IEEE 802.1 January 2023 Interim Session

Source Flow Control Design: Management

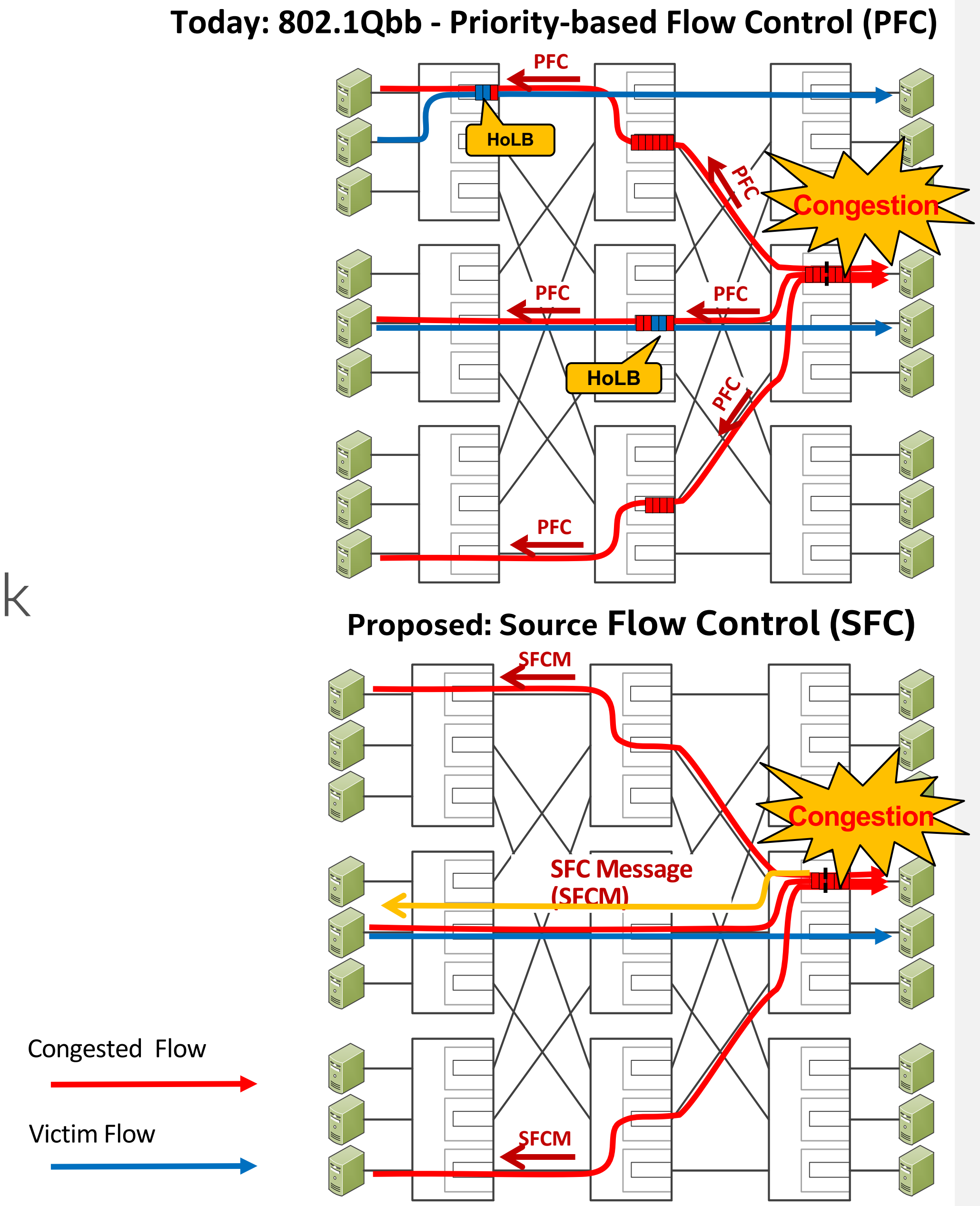
P802.1Qdw contribution

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

intel®

SFC High Level Concept

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



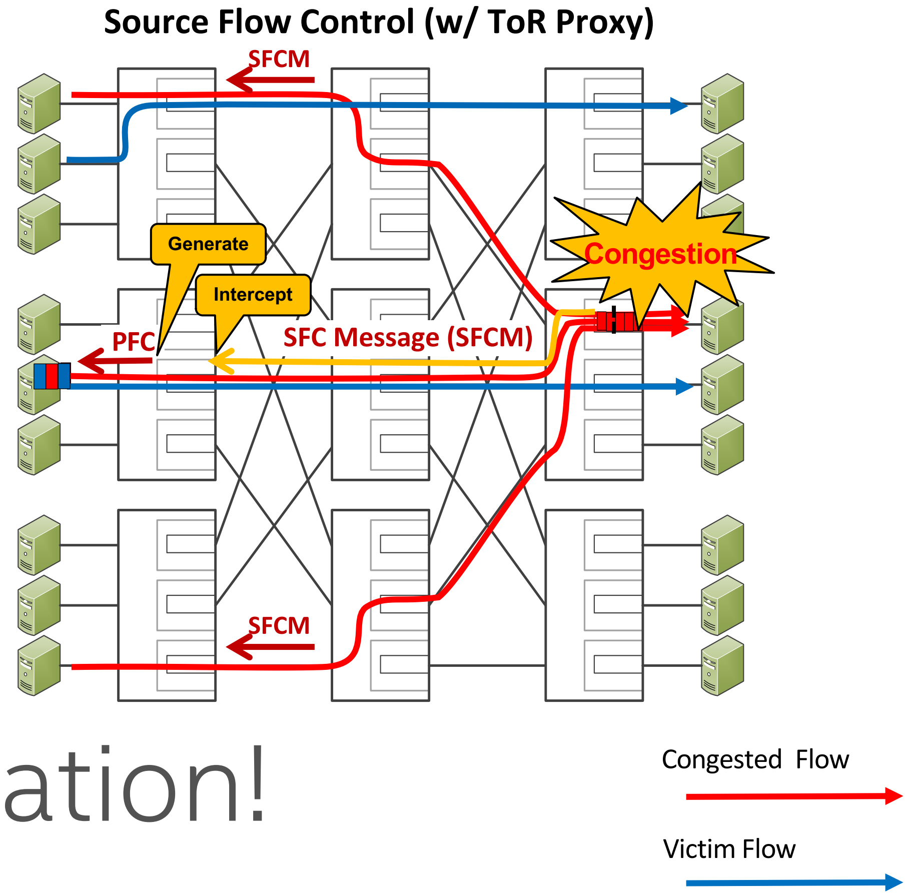
Today: 802.1Qbb - Priority-based Flow Control (PFC)

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
    - HolB possible at sender NICs but not in switches



Not the focus of this presentation!

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

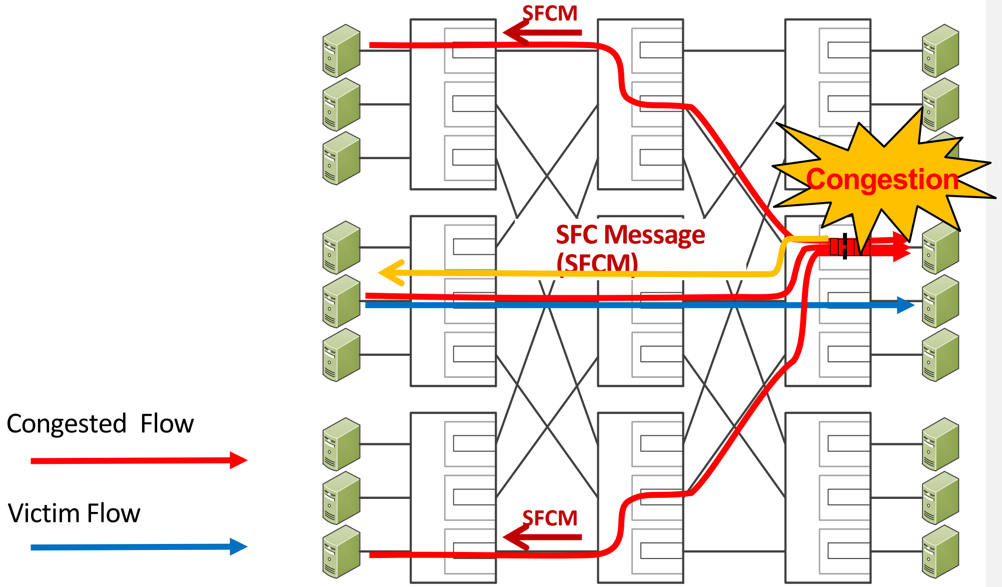
Managed Objects

Overview

- Goals
  - Minimize configuration to absolutely required parts
  - Separate base model for native SFC from SFC-P specific details
- Management hierarchy
  - SFC Entity (per bridge/end station)
  - SFC Instance (per VLAN/virtual network)
  - SFC Instance Port (per instance per port)
    - Multiple SFC Instance Port rows can exist per physical port
  - SFC Queue (per queue)
    - One queue is assigned to zero or more SFC Instance Ports.
- Open issues: SFC-P/caching options, counters

Key Management Parameters

- Enable/disable feature
- Proxy mode enablement
- SFCM packet header fields
  - DSCP
  - UDP dst port
  - Source IP address
  - IP connectivity requires configuration parameters per-VLAN/virtual network



SFC Entity Managed Row Elements

Name	**Data type**	**Operations supported**	**Conformance**	**AutoCfg**
sfcMainEnable	Boolean	RW	BE	No
sfcSFCMSuppressionInterval	unsigned integer [0..1024]	RW	BE	No
sfcSFCMDestUDPPort	UDP port number	RW	IETF RFC 798	No

SFC Entity Managed Row Elements

- sfcMainEnable (Boolean)
  - Enable SFC feature globally on this device.
- sfcSFCMSuppressionInterval (unsigned integer [0..1024])
  - Suppression interval in microseconds. The interval during which SFCM to the same destination are suppressed, only one SFCM per destination per interval will be sent.
- sfcSFCMDestUDPPort (UDP port number)
  - Destination UDP address used for sending and accepting SFCM

SFC Instance Managed Row Elements

Name	Data type	Operations supported	Conformance	AutoCfg
sfcEnable	Boolean	RW	BE	No
sfcSFCMTransmitDiffServCodePoint	unsigned integer [0..64]	RW	BE	No
sfcSFCMTruncationLength	unsigned integer [0..512]	RW	BE	No
sfcProxyMode	Enum (off, on, auto)	RW	B	No
Opt: sfcEnableCaching	TBD			

- There does not seem to be a IETF standard on how to model the relation between IP networks, VLANs, and ports
  - This draft has expired: https://datatracker.ietf.org/doc/html/draft-ietf-netmod-sub-intf-vlan-model-07

SFC Instance Managed Row Elements

- sfcEnable (Boolean)
  - Enable SFC feature in general. Masks all other configuration parameters if false
- sfcSFCMTransmitDiffServCodePoint (unsigned integer [0..64])
  - DiffServ Code Point (DSCP) to use in IP header of SFCM
- sfcSFCMTruncationLength (unsigned integer[0..512])
  - Length to which the original packet in the SFCM is truncated to
- sfcProxyMode (Enum (off, on, auto))
  - Defines the proxy mode behavior: off, on to override automatic configuration and auto to accept negotiation from DCBX

SFC Instance Port Managed Row Elements

Name	Data type	Operations supported	Conformance	AutoCfg
sfcSFCMSourceAddressIPv4	IPv4 address	RW	IETF RFC 791, BE	Opt
sfcSFCMSourceAddressIPv6	IPv6 address	RW	IETF RFC 8200, BE	Opt
sfcAdminPortMode	Enum (disable, native, proxy-mode, auto)	RW	BE	No
sfcOperPortMode	Enum (disable, native, proxy-mode)	R	BE	Yes

SFC Instance Port Managed Row Elements

- sfcSFCMSourceAddressIPv4/6 (IPv4/IPv6 address)
  - IP address to use a source address for the SFCM. The IP address of the port of the congested queue should be used as source if available, the VLAN/VRF port otherwise.
- sfcAdminPortMode (Enum (disable, native, proxy-mode, auto))
  - Configures the port mode
    - disable: no send/accept SFCM or translate PFC messages, no DCBX override
    - native: do send/accept native SFCM, no DCBX override
    - proxy-mode: do send/accept PFC and translate to/from SFCM, no DCBX override. Only valid if sfcProxyMode in the SFC Instance is enabled. If any SFC instance is in this mode on a port, all SFC instances must use this mode on this port.
    - auto: allow any port mode through DCBX and subject to sfcProxyMode, disable port by default.

SFC Instance Port Managed Row Elements

- sfcOperPortMode (Enum: disabled, native, proxy-mode)
  - Displays the port’s mode:
    - disabled: no send/accept SFCM or translate PFC messages
    - native: do send/accept native SFCM
    - proxy-mode: do send/accept PFC and translate to/from SFCM.

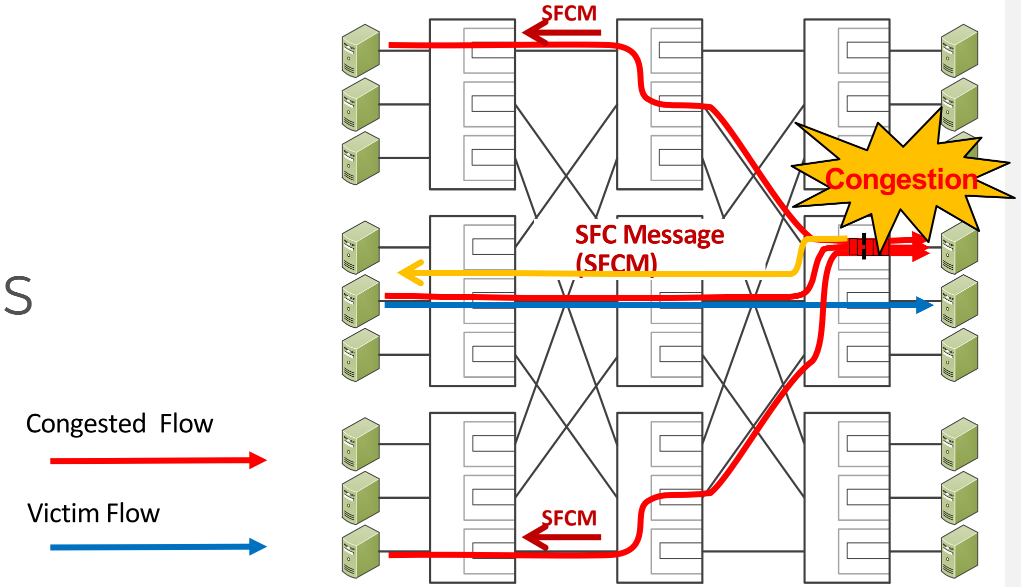
SFC Queue Managed Row Elements

Name	Data type	Operations supported	Conformance	AutoCfg
sfcMonitorQueue	Boolean	RW	BE	No

- sfcMonitorQueue (Boolean)
  - Monitor and send SFCM for traffic of the associated VLANs in this queue
  - For native SFC, the flow classification and traffic class -> priority -> queue mapping is not relevant.
  - SFC requires that all traffic assigned to a monitored queue is part of an SFC-enabled VLAN
  - It is the task of the queueing system to specify when a queue is congested and for how long senders should be paused to alleviate the congestion.

Operational Model for SFC

- Bridges use SFC to control congestion
- SFC operates between a congested bridges and senders of congesting flows
  - End stations must adhere to SFCM pause specification
  - There are no SFC parameter interactions between congested bridges
  - The SFC trigger mechanism and pause duration calculation is specific to a bridge
    - Not required to be part of the specification, although it might be beneficial to do so.



Conclusion

- We proposed a managed object model for native SFC
  - Key features and fundamental structure is in place
- Future topics
  - Yang model of L3 (virtual) networks to L2 port mapping
    - This seems to be an open issue
  - Configuration of queue to DSCP mapping for generating SFCM
    - SFC-P
    - Caching
  - Configuration of other caching-related parameters

intel®