Source Flow Control Project Proposal

Paul Bottorff (HPE)
Paul Congdon (Huawei)
JK Lee (Intel)
Lily Lv (Huawei)

802.1 March Plenary, electronic
March 14, 2022

Background

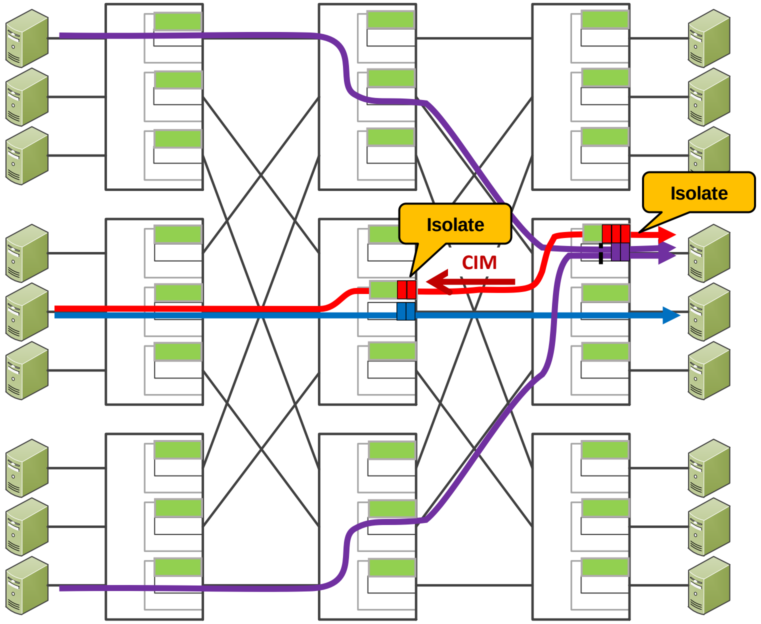
- Motivation
  - Further enable the success of Ethernet in the low-latency, low-loss, high-reliability Data Center Networks supporting RDMA (RoCE) and AI/HPC workloads.

- Previous presentations
  - Public presentations at P4 Workshops (Apr'20, May'21) and Open Fabrics Alliance (Mar'21)
    - https://opennetworking.org/wp-content/uploads/2020/04/JK-Lee-Slide-Deck.pdf (slide 12)
    - https://www.openfabrics.org/wp-content/uploads/2021-workshop-presentations/503_Lee_flatten.pdf
    - https://opennetworking.org/wp-content/uploads/2021/05/2021-P4-WS-JK-Lee-Slides.pdf (slide 14)
  - Previous Nendica/TSN presentations
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf - 9/16/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0061-00-ICne-source-remote-pfc-test.pdf – 10/14/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0067-00-ICne-source-remote-pfc-status-update.pdf - 11/04/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0077-00-ICne-consideration-of-spfc-sfc-issues-when-leveraging-qcz.pdf - 12/16/2021
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0079-00-ICne-spfc-sfc-next-steps.pdf - 12/23/2021
    - https://www.ieee802.org/1/files/public/docs2022/new-congdon-SFC-overview-0122-v01.pdf - 01/19/2022
    - https://mentor.ieee.org/802.1/dcn/22/1-22-0001-01-ICne-sfc-q-changes.pdf - 01/27/2022
    - https://mentor.ieee.org/802.1/dcn/22/1-22-0005-00-ICne-new-bottorff-sfc-0222-v5.pdf - 02/24/2022
  - IETF Awareness
    - Topic raised at IEEE 802 / IETF Coordination call – 10/25/2021
    - https://datatracker.ietf.org/meeting/112/materials/slides-112-iccrg-source-priority-flow-control-in-data-centers-00 - 11/08/2021
    - Upcoming at the IETF-113 – HotRFC session – 03/20/2022, Scheduled side-meeting discussion – 03/23/2022

- Nendica vetting and technical design team collaboration

Source Flow Control High Level Concept

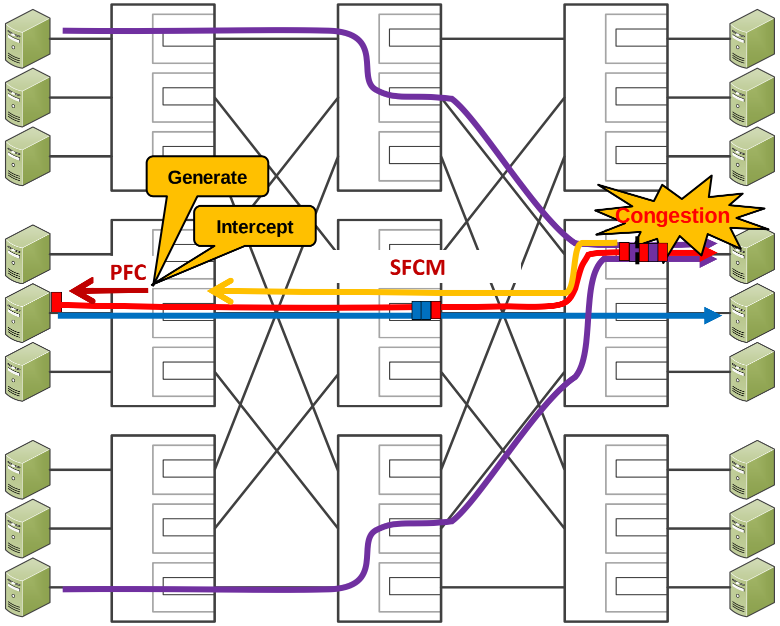
P802.1Qcz - Congestion Isolation



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).

Source Flow Control (w/ ToR Proxy)



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

SFC verses Congestion Notification (Qau)

Differences

- Qau is a L2 protocol, SFC is L3
- Qau is congestion control, SFC is flow control
- Qau defines a comprehensive control algorithm with many parameters, SFC uses PFC
- Qau CNM carries Quantized Feedback for a Reaction Point, SFC carries ‘pause’ duration for PFC
- SFC allows a ToR to proxy SFCM processing

Similarities

- Congestion points monitor queues for congestion
- Congestion points send signaling messages back to source
- Flow information (from received congesting frame) is provided in signaling messages

SFC verses Congestion Isolation (Qcz)

Differences

- Qcz uses an additional traffic class to isolate frames
- Qcz signals to upstream neighbor (L2 or L3), SFC signals to end-station (also via ToR Proxy) using L3 message
- Qcz does not directly rate control the sending host, SFC pauses the sending host
- SFC allows a ToR to proxy SFCM processing

Similarities

- Both schemes support L3 message formats
- Congestion points monitor queues for congestion
- Congestion points send signaling messages backward toward source
- Flow information (from received congesting frame) is provided in signaling messages

Need for the Project

- Congestion, in particular incast, is detrimental to RoCE performance in HPC/AI Data Center Networks due to packet loss
- PFC use prevails as a means for lossless operation, however, side effects of PFC are problematic (e.g. congestion spreading, head-of-line blocking, PFC storms, and deadlocks)
- The delay for end-to-end congestion control using ECN markings is too long for existing switch buffering. Need sub-RTT reaction.
- Use of PFC at the source edge has less negative impact and supports early adoption.

See: https://www.openfabrics.org/wp-content/uploads/2021-workshop-presentations/503_Lee_flatten.pdf
https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf

Proposed Scope of Work

- Amendment to 802.1Q with a new feature clause, associated management, YANG, and minor modification to existing clauses. Leveraging concepts and mechanisms from Qcz. Expected to include:
  - Configuration elements enabling/disabling the feature (system wide)
  - Specification of SFC messages and how to generate them
  - Specification of monitoring queues for congesting flows?
  - Specification of SFCM suppression (timeout) mechanism
  - Mechanism and configuration of SFCM ToR proxy capability
  - YANG support

See: https://mentor.ieee.org/802.1/dcn/22/1-22-0001-01-ICne-sfc-q-changes.pdf

Design Team Progress

Team

Paul Bottorff (HPE), Paul Congdon (Huawei), Claudio Desanti (Dell) , Uri Ezur (Intel), JK Lee (Intel), Lily Lv (Huawei)

Current List of Topics

1. UDP port number for SFCM
2. How to secure SFCM
3. Contents of SFCM
4. Identifying the source priority/traffic-class to pause
5. Operation in overlay networks (VxLAN, Geneve)
6. Calculation of pause interval
7. SFCM suppression
8. Multicast considerations
9. Source ToR intercept of SFCM packets
10. Consideration of DCBX enhancements

Next steps

- Presentation of design team discussion and analysis in TSN
- Continue drafting proposed text for PAR and CSD (shared as a contribution within TSN)
- Motion at the March 2022 Plenary to authorize PAR & CSD development for pre-circulation before the July 2022 Plenary