P802.1Qcz
Design Team Topics

IEEE 802 Plenary

Bangkok

November 2018

Paul Congdon (Huawei/Tallac)

Previous Design Team Activity

- Prior to May 2018, an informal design team held conference calls to discuss CI design topics.
- This team no longer convenes now that we have a project
- There were several useful discussions worth reviewing.
- Participants
  - Kevin Shen, Paul Congdon, Sam Sun (Huawei),
  - Ilan Yerushalmi (Marvell),
  - Barak Gafni (Mellanox),
  - Martin White (Cavium),
  - Sowmini Varadhan (Oracle),
  - Jose Duato (Polytechnic University of Valencia),
  - Jesús Escudero Sahuquillo (University of Castilla-La Mancha)

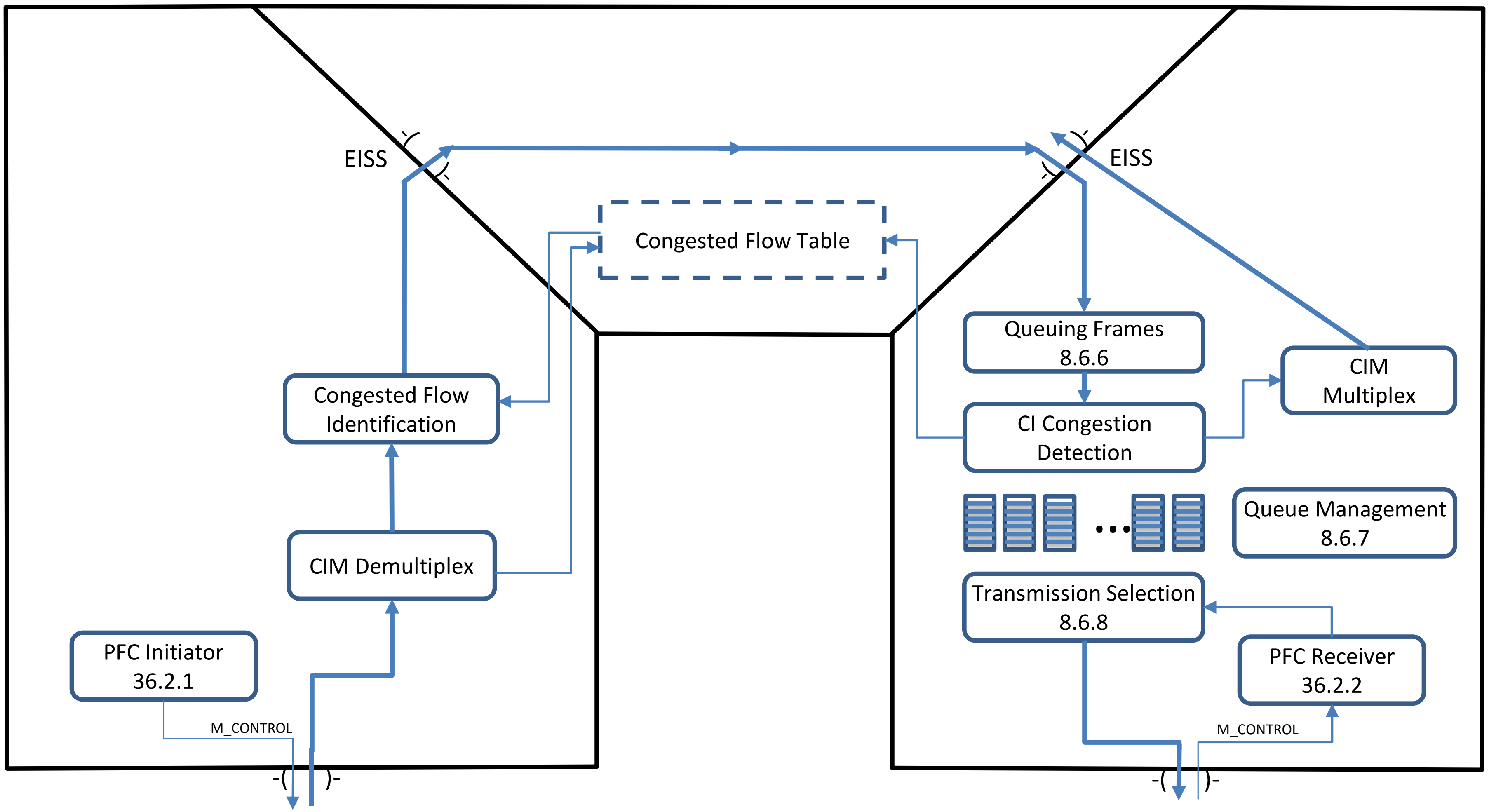
Design team discussion topics and resolution proposals

- Congested to non-congested transition (Note: to be discussed in a separate contribution)
- Congested flow packets in non-congested queue upstream and interaction with PFC
- Asynchronous upstream ageing
- Re-use 802.1Qau CNM message format
- Specifying order preservation after isolation
- Neighbor capability discovery
- Operation in hierarchical networks (e.g. VXLAN)
- Multicast Operation
- Congested flows changing paths
- Flow remaining congested upstream after downstream un-congests
- Recovery From Loss Of CIM (forward and backward)

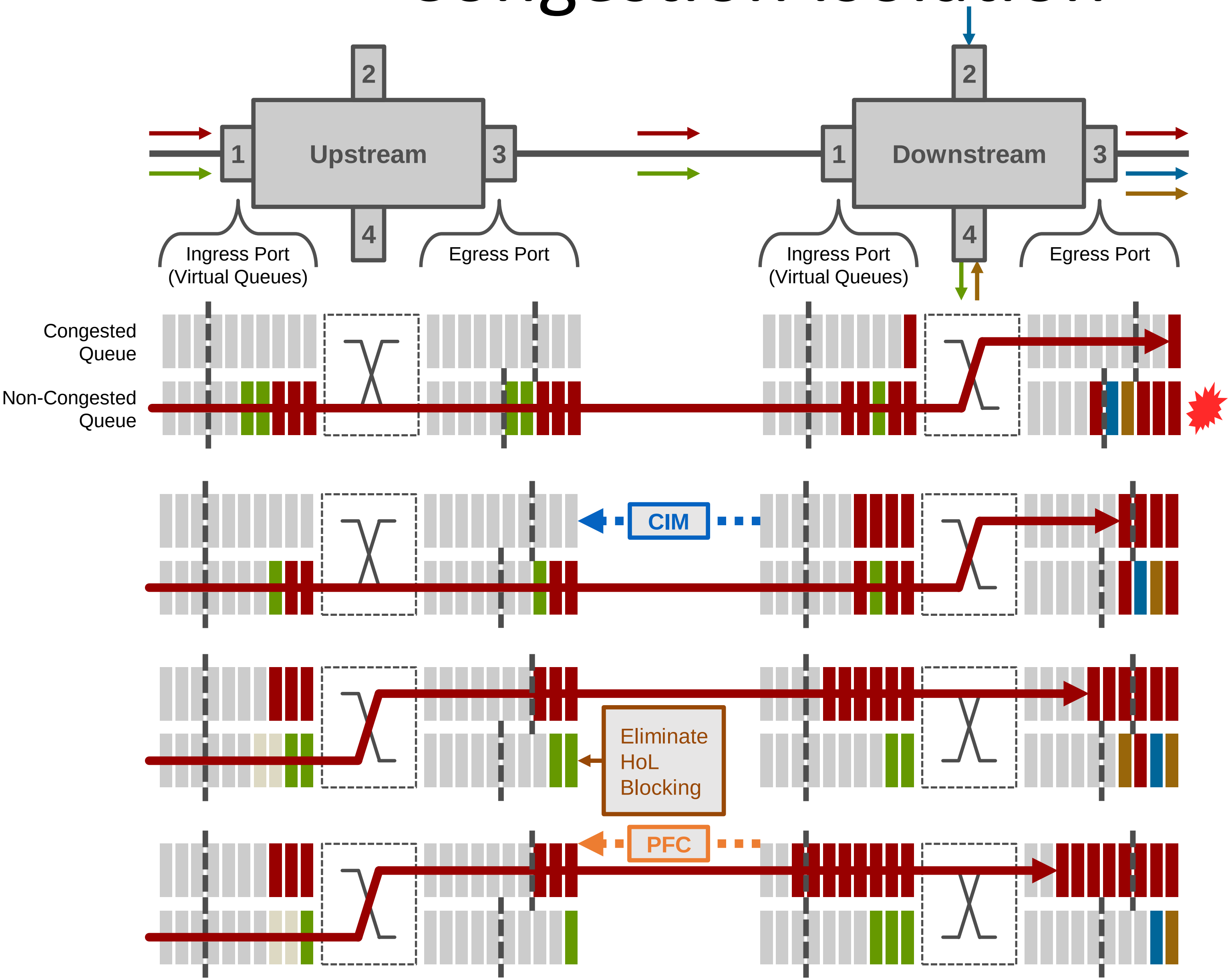
Agenda

- Reference diagram
- PFC interaction challenges
- Need for ‘deallocate’ CIM
- Recovery From Lost CIMs
- Summary of Congestion Isolation Critical Processes
- Next steps

Proposed Reference Diagram



Congestion Isolation



1. Identify the flow causing congestion and isolate locally

2. Signal to neighbor when congested queue fills

3. Upstream isolates the flow too, eliminating head-of-line blocking

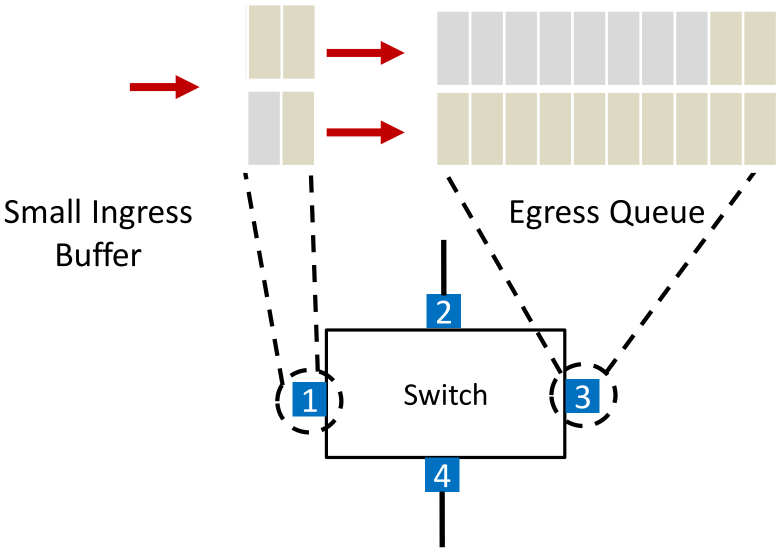
4. Last Resort! If congested queue continues to fill, invoke PFC for lossless

Lossless vs Lossy

- Priority-based Flow Control (PFC) is required for fully ‘lossless’ mode of operation
- The goal is, that if needed, PFC should be issued primarily on the congested traffic class. However…
  - Due to CIM signaling delays, packets from the same flow may exist in both un-congested and congested upstream traffic classes after isolation.

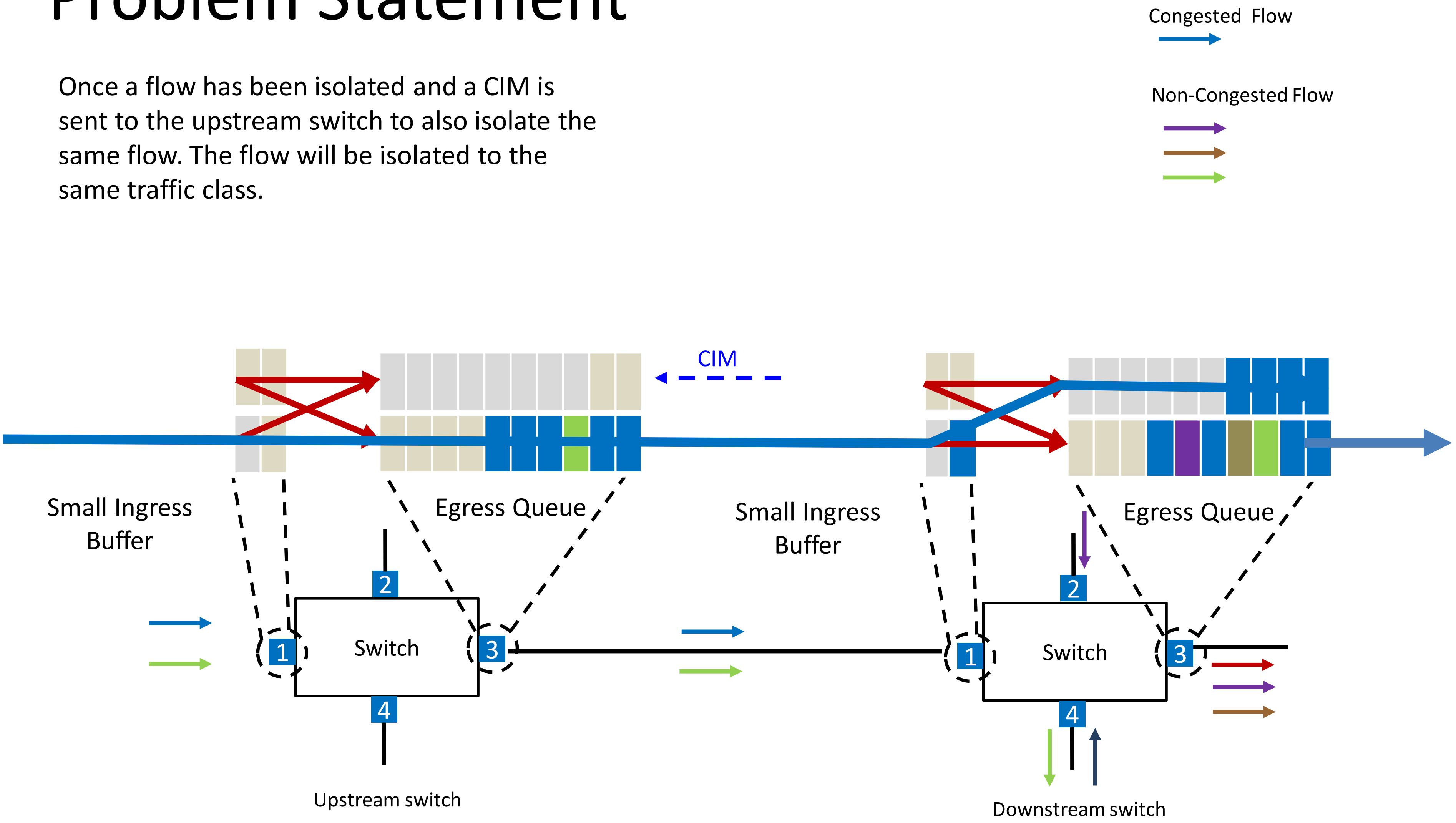
A Lossless Bridge can’t drop internally

1. 802.1 Bridge architecture is modeled as a pure egress buffered switch
2. Many different implementations exists
   a) Input buffered Virtual input queues
   b) Shared memory
   c) Other
3. When and how to trigger PFC on ingress w vary based on implementation, but the following is true:
   a) In order to receive a packet at ingress you must have buffer space
   b) In order to relay from ingress to egress there must be space in egress.
   c) If no space exists at egress, then the packet remains at ingress to be lossless. PRC may be triggered on the received traffic class
   d) Changing traffic classes during forwarding does not change these requirements.



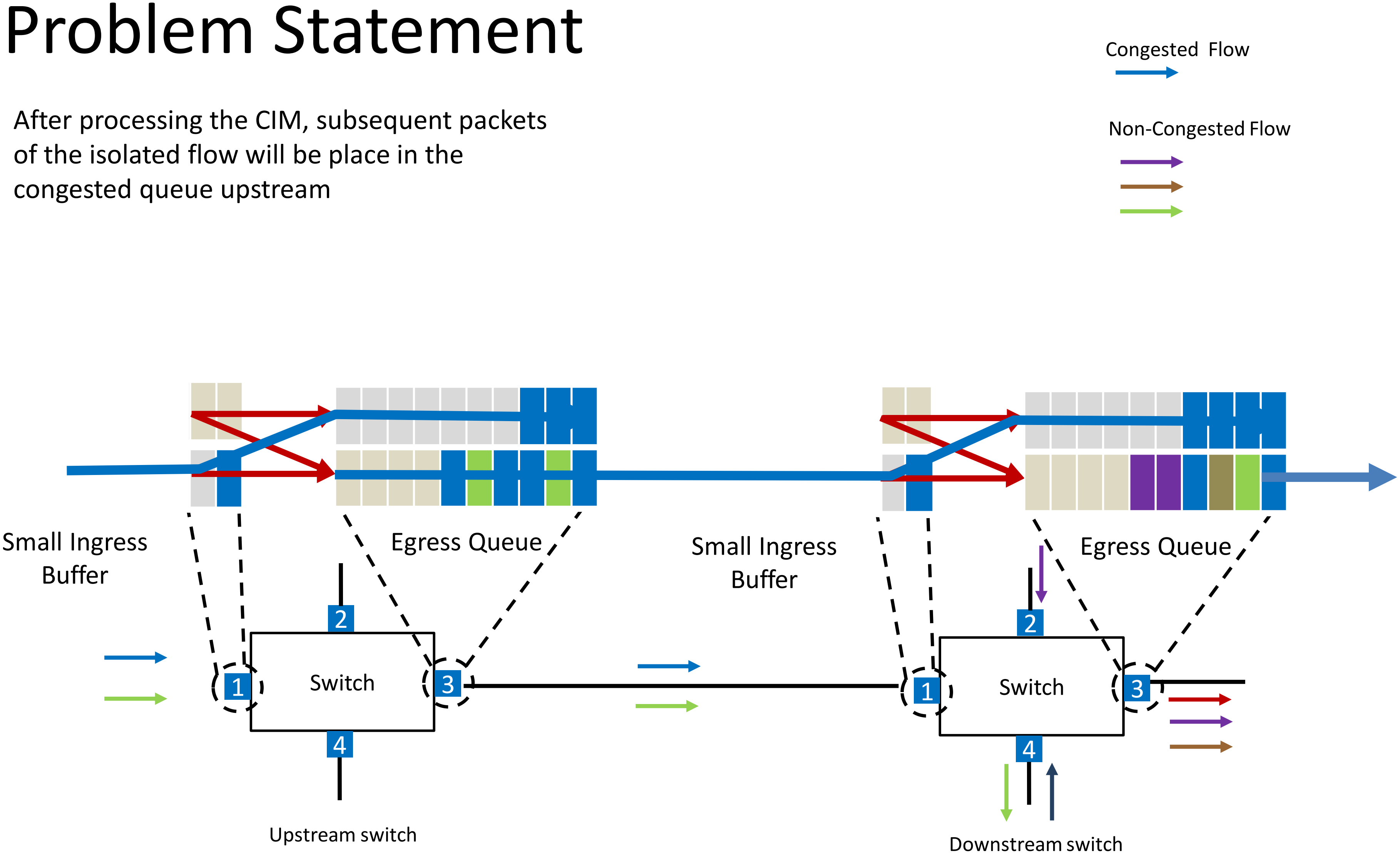
Problem Statement

Once a flow has been isolated and a CIM is sent to the upstream switch to also isolate the same flow. The flow will be isolated to the same traffic class.



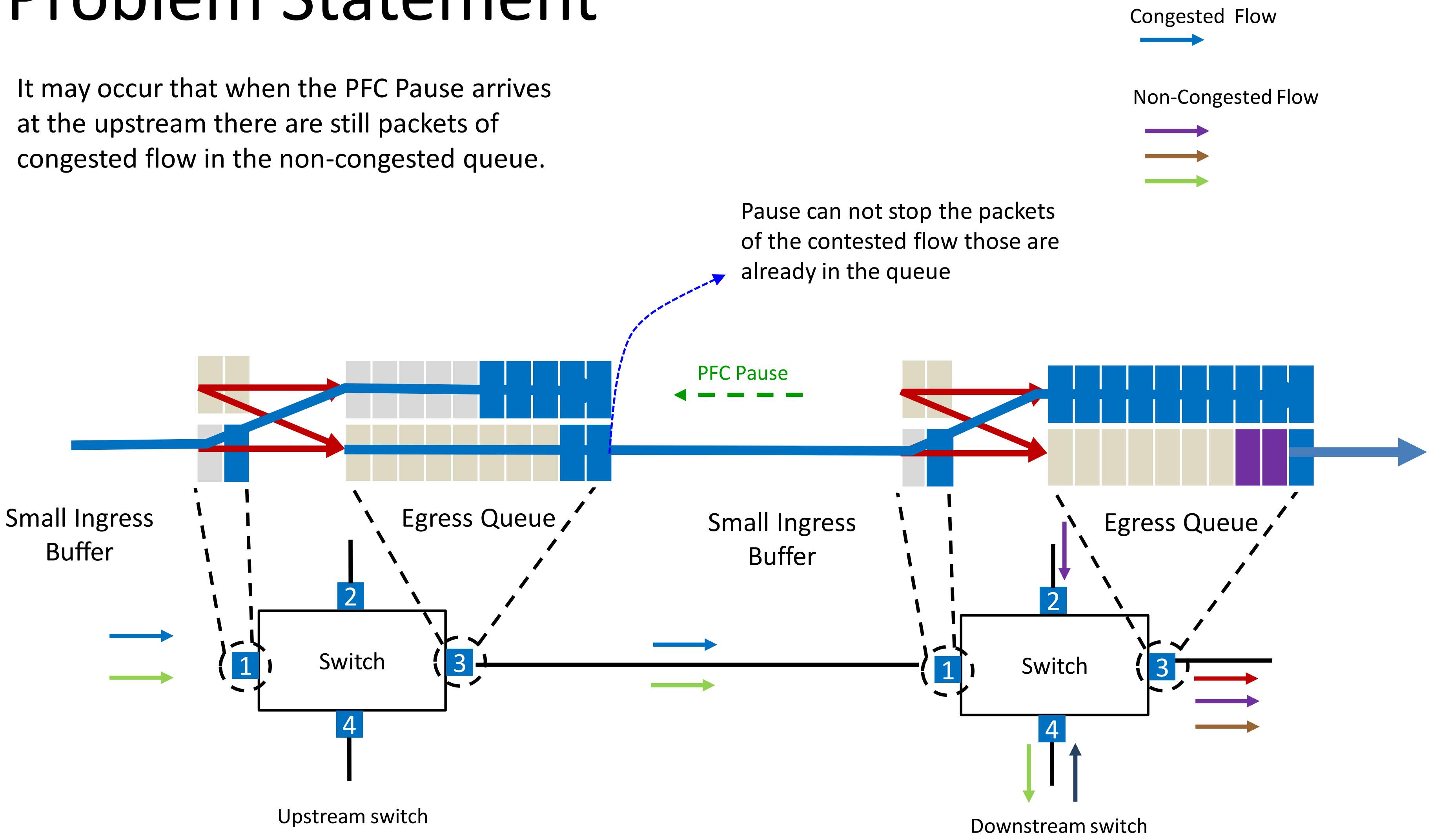
Problem Statement

After processing the CIM, subsequent packets of the isolated flow will be place in the congested queue upstream



Problem Statement

It may occur that when the PFC Pause arrives at the upstream there are still packets of congested flow in the non-congested queue.

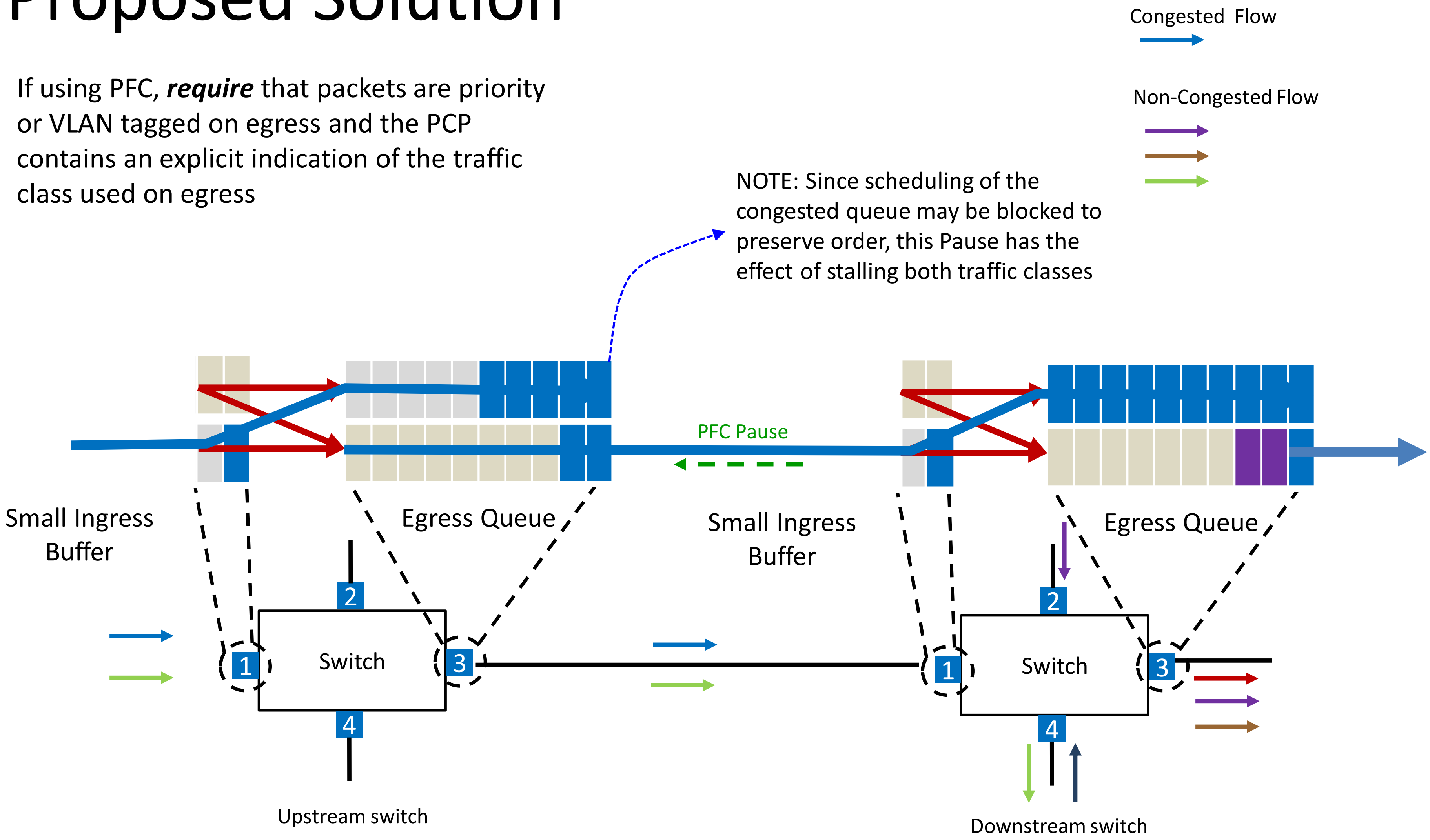


Observation

- Previous discussion of the complexity of interacting with PFC had assumed the downstream switch can not identify the traffic class used upstream to egress the packet.
- As long as the downstream switch ‘knows’ what traffic class the upstream switch egressed the packet, the downstream switch can Pause the correct traffic class

Proposed Solution

If using PFC, ***require*** that packets are priority or VLAN tagged on egress and the PCP contains an explicit indication of the traffic class used on egress



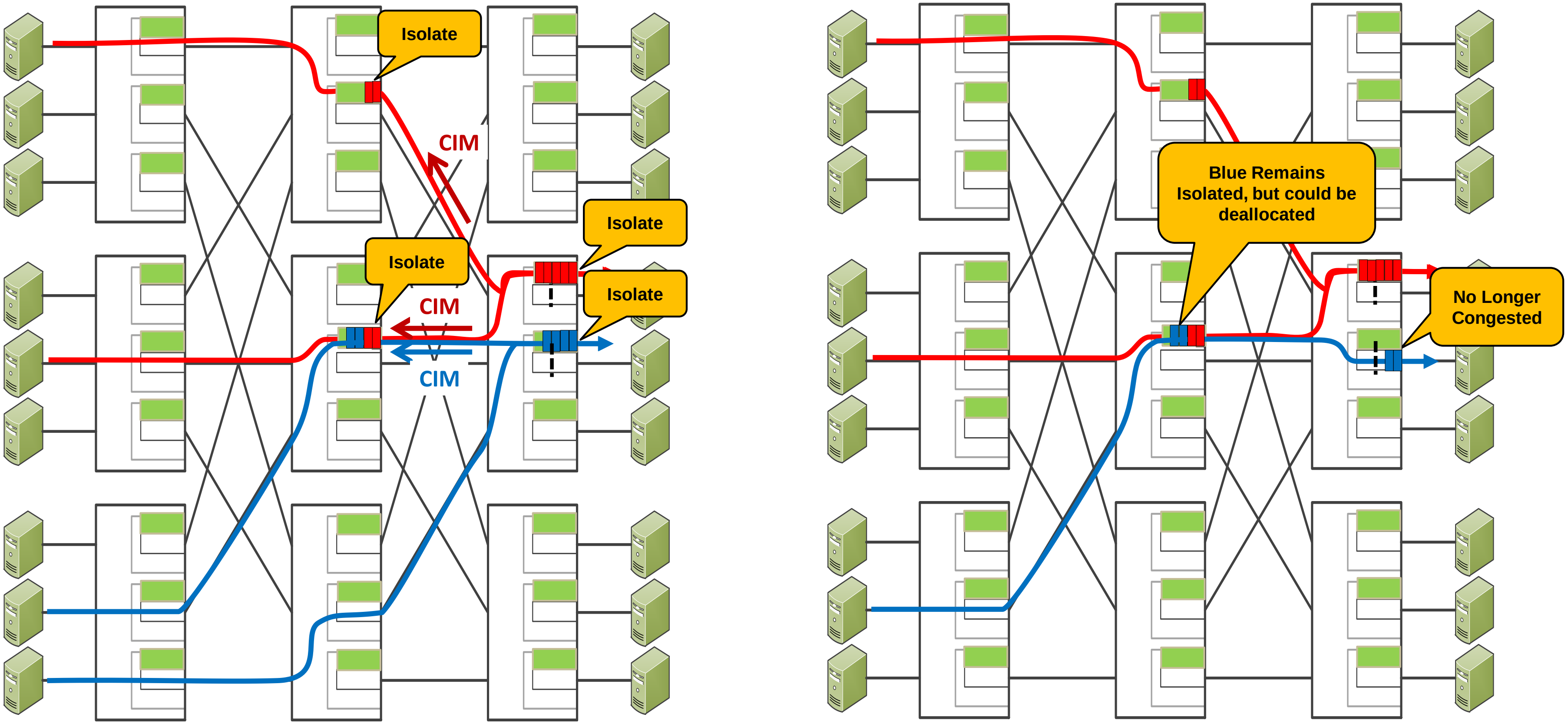
Need for ‘deallocate’ CIM

Congested Flow 🟥

Congested Queue 🟩

2^{nd} Congested Flow 🟦

Normal Queue □



Two congested egress queues

Congestion subsides on one queue

Need for 'deallocate' CIM

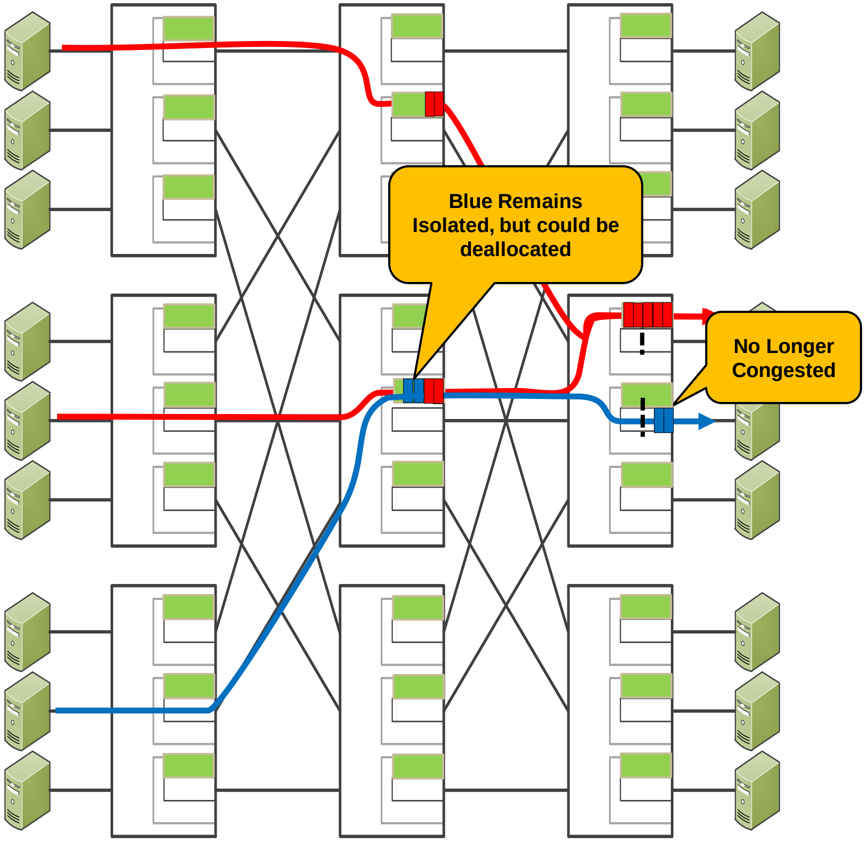
Congested Flow →

Congested Queue 🟩

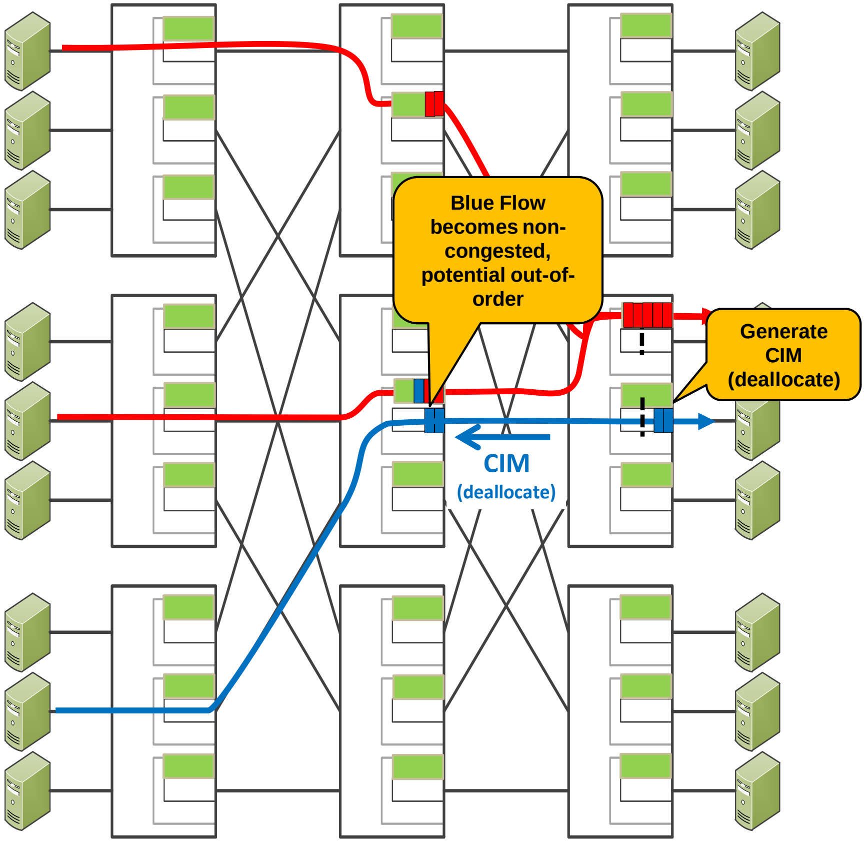
2nd Congested Flow →

Normal Queue □

Congestion subsides on one queue



Signal 'Deallocate' to free upstream flow


Blue Flow becomes non-congested, potential out-of-order
Generate CIM (deallocate)
CIM
(deallocate)
!
!

Need for ‘deallocate’ CIM

- Problem statement: Flow remains congested upstream longer than necessary
  - If congestion subsides in downstream switch, the upstream switch does not know to transition the flow from congested to non-congested state
- Some Options:
  - Downstream switch sends ‘deallocate’ CIM when a flow is transitioned from congested to non-congested state.
- Concerns:
  - Single ‘deallocate’ CIM is subject to loss and may need additional reliability.
  - Upstream switch may have residual packets in congested queue, creating out-of-order problem
- Resolution: Signal CIM deallocate when flow entry is removed. Receiver of CIM deallocate may need further logic to process (e.g. know if packets are in congest queue) or may need to block scheduling of non-congested queue to preserve order.

Recovery From Lost CIM

- Problem statement:
  - Downstream switch moves a flow to the congested queue → CIM is lost → Upstream switch keeps classifying the flow to the uncongested queue → Downstream switch has to pause uncongested queue → Head-of-line blocking on uncongested queue
- Some Options:
  - Send multiple (such as 3) CIMs in succession to provide redundancy.
  - Send CIMs periodically. Upstream send an ACK when it has received a CIM. Downstream stops after receiving ACK.
  - Send CIMs periodically. Upstream mark the subsequent packets when it has received a CIM.
  - Use Qau algorithm for generating CIMs (30.2.1), increasing CIMs as congestion worsens, automatic refresh
- Proposed Resolution: the congested queue should be capable of triggering CIMs
  - 802.1Qau defines the logic to control the rate of CIM generation, but doesn’t specify its granularity: per bridge or per congestion point (i.e. per egress queue)

Recovery From Lost ‘deallocate’ CIM

- Problem statement: loss of CIM may result in packet loss
  - Downstream switch moves a flow from congested to uncongested → CIM ‘deallocate’ is lost → upstream switch continues to isolate a flow that is not congested.
  - There is no natural trigger for the Downstream switch to re-send the CIM once it has ‘deallocated’ the flow.
- Some Options:
  - Always send multiple times (e.g. 3 ‘deallocate’ CIMs)
  - Send CIMs periodically. Upstream send an ACK when it has received a CIM. Downstream stops after receiving ACK.
  - Send CIMs periodically. Upstream mark the subsequent packets when it has received a CIM.
  - Just end once and hope for the best
- Proposed Resolution: Just send once and hope for the best
  - Upstream switch needs to have its own mechanism for removing a congested flow entry

Congestion Isolation Critical Processes

1. Detecting flows causing congestion
2. Creating flows in the congested flow table
3. Signaling congested flow identify to neighbors
4. Isolating congested flows without ordering issues
5. Interaction with PFC generation
6. Detecting when congested flows are no longer congested
7. Signaling congested to non-congested flow transitions to neighbors
8. Un-isolating previously congested flows without ordering issues

Next Steps

- Procedure for reconvening Design Team?
- Design topics to focus more detail on?
- Reaction to a draft contribution