (PFC)
Small	• 5.4.X VLAN Bridge Conformance
Small	• 5.X SFC End-station Conformance
Small	• 8.6.6 Queuing of frames
Medium	• 12 Bridge management
Small	• 36 Priority-based Flow Control (PFC)
Large	• 48.X YANG
Large	• XX Source Flow Control
Small	• Annex A PICs
Small	• Annex B End-station PICs
Small	• Annex D 802.1 LLDP TLVs

1.3 Introduction

- NOTE: Current text related to PFC – should this be augmented to include SFC invocation?

- This standard specifies protocols, procedures, and managed objects to support Priority-based Flow Control (PFC). These allow a Virtual Bridged Network, or a portion thereof, to enable flow control per traffic class on IEEE 802 point-to-point full-duplex links. To this end, it
  - bh) Defines a means for a system to inhibit transmission of data frames on certain priorities from the remote system on the link.
  - bi) Defines a means for a system to inhibit transmission of data frames on certain priorities from the remote system on a remote link supporting Source Flow Control

1.3 Introduction

- Since CI and SFC are separate mechanisms, assume we augment the introduction with a new base paragraph and bullets

- This standard specifies protocols, procedures and managed objects that support Source Flow Control (SFC) within data center environments. This is achieved by enabling systems to individually identify flows creating congestion and signaling to end stations supporting SFC to avoid frame loss. Source Flow Control may invoke Priority-based Flow Control (PFC) on Bridges attached to stations that do not support SFC but support only PFC. To this purpose it:
  - Xx) TBD

Definitions

- Some Proposed Needed Terms:
  - **Source Flow Control Aware System:** A Bridge Component conforming to the source flow control provisions of this standard.
  - **Source Flow Control Message (SFCM):** A message transmitted by a Source Flow Control Aware System, conveying the need to pause a traffic class associated with a flow.

- Is this needed?
  - **Source Flow Control Point (SFCP):** A Virtual Local Area Network (VLAN) Bridge that monitors a set of queues for Congesting Flows and generates Source Flow Control Messages.
  - NOTE: an end-station doesn’t need SFCP – correct?

- Edit existing definition
  - **3.2 Congesting 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 or a Source Flow Control Aware System.

Abbreviations

- Some Possible Abbreviations:

SFC	Source Flow Control
SFCM	Source Flow Control Message

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

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

  g) Support enabling PFC on up to eight priorities per port.

  h) Support the IEEE8021-PFC-MIB (17.7.17).

  i) Support automatic configuration of PFC buffer requirements for lossless operation.

  j) Support invoking PFC at source ToR proxy as the result of processing received SFC messages.

Conformance

- 5.4.X VLAN Bridge requirements for source flow control
  - A VLAN-aware Bridge implementation that conforms to the provisions of this standard for source flow control (XX, XX+1) shall:
    a) Support, on one or more Ports, the creation of at least one Source Flow Control Point (xx.x.x);
    b) Support per-stream classification and metering for SFC as specified in (8.6.5.xx.xx).
    c) Support, at each Source Flow Control Point, the generation of Source Flow Control Messages (xx.x);
    d) Support the ability to configure the variables controlling the operation of each Source Flow Control Point (xx.x.x);
  - A VLAN Bridge implementation that conforms to the provisions of this standard for source flow control may:
    f) Conform to the required capabilities of IEEE Std 802.1AB
    g) Support DCBX (Clause 38).
    h) Support invoking PFC on links attached to stations only supporting PFC (xx.xx)
    i) Support Topology Recognition (47.5).
    j) Support the Source Flow Control YANG model (xx.x.x)

- 5.XX End station requirements for source flow control
  - TBD shall and may statements like above

8.6.6 Queuing of frames

- We need to include text (like Congestion Notification and Congestion Isolation) about SFCM frame generation out of the Forwarding process. This could be done as follows:
  - In a congestion aware Bridge (Clause 30) ~~or~~, a congestion isolation aware Bridge (Clause 47) or a source flow control 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 or a SFCM. The CNM ~~and~~, CIM and SFCM are injected back into the Forwarding Process (8.6.1) as if it had been received on that Bridge Port.

Clause 12 Management

- Insert Source Flow Control objects into the existing list of managed objects in 12.1.1

  j) The ability to create and delete the functional elements of source flow control and to control their operation

- Insert Source Flow Control in the list of VLAN Bridge Objects of 12.2

  p) The source flow control entities (12.xx).

- New clause 12.xx Source Flow Control managed objects with potential objects:

  a) SFC component managed object

  b) Source Flow Control Point (SFCP) component managed object

? Do we need to define a stream table similar to the CI Stream Table

SFC Managed Object

- A per-bridge object that includes:
  - sfcMasterEnable
    - Boolean that turns on or off the functionality
  - sfcmTransmitPriority
    - Integer between 0-7 that determines the traffic class used to send SFC messages upstream. Default 7
  - sfcmMacAddress
    - The MAC address, belonging to the system transmitting the SFCM (xx.x.x), used as the source_address of SFCMs sent from the SFCP
  - sfcmIPv4Address
    - The IPv4 address, belonging to the system transmitting the IPv4 layer-3 SFCM (XX.X.X), used as the IPv4 source address in the IPv4 header (IETF RFC 791) of IPv4 layer-3 SFCMs sent from the SFCP.
  - sfcmIPv6Address
    - The IPv6 address, belonging to the system transmitting the IPv6 layer-3 SFCM (XX.X.X), used as the IPv6 source address in the IPv6 header (IETF RFC 8200) of IPv6 layer-3 SFCMs sent from the SFCP.
  - sfcmUDPPort
    - The destination UDP port number in the UDP header (IETF RFC 768) of IPv4 and IPv6 layer-3 SFCMs sent by peers. The value is also used as the UDP source port number in layer-3 SFCMs sent to peers. The UDP port number must be selected from the range of dynamic port numbers, between 49152 and 65535, as specified in IETF RFC 6335. The port number must be currently available for use by the implementation and consistently configured for all source flow control aware systems in the data center. For example, an implementation may use UDP port 58622, if it is not currently being used by any other application in the system.
    - NOTE: Can we get IETF to allocate a well-known port now that it needs to be consistent across the entire data center?
  - sfcMaxSFCM
    - The maximum number of times a SFCM PDU will be sent for a congesting flow. The default value is 3. A larger value provides more resilience to lost SFCMs but generates more traffic on the network.
  - sfcMonitoredQueues
    - A bit mask of traffic classes that are to be monitored for congestion
  - sfcMinHeaderOctets
    - The minimum number of octets that the SFCP is to return in the Encapsulated MSDU field (xx.x.x) of each SFCM it generates (xx.xx.x). Default value 64.
- Do we need to age/remove entries in SFC Proxy bridge by SFCM?
  - sfcMaxFlowLife
    - An unsigned integer specifying the maximum number of centiseconds that a congesting flow entry, created by the receipt of a SFCM, can remain in the SFC stream table.

Source Flow Control Point (SFCP) Managed Object

- A per SFCP object that includes:
  - sfcTransmittedSFCMs
    - A counter of the number of SFCMs sent
  - Other per-port and/or per-queue objects?

36. Priority-based Flow Control (PFC)

- Updates to the overview section to describe SFC invocation?

New Clause - Principles of Operation

- Introduce the concepts essential to source flow control, including:
  - Reference diagram
  - Problems being solved
  - Requirements and objectives for the solution
  - Methods for identifying congested flows (reference to CI?)
  - Allocating and deallocating congesting flows in a congested flow table
  - Role and operation of signaling
  - Relationship and comparison with Congestion Notification and Congestion Isolation

- Start off with a reference diagram showing the problem

SFC verses Congestion Notification (Qau)

Differences

- Qau is a L2 protocol, SFC is L3
- Qau is congestion control, SFC is flow control
- Qau defines a comprehensive control algorithm with many parameters, SFC uses PFC
- Qau CNM carries Quantized Feedback for a Reaction Point, SFC carries ‘pause’ duration for PFC
- SFC allows a ToR to proxy SFCM processing

Similarities

- Congestion points monitor queues for congestion
- Congestion points send signaling messages back to source
- Flow information (from received congesting frame) is provided in signaling messages

SFC verses Congestion Isolation (Qcz)

Differences

- Qcz uses an additional traffic class to isolate frames
- Qcz signals to upstream neighbor (L2 or L3), SFC signals to end-station (also via ToR Proxy) using L3 message
- Qcz does not directly rate control the sending host, SFC pauses the sending host
- SFC allows a ToR to proxy SFCM processing

Similarities

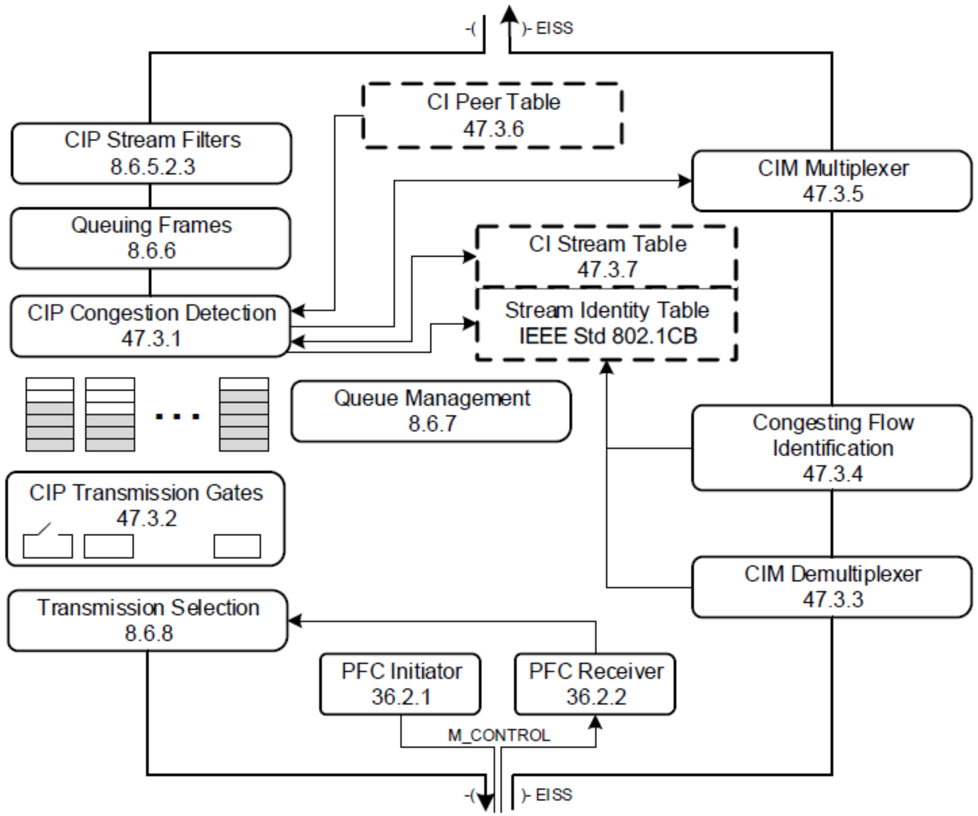
- Both schemes support L3 message formats
- Congestion points monitor queues for congestion
- Congestion points send signaling messages backward toward source
- Flow information (from received congesting frame) is provided in signaling messages

Source Flow Control Entity Operation

- Describes the architecture of the SFCP in the Forwarding Process, including: NOTE: do we combine with existing Congestion Isolation architecture diagram?
  - Source Flow Control aware Bridge Forwarding Process diagram (see next slide)
  - Source Flow Control Point (CIP)
  - Source Flow Control Flow Identification and Table
- NOTE: SFC proxy bridges will use an egress stream filter (i.e. ACL) to trap SFCM

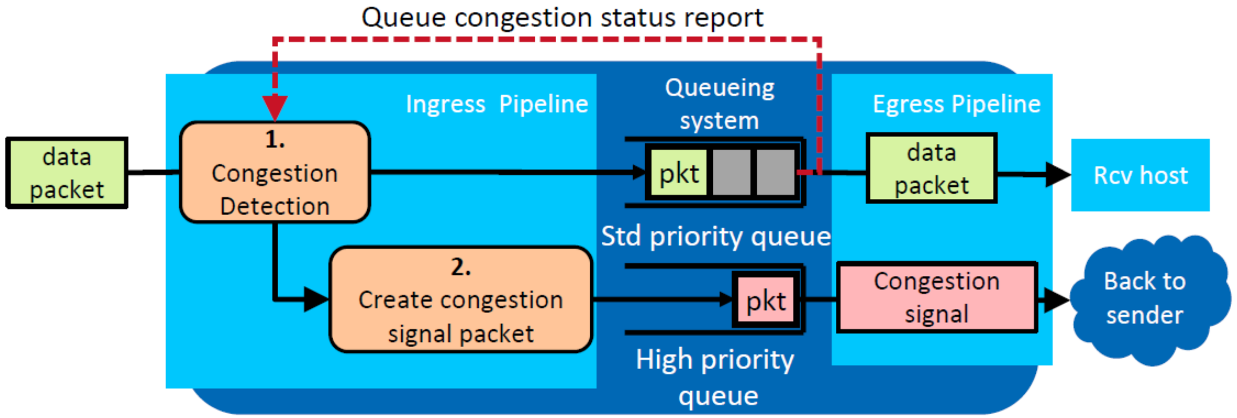
Leveraging the Qcz reference architecture

- Believe it or not, these figures are similar…



EISS
CIP Stream Filters 8.6.5.2.3
CI Peer Table 47.3.6
CIM Multiplexer 47.3.5
Queuing Frames 8.6.6
CI Stream Table 47.3.7
CIP Congestion Detection 47.3.1
Stream Identity Table IEEE Std 802.1CB
Queue Management 8.6.7
Congesting Flow Identification 47.3.4
CIP Transmission Gates 47.3.2
CIM Demultiplexer 47.3.3
Transmission Selection 8.6.8
PFC Initiator 36.2.1
PFC Receiver 36.2.2
M_CONTROL
EISS

Figure 47-2—Congestion Isolation reference diagram



- Above figure is from https://datatracker.ietf.org/meeting/112/materials/slides-112-iccrg-source-priority-flow-control-in-data-centers-00
- Congestion detection above (1) is similar to 47.3.1, but perhaps with different thresholds
- Creating SFCM frame above (2) is similar to input to CIM Multiplexer 47.3.5, but with different parameters to CIM creation (e.g. Dest IP address)
- CI Peer Table 47.3.6 is used to identify upstream bridge/router – not needed by SFC – address is in frame.
- CI Stream Table 47.3.7 could be used by Source Flow Control.
- CIM Demultiplexer 47.3.3 could be used to intercept SFCMs?

Source Flow Control Protocol

- Specifies how Source Flow Control Aware systems participate in the source flow control protocol and procedures, including:
  - Variables controlling operation
  - Associated variables and procedures that control the generation of SFCMs and the congested flow table
  - The Source Flow Control Protocol
    - Generating SFCMs
    - Creating and Deleting entries in the Source Flow Control Flow Table
    - For Proxy SFC devices, processing received SFCMs
  - Encoding of SFCM PDUs
  - DCBX operation for proxy SFC devices
  - UML and YANG model

Backup

Progress To Date

- Public presentations of the concept and data at P4 Workshops (Apr’20, May’21) and Open Fabrics Alliance (Mar’21)
    - https://opennetworking.org/wp-content/uploads/2020/04/JK-Lee-Slide-Deck.pdf (slide 12)
    - https://www.openfabrics.org/wp-content/uploads/2021-workshop-presentations/503_Lee_flatten.pdf
    - https://opennetworking.org/wp-content/uploads/2021/05/2021-P4-WS-JK-Lee-Slides.pdf (slide 14)
- Previous Nendica/TSN presentations
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf - 9/16/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0061-00-ICne-source-remote-pfc-test.pdf – 10/14/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0067-00-ICne-source-remote-pfc-status-update.pdf - 11/04/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0077-00-ICne-consideration-of-spfc-sfc-issues-when-leveraging-qcz.pdf - 12/16/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0079-00-ICne-spfc-sfc-next-steps.pdf - 12/23/2021
    - https://www.ieee802.org/1/files/public/docs2022/new-congdon-SFC-overview-0122-v01.pdf - 01/19/2022
- IETF Awareness
    - Topic raised at IEEE 802 / IETF Coordination call – 10/25/2021
    - https://datatracker.ietf.org/meeting/112/materials/slides-112-iccrg-source-priority-flow-control-in-data-centers-00 - 11/08/2021