Lossless Bridges and CI

Paul Congdon

Important assertions about CI

- There are various degrees of conformity that can be specified and agreed upon
  - If lossless operation is NOT a requirement, CI works without enabling PFC
  - CI can perform local isolation only, without signaling
  - CI can coordinate isolation with upstream neighbors – best performance
- CI is designed to support higher layer end-to-end congestion control
  - CI is NOT an improvement on PFC
  - CI is NOT an improvement on QCN (Congestion Notification)
  - Congestion isolation provides necessary time for the end-to-end congestion control loop.
- To create a fully lossless network, PFC is needed as a last resort
  - CI has been shown to reduce both the number of pause frames and duration of pause
- A bridge that has been designed to support lossless operation shall not drop a packet internally because of congestion.

Bridge Forwarding

Flow Isolation / Re-Classification

Congestion Detection

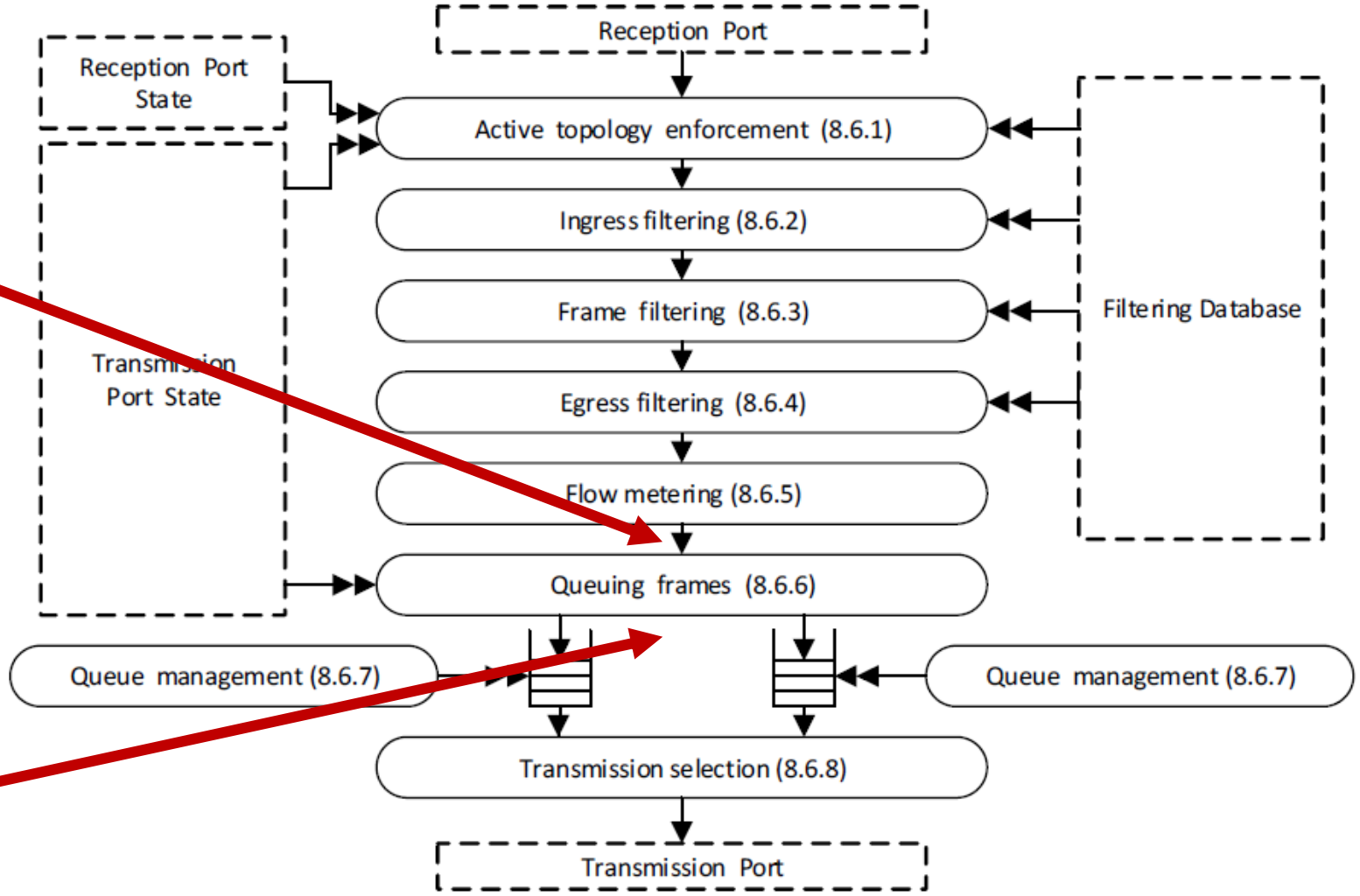
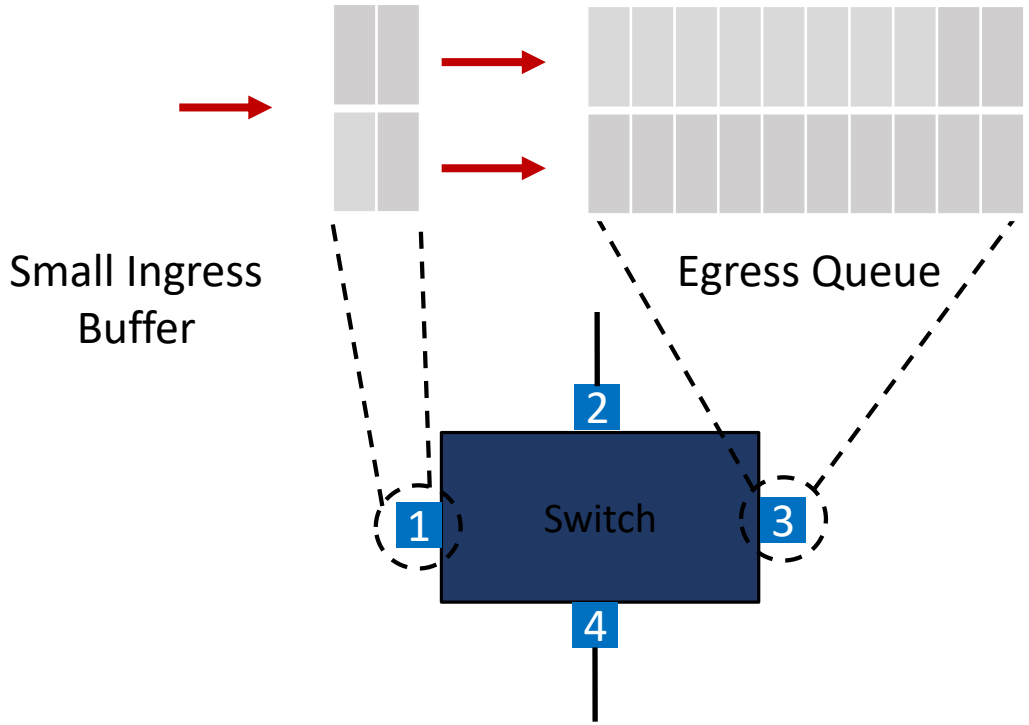


Figure 8-11—Forwarding process functions

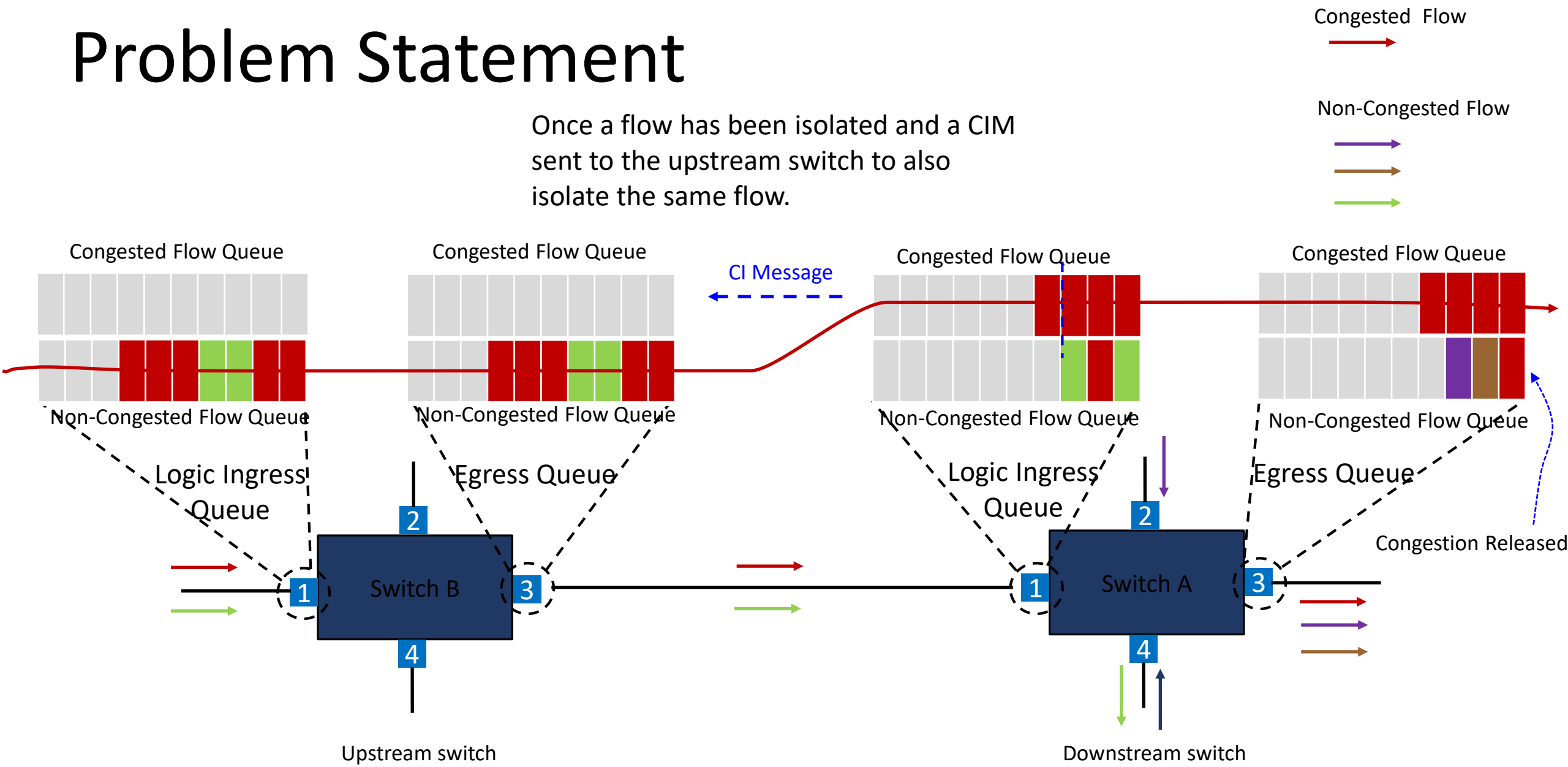
A Lossless Bridge can’t drop internally

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 will 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
   d) Changing traffic classes during forwarding does not change these requirements.

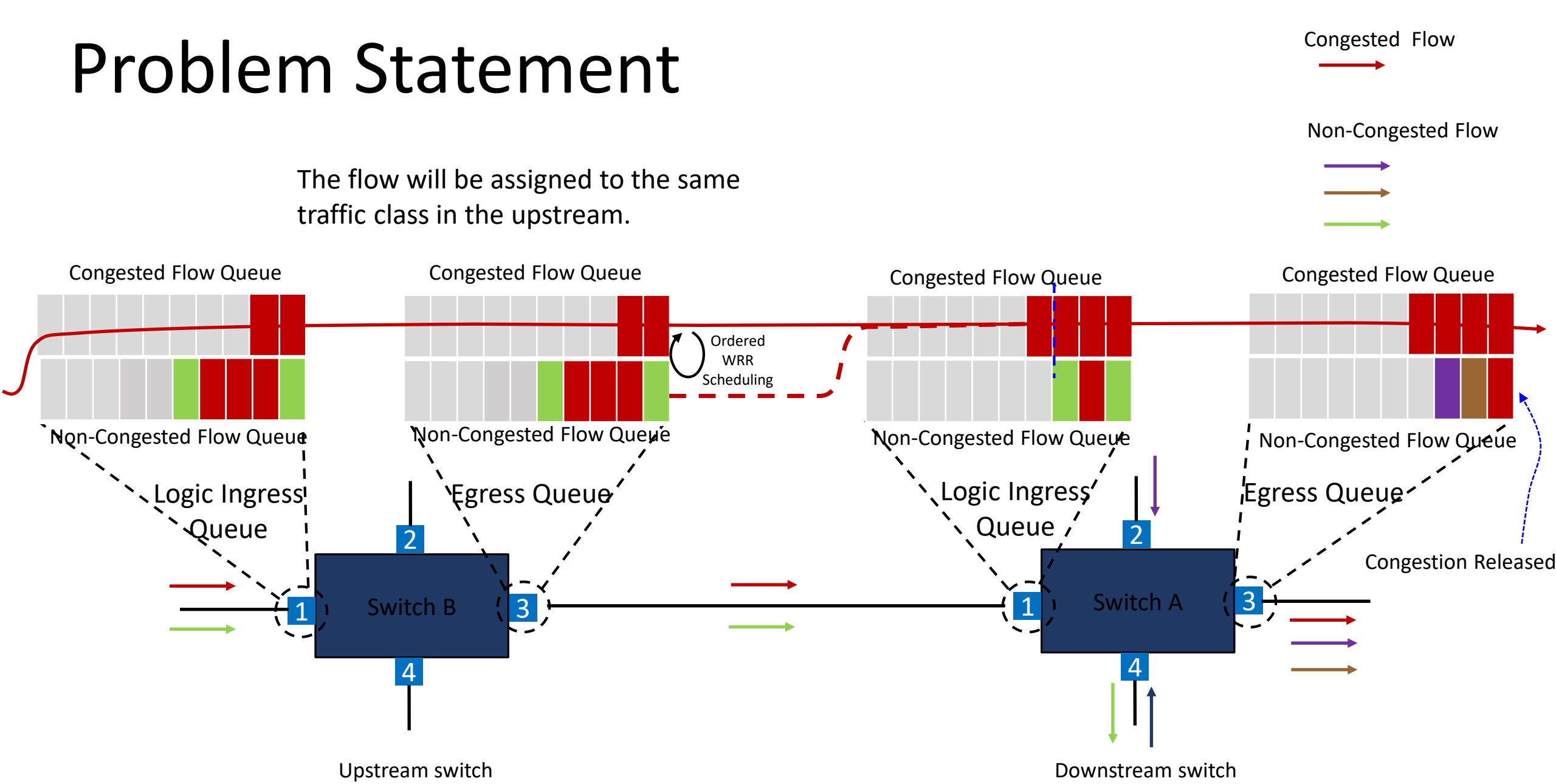


Problem Statement

Once a flow has been isolated and a CIM sent to the upstream switch to also isolate the same flow.

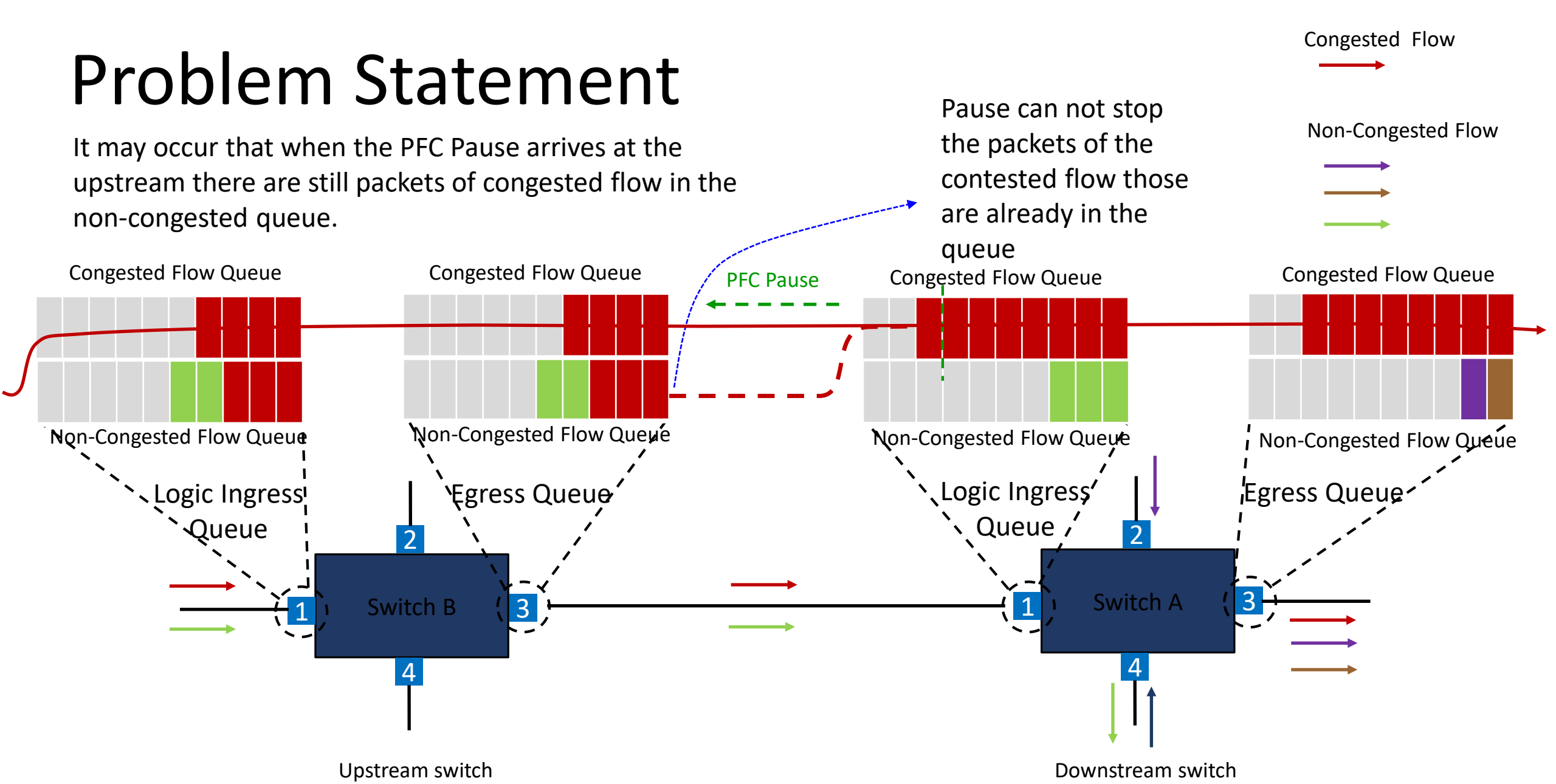


Problem Statement



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.

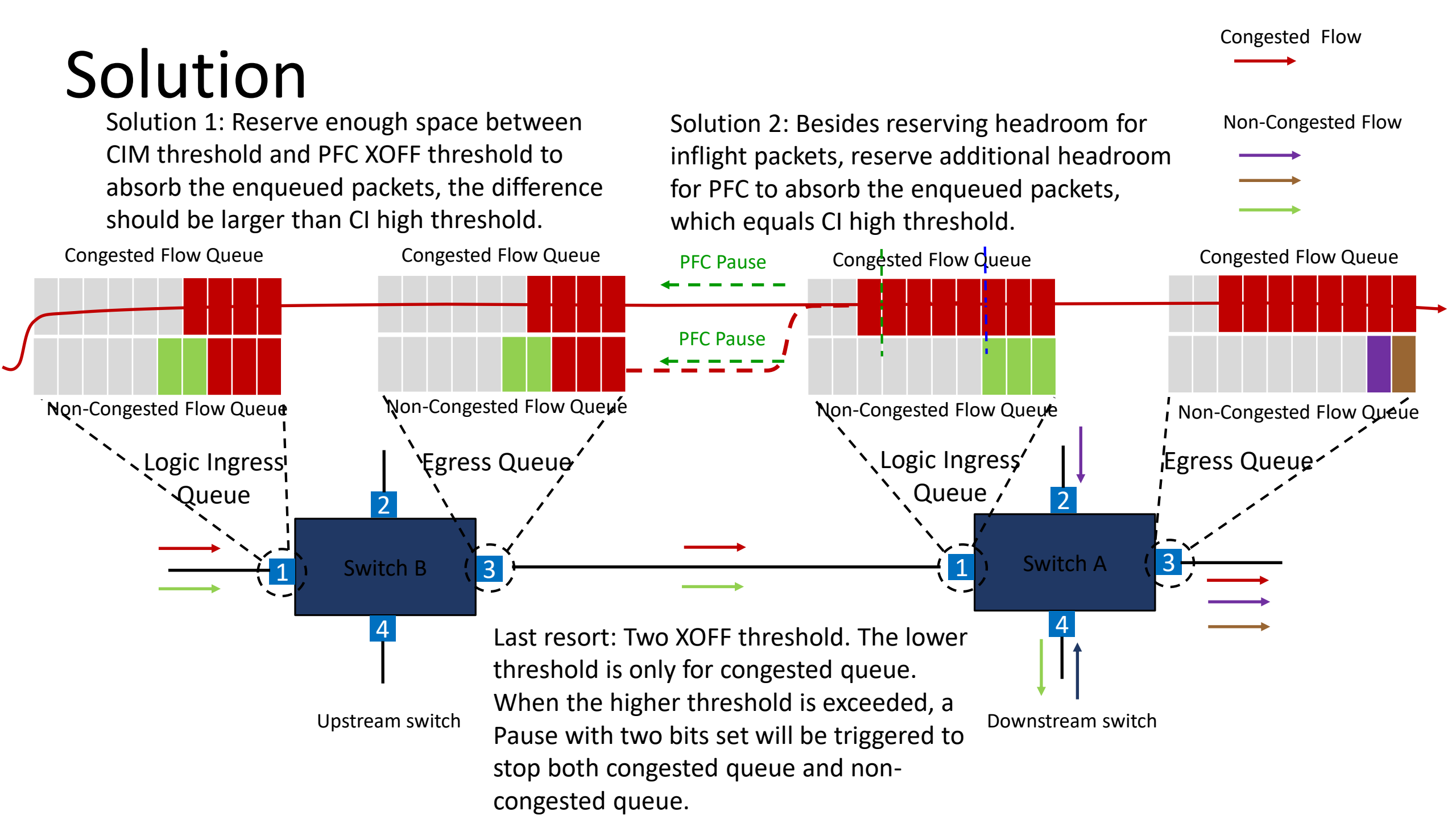


For simplicity, assume there is no other congested flow, so the bytes buffered in the egress congested queue is equal to the logic ingress congested queue.

Solution

Solution 1: Reserve enough space between CIM threshold and PFC XOFF threshold to absorb the enqueued packets, the difference should be larger than CI high threshold.

Solution 2: Besides reserving headroom for inflight packets, reserve additional headroom for PFC to absorb the enqueued packets, which equals CI high threshold.



Last resort: Two XOFF threshold. The lower threshold is only for congested queue. When the higher threshold is exceeded, a Pause with two bits set will be triggered to stop both congested queue and non-congested queue.