Congestion Isolation Proposal Analysis

Vahid Tabatabaee (Broadcom)
Mohan Kalkunte (Broadcom)
Hugh Holbrook (Arista)

3/7/2018

Outline

- Background
- Analysis
- Summary and Conclusion

Introduction

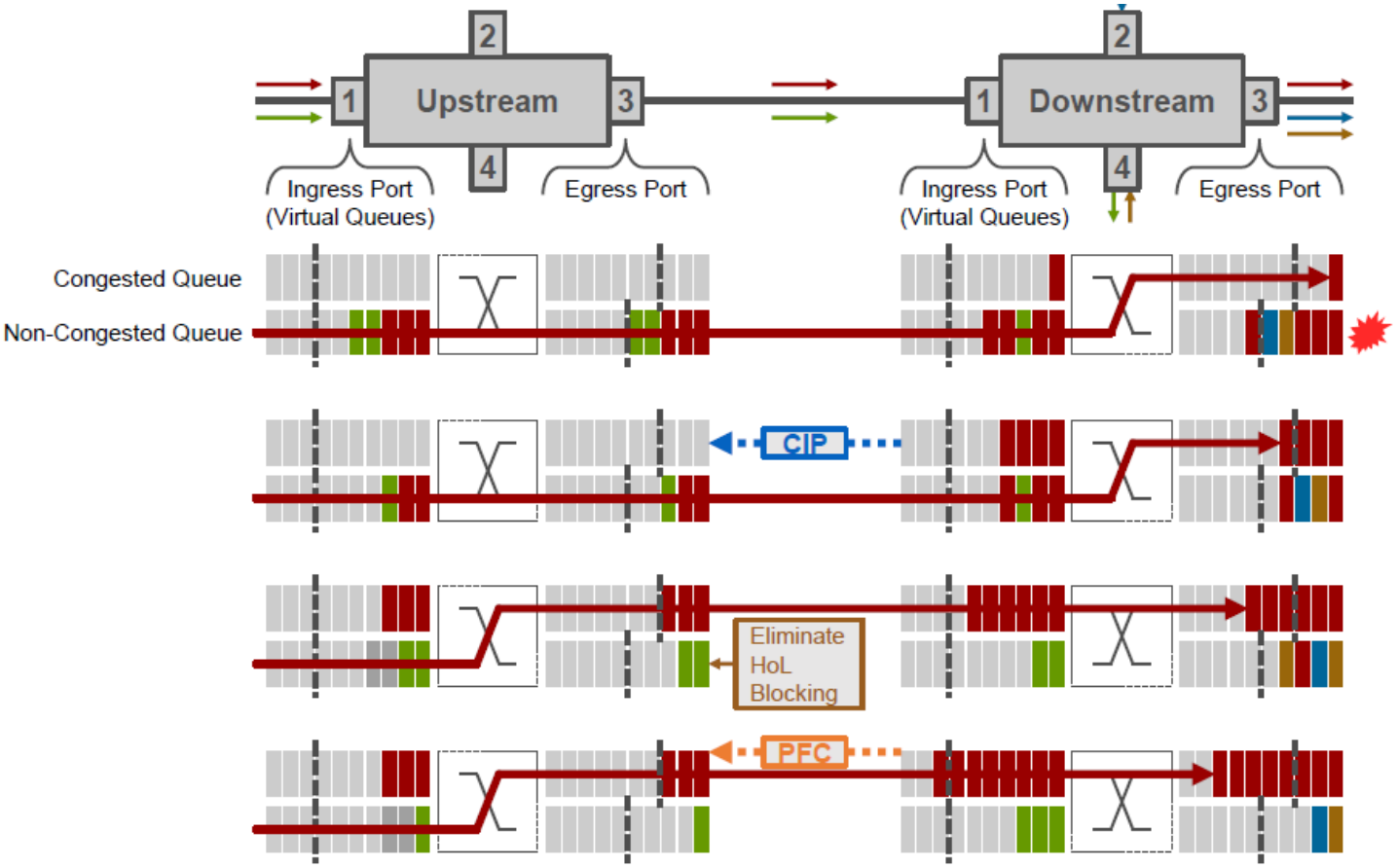
Definition: An approach to isolate flows causing congestion and signal upstream to isolate the same flows to avoid head-of-line blocking.

- Congestion Isolation (CI) proposal is a link-level flow control mechanism that has higher flow granularity than PFC
- Goal of the CI is to enhance PFC performance issues due to HOL blocking
- However, there are other factors that should be considered in evaluation of a link-level flow control mechanism
  - Fairness across flows
  - Required switch buffer to support lossless
  - Interaction with other flow and congestion control mechanisms
    - PFC
    - ECN based end-to-end congestion control mechanisms such as DCTCP

Goal

- Highlight some potential issues with the CI proposal

Review of the CI proposal



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. If congested queue continues to fill, invoke PFC for lossless

Outline

- Background
- Analysis
- Summary and Conclusion

Congested flow identification

- Background
  - Congested flows are identified by sampling packets that arrive to congested queues
    - In this way, the probability of identifying high-rate flows is higher
  - PFC role is to avoid packet drops due to sudden burst of short-lived flows or transient behavior of long-lived flows
    - The steady state rate of long-lived flows are controlled by ECN based congestion management mechanism

- Issues
  - Ineffective in controlling sudden burst of short-lived flows
    - Sampling based method does not identify short-lived flows accurately
      - Cause unfairness in throughput and latency across flows
      - Hence negatively impacting tail latency and 99% FCT
    - Latency in congested flow identification degrades effectiveness for short-lived flows
      - There is a delay in identifying and reporting congested flows to upstream node
        - First the uncongested queue should pass the threshold
        - The flow packets should be sampled
        - CIP should be transmitted and received at the upstream node
      - Short-lived flows may be finished by the time they are identified and reported
        - 100KB flow takes ~8 usec. on a 100G port

Congested to uncongested flow transition

- Background
  - Congested to uncongested transition is done by using inactivity timeout
- Issues
  - Interaction with PFC
    - Congested flow packets may stay for a long time in the congested queue due to PFC-XOFF
    - Therefore, using inactivity timeout on packet arrivals does not avoid packet reordering
  - Interaction with ECN
    - A congested flow rate can be controlled and reduced by e2e congestion management
    - However, since sources use traffic pacing the inactivity timer may not timeout
    - The impact is HOL blocking of the well behaved flows

Buffer requirement

- Background
  - There are two main buffer components in the switch when PFC is used
    - **Burst absorption buffer:** Provides good throughput and limit PFC trigger
    - **Headroom buffer:** Absorbs packets-in-flight after PFC is triggered to avoid packet drops

- Issues
  - CI proposal requires larger burst absorption buffer
    - CI works properly only if the PFC ingress port thresholds is multiple times larger than egress queue thresholds
      - This is to ensure CIP messages are sent to the upstream nodes well before sending PFC-XOFF
      - The required buffer increases as the switches' radix increase
  - CI proposal requires larger headroom buffer
    - PFC can be triggered for both congested and uncongested priorities and both need reserved headroom buffer
    - More importantly it is not clear how the congested priority headroom should be sized
      - For PFC, headroom is sized based on the cable length and PFC response time
      - However, with CI upstream switch can send congested flow packets after its congested queue is stopped
        - This is because congested flow packets can be in the non-congested queue
          - Packets from already identified congested flows can still be in the non-congested queue
          - New congested flows can be identified with packets in non-congested queue

Outline

- Introduction
- Analysis
- Summary and Conclusion

Summary of Issues

CI Proposal Issue	Performance Impact
Sampling based congested flow detection	Short-lived flows are not stopped effectively
	Unfairness across flows
	Poor tail latency and FCT
	Lag in detecting flows that cause congestion
Timeout based transition from congested to uncongested flows	Packet reordering due to interaction with PFC
	HOL blocking for flows that their rate is controlled by an ECN based congestion management + traffic pacer
PFC ingress thresholds need to be larger than CI egress thresholds	Increases burst absorption buffer requirement
Congested flow packets can be in non-congested queue when PFC is triggered	Increases headroom buffer requirement

Conclusion

- Concerns about technical and economical feasibility
  - Not clear if it improves 99% FCT for short-lived flows
  - Can cause other performance issues that covered in this presentation
  - Has significant impact on the switch buffer requirement
  - Compromises simplicity of Ethernet