Proposal for the P802.1Qdw SFC

Lihao Chen (lihao.chen@huawei.com)

HUAWEI

Previous contributions from the author

- 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.

- November Plenary
  > https://www.ieee802.org/1/files/public/docs2024/dw-chen-text-contribution-overview-1124-v02.pdf
  > https://www.ieee802.org/1/files/public/docs2024/dw-chen-individual-text-1124-v02.pdf
  > The SFC procedures were presented.

Proposal

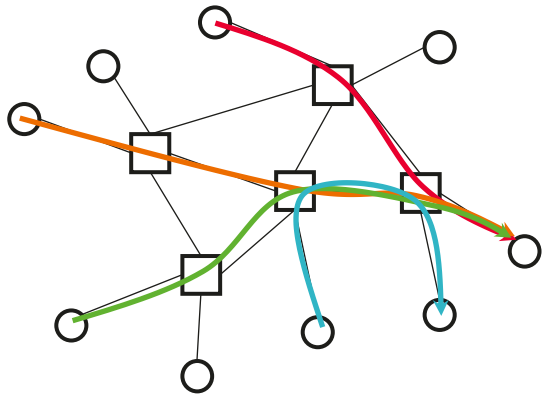
- The PAR and CSD of IEEE P802.1Qdw were agreed.
- The author believes this project should proceed, as it is a crucial and indispensable addition to the 802.1 toolset for handling congestion and flow control.
  > SFC and its benefits (qualitatively). - Page 4
  > Simulations prove benefits (quantitatively). - Page 5

- Proposal for drafting the text.
  > Structure aspects, refer to previous standards and text contributions. - Page 6
  > High level technical aspects, follow the scope to provide a straightforward solution. - Page 7, 8, 9
  > Protocol specification. - Page 10, 11, 12
  > The informative math in the Annex. - Page 13
  > Regarding the details of text drafting, follow the regular procedures, i.e., editor assignment, drafting, comments, comment resolution, ballots.

The SFC benefits for IEEE 802 congestion and flow control

- The differences and similarities compare to CN (Qau), PFC (Qbb), and CI (Qcz):
  > Page 4 and 5 in https://www.ieee802.org/1/files/public/docs2022/new-congdon-SFC-proposal-0322-v01.pdf
  > Clause 52.2.5, 52.2.6, and 52.2.7 in https://www.ieee802.org/1/files/public/docs2024/dw-chen-individual-text-1124-v02.pdf
- A comparison of PFC, DCQCN, SFC in P18 https://www.ieee802.org/1/files/public/docs2022/new-bottorff-sfc-0322-v6.pdf
- Besides the commonly used end-to-end congestion control algorithms and hop-by-hop flow control (i.e., PFC), a new edge-to-edge flow control scheme is needed.
  > Page 4 in https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf

- SFC Key features:
  > **Fastest reaction** to stop Reaction Point Egress.
    - Crucial for ultra-high-bandwidth Ethernet.
    - “Precise PFC” mitigates PFC side effects, e.g., PFC storm, HoL, deadlock.
  > **Non-scenario-specific**. Congestion cased by many-to-one traffic is very common.
  > **Easy adoption**. No mandatory math, and end stations don't necessarily need to be modified.



SFC Simulation results

- Simulations show the benefits of SFC.
  - > https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf
  - > https://mentor.ieee.org/802.1/dcn/21/1-21-0061-00-ICne-source-remote-pfc-test.pdf
  - > https://www.ieee802.org/1/files/public/docs2022/new-blendin-SFC-sim-0522-v01.pdf
  - > https://www.ieee802.org/1/files/public/docs2022/new-blendin-SFC-Simulation-Results-0722-v01.pdf

- New simulations is on the way.
  - > Can be used for the subsequent verification of the detailed design of the SFC protocol.

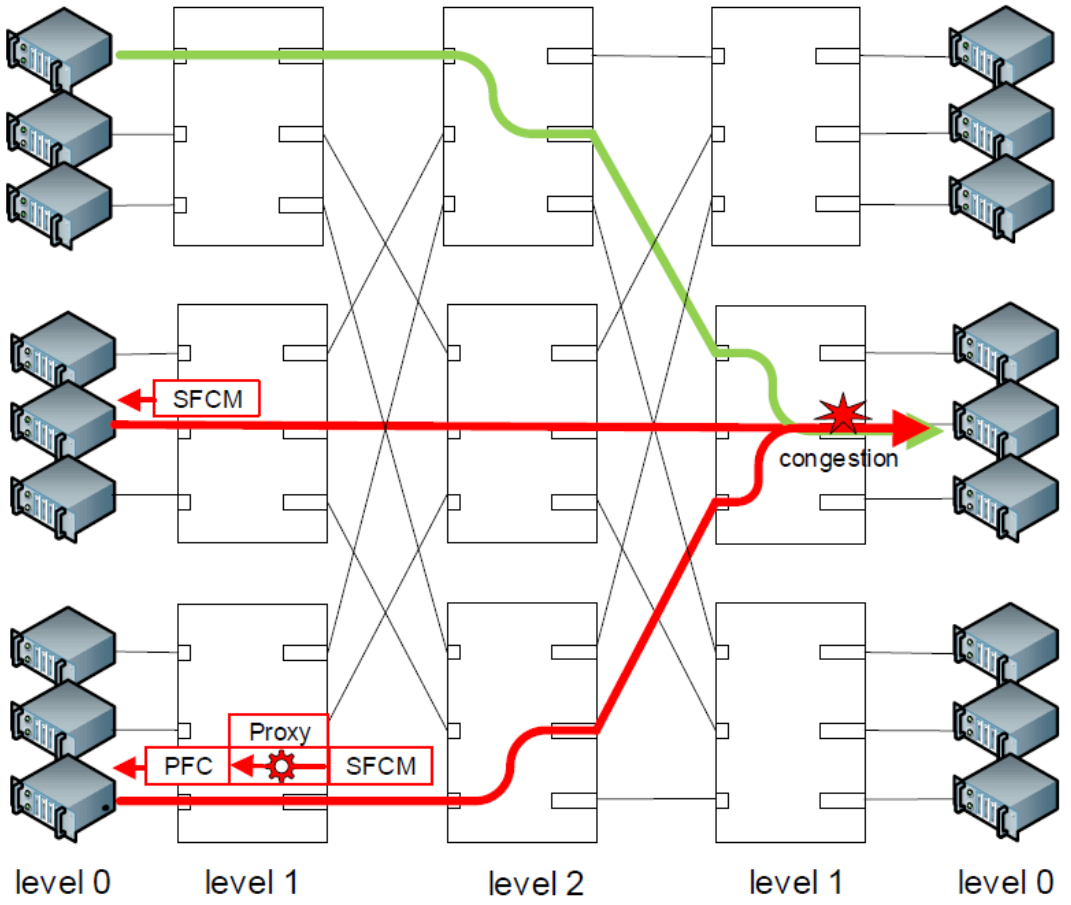
Proposal for drafting the text - Structure aspects

- Refer to previous standards and text contributions.
  > IEEE 802.1Q-2022 Clause 30-33 (Quantized) Congestion Notification.
  > IEEE 802.1Qcz-2023 Congestion Isolation.
  > https://www.ieee802.org/1/files/public/docs2022/dw-congdon-individual-text-1122-v01.pdf
  > https://www.ieee802.org/1/files/public/docs2024/dw-chen-individual-text-1124-v02.pdf

Contents	QCN (in 802.1Q-2022)	CI (in 802.1Qcz-2023)	SFC (in individual text)
Objectives	Clause 30	Clause 49.1	Clause 52.1
Principles	Clause 30	Clause 49.2	Clause 52.2
Entity operation	Clause 31	Clause 49.3	Clause 52.3 and 52.4
Protocol (variables, procedures, encodings)	Clause 32	Clause 49.4	Clause 52.5

Proposal for drafting the text - High level technical aspects

- Follow the scope to provide a straightforward solution.
  - > This amendment specifies procedures, managed objects, and a YANG data model for **the signaling and remote invocation of flow control at the source of transmission** in a data center network. This amendment specifies enhancements to the Data Center Bridging Capability (DCBX) protocol to advertise the new capability. 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**. This amendment also addresses errors and omissions in the description of existing IEEE Std 802.1Q functionality.
  - > 52.1 SFC objectives
  - > 52.2 SFC principles



HUAWEI

Proposal for drafting the text - SFC bridge component

- Provide a basis for specifying the externally observable behavior of SFC entity operations.

  > 52.3 bridge component

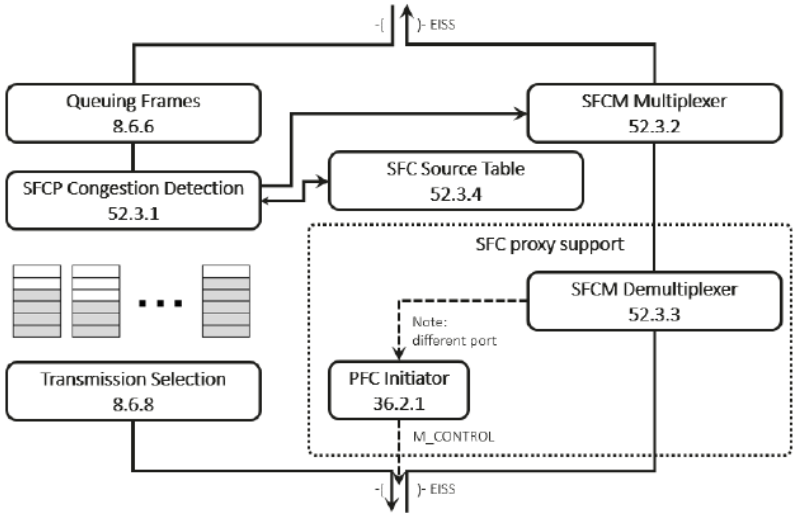


Figure 52-2—Bridge component SFC reference diagram

SFC

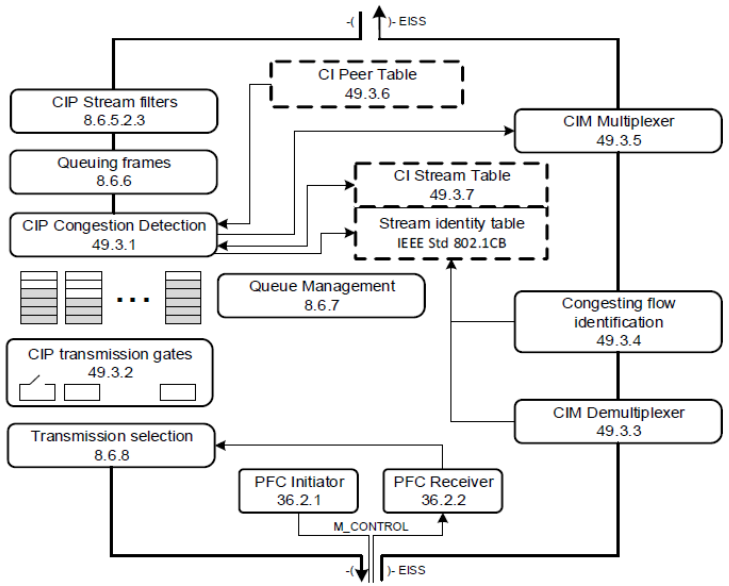


Figure 49-2—Congestion Isolation reference diagram

CI

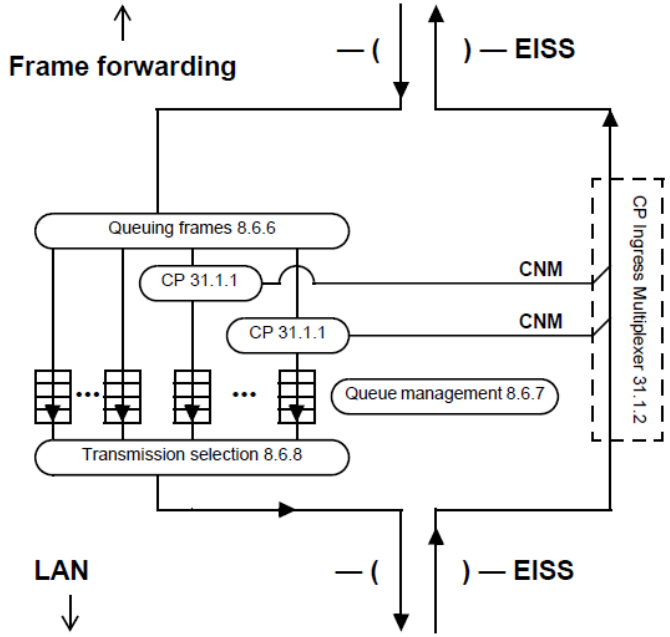


Figure 31-1—CPs and congestion-aware queues in a Bridge

CN

HUAWEI

Proposal for drafting the text - SFC end station component

- Provide a basis for specifying the externally observable behavior of SFC entity operations.

  > 52.4 end station component

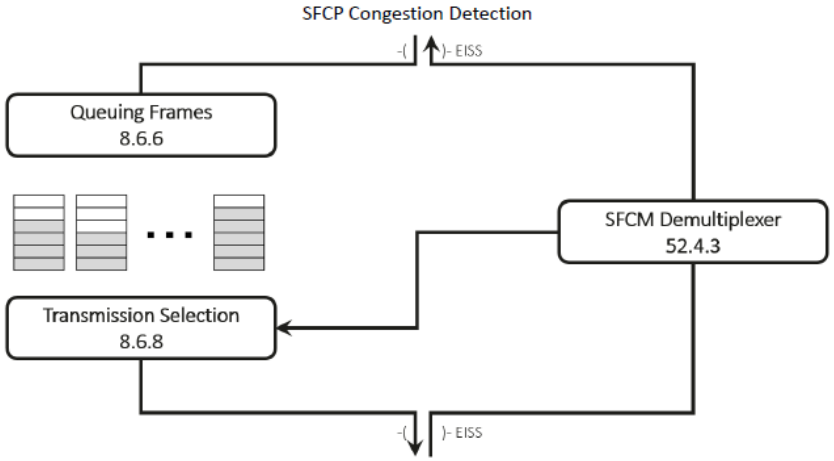


Figure 52-3—End station SFC reference diagram

SFC

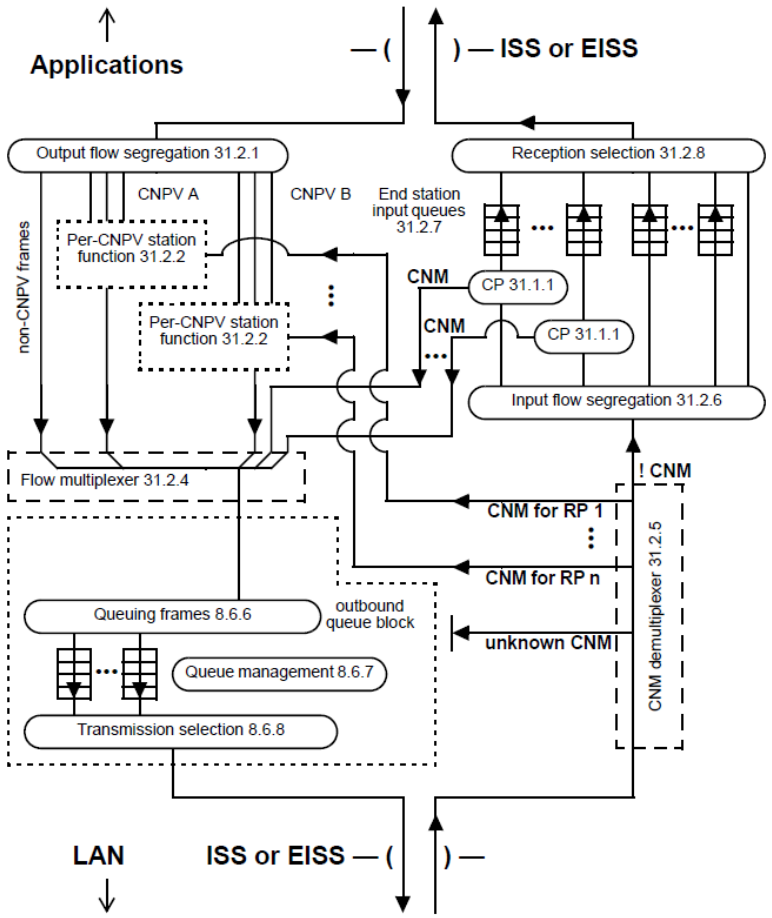


Figure 31-2—Congestion-aware queue functions in an end station

CN

HUAWEI

Proposal for drafting the text - Protocol aspects

- SFC Protocol specification is the meat of clause 52 SFC.
  - > 52.5.1 Variables - specified based on the need of SFCP procedures and SFCM PDU encodings.
  - > 52.5.2 SFCP (SFC Point) Procedures - see next pages
  - > 52.5.3 Encoding of the SFCM (SFC Message) PDU
    - Layer 2 encapsulation: use EtherType 89-A2 with subtype value 1 (CIM uses subtype value 0).
    - IPv4 and IPv6 encapsulation: similar to CIM.

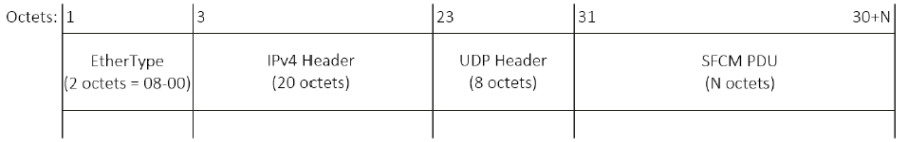


Figure 52-5—IPv4 SFCM Encapsulation

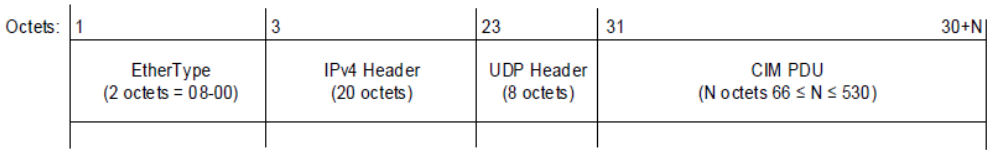


Figure 49-4—IPv4 layer-3 CIM encapsulation

  - SFCM PDU: version, type, length, and value.
  - > 52.5.4 LLDP SFC TLV and procedures - TBD

HUAWEI

Proposal for drafting the text - Protocol procedures

- The sender side (sender of SFCM):

---

sfcInitialize()

Frames given to SFCP Congestion Detection by the Queuing Frames entity in an **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()

Frames given to CIP Congestion Detection by the Queuing Frames entity in an **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.

Proposal for drafting the text - Protocol aspects (receiver side)

- The receiver side (receiver of SFCM):

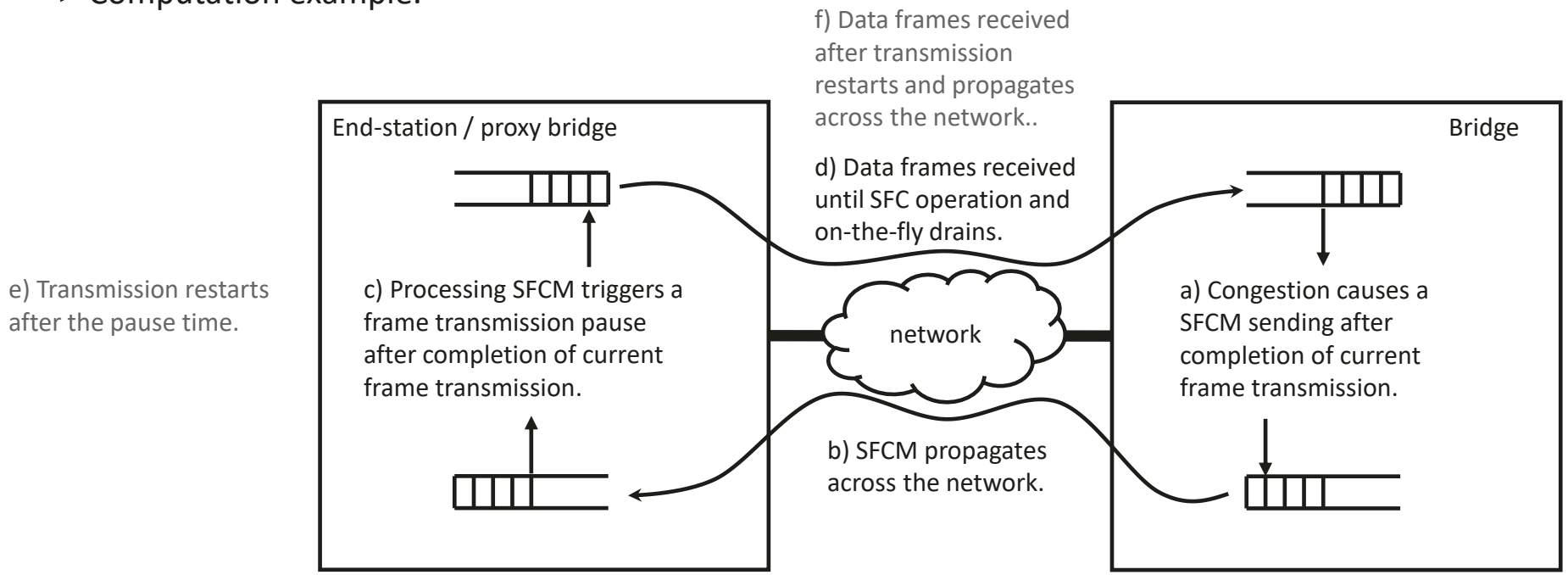
The SFCM Demultiplexer identifies SFCMs and invoke
processSfcmPdu()

SFCM reaches 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 (proxy mode bridge).

- The author's opinion is that the SFC procedure is simple and unidirectional. Therefore, there is no need to define a state machine.
  > E.g., the outcome of executing processSfcmPdu() on a SFC bridge is either forwarding the SFCM or invoking a PFC, and it is solely based on the input of SFCM. And executions of processSfcmPdu() are independent.

Proposal for drafting the text - Math in the Annex (informative)

- Buffer requirements for SFC. Can refer to Annex N (802.1Q) Buffer requirements for PFC.
  - > Overview.
  - > Delay model.
  - > Computation example.



HUAWEI

Questions?

SFCP Procedures overview (SFCM sender side)

sfcInitialize()

Frames given to SFCP Congestion Detection by the Queuing Frames entity in an **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().

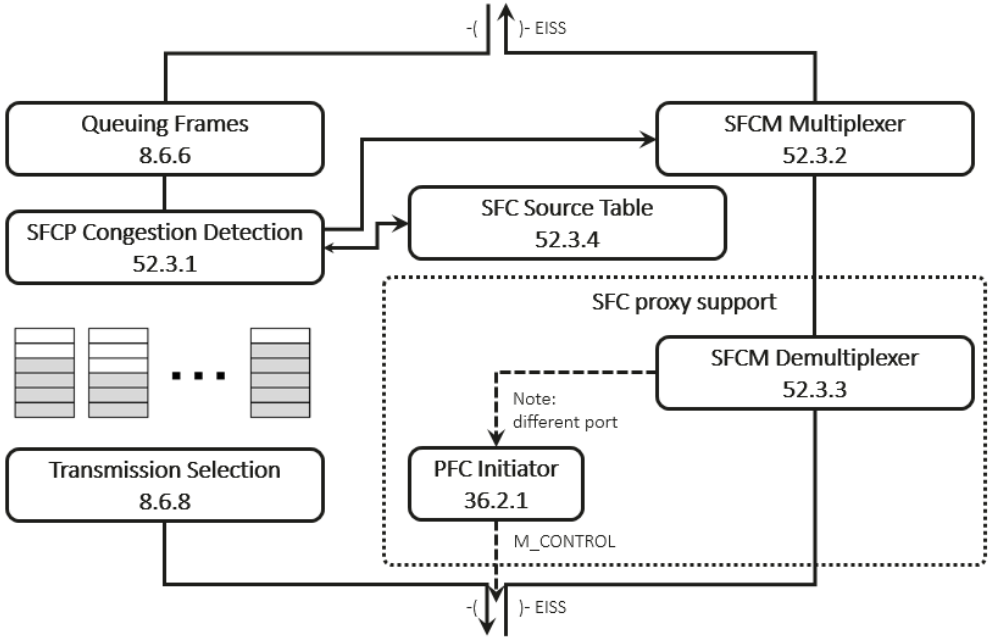


Figure 52-2—Bridge component SFC reference diagram

HUAWEI

SFCP Procedures overview (SFCM receiver side)

The SFCM Demultiplexer identifies SFCMs and invoke **processSfcmPdu()**

SFCM reaches 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 (proxy mode bridge).

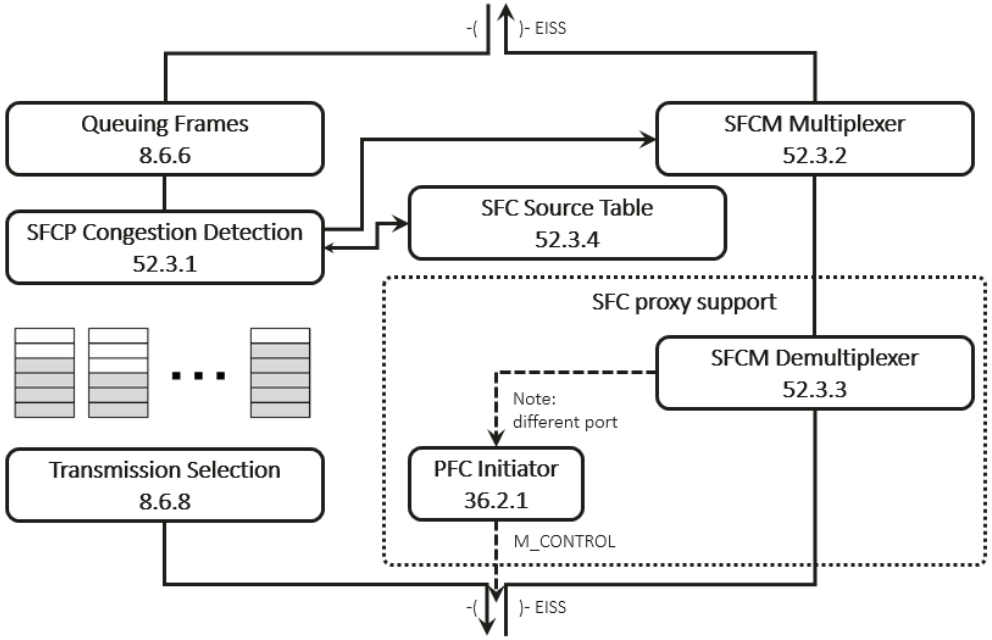


Figure 52-2—Bridge component SFC reference diagram