P802.1Qdw text contribution overview

Lihao Chen (lihao.chen@huawei.com)

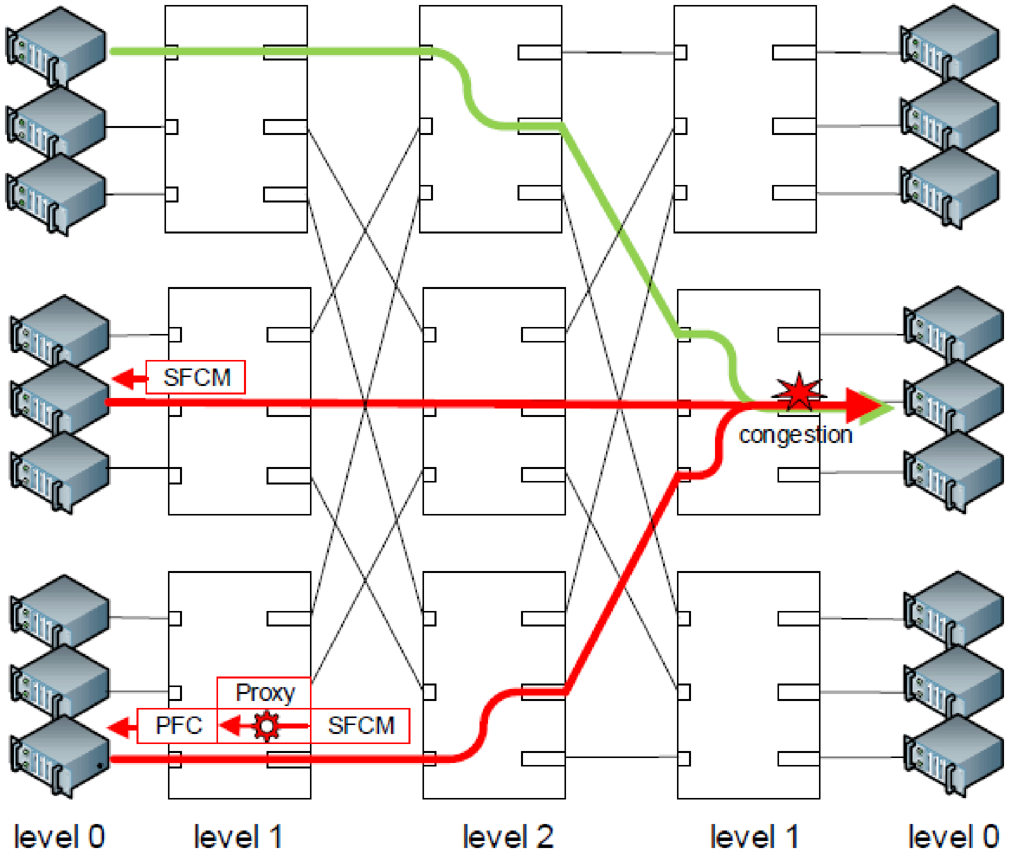
Previous contributions

- July Plenary
  - > https://www.ieee802.org/1/files/public/docs2024/dw-chen-recap-restart-0724-v01.pdf

- September Interim
  - > https://www.ieee802.org/1/files/public/docs2024/dw-chen-text-contribution-overview-0924-v01.pdf
  - > https://www.ieee802.org/1/files/public/docs2024/dw-chen-individual-text-0924-v01.pdf
  - > The whole document of the text contribution was presented.

Recap

- Discussion started: 17 Jan 2022
  > HPE, Huawei, Intel, …
  > Congestion, in particular incast congestion (AI DC networks).
- PAR approved: 21 Sep 2022
  > Scope: … for the signaling and remote invocation of flow control at the source of transmission in a data center network… to allow bridges at the edge of the network to intercept and convert signaling messages to existing Priority-based Flow Control (PFC) frames…
  > Precise PFC, quick reaction, easy adoption.
- Why SFC? How about other 802.1Q tools?
  > PFC – slow react on root cause of incast. HoLB, spread, deadlock.
  > ECMP (load balancing) – has nothing to do with incast congestion.
  > CI – has no reaction on root cause of incast.
  > QCN – NIC based rate-limiters, L2 addressing. The idea of QCN and SFC have similarities, i.e., upstream signaling from congestion point. But SFC is flow control.



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

Contributor’s Notes

- Clause 52, the meat of this text contribution, follows the structure of Congestion Notification (Clause 30-33) and Congestion Isolation (Clause 49):
  - > SFC Objectives and Principles
  - > SFC Entity (bridge and end station) operations
  - > SFC Protocol (Variables, Procedures, Encoding of PDUs)

- This text contribution (compares to the previous),
  - > Add 52.5.2.3 condTransmitSfcmPdu() procedure, and 52.5.1.2.1 sfcmMinInterval correspondingly.
  - > Reconstruct 52.5.2.4 pauseTimeCalc()() into buildAndSendSfcm().
  - > Add 52.5.2.6 addSfcSource() along with 52.3.4 SFC Source Table and 52.5.2.6 periodicTableCleanup().
  - > Add Layer-2 and IPv6 SFCM PDU encapsulation, and modify 52.5.3 Encoding of the SFCM PDU accordingly.

SFCP Procedures overview (SFCM sender side)

sfcInitialize()

->
EM_UNITDATA.request
Called by Queuing Frames. Check if the target queue of the frame is a monitored queue. (sfcMonitorQueues)

->Yes!
Check if the frame has caused congestion in the monitored queue. (by any methods)

->Yes!
Call **addSfcSource()**, add an entry indexed by the source address of the congesting flow for the SFC Source Table if the index does not exist.
Call **condTransmitSfcmPdu()**. Check if the condition sfcmMinInterval is met.

->Yes!
Call **buildAndSendSfcm()**. Fill the SFCM PDU with the information from SFC entity variables(52.5.1), either configured or from the SFC Source Table.

periodicTableCleanup().

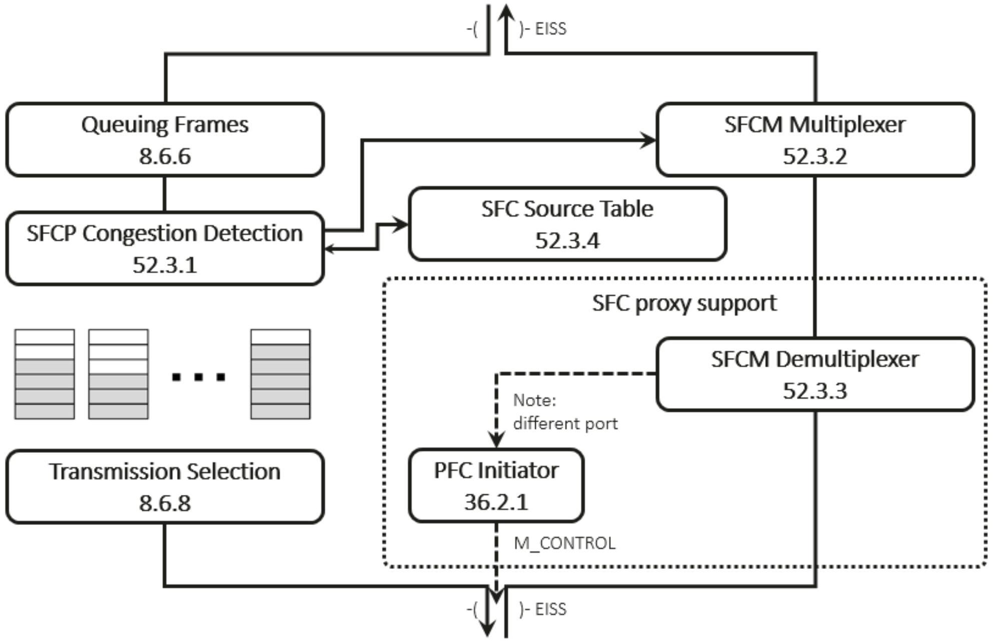


Figure 52-2—Bridge component SFC reference diagram

SFCP Procedures overview (SFCM receiver side)

processSfcmPdu()
The SFCM reaches its destination?

->Yes!
According to the information provided by the SFCM PDU,
->Execute the PAUSE (End station).
->Invoke a PFC (proxy mode bridge).

->No!
->forward the SFCM (bridge).

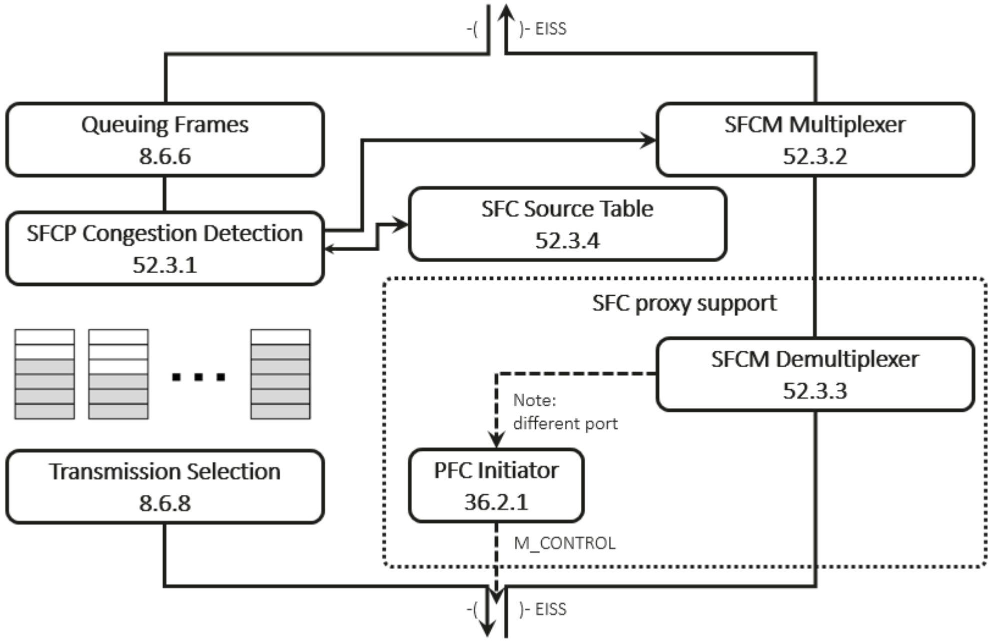


Figure 52-2—Bridge component SFC reference diagram

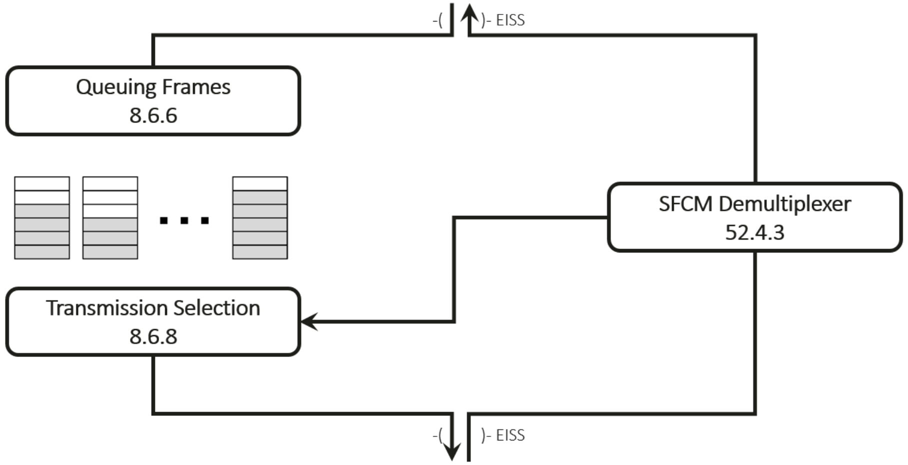


Figure 52-3—End station SFC reference diagram

Next steps

- Unfinished SFCP procedures and SFCM PDU.
- Management objects, YANG data models, and enhancements to DCBX protocol to advertise the new capability.

- Give more quantized analysis on SFC.
- Thoughts and feedbacks?

Questions?

Back up - SFCP & CIP Procedures comparison

sfcInitialize()

->

EM_UNITDATA.request

Monitored queue? Cause congestion?

->

addSfcSource() <-> SFC Source Table.

condTransmitSfcmPdu() <-> Time elapse > sfcmMinInterval.

->

buildAndSendSfcm(). <-> SFCP entity managed object & SFC Source Table.

periodicTableCleanup().

One frame can trigger SFC to the source.

ciInitialize()

->

EM_UNITDATA.request

Monitored queue? Cause congestion? stream_handle is present?

->

addCongestingFlow(), delCongestingFlow(), flushCongestingFlows() <-> CI Stream Table

condTransmitCimAddPdu() <-> ciCIMCount<cipMaxCIM

transmitCimDelPdu()

->

buildAndSendCim() <-> CIP entity managed object & CI Peer Table & CI Stream Table

periodicTableCleanup()

One frame can trigger the CI to the peer.

Need to store the flow information to identify and change enqueuing.