SFC Design Team Topics

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

June 2022

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
11. Mitigating HoLB at Source ToR

Topic 1: UDP port number for SFCM

No global well known UDP port has been assigned by IETF. Qcz uses locally assigned UDP port for L3 CIM. The CI Peer Table configures UDP port to be used for L3 CIM. This is obtained through LLDP

- Issue: ability to determine UDP port for distant L3 CIM receiver. Better to have well known UDP port used by all systems.

Solution:

- SFC is intended to be used in a data center network which is a closed environment within a single administrative domain.
- It is possible to configure the DCN, setting a UDP port number for SFCM by administration.
- Alternatively, request IETF to assign a dedicated UDP port number to SFCM.

Topic 2: How to secure SFCM

Qcz CIM security can use MACSec because it is hop-by-hop. How to secure edge-to-edge sPFC messages?

Explanation/Solution:

- Refer to IETF 112 ICCRG presentation. https://datatracker.ietf.org/meeting/112/materials/slides-112-iccrg-source-priority-flow-control-in-data-centers-00
- Securing SFCM is no different that securing other switch-to-switch protocols within the DCN, including:
  - BGP, LLDP
  - End-to-end packet marking (i.e. ECN)
- ACLs can be used at domain boundaries to block signaling packets
- For IPsec supported environment, Layer 3 SFC message can also be protected by IPsec.

Topic 3: Contents of SFCM

What needs to be in the SFCM? Should it include Qau ‘quantized’ parameters?

Explanation/Solution:

- Qau specifies ‘quantized’ parameter F_b. CNM message carries F_b to host as input of rate calculation.

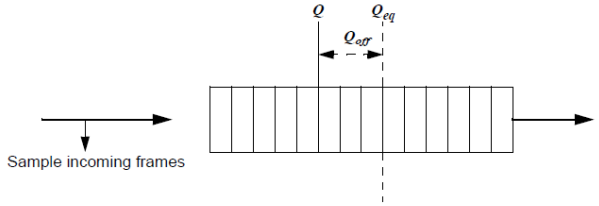


Figure 30-1—Congestion detection in QCN CP

Let Q denote the instantaneous queue size and Q_{old} denote the queue size when the last feedback message was generated. Let $Q_{off} = Q - Q_{eq}$ and $Q_\delta = Q - Q_{old}$.

Then F_b is given by the formula

$$F_b = -(Q_{off} + wQ_\delta)$$

(From 802.1Q -2018 30.2.1 CP algorithm)

- SFC proxy mode generates a PFC frame and does not need F_b. Pause time is needed.
- SFCM is sent to the sending host and is interpreted as if a PFC frame was received, F_b is not needed.
- Source IP address of offending flow is needed to generate SFCM
- Offending flow information is needed so source can map SFCM to appropriate traffic class. This includes DSCP
- A congestion locator such as Topology Recognition level to identify ‘incast’ congestion verses ‘in-network’ congestion.
- An optional PTP timestamp when the message is sent to assist in pause duration adjustments at the source.

Topic 4: Identifying the source priority/TC to pause

The priority/TC used to send the packet at the source may be different than the priority/TC received at the congestion point. Which priority/TC to pause?

Explanation/Solution:

- SFCM includes information to identify the flow which should be paused, as well as pause time.
- Because of the provided flow information in the SFCM, the source knows which queue(priority) needs to be paused.
- Flow-control can be generated at the source accordingly.

Topic 5: Operation in overlay networks (VxLAN, Geneve)

Data Center Networks typically deploy hierarchical overlay networks. Visibility of the inner-flow may be impractical. Which flow is triggering pause?

Explanation/Solution:

- The outer header carries the priority/TC across the data center network
- The outer header carries 'entropy' from the inner flow to enable load balancing. This entropy may be sufficient to distinguish the inner flow by the SFCM receiver.
- The edge device (hypervisor) is responsible for mapping QoS from inner to outer headers
- If knowledge of the encapsulation technic is necessary, a pre-defined set of encapsulations (e.g. VxLAN, Geneve) may be sufficient.
- The SFCM can be configured to define the amount of payload to carry from the original packet. **NOTE:** requires DCN consistent configuration.

Topic 6: Calculation of pause interval

How should the pause interval be determined?

Explanation/Solution:

- Calculation of the Pause Interval (PI) has three components:
  - Time To Drain (TTD) the Overloaded FIFO
  - Time To Source (TTS) from the Overloaded FIFO
  - Time From Source (TFS) to the Overloaded FIFO
- TTD can be calculated from the number of octets in the FIFO and the current FIFO bandwidth
- TTS is the latency for delivering a SFCM from the overload FIFO to the source FIFO
- TFS is the latency for delivery of traffic from the source FIFO to the overload FIFO
- Ideally the Pause Interval seen at the Overload Point will be sufficient to drain the FIFO. Assuming the drain time is TTD = TTS+TFS+PI_n and solving for PI_n we have PI_n = TTD - (TTS + TFS) which is independent from the sourced bandwidth. **NOTE:** The above equations assume an SFC is generated for every packet exceeding a threshold – any SFCM filtering/suppression algorithms may alter this.
- If 0 is used for (TTS + TFS) there is no over-run risk, however the throughput is reduced proportionate to (TTD-(TTS+TFS))/TTD
- **NOTE:** calculation of Pause Interval should be implementation dependent, just like PFC.

Topic 7: SFCM suppression

What mechanism is needed to avoid sending too many SFCM packets?

Explanation/Solution:

- Congestion Isolation has a mechanism for reducing CIM messages, but it requires state held in a stream table. The stream table only needs to hold flows that have been identified as congesting
- SFC only needs the stream information to perform a mapping to the priority/TC to pause
- SFCMs must be generated faster than the Pause Interval apart for each congested flow.
- Alternative ways to suppress messages?
  - (e.g. store a simple SFCM message record with an age relative to the pause interval. Before sending a SFCM message, make sure there is no record of a previous message. Remove old SFCM message records periodically).

Topic 8: Multicast considerations

Does SFC work when the flow is multicast?

Explanation/Solution:

- SFC, as a flow control technique, works on multicast flows as well as unicast flows
- Multicast flows may experience congestion at multiple congestion points, causing a greater number of SFCM packets to be received at the sender.
    - The receiver only needs to pay attention to the longest pause interval received over the current running pause interval. Other SFCMs may be discarded.
    - There is only a single pause interval active at a time by the SFCM receiver
    - Perhaps some sort of discard policy?
- Is there a fairness issue to consider? All flows are paused fairly in time duration; the paused bytes are proportional to their flow rates. SFC aims to push the buffering location to sender-side per-flow queues and rapidly squelch transient flows. That should not 'hurt' fairness in byte metrics or in Jain's fairness index. It's similar to VoQ, which doesn't hurt fairness.
- Multicast has some use with RoCE in the DCN for collective operation and communication, but it is less common. We believe no special treatment is required for multicast at this time.

Topic 9: Source ToR intercept of SFCM packets

What is the specified mechanism for intercepting SFCM messages?

Explanation/Solution:

- The source ToR can intercept SFCM packets sent to the attached station using an egress ACL (i.e. stream filter) matching the UDP port of the SFCM
- If the attached station is SFC aware, the ACL should not be instantiated.
- Additions to DCBX could be used to allow an SFC aware station to advertise support to the source ToR and allow auto-configuration of the ACL.

Topic 10: Consideration of DCBX enhancements

An SFC aware end-station can advertise SFC support using DCBX

Explanation/Solution:

- Source ToR can use DCBX indication to determine whether to install an egress ACL for SFCM and convert to PFC
- DCBX already includes PFC configuration and capability negotiation. SFC capability would be similar.

Topic 11: Mitigating HoLB at Source ToR

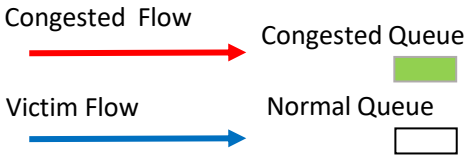
PFC still has a HoLB issue since it is per-traffic class. Asserting PFC at the source should have less negative impact, but is traffic dependent.

Explanation/Solution:

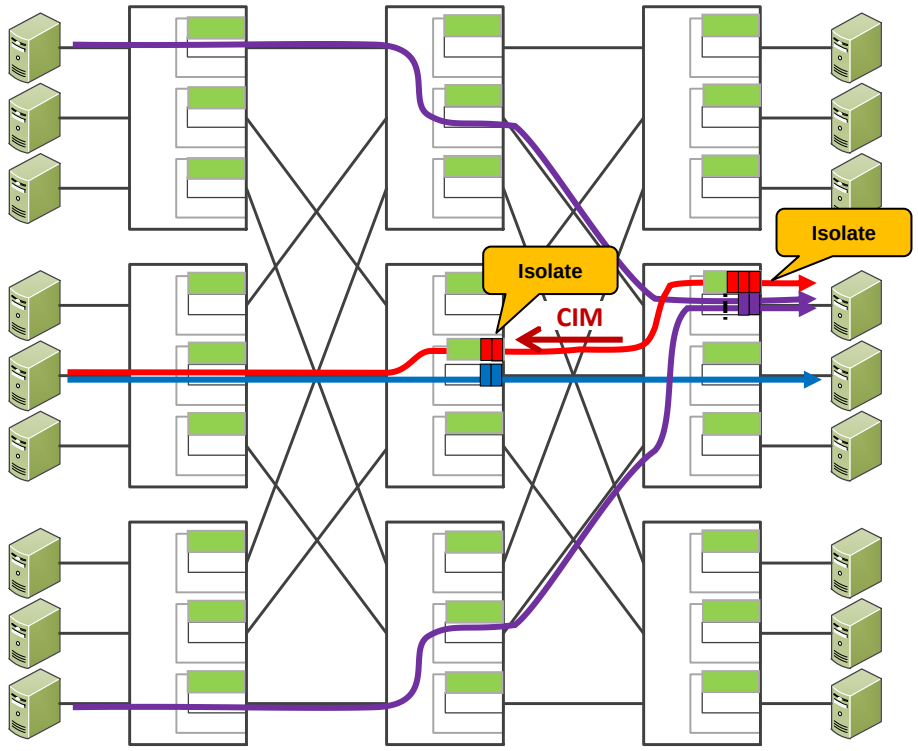
- Topic of future contribution

Backup

Future 802.1 Congestion Management Tools



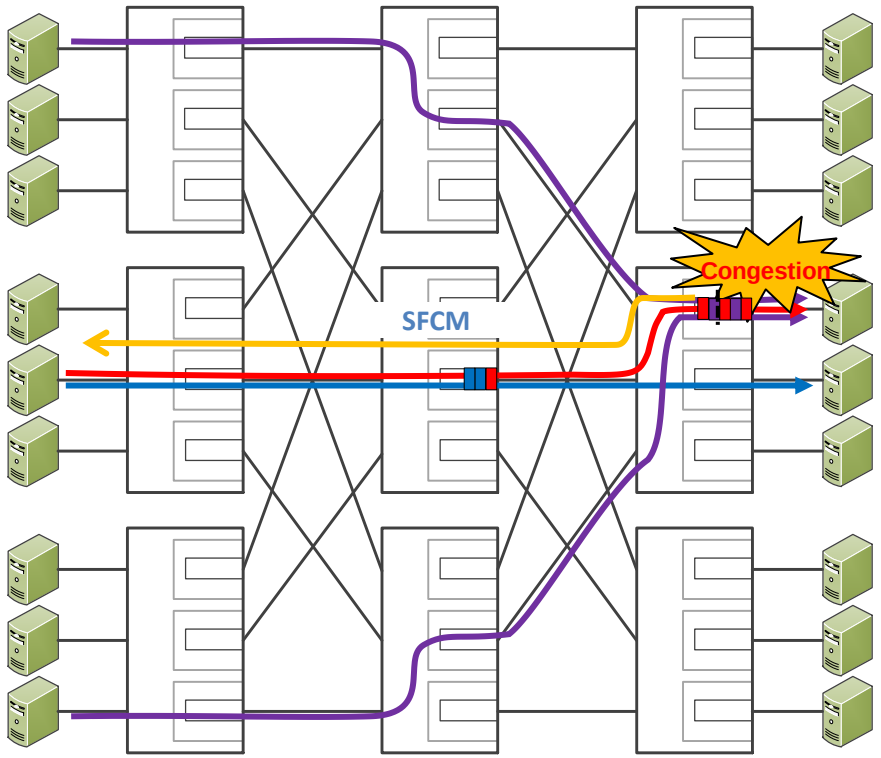
P802.1Qcz - Congestion Isolation



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



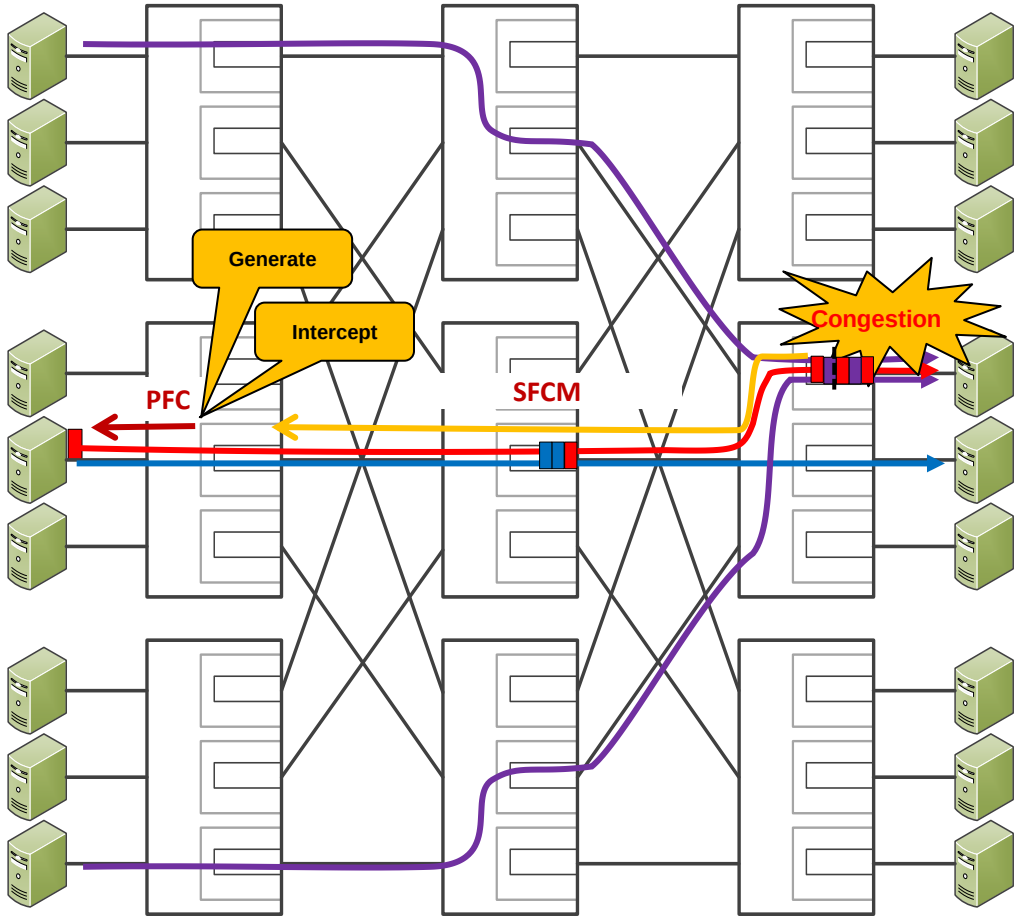
Details

- Can be combined with Congestion Isolation
- If congestion persists, Edge-to-Source signaling using L3 message
- Somewhat like a L3 version of 802.1Qau (L3-QCN), but no Reaction Point (RP) rate controller defined – instead, this is Flow Control
- Optional source Top-of-Rack switch involvement (see next slide)

Top-of-Rack Source Flow Control (proxy)

- Important use case for early deployment.
- ToR intercepts SFCM at egress port connected to non-supporting host using an egress stream_filter matching SFCM UDP port number
- ToR generates traditional PFC frame from SFCM

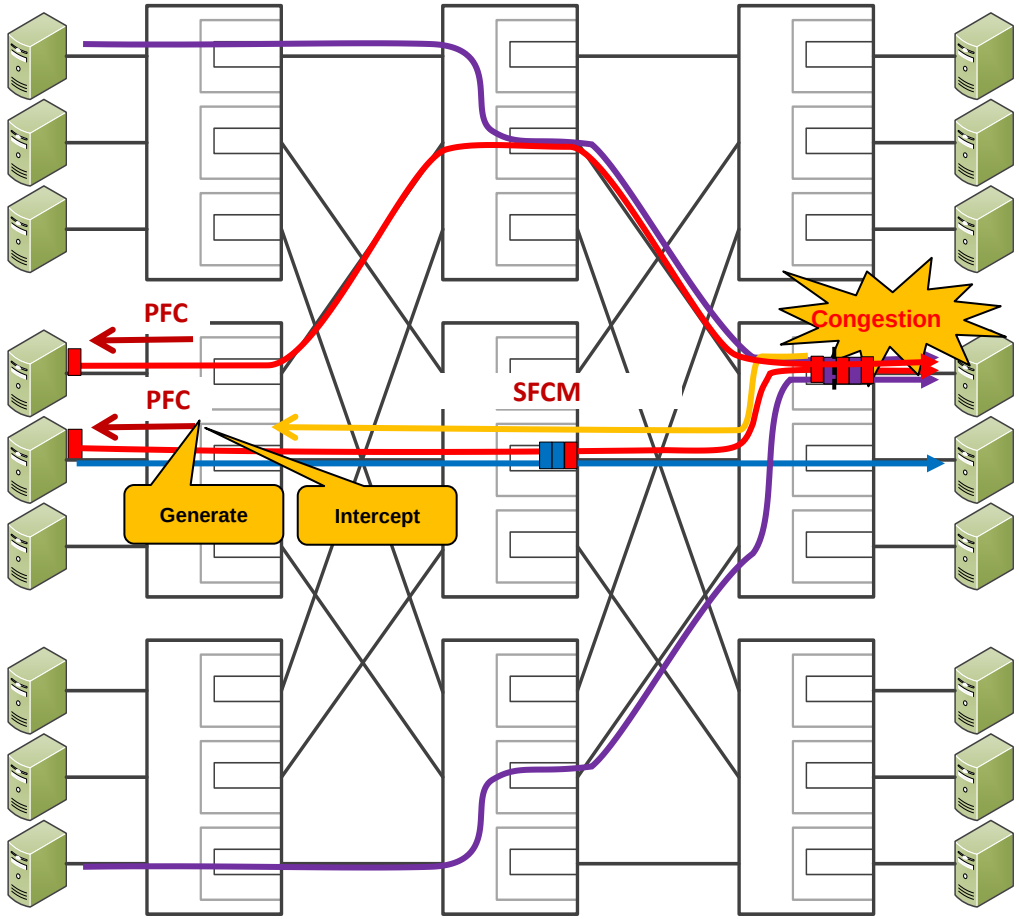
Source Flow Control (ToR Proxy)



SFCM Caching Concept

- Source ToR associates an active pause interval with a particular destination IP address
- New flows to the same destination IP address should be subject to the active pause interval
- Source ToR generates traditional PFC frame on other ports with new flows to the same destination IP address
- **NOTE:** only effective if SFCM was generated by a Destination ToR indicating ‘incast’ congestion.

Source Flow Control (ToR Proxy)



Contributions To Date

- 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/22/1-22-0005-00-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

Notes – From 2/10 Nendica

- Norm and Paul suggest a short list of encapsulation types should be configured. << Why is this necessary? Unclear to me why we need to configure anything related to overlay networks and leave that up to the edges to deal with >>
- Need a diagram that describes how ‘caching’ might work. Could use something like Slide 16, but show new flows from a different port on the same ToR headed the same destination server – as the example