Data Center Congestion Management Initiatives in IEEE 802.1

Paul Congdon (Tallac Networks)

Disclaimer

- This presentation should be considered as the personal view of the presenter not as a formal position, explanation, or interpretation of IEEE.

- Per IEEE-SA Standards Board Bylaws, December 2017
  - “At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position of IEEE.”

Three Initiatives of Interest

Motivated to enable low-latency, low-loss, high-reliability Ethernet-based Data Center Networks supporting RDMA and AI/HPC workloads.

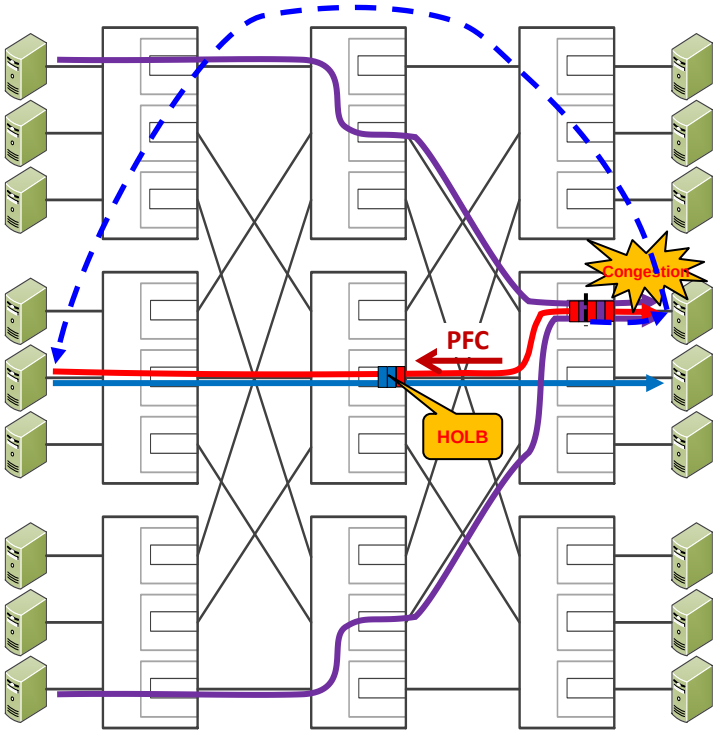
1. P802.1Qcz – Congestion Isolation
2. P802.1Qdt – PFC Enhancements
3. Source Flow Control

These are all ‘amendments’ to IEEE Std 802.1Q

See: Intelligent Lossless Data Center Networks (https://ieeexplore.ieee.org/servlet/opac?punumber=9457236)

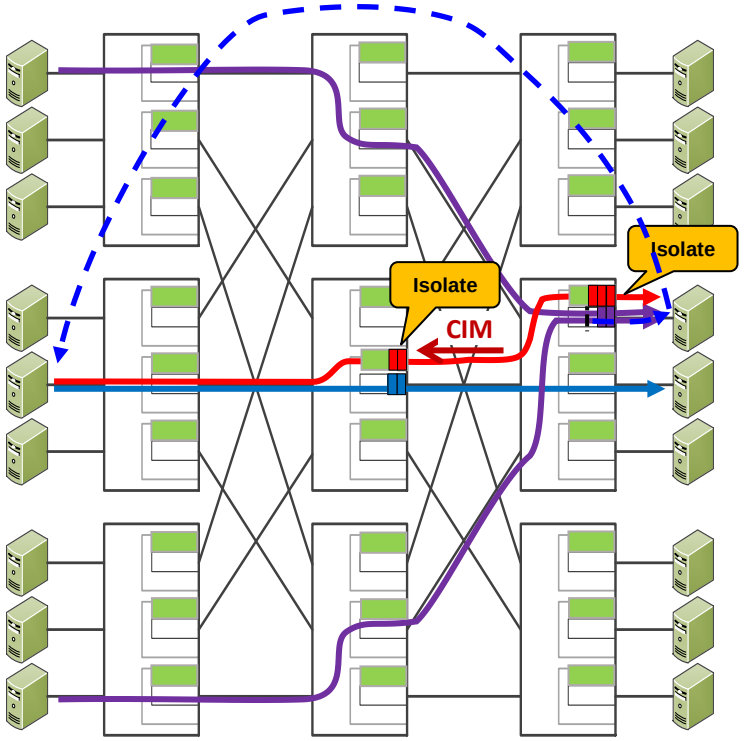
P802.1Qcz - Congestion Isolation

Today – Without Congestion Isolation



1. End-to-end congestion control using ECN marking
2. Priority-based Flow Control (PFC) as last-ditch effort to avoid drops

Congestion Isolation

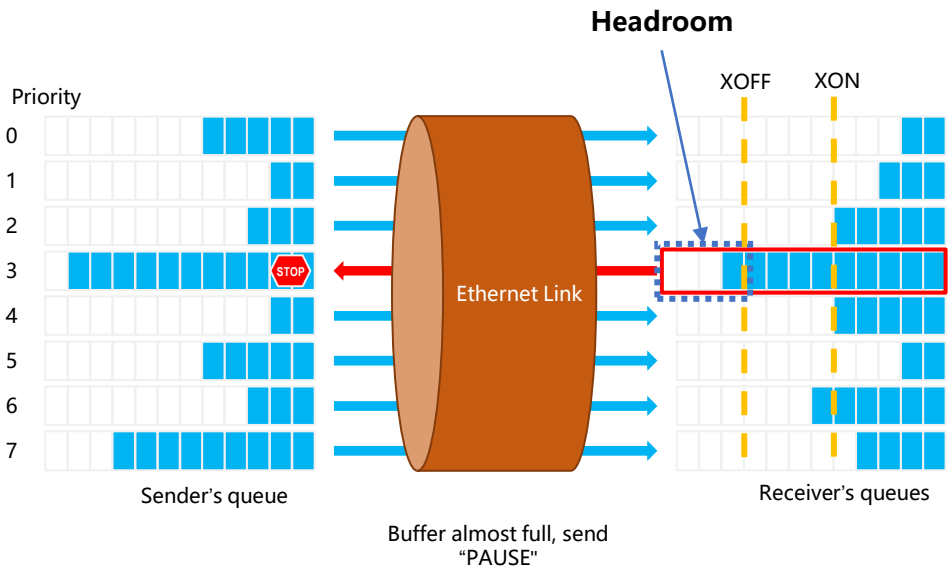


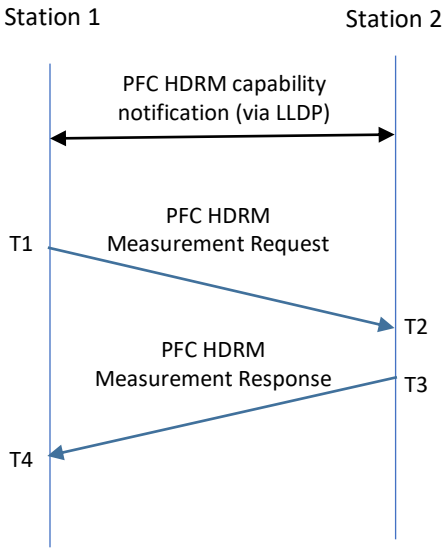
1. Move congesting flows to a separate queue and signal your upstream neighbor

Standard Status: the IEEE 802 equivalent of 'IESG Last Call'

P802.1Qdt – PFC Enhancements

Objective: Automatically calculate minimum PFC buffer requirements (i.e. headroom) for lossless operation, without user intervention. Additionally – protect PFC frames using MACsec encryption





Headroom needed = (Port speed * (T4-T1-(T3-T2)) + 2*(Max Frame) + (PFC Frame)) * Alpha

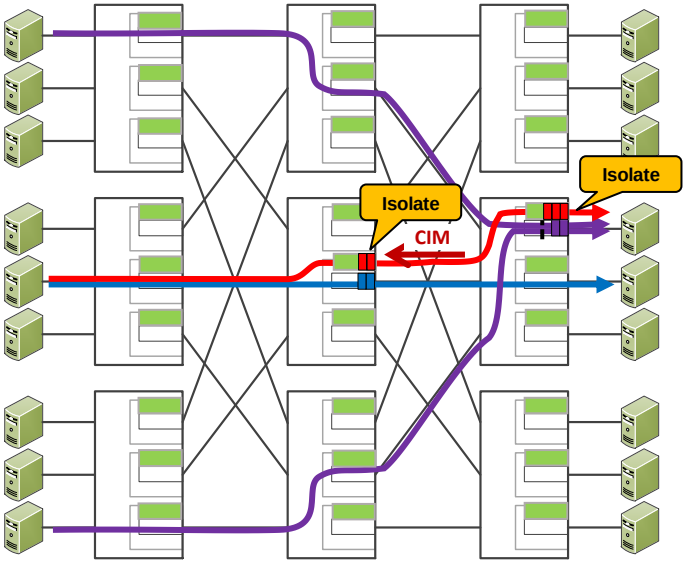
NOTE: Alpha is implementation dependent, based on internal buffer chunk size

1. Re-use the Precision Time Protocol (PTP) to measure cable delay
2. Exchange internal delay values using LLDP

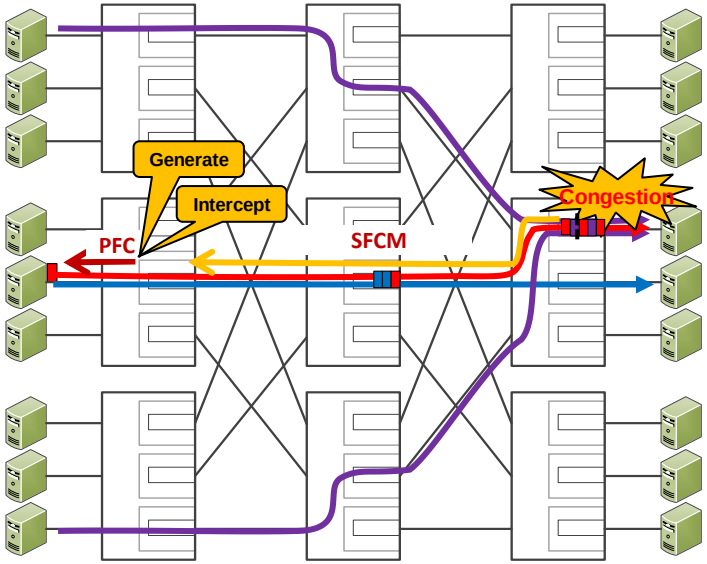
Status: New project, just forming now.

Source Flow Control

P802.1Qcz - Congestion Isolation



Source Flow Control (w/ ToR Proxy)



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).

Details

- Can be combined with Congestion Isolation
- Edge-to-Source signaling using L3 message
- Like an L3 version of 802.1Qau (L3-QCN), but no Reaction Point (RP) rate controller defined – this is Flow Control
- Optional source Top-of-Rack switch involvement

Status: New project proposal

Join us for further discussion

- Non-WG IETF Mailing list rdma-cc-interest@ietf.org
  - Subscribe at: https://www.ietf.org/mailman/listinfo/rdma-cc-interest

- Side Meeting: Wednesday 10:00AM – 11:30 AM – Green Room 1
  - NOTE on side meetings:
    - Open to all
    - Meeting minutes will be posted to rdma-cc-interest@ietf.org
    - Not under NDA of any form
- You're invited to join a Microsoft Teams meeting

  Join on your computer or mobile app
  Click here to join the meeting