

IEEE 802.1 July 2022 Plenary Session

Source Flow Control Design: Caching

Jeremias Blendin

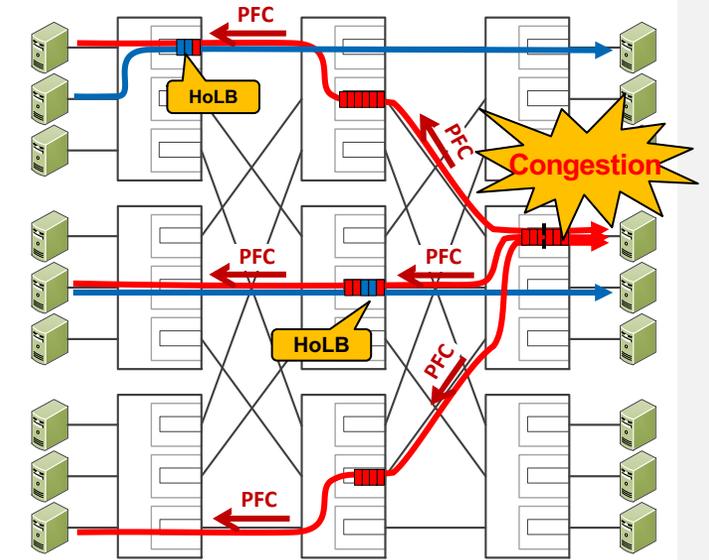
Contributors: Jeongkeun "JK" Lee, Yanfang Le, Paul Congdon



SFC High Level Concept

- Source Flow Control
 - Signal from switch directly to traffic source: per-flow pausing
 - Removes head-of-line blocking from network
 - Simplify deployments compared to PFC
 - Does not require complex buffer tuning
 - Completely remove risk of deadlocks

Today: 802.1Qbb - Priority-based Flow Control (PFC)



Proposed: Source Flow Control (SFC)

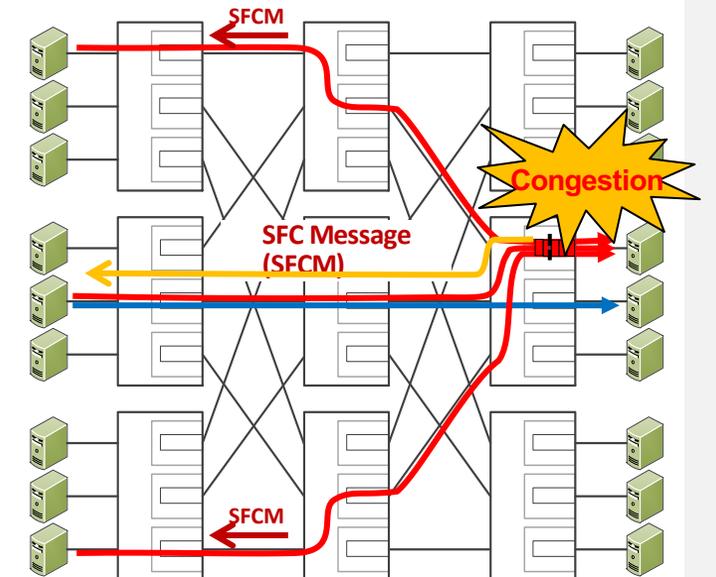
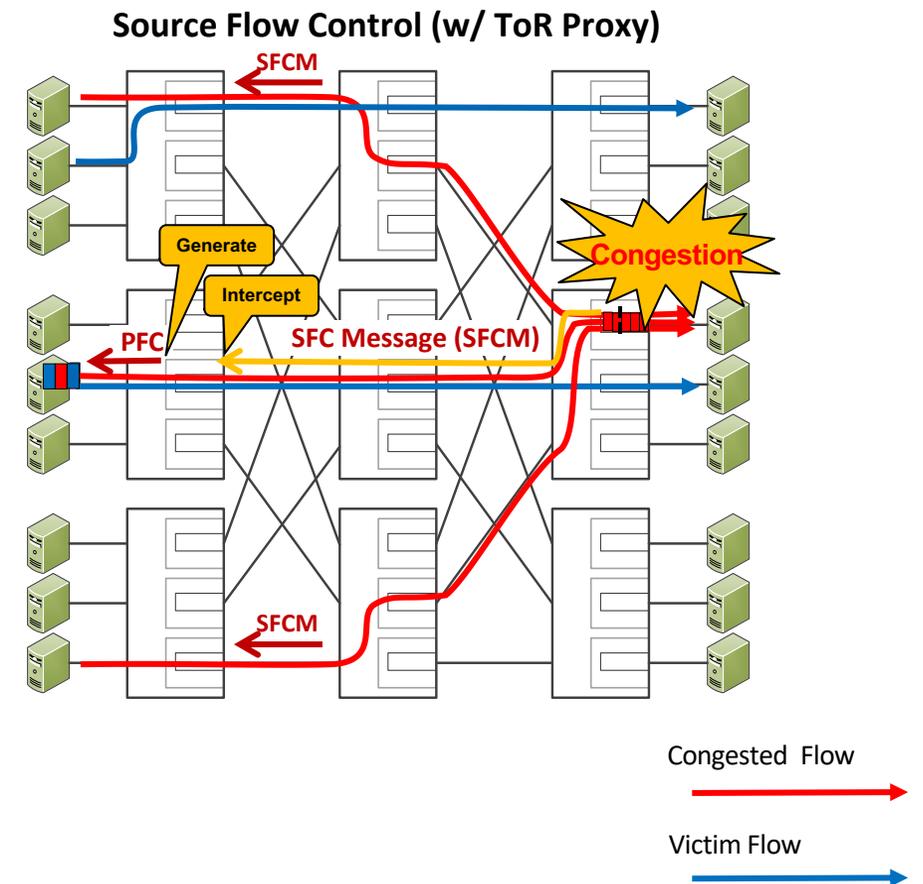


Figure source: <https://mentor.ieee.org/802.1/dcn/21/1-21-0068-01-ICne.pdf>

SFC w/ ToR Proxy (SFC-P)

■ SFC with ToR Proxy

- Works with today's RDMA NICs
- SFC proxy converts SFC message to PFC frame at sender ToR
- Removes congestion from network
 - HoB possible at sender NICs but not in switches



Design Discussion

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.

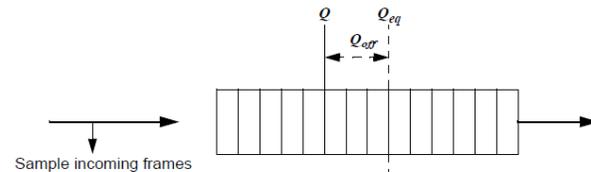


Figure 30-1—Congestion detection in QCN CP

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.

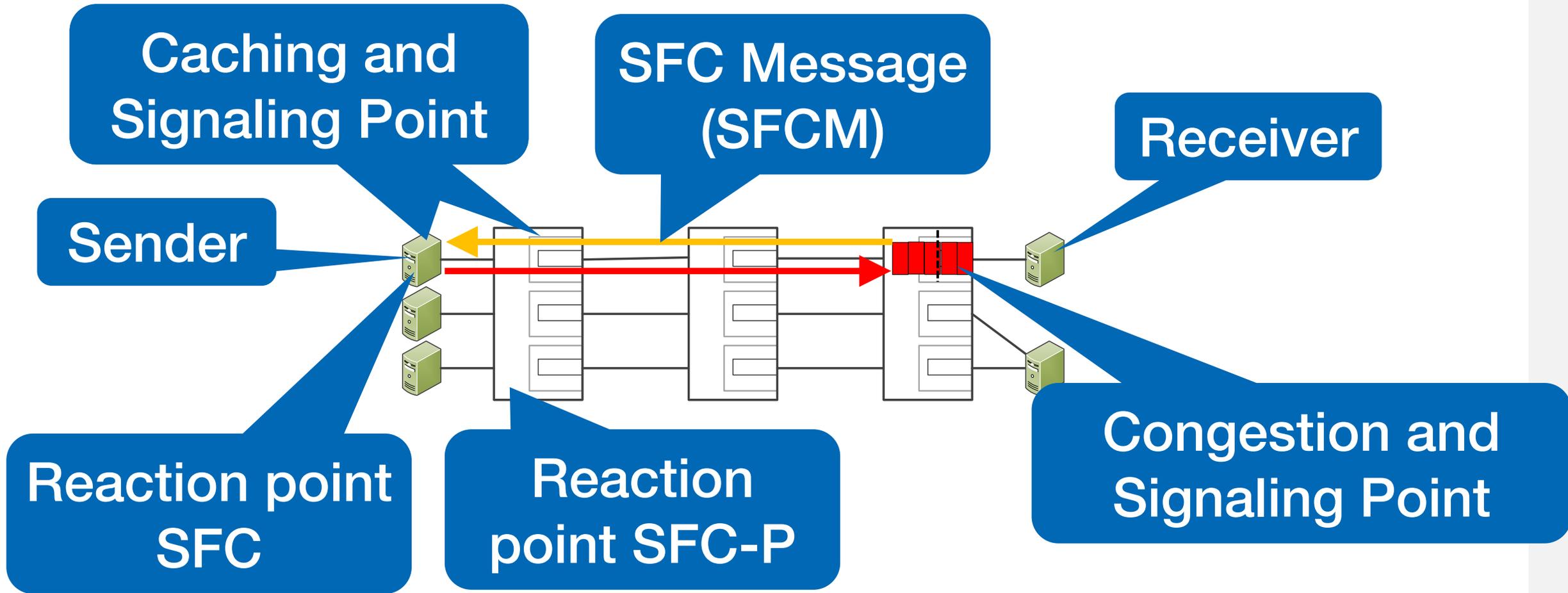
Topic 4: Identifying the source priority/TC to pause

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.
- PFC can be generated to the source accordingly.

Terminology used in this Slide Deck



Terminology based on QCN (802.1Qau) and SFC Design Team: "SFC Design Team Topics", IEEE March 2022

SFC Message Contents: What to Pause?

- Baseline: Identify specific 5-tuple
 - Easy to implement: first X bytes of original packet
 - Enables per-flow pausing, works for all use cases
 - Simple, yet effective
- SFC Caching
 - Enable cache points to identify all current and future traffic within the pause period going to the Congestion Point
 - Discuss details later

Minimal SFC Message Contents

SFC Header

version,
header length,
etc

Pause
duration
[us]

x bytes of
original
packet
header

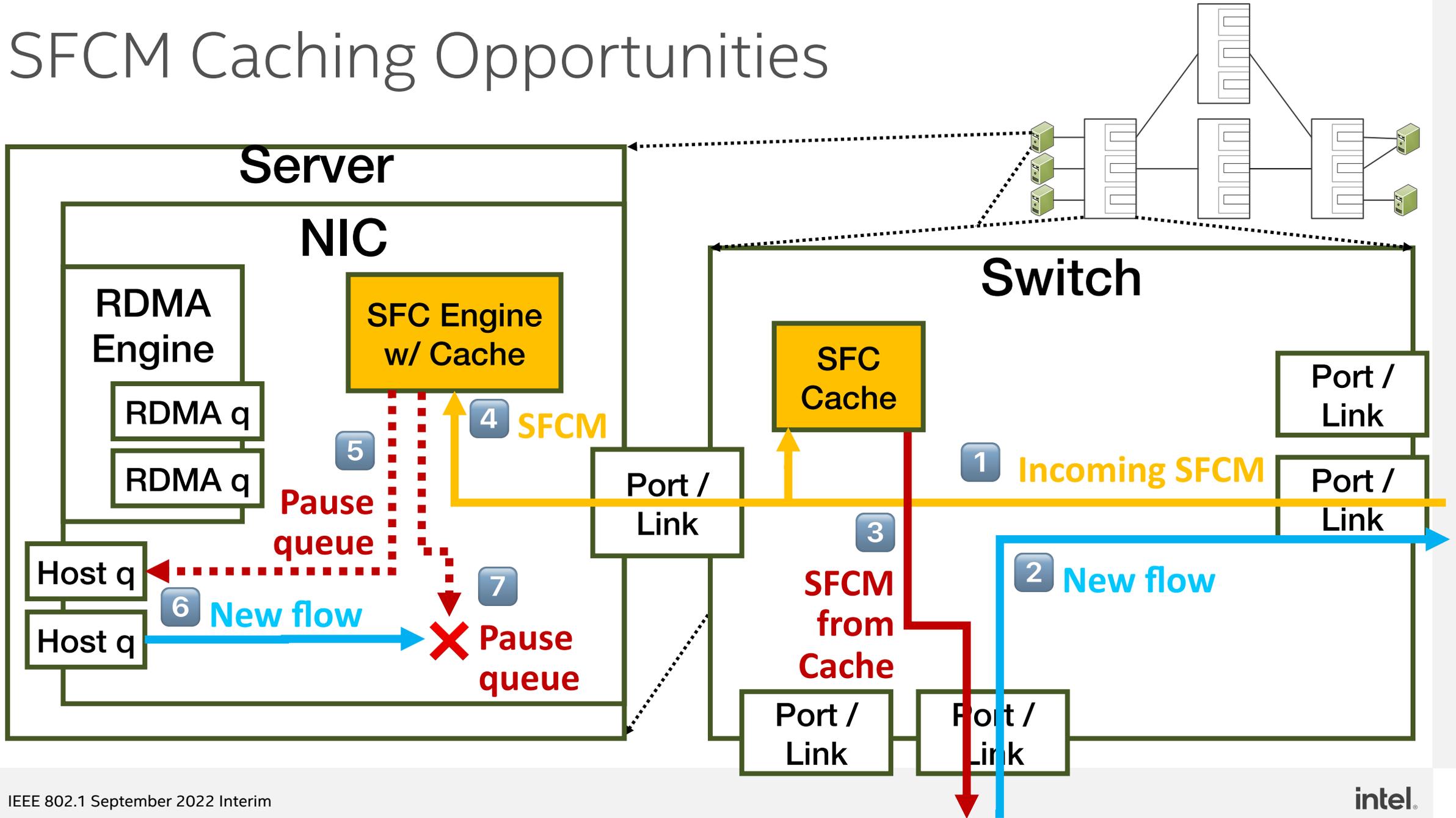
TCP
IP
...
Ethernet

SFC Caching

Congestion Point Locator

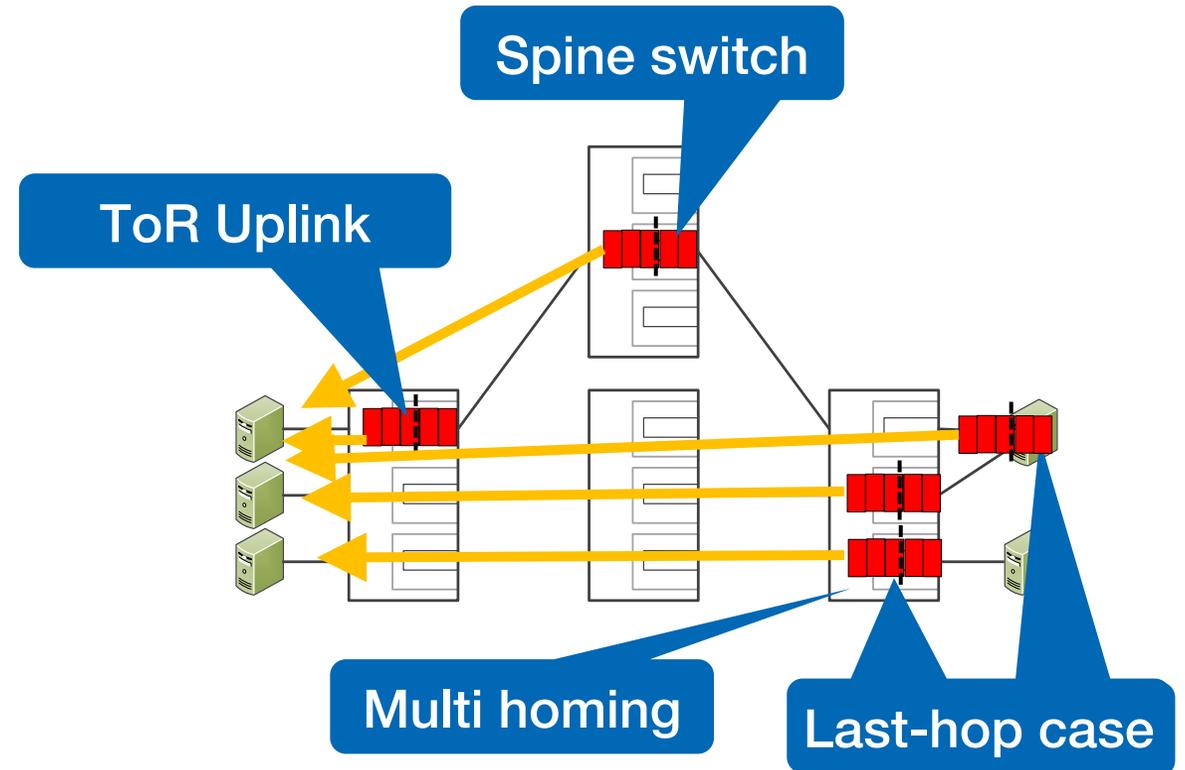
- For incast scenarios, caching is important
- Cache points (details on next slide)
 - Sender ToR
 - Sender NIC
- Use Congestion Point Locators
 - Specify traffic patterns to pause
 - The original message is not sufficient
 - When tunneling is used: which header should be used?
 - IPv6: destination host might have a /64 prefix assigned
 - Multiple DSCP values might map to congested queue

SFCM Caching Opportunities



Congestion Point Locators

- Specification of congested queue
 - Enable senders to identify traffic going to the congestion point within the pause period
- Last-hop case
 - Congestion point is part of all paths to the receiver
 - Covers incast use cases
- Other cases
 - Congestion point is only part a subset of paths to the receiver



Congestion Point Locator: Last-hop case

- Port identification
 - Destination prefix of receiver
- Queue identification
 - TC as is used by PFC
 - Simple: can use 8bit one hot encoding
 - Requires consistent traffic to TC mappings in reaction point and signaling point
 - List of PCP/DSCP values that map to the queue on the congested switch
 - PCP/DSCP to TC mappings might be different on different switches
 - No TC mapping synchronization between reaction and signaling point required
 - Complex header format (list with up to 64 6bit values)

Our Thoughts SFC Message Contents

SFC Message

version,
header
length, etc

Pause
duration
[us]

Possible options

Cache:
Use
packet
header

Cache:
prefix1
TC
bitmap

Cache:
prefix2
DSCP
values

x bytes
of
original
packet
header

TCP
IP
...
Ethernet

intel®