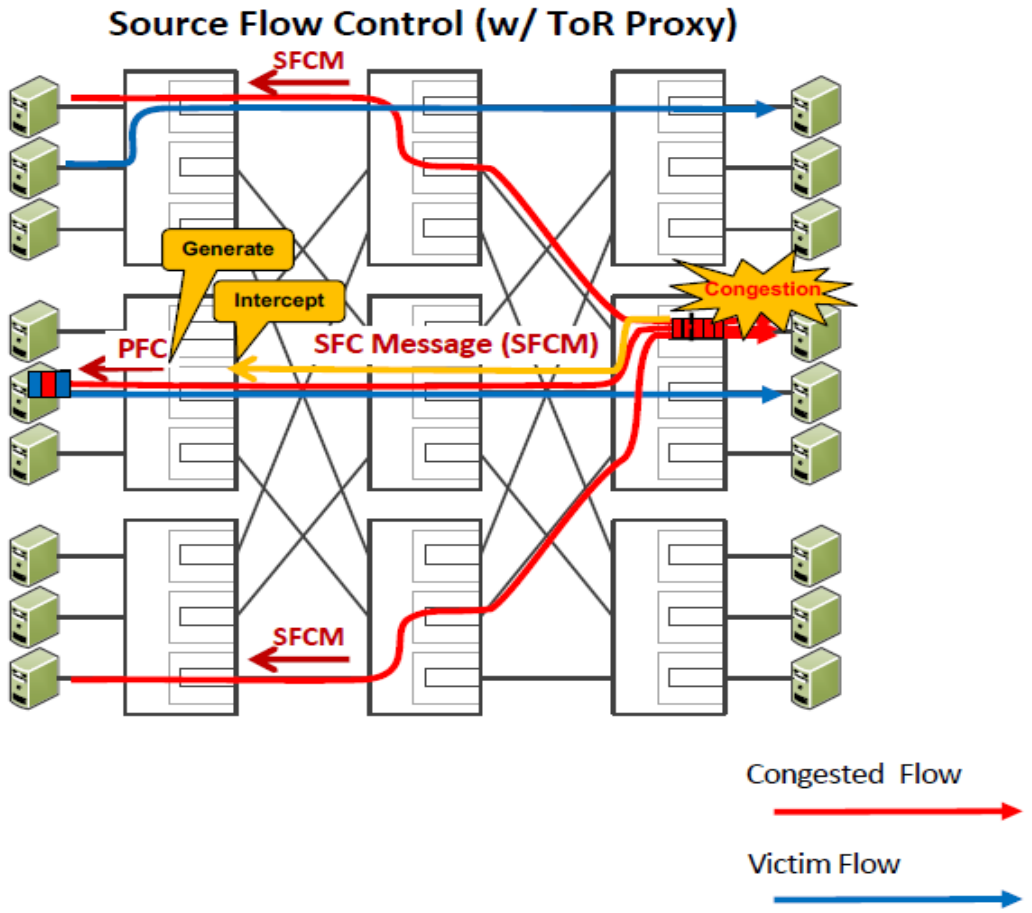
Issue to be Considered in SFC

Lily Lv (Huawei)

Fei Chen (Huawei)

SFC Proxy Mode Recap

- **What is SFC-P**
  - “This amendment specifies the optional use of existing stream filters to allow bridges at the edge of the network to intercept and convert signaling messages to existing Priority-based Flow Control (PFC) frames.”
- **Importance of SFC-P**
  - Fast way for deployment leveraging current PFC implementation on datacenter network.
  - No dependency on host upgrades
- **Implied principle of SFC-P implementation**
  - No change on host-side implementation
  - No change on PFC implementation (except the way to invoke PFC)



Key to Design SFCM for SFC-P

Point 1: SFCM can be captured by sTOR

Point 2: SFCM can tell sTOR which port to initiate PFC

Point 3: SFCM can tell sTOR PFC parameters (priority, pause time)

Previous consideration on above points.

Point 1: SFCM can be captured by sTOR

✓SFCM is sent to source host and sTOR is on the path.

Point 2: SFCM can tell sTOR which port to initiate PFC

✓sTOR knows egress port of SFCM destination by looking up internal address table.

Point 3: SFCM can tell sTOR PFC parameters (priority, pause time)

✓SFCM PDU contains ‘pause duration in us’.

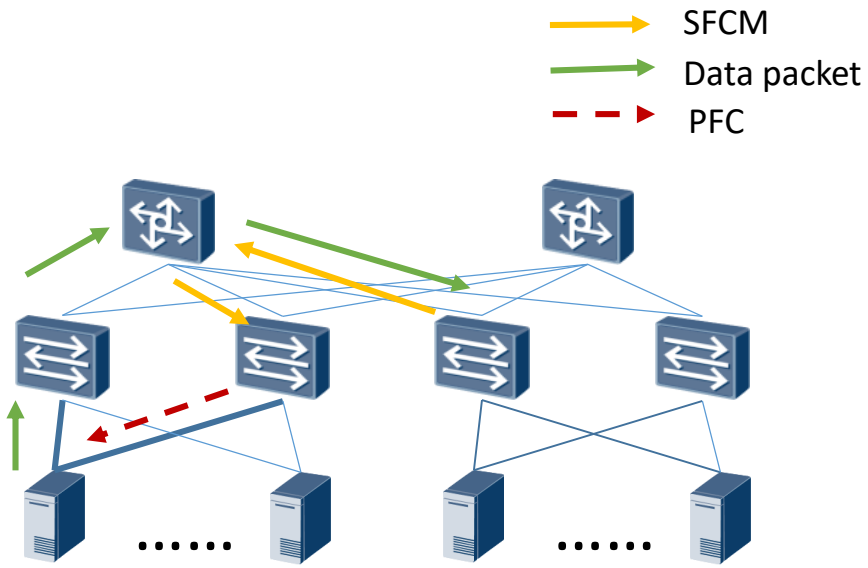
✓SFCM PDU contains ‘priority’ which is “the priority parameter of the EM_UNITDATA.request (6.8.1) of the frame that triggered the creation of the SFCM”

(https://www.ieee802.org/1/files/public/docs2022/dw-congdon-individual-text-1122-v01.pdf)

Re-visit Previous Consideration

- **Multi-homing**

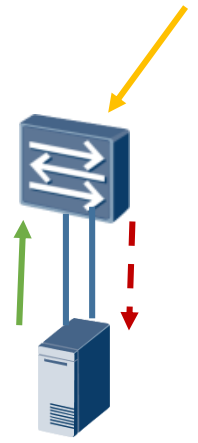
✘ SFCM is sent to source host and sTOR is on the path.



  - Virtualization technologies, such as stacking, allows server to connect to more than one switches in order to increase reliability. It works as if server connects to one virtual switch.
  - SFCM may be sent back to host server via another sTOR.
  - That will initiate PFC on the wrong port of server.

- **Link aggregation**

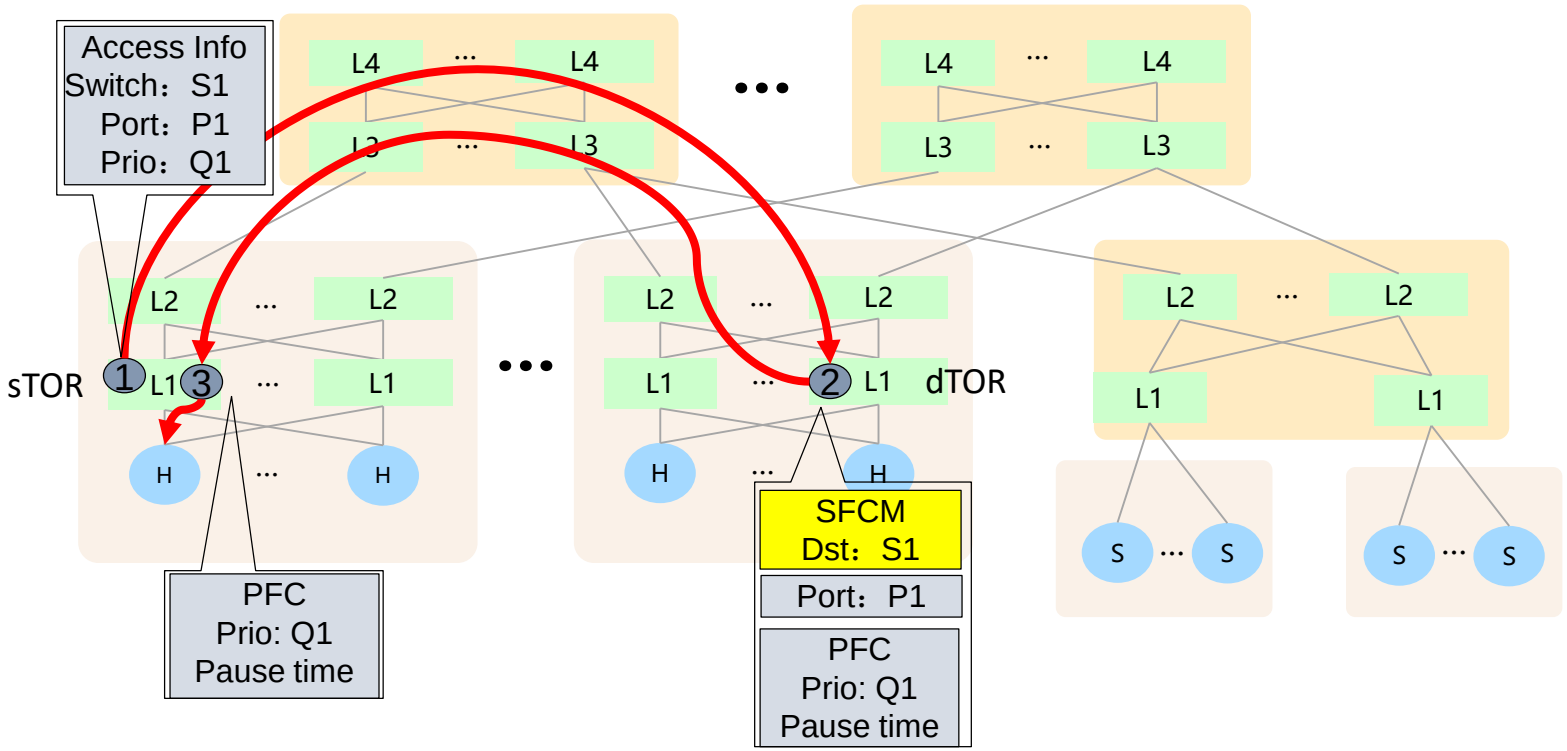
✘ sTOR knows egress port of SFCM destination by looking up internal address table.



  - In load sharing mode, the aggregated ports are all active.
  - sTOR may choose a wrong port to send pause frame, thus it has no effect to mitigate congestion.

Proposed Solution (1/2)

- **General idea: insert network access point information into data packets when packets enter the network**
  - **Network access point is STOR**
  - **Access point information includes switch ID, port ID, priority ID**



Proposed Solution (2/2)

- **Behavior of sTOR**
  - Insert access point information into data packets (see subsequent slides for more detail)
  - When receiving SFCM, generate PFC according to SFCM content
- **Behavior of congestion point (CP is not limited to dTOR)**
  - Generate SFCM
    - Extract access point information from data packet
    - Structure SFCM to be sent to sTOR instead of source host
    - Structure SFCM containing port ID, priority ID which are from access point information
  - Send SFCM back to sTOR
  - Remove the inserted access point information
- **Behavior of dTOR**
  - Remove the inserted access point information

Access Point Information Design (1/2)

- **What is in access point information?**
  - Switch ID: a network domain unique ID
    - It is assigned when configuring SFC-P
    - If it is a IP address, switch ID can be used for forwarding directly
    - If it is a number, an ID->forwarding address table should be maintained in switch
  - Port ID: a switch domain unique ID
    - Port number on the switch
  - Priority ID: priority 1~8
    - Packet header may include the priority information, like DSCP or PRI

SFC-P-Tag

SFC-P-Tag maximum size (6B)

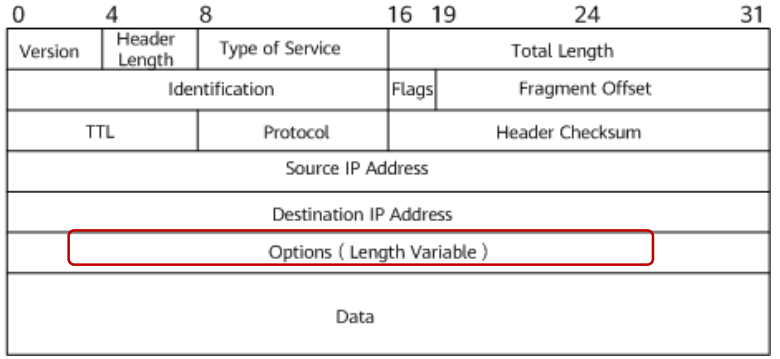
32 bits	8 bits	8bits
Switch ID	Port ID	Priority ID

SFC-P-Tag compact size (3B)

16 bits	8 bits
Switch ID	Port ID

Number of supported switch: 65535
Number of supported port: 255

- **How to insert it into data packet?**
  - Option 1： Insert it in IP header



Options：0～40B

  - Option 2： Utilize INT(In-band Network Telemetry) technology

There is already congestion control mechanism using INT, such as HPCC

3. System Overview https://datatracker.ietf.org/doc/draft-miao-tsv-hpcc/01/

Figure 1 shows the end-to-end system that HPCC++ operates in. During the traverse of the packet from the sender to the receiver, each switch along the path inserts inband telemetry that reports the current state of the packet's egress port, including timestamp (ts), queue length (qLen), transmitted bytes (txBytes), and the link bandwidth capacity (B), together with switch_ID and port_ID. When

Access Point Information Design (2/2)

- **How to generate SFCM with access point information?**
  - Extract switch ID, port ID and Queue ID from data packet
  - Convert switch ID to switch address if it is not the ip address
    - Switch should maintain a switch ID -> switch address table.
  - Fill the converted address in SFCM destination field
  - Add port ID and priority ID into SFCM payload
- **Overhead consideration**
  - 6B/MTU 1500B =0.4%
  - Optimized way to insert SFC-P-Tag：
    - Only tag the elephant flow
    - Tag the packet every half RTT

SFCM Design

Add one type in Option TLVs.

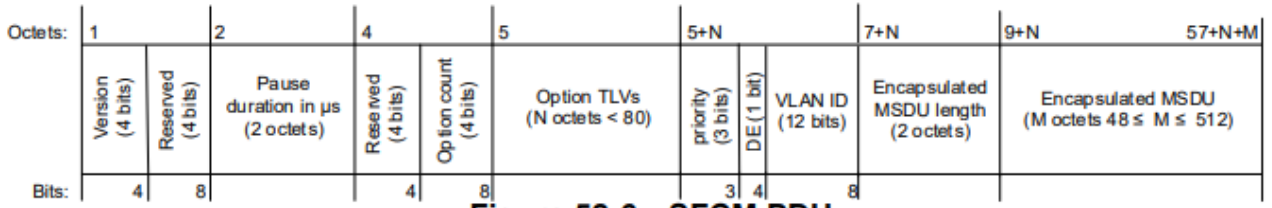


Figure 52-6—SFCM PDU

Table 52-1—Option TLV definitions

Type	MSDU required	Description	Length (octets)	TLV value layout			
0	1	DSCP in MSDU	0	N/A			
1	0	DSCP / IP Prefix	6 for IPv4 18 for IPv6	DSCP (8-bits)	Address family (1 bit)	Address prefix (7 bits)	IP Address (4 or 16 octets)
2	0	TC / IP Prefix	6 for IPv4 18 for IPv6	TC (8-bits)	Address family (1 bit)	Address prefix (7 bits)	IP Address (4 or 16 octets)
7-126	0						
127		Organizationally specific	3 < n < 64	See Figure 52-8			

Example:

Type	MSDU required	Description	Length	TLV value layout		
3	1	Switch ID/Port ID/Priority ID	6	Switch ID(32 bits)	Port ID(8 bits)	Priority ID(8bit)

Summary

- **SFC-P is a fast way for the new feature deployment.**
- **Multi-homing and link aggregation are common for server-TOR connection. Current SFC-P design consideration has an issue in such scenarios.**
- **Propose a solution to solve the issue.**
- **Propose to add SFC-P-Tag (switch ID/port ID/Queue ID) in SFCM as option TLV.**

Thanks

Link Aggregation

that priority is paused (i.e., if Priority_Paused[n] is TRUE (see 36.1.3.2) on that port. When Transmission Selection is running above Link Aggregation, a frame of priority n is not available for transmission if that priority is paused on the physical port to which the frame is to be distributed.

The PFC Receiver entity acts per physical port. When Transmission Selection is running above Link Aggregation, each PFC Receiver entity processes the M_CONTROL.indication primitives as specified in 36.1.3.2, and maintains and makes available to Transmission Selection the vector of the Priority_Paused[n] variables, indicating the state of each of the eight priorities of that physical link, as shown in Figure 36-4.

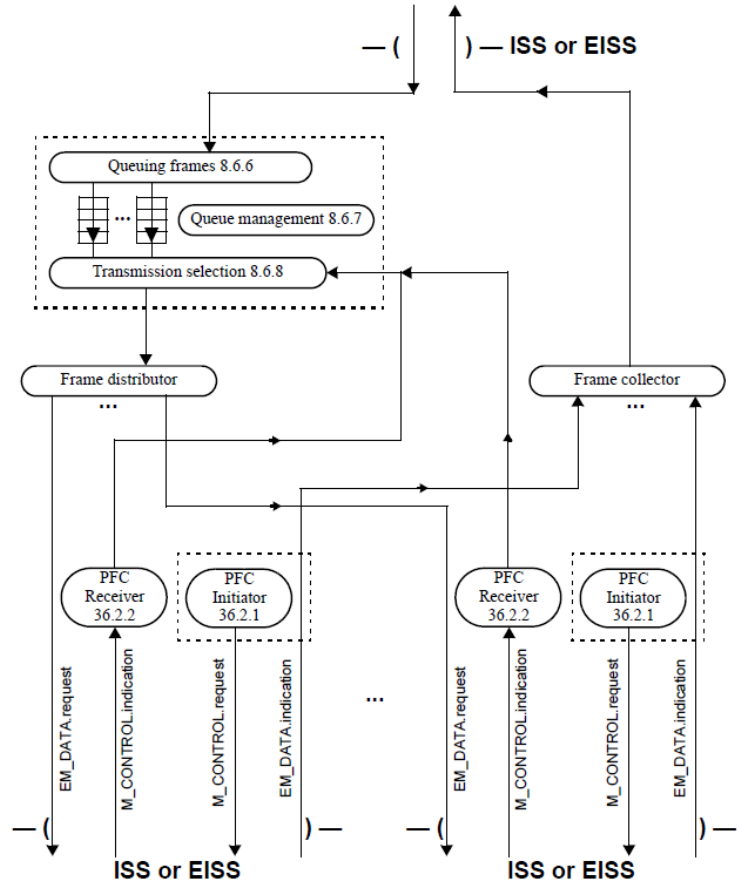


Figure 36-4—PFC-aware system queue functions with Link Aggregation