# P802.1Qcz D0.4

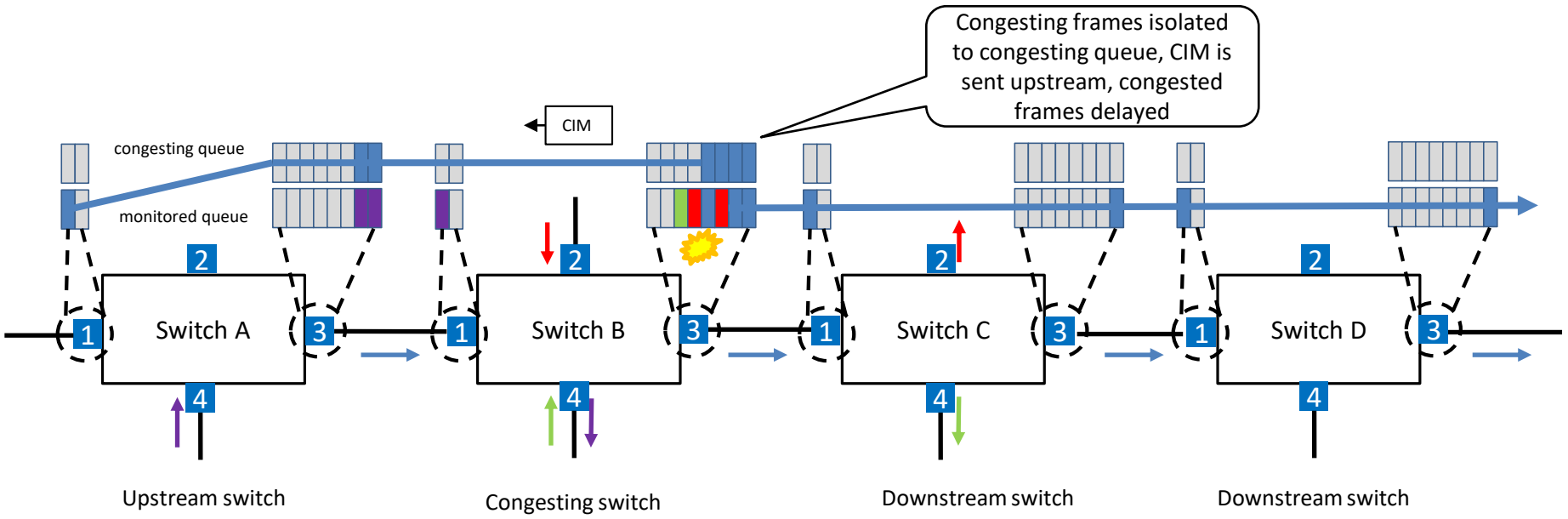
Paul Congdon

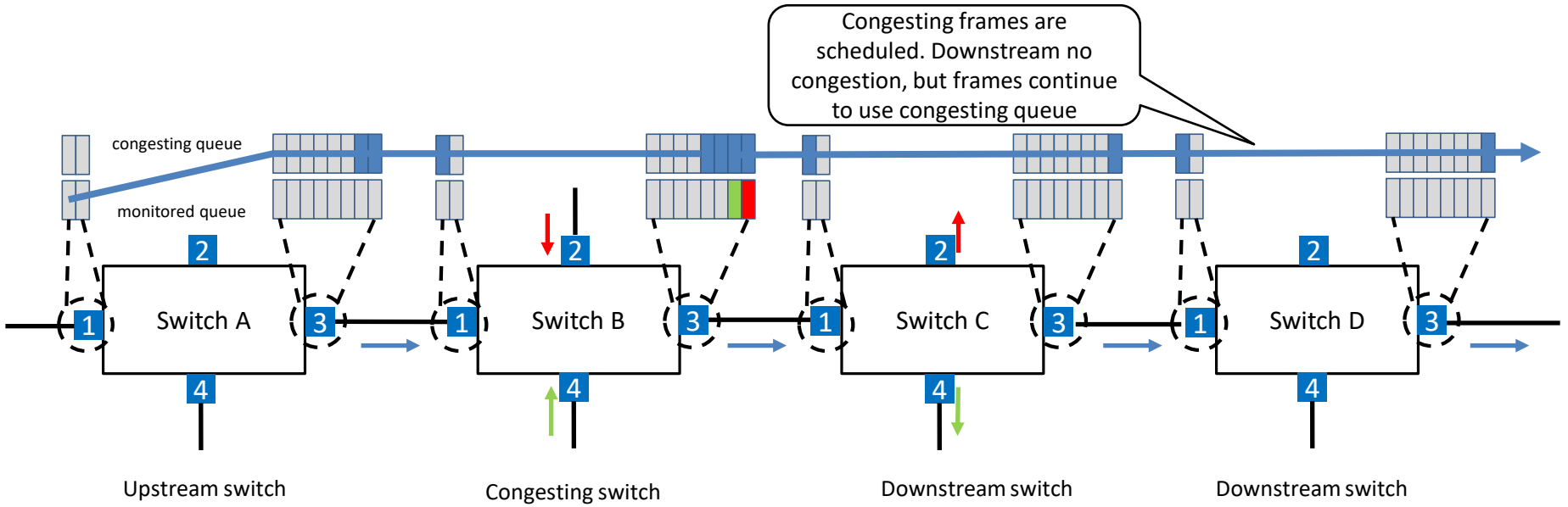
November 2019

# Agenda

- Final disposition posted:
  - http://www.ieee802.org/1/files/private/cz-drafts/d0/802-1Qcz-d0-3-dis-v01.pdf
- Draft d0.4 with latest changes for review:
  - http://www.ieee802.org/1/files/private/cz-drafts/d0/802-1Qcz-d0-4.pdf
- Major changes to discuss/review
  - Handling congesting flow downstream of the congestion root
  - New stream_handle specification type
  - CI stream table and 802.1CB stream identity table
  - Multiple Monitored/Congesting Queue pairs
  - Keeping track of why a flow as been considered congesting
  - End-station support
  - CIM Encapsulations
  - CIM PDU and the new Add/Del CIM indicator
  - Sending a CIM when a flow is removed from stream identity table
  - CIM transmission conditions
  - Knowing position in the topology using LLDP TLV

# Congesting Flow Stickiness



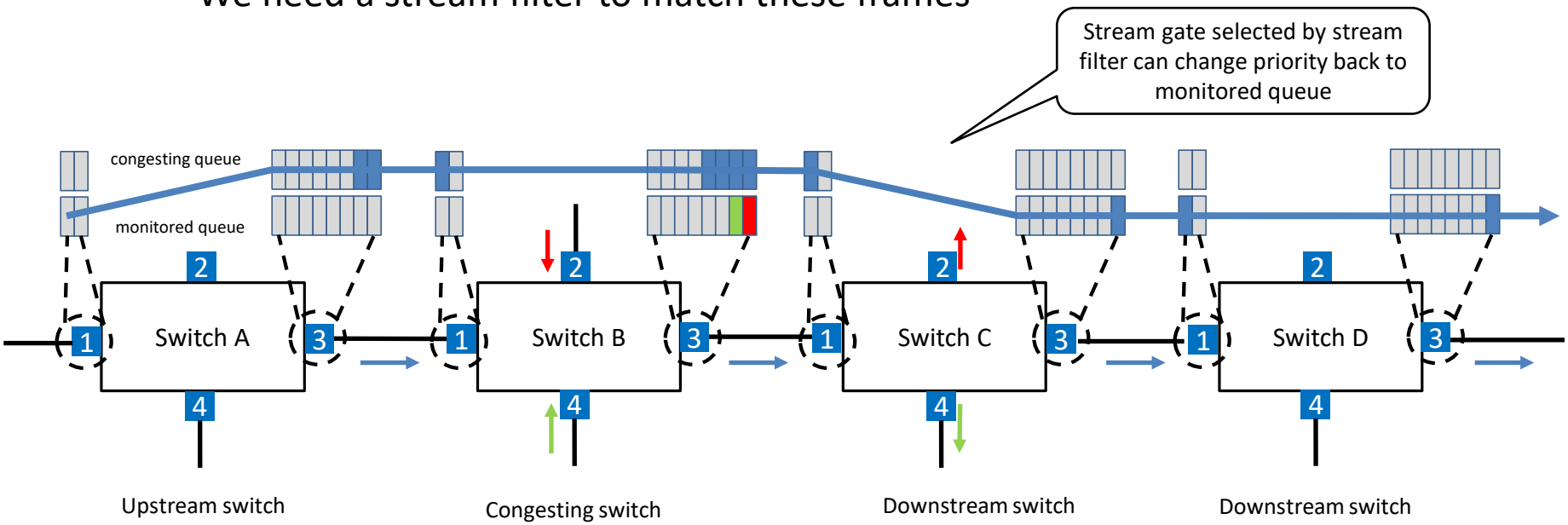



# Solution to congesting stickiness

- Downstream switches should change the priority of these frames back to use the monitored queue.
  - We need a stream filter to match these frames
- Downstream switches do NOT have an entry in the stream identity table for these streams
  - Entries are created by detecting local congestion – not present downstream
  - Entries are created by receipt of CIM – not sent downstream
- We need a new stream_handle specification to match the absence of a stream_handle

# Solution to congesting stickiness

- Downstream switches should change the priority of these frames back to use the monitored queue.
  - We need a stream filter to match these frames




- However, downstream switches do NOT have an entry in the stream identity table for these streams
  - Entries are created by detecting local congestion – not present downstream
  - Entries are created by receipt of CIM – not sent downstream
- We need a new stream_handle specification to match the absence of a stream_handle!

# Proposed changes

**8.6.5.3 Stream Identification**

*Modify item b) in list as follows:*

b) A stream_handle specification, either:

1) A single value, as specified in IEEE Std 802.1CB.
2) A wildcard, that matches any stream_handle.
3) A null, that matches when no stream_handle is provided.

**12.31.2.2 stream_handle specification data type**

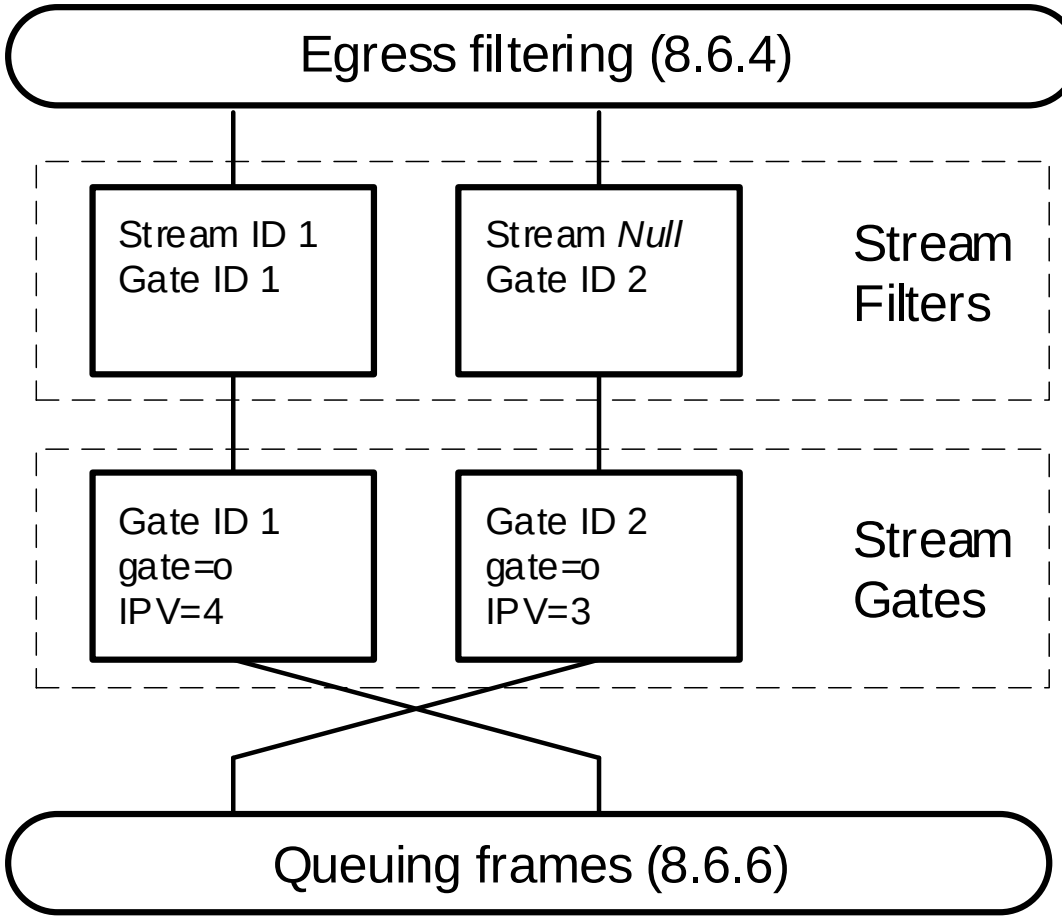
*Change the opening list as follows:*

The stream_handle specification data type allows either of the following to be represented:

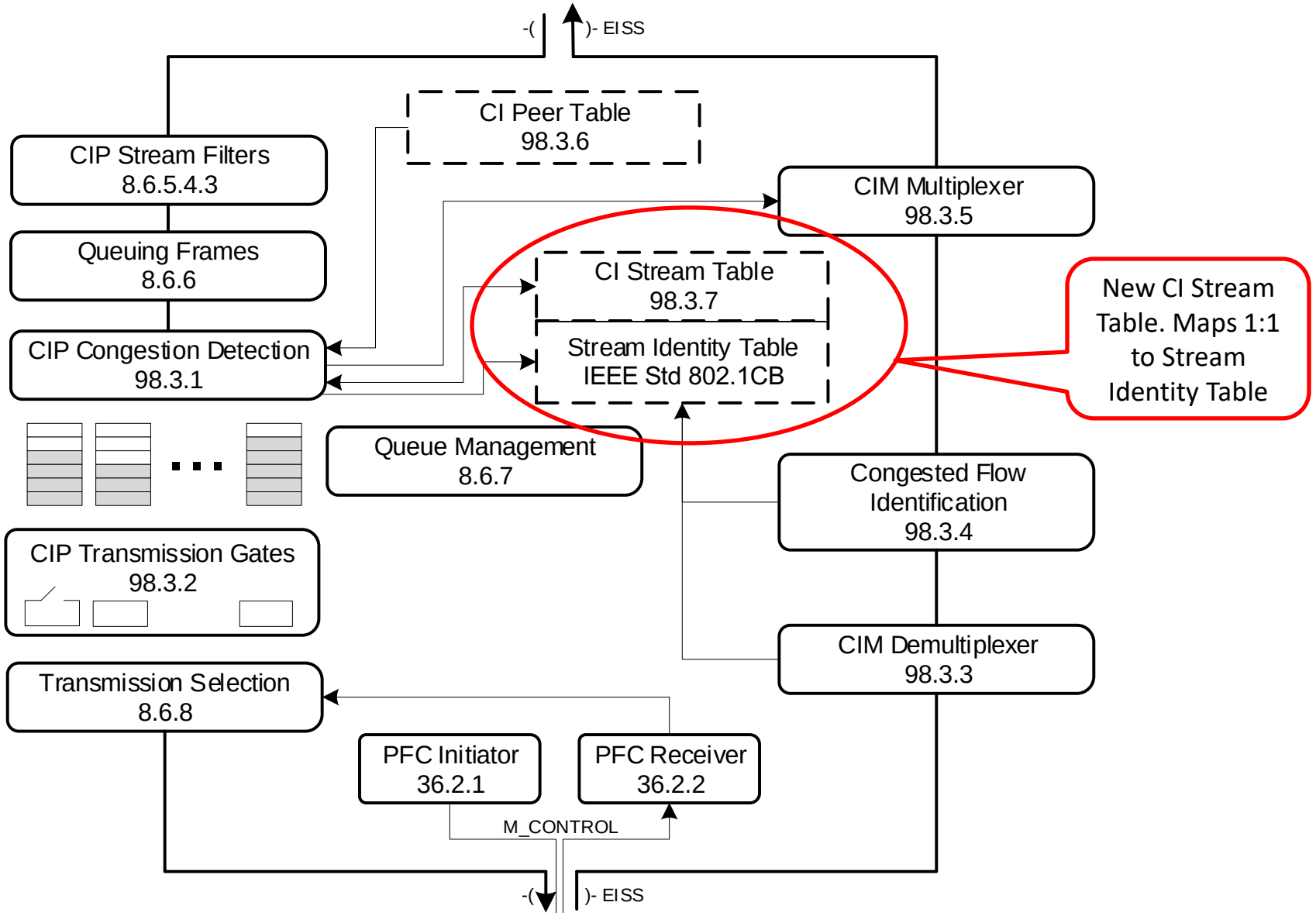
a) A stream_handle value, represented as an integer.
b) The wild card value.
c) The null value.

**Must change MIBs and YANG as well**

# New Figure 8-16 for CI



# 802.1Qcz Reference Diagram



# 12.32.3 CI Stream Table

**Table 12-3—CI stream table entry**

| Name | Data type | Operations supported | Conformance | References |
|---|---|---|---|---|
| ciStreamIdHandle | stream_handle value | RW | IEEE Std 802.1CB | 98.4.1.5.1 |
| ciCIMCount | integer | RW | | 98.4.1.5.2 |
| ciStreamCreateMask | 2-bit mask | RW | | 98.4.1.5.3 |
| ciQueueKey | integer | RW | | 98.4.1.5.4 |
| ciDestination_address | MAC address | RW | | 98.4.1.5.5 |
| ciSource_address | MAC address | RW | | 98.4.1.5.6 |
| ciVlan_identifier | 12-bit VID | RW | | 98.4.1.5.7 |
| ciMsdu | octet string (size 64-512) | RW | | 98.4.1.5.8 |

# 98.4.1.2.5 cipQueueMap[]

- Managed object that provides the mapping between monitored queues and congested queues.
- Used by ciInitialize() to set-up Stream Filters and Stream Gates.
- An array of 8 entries – one for each traffic class.
- Contents in range -8..8:
  - 0 – not participating in CI
  - Negative – maps to a congested queue (|value|-1)
  - Positive – maps to a monitored queue (value-1)
- Example; traffic class 3 is the congesting class for monitored class 4

| Index | Value |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 0 |
| 3 | -5 |
| 4 | 4 |
| 5 | 0 |
| 6 | 0 |
| 7 | 0 |

# 98.4.1.5.3 ciStreamCreateMask

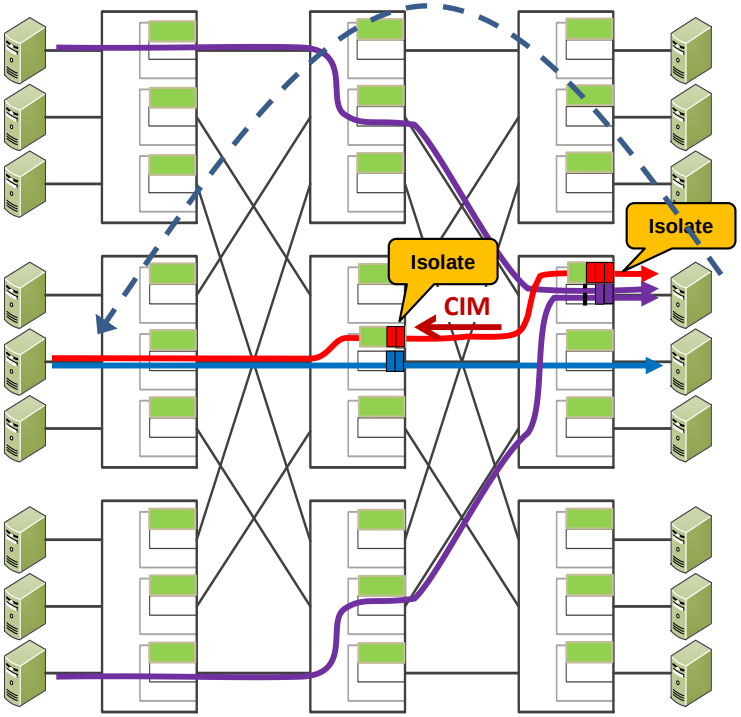
- 2-bit attribute of entries in the CI Stream Table
- LSB = entry created because of local congestion
- MSB = entry created because of receipt of CIM
- Actions:
  - Clear LSB when local congestion subsides
  - Set MSB when CIM received/refreshed
  - Clear MSB when receiving a CIM Del or not refreshed
  - Delete entries when LSB and MSB are clear

# End-station Support

- End-station queuing, stream filters and stream gates aren’t really defined for end-stations?
- Propose allowing optional sending/receiving CIM messages
- Consider future amendment to 802.1DC
- Specific text changes are TBD

# Different CIM PDU encapsulations

Typical Environment for CI



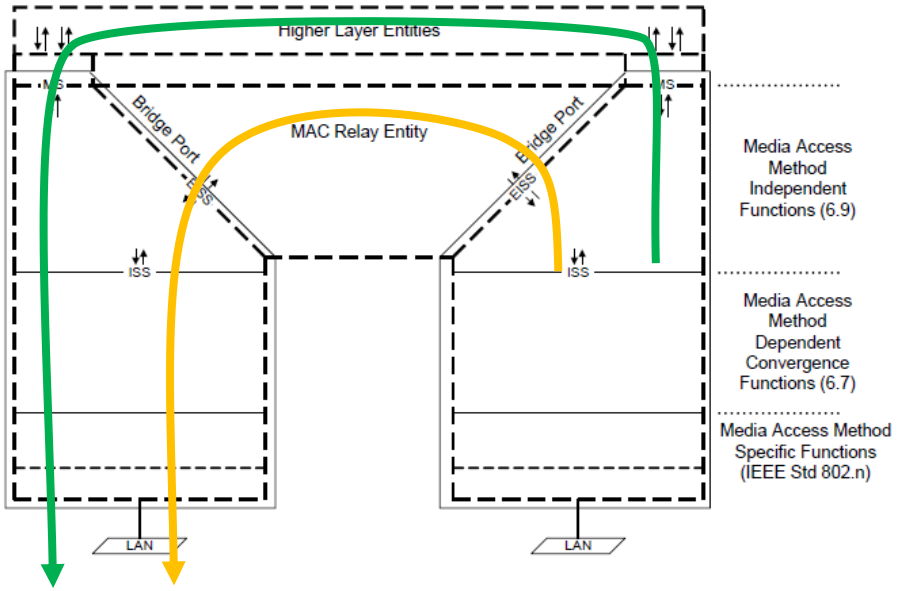

Congested Flow

Victim Flow

Congested Queue

Normal Queue

- DCN is primarily Layer-3
- CIMs are sent hop-by-hop
- CIM Multiplexer uses relay to deliver




Layer-3 CIM

Layer-2 CIM

# CIM encapsulations

**Table 98-1—Layer-2 CIM Encapsulation**

| | Octet | Length |
|---|---|---|
| PDU EtherType (XX-XX) | 1 | 2 |
| Version | 3 | 4 bits |
| Subtype | 3 | 4 bits |
| CIM PDU | 4 | 66-514 |

**Table 98-2—IPv4 Layer-3 CIM Encapsulation**

| | Octet | Length |
|---|---|---|
| PDU EtherType (08-00) | 1 | 2 |
| IPv4 Header (RFC 791) | 3 | 20 |
| UDP Header (RFC 768) | 23 | 8 |
| CIM PDU | 31 | 66-514 |

**Table 98-3—IPv6 Layer-3 CIM Encapsulation**

| | Octet | Length |
|---|---|---|
| PDU EtherType (86-DD) | 1 | 2 |
| IPv6 Header (RFC 2460) | 3 | 40 |
| UDP Header (RFC 768) | 43 | 8 |
| CIM PDU | 51 | 66-514 |

# The CIM PDU

**Table 98-4—CIM PDU**

| | Octet | Length |
|---|---|---|
| Version | 1 | 4 bits |
| Reserved | 1 | 3 bits |
| Add/Del | 1 | 1 bit |
| destination_address | 3 | 6 |
| source_address | 9 | 6 |
| vlan_identifier | 15 | 12-bits |
| Encapsulated MSDU length | 17 | 1 |
| Encapsulated MSDU | 18 | 64-512 |

**Note:** New Add/Del bit informs peer if congesting flow is being added or removed from Stream Identity Table and CI Stream Table

# CIM Del Option

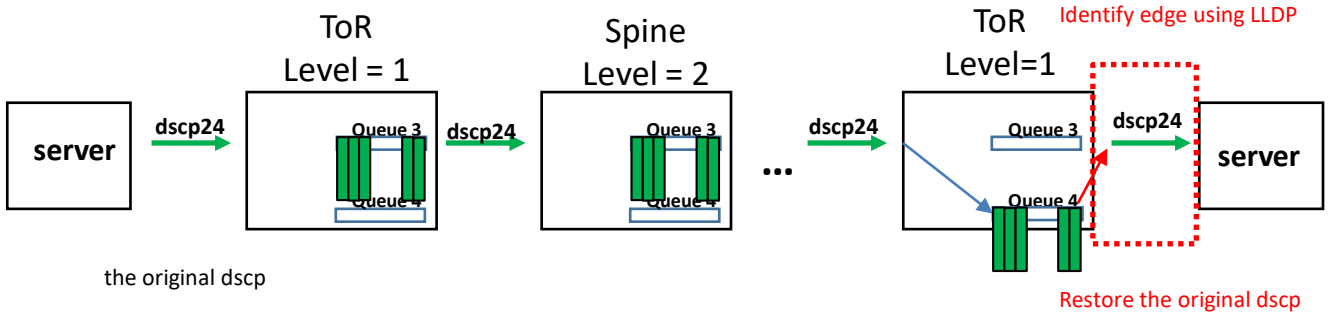
- Topics has been discussed in detail: http://www.ieee802.org/1/files/public/docs2018/cz-congdon-ci-design-topics-1118-v01.pdf
- Purpose:
  - Prevent flows from remaining congesting upstream longer than needed.
  - Allows entries to remain congested upstream when necessary downstream.
- Challenges:
  - CIM Del can be a ‘one shot’ message. If lost, no gain and requires ageing of entries in CI Stream Table.
  - Receipt of CIM Del requires a look-up in CI Stream Table to find and potentially delete entry.
- Proposal:
  - Specify this functionality is optional.
  - Allow optional ageing of CI Stream Table entries with MSB of ciStreamCreateMask set (98.4.1.5.3).

# 98.4.2.3 condTransmitCimAddPdu()

- Original text left it to the implementation to decide if it needed to transmit a CIM.
- Sending multiple CIMs helps reliability but increases network overhead.
- New text defines the following condition:
  - If we have sent less than cipMaxCIM CIMs and the AQM indicates that the congesting queue is congesting.
- Requires a CIM count in the CI stream table

# Knowing the position in the topology (Topology Recognition)

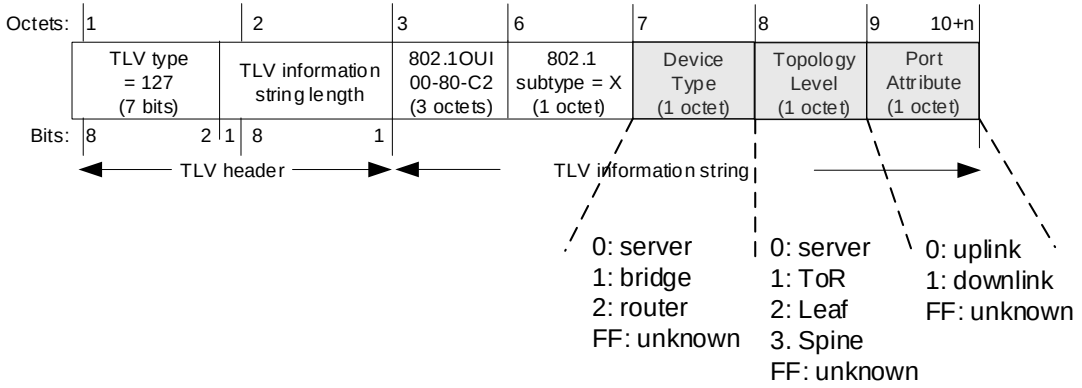
- Some upper layer services (applications) expect the same DSCP (or 802.1p) at both destination end-point and the source end-point.
- TOR bridge can recover the original DSCP (or 802.1) at edge if it knows where it is in the topology.



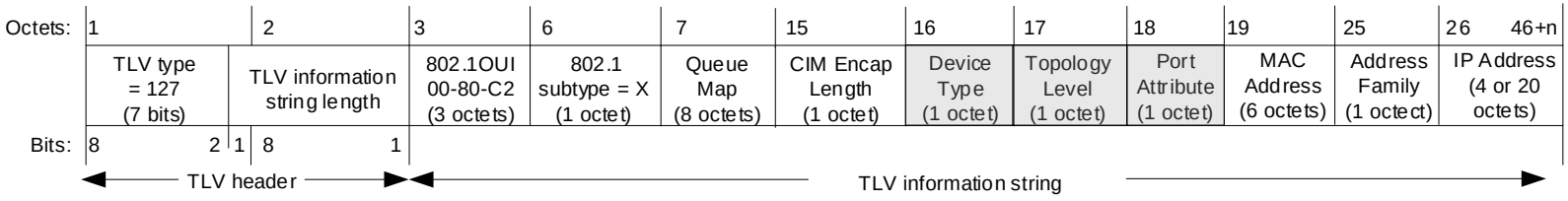

- Knowing position in topology also helps identify congestion type (in-network verses incast)
- Knowing position help prevent PFC deadlocks in lossless networks. See http://www.ieee802.org/1/files/public/docs2019/new-yu-pdf-0919-v00.pdf

# LLDP TLV Proposal

## 1st Choice – Independent 802.1 Organizationally Specific TLV




## 2nd Choice – Include in existing Congestion Isolation TLV




Note: if Topology Recognition is not supported, all new values are 0xff