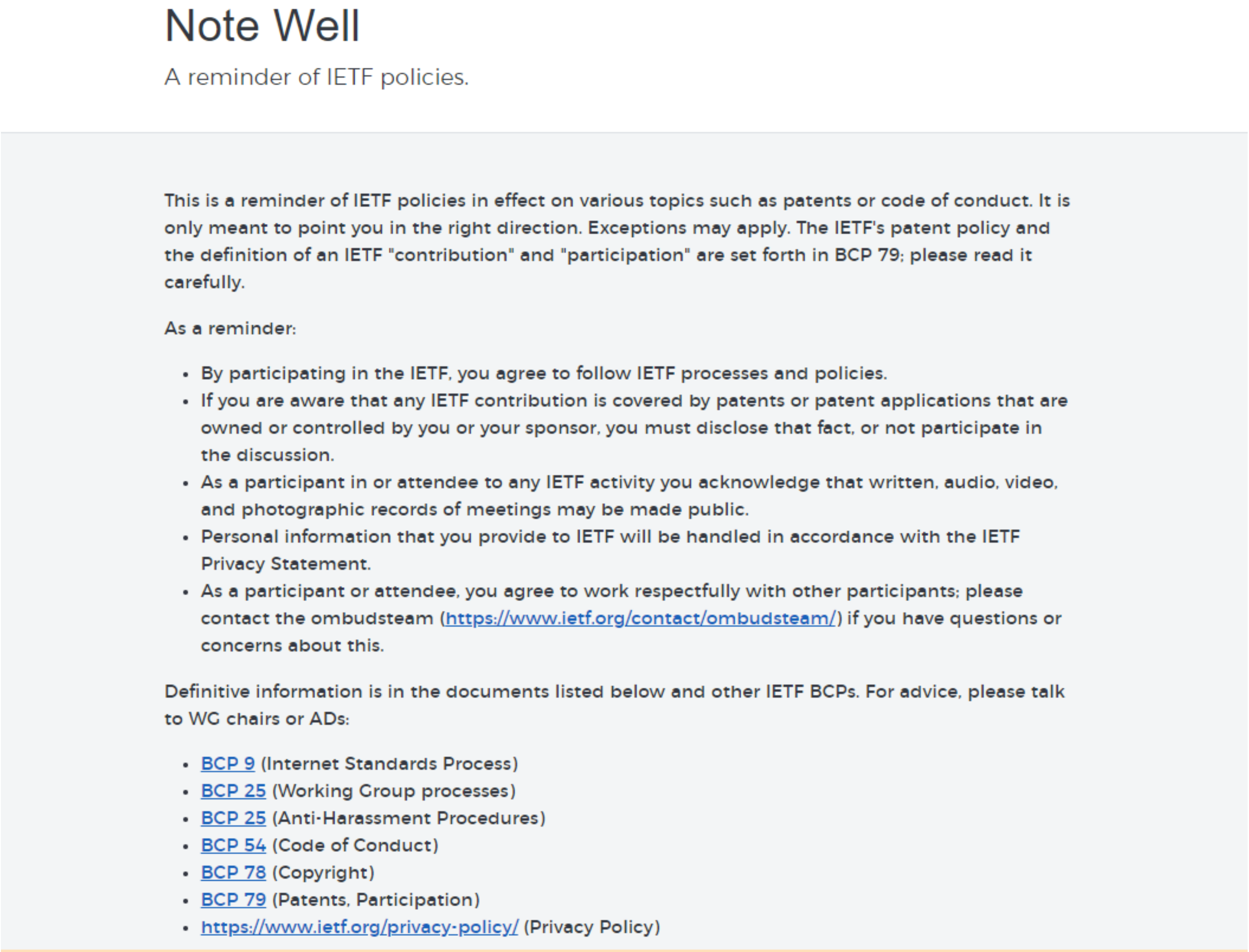
Data Center Congestion Management Initiatives

Paul Congdon (Tallac Networks)

IETF Note Well

see: https://www.ietf.org/about/note-well/



Note Well

A reminder of IETF policies.

This is a reminder of IETF policies in effect on various topics such as patents or code of conduct. It is only meant to point you in the right direction. Exceptions may apply. The IETF's patent policy and the definition of an IETF "contribution" and "participation" are set forth in BCP 79; please read it carefully.

As a reminder:

- By participating in the IETF, you agree to follow IETF processes and policies.
- If you are aware that any IETF contribution is covered by patents or patent applications that are owned or controlled by you or your sponsor, you must disclose that fact, or not participate in the discussion.
- As a participant in or attendee to any IETF activity you acknowledge that written, audio, video, and photographic records of meetings may be made public.
- Personal information that you provide to IETF will be handled in accordance with the IETF Privacy Statement.
- As a participant or attendee, you agree to work respectfully with other participants; please contact the ombudsteam (https://www.ietf.org/contact/ombudsteam/) if you have questions or concerns about this.

Definitive information is in the documents listed below and other IETF BCPs. For advice, please talk to WG chairs or ADs:

- BCP 9 (Internet Standards Process)
- BCP 25 (Working Group processes)
- BCP 25 (Anti-Harassment Procedures)
- BCP 54 (Code of Conduct)
- BCP 78 (Copyright)
- BCP 79 (Patents, Participation)
- https://www.ietf.org/privacy-policy/ (Privacy Policy)

Disclaimer

- This presentation should be considered as the personal view of the presenter not as a formal position, explanation, or interpretation of IEEE.

- Per IEEE-SA Standards Board Bylaws, December 2017
  - “At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position of IEEE.”

Join us for further discussion

- Non-WG IETF Mailing list rdma-cc-interest@ietf.org
  - Subscribe at: https://www.ietf.org/mailman/listinfo/rdma-cc-interest
- Side Meeting: Wednesday 10:00AM – 11:30 AM – Green Room 1
  - NOTE on side meetings:
    - Open to all
    - Meeting minutes will be posted to rdma-cc-interest@ietf.org
    - Not under NDA of any form
- You're invited to join a Microsoft Teams meeting

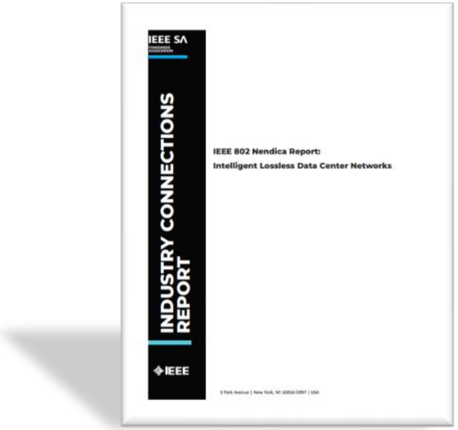
  Join on your computer or mobile app
  Click here to join the meeting

The Case for Low-latency, Low-Loss, High-Performance, Large-Scale DCNs

- More and more latency-sensitive applications are being deployed in data centers
  - Distributed Storage
  - AI / Deep Learning
  - Cloud HPC
  - High-Frequency Trading
- RDMA is operating at larger scales thanks to RoCEv2
  - Chuanxiong Guo, et. al., Microsoft, ”RDMA over Commodity Ethernet at Scale”, SIGCOMM 2016
  - Y Zhu, H Eran, et. al., Microsoft, Mellanox, “Congestion control for large-scale RDMA deployments”, SIGCOMM 2015
  - Radhika Mittal, et. al., UC Berkeley, Google, “TIMELY: RTT-based Congestion Control for the Datacenter”, SIGCOMM 2015
- The scale of Data Center Networks continues to grow
  - Larger, faster clusters are better than more smaller size clusters
  - Server growth continues at 25% - 30% putting pressure on cluster sizes and networking costs

Nendica Reports

- IEEE 802 “Network Enhancements for the Next Decade” Industry Connections Activity



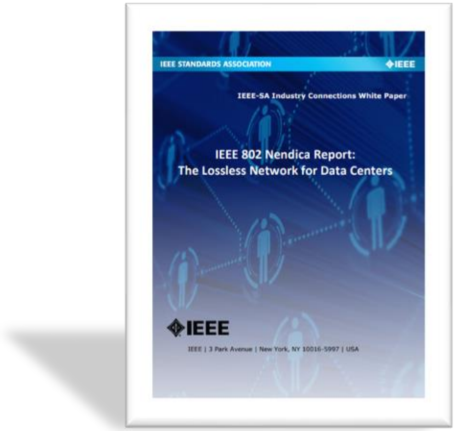
IEEE SA
STANDARDS ASSOCIATION

INDUSTRY CONNECTIONS REPORT

IEEE 802 Nendica Report:
Intelligent Lossless Data Center Networks

IEEE

3 Park Avenue | New York, NY 10016-5997 | USA



IEEE STANDARDS ASSOCIATION IEEE

IEEE-SA Industry Connections White Paper

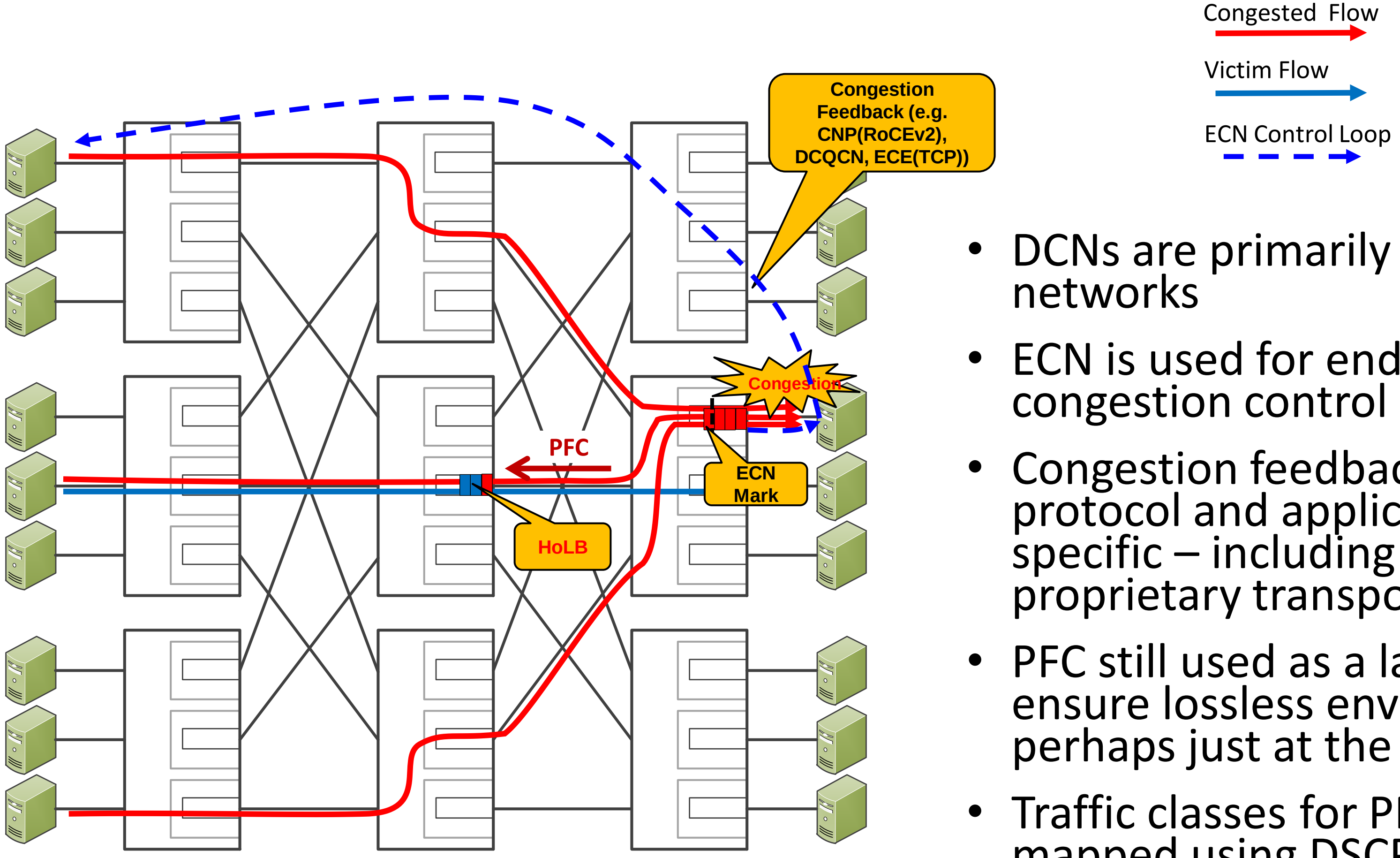
IEEE 802 Nendica Report:
The Lossless Network for Data Centers

IEEE

IEEE | 3 Park Avenue | New York, NY 10016-5997 | USA

- Two Published Reports on Data Center Networks:
  - 2021-06-22: IEEE 802 Nendica Report: Intelligent Lossless Data Center Networks (ISBN: 978-1-5044-7741-3)
  - 2018-08-17: IEEE 802 Nendica Report: The Lossless Network for Data Centers (ISBN: 978-1-5044-5102-4)

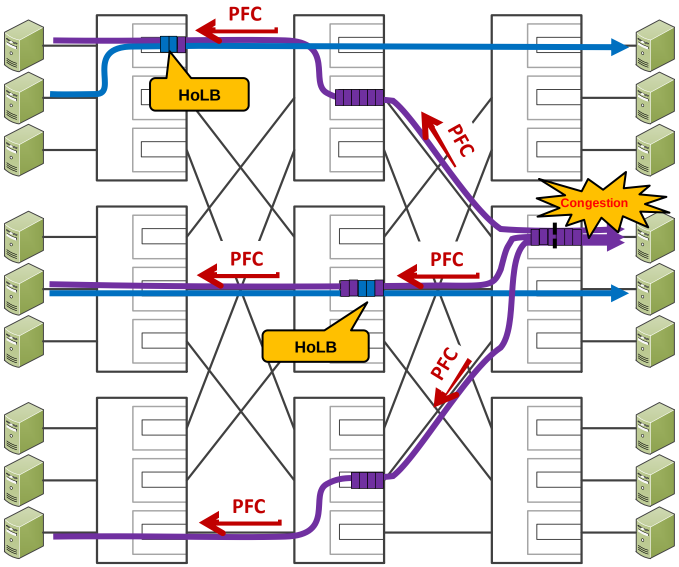
Has the DCN state-of-the-art changed?



- DCNs are primarily L3 CLOS networks
- ECN is used for end-to-end congestion control
- Congestion feedback can be protocol and application specific – including new proprietary transports
- PFC still used as a last resort to ensure lossless environments – perhaps just at the edge.
- Traffic classes for PFC are mapped using DSCP as opposed to VLAN tags – It’s L3!

Existing 802.1 Congestion Management Tools

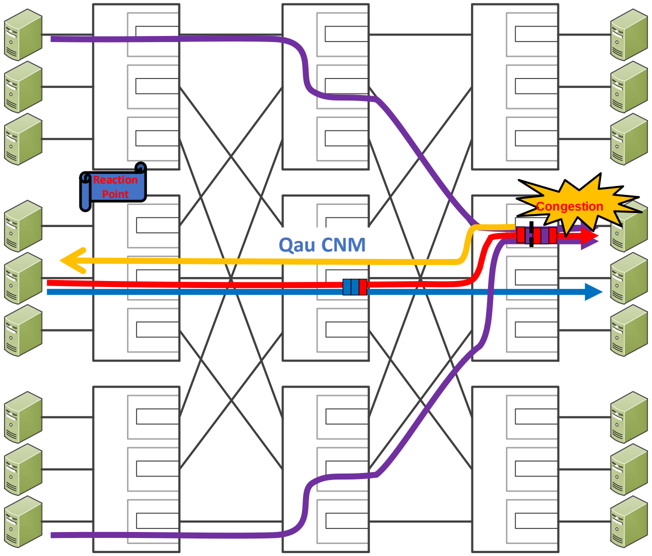
802.1Qbb - Priority-based Flow Control



Concerns with over-use

- Head-of-Line blocking
- Congestion spreading
- Buffer Bloat, increasing latency
- Increased jitter reducing throughput
- Deadlocks with some implementations

802.1Qau - Congestion Notification

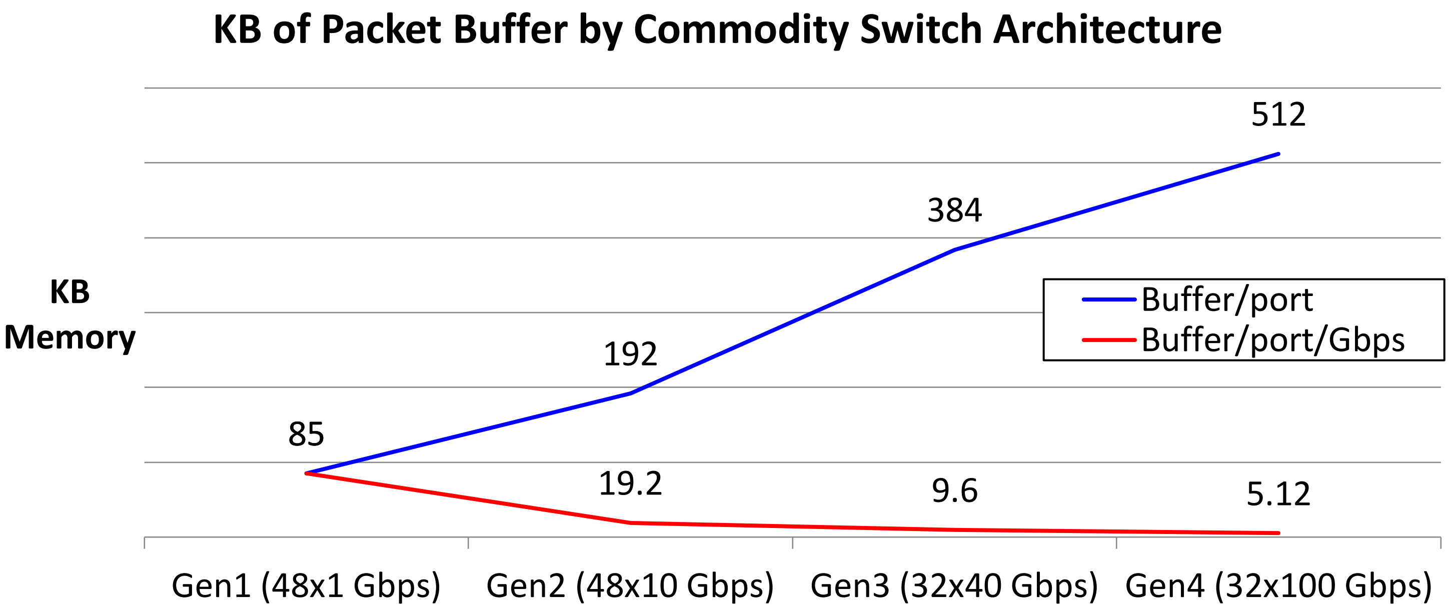


Concerns with deployment

- Layer-2 end-to-end congestion control
- NIC based rate-limiters (Reaction Points)
- Designed for non-IP based protocols
  - FCoE
  - RoCE – v1

Challenges going forward

- Scaling the high-performance data center
  - More hops => more congestion points
  - Faster links => more data in flight
- Switch buffer growth is not keeping up



Source: "Congestion Control for High-speed Extremely Shallow-buffered Datacenter Networks". In Proceedings of APNet'17, Hong Kong, China, August 03-04, 2017, https://doi.org/10.1145/3106989.3107003

Three Initiatives of Interest

Motivated to enable low-latency, low-loss, high-reliability Ethernet-based Data Center Networks supporting RDMA and AI/HPC workloads.

1. P802.1Qcz – Congestion Isolation
2. P802.1Qdt – PFC Enhancements
3. Source Flow Control

These are all ‘amendments’ to IEEE Std 802.1Q

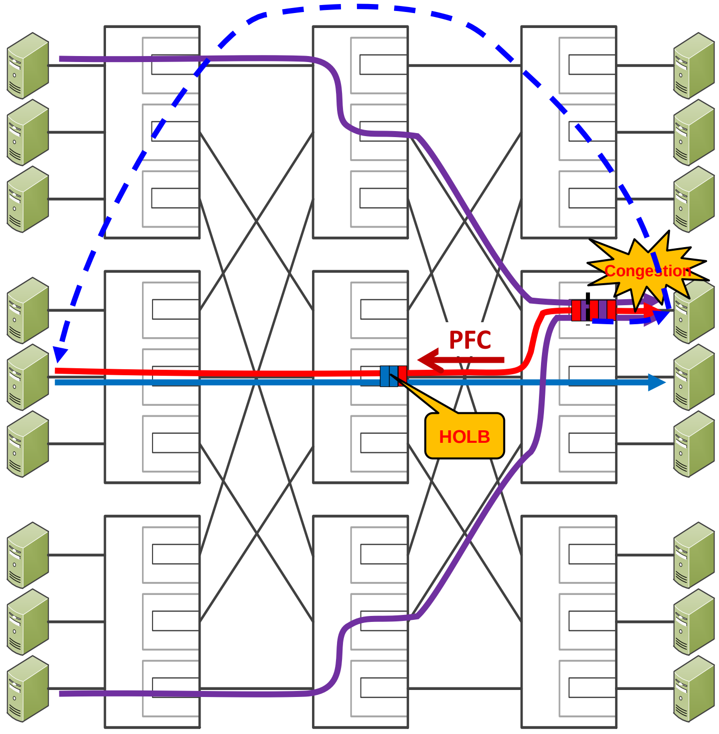
P802.1Qcz – Congestion Isolation

- Project Initiation
  - November 2017 – IEEE 802.1 agreed to develop a Project Authorization Request (PAR) and Criteria for Standards Development (CSD) to amend IEEE 802.1Q with “Congestion Isolation”
  - Amendment to IEEE 802.1Q-2018 to Support the isolation of congested data flows within ***data center environments***, such as high-performance computing, distributed storage and central offices re-architected as data centers.
  - Motivation discussed in draft report of “802 Network Enhancements For the Next Decade”
    - https://mentor.ieee.org/802.1/dcn/18/1-18-0007-03-ICne-draft-report-lossless-data-center-networks.pdf
- Project Status
  - Sep 2018 – Project approved
  - Aug 2019 – Initial draft for ballot
  - Jan 2021 – Standards Association ballot
  - Jan 2022 – Completed SA ballot recirc, waiting on Q-Rev completion

- So, what is Congestion Isolation?

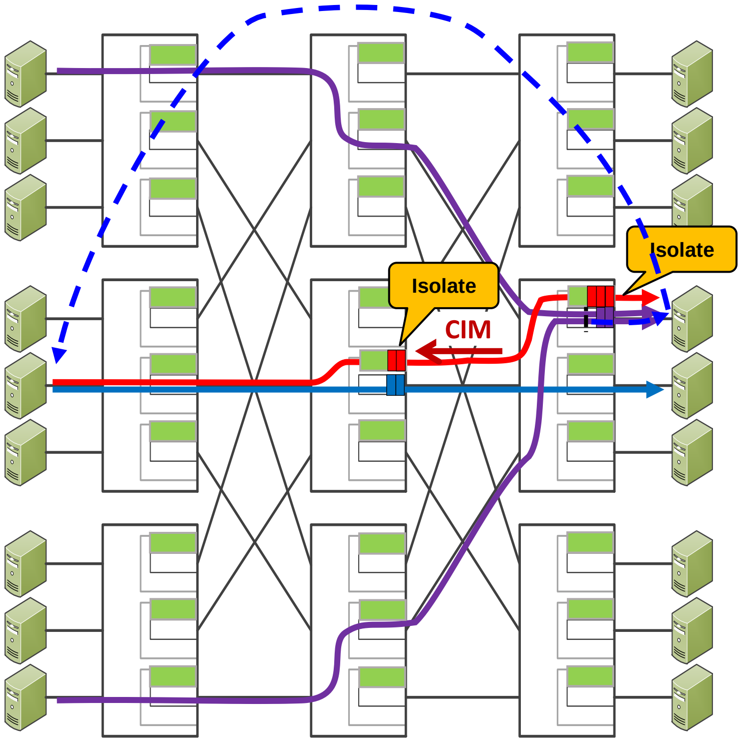
P802.1Qcz - Congestion Isolation

Today – Without Congestion Isolation



1. End-to-end congestion control using ECN marking
2. Priority-based Flow Control (PFC) as last-ditch effort to avoid drops

Congestion Isolation



1. Move congesting flows to a separate queue and signal your upstream neighbor
2. Upstream neighbor moves congesting flows to separate queue

Congestion Isolation - Goals

- Work in conjunction with higher-layer end-to-end congestion control (ECN, BBR, etc)
- Support larger, faster data centers (Low-Latency, High-Throughput)
- Support lossless and low-loss environments
- Improve performance of both TCP and UDP based flows
- Reduce pressure on switch buffer growth
- Reduce the frequency of relying on PFC for a lossless environment
- Significantly reduce HOLB caused by over-use of PFC

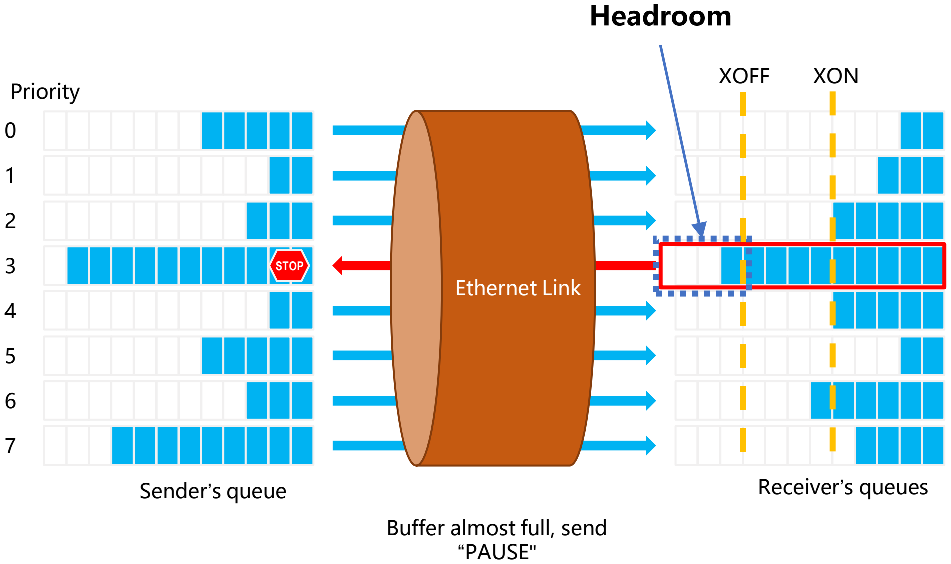
P802.1Qdt – PFC Enhancements

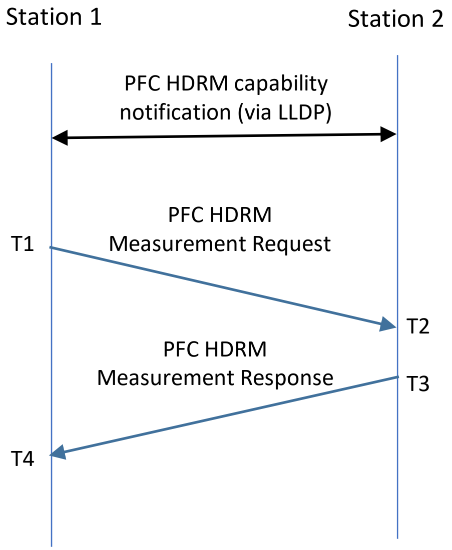
- Project Initiation
  - Multiple contributions:
    - https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-0121-v02.pdf - Adaptive PFC Headroom
    - https://www.ieee802.org/1/files/public/docs2021/new-congdon-a-pfc-h-Q-changes-0521-v01.pdf - Consideration of Adaptive PFC Headroom in 802.1Q
    - https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-and-PTP-0602-v03.pdf - Adaptive PFC Headroom and PTP
    - https://www.ieee802.org/1/files/public/docs2021/cz-finn-pfc-headroom-0629-v01.pdf - Determining Priority Flow Control Headroom
    - https://www.ieee802.org/1/files/public/docs2021/new-lv-PFC-Headroom-Project-Proposal-0721-v01.pdf - PFC Headroom Measurement and Calculation Project Proposal
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0048-00-ICne-pfc-headroom-with-macsec.pdf - Incorporating MACSec into PFC Headroom Calculation
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0050-00-ICne-pfc-enhancements-project-proposal.pdf - PFC Enhancements Project Proposal
    - https://mentor.ieee.org/802.1/dcn/21/1-21-0052-00-ICne-pfc-enhancements-next-steps.pdf - PFC Enhancements Project Proposal
  - The need to protect PFC frames with MACSec highlighted by Microsoft Azure
  - March 2020 – IEEE 802.1 agreed to develop a Project Authorization Request (PAR) and Criteria for Standards Development (CSD) to amend IEEE 802.1Q with “PFC Enhancements”
- Project Status
  - January 2021 – Initial proposal
  - March 2020 – Approved to forward PAR to IEEE NesCom
  - April 2022 – Anticipated initial draft ‘individual contribution’

- So what are the PFC Enhancements?

P802.1Qdt – PFC Enhancements

Objective: Automatically calculate minimum PFC buffer requirements (i.e. headroom) for lossless operation, without user intervention. Additionally – protect PFC frames using MACsec encryption



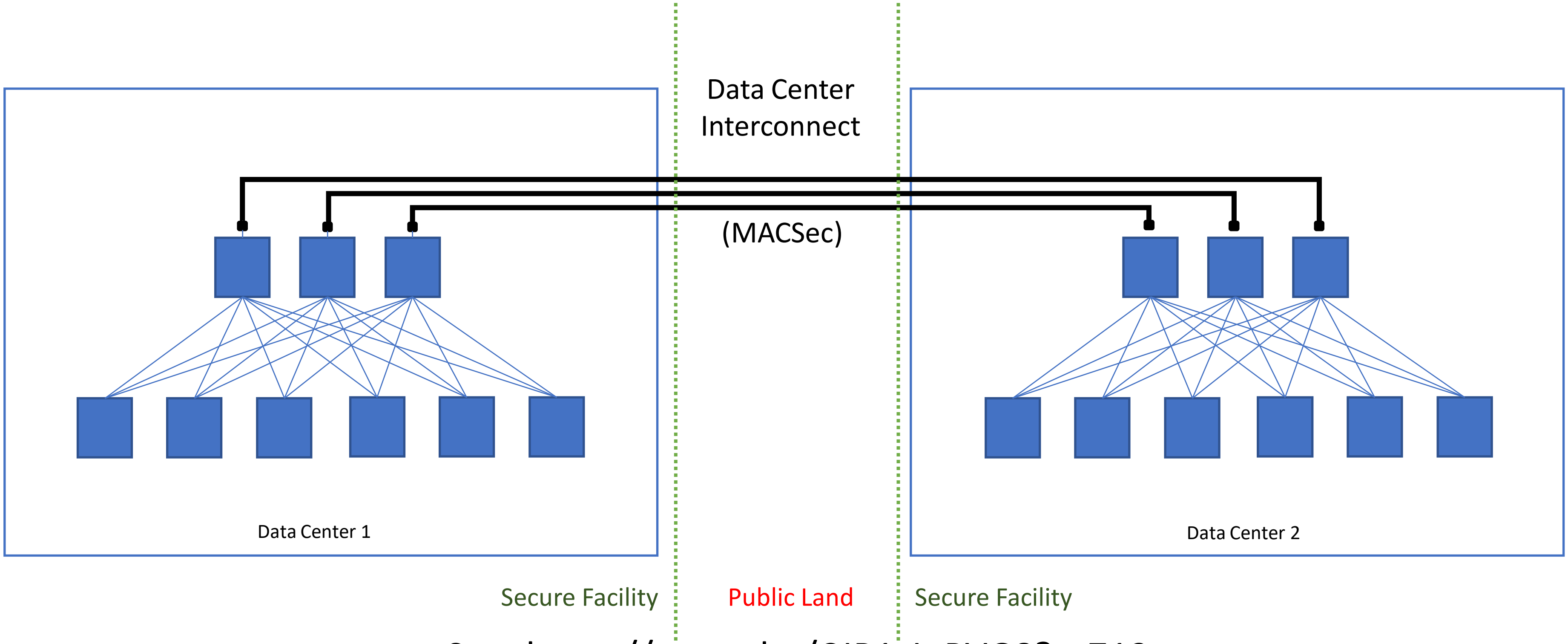

Station 1
Station 2
PFC HDRM capability notification (via LLDP)
T1
PFC HDRM Measurement Request
T2
PFC HDRM Measurement Response
T3
T4

Headroom needed = (Port speed * (T4-T1-(T3-T2)) + 2*(Max Frame) + (PFC Frame)) * Alpha

NOTE: Alpha is implementation dependent, based on internal buffer chunk size

1. Re-use the Precision Time Protocol (PTP) to measure cable delay
2. Exchange internal delay values using LLDP via DCBX

A Use Case To Consider with MACSec



Data Center Interconnect

(MACSec)

Data Center 1

Data Center 2

Secure Facility

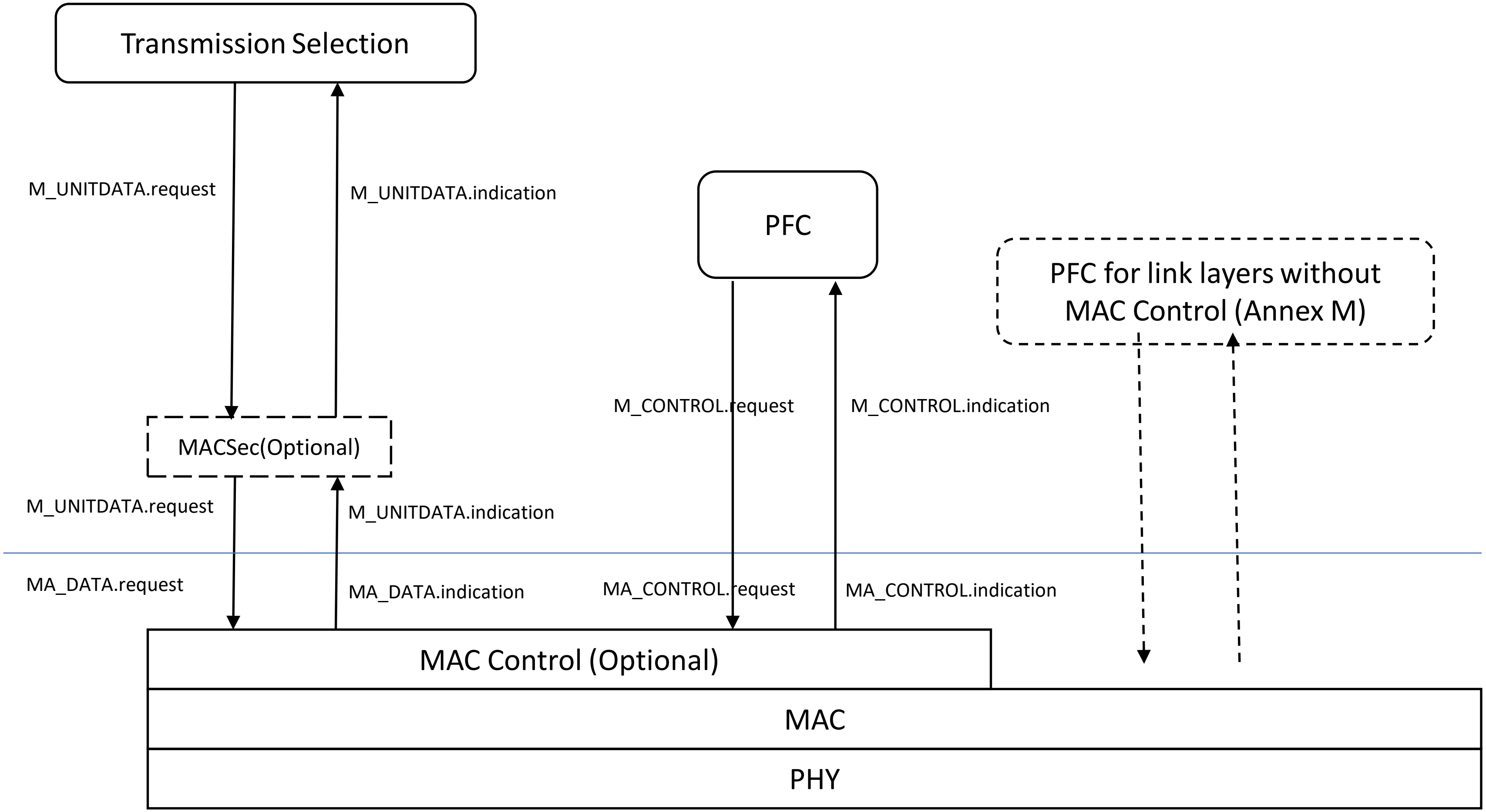
Public Land

Secure Facility

See: https://youtu.be/CJP1rJnPVG8?t=712

NOTE: The RDMA protocol over Ethernet (RoCEv2) necessitates the use PFC to avoid frame loss. It is desirable to protect PFC frames when they traverse data center interconnect links

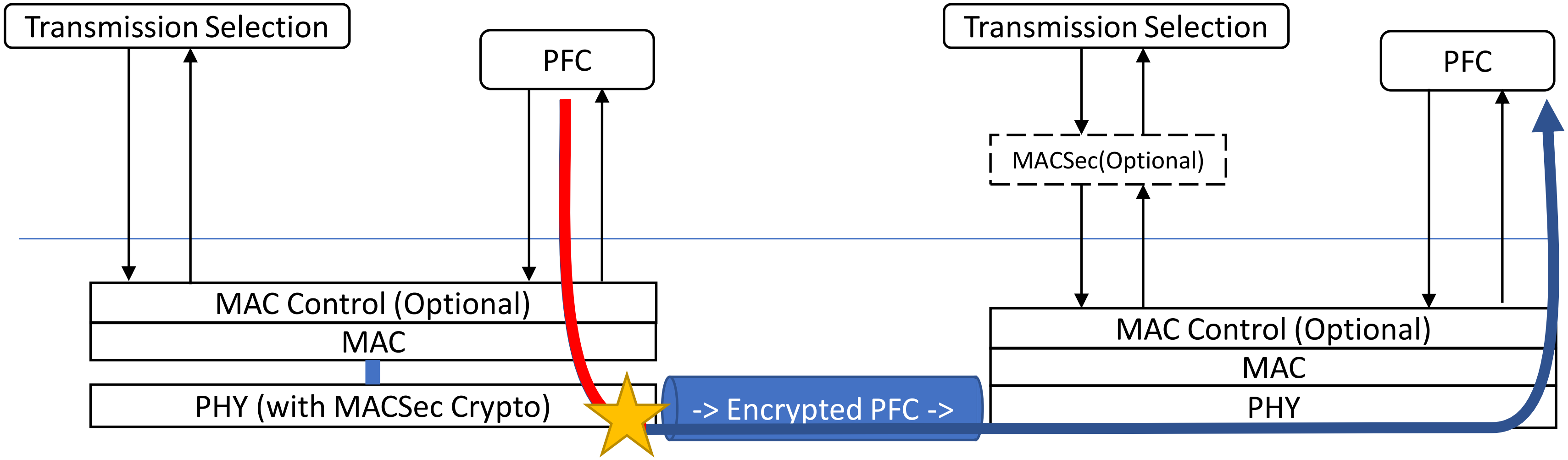
Current Protocol Layers



NOTE: Figure indicates that PFC Frames are not encrypted

Interoperability issue in the field

- Early implementations of MACSec were implemented external to the MAC (i.e. within a PHY as a ‘bump in the wire’).
  - These early implementations encrypt everything coming out of the MAC
  - These early implementations were never compliant with 802.1AE
  - These early implementations do not run MKA and may suffer outages



PFC Enhancements - Goals

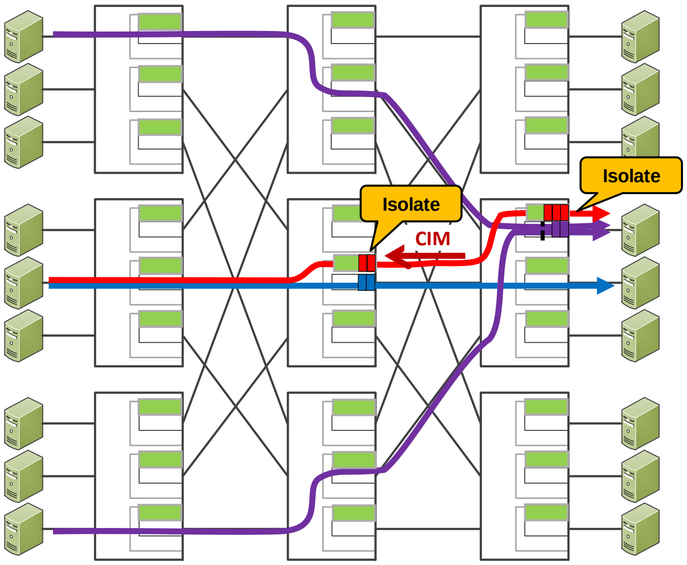
- Reduce the complexity of deploying PFC
    - Manual configuration is complex and is different for each vendor solution
    - Consistent settings across a large-scale data center network is tedious
    - Vendor provided default values waste buffer resources, and do not work in certain circumstances (e.g. long distance data center interconnection)
- Specify a wire protocols (e.g. capability exchange) and a headroom measurement mechanism.
- Address inconsistent and unclear specification of PFC and MACSec operation

Background - Source Flow Control

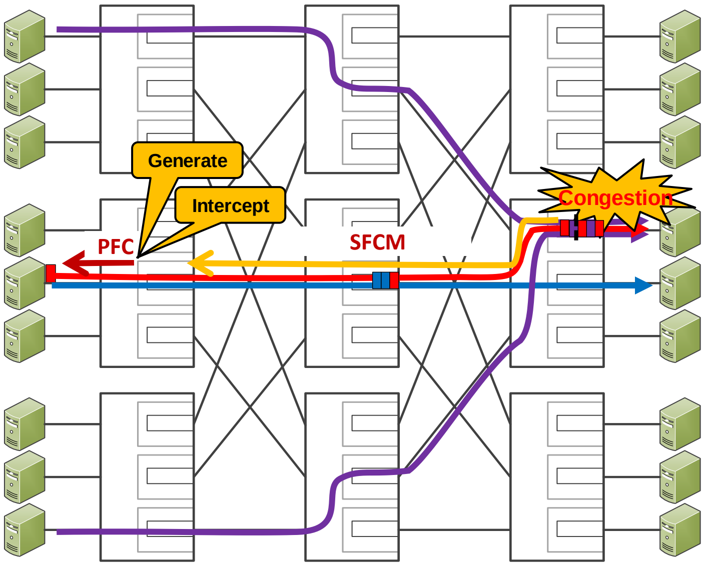
- Project Initiation
    - Multiple contributions:
        - Public presentations at P4 Workshops (Apr’20, May’21) and Open Fabrics Alliance (Mar’21)
            - https://opennetworking.org/wp-content/uploads/2020/04/JK-Lee-Slide-Deck.pdf (slide 12)
            - https://www.openfabrics.org/wp-content/uploads/2021-workshop-presentations/503_Lee_flatten.pdf
            - https://opennetworking.org/wp-content/uploads/2021/05/2021-P4-WS-JK-Lee-Slides.pdf (slide 14)
        - Previous Nendica/TSN presentations
            - https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf - 9/16/2021
            - https://mentor.ieee.org/802.1/dcn/21/1-21-0061-00-ICne-source-remote-pfc-test.pdf – 10/14/2021
            - https://mentor.ieee.org/802.1/dcn/21/1-21-0067-00-ICne-source-remote-pfc-status-update.pdf - 11/04/2021
            - https://mentor.ieee.org/802.1/dcn/21/1-21-0077-00-ICne-consideration-of-spfc-sfc-issues-when-leveraging-qcz.pdf - 12/16/2021
            - https://mentor.ieee.org/802.1/dcn/21/1-21-0079-00-ICne-spfc-sfc-next-steps.pdf - 12/23/2021
            - https://www.ieee802.org/1/files/public/docs2022/new-congdon-SFC-overview-0122-v01.pdf - 01/19/2022
            - https://mentor.ieee.org/802.1/dcn/22/1-22-0001-01-ICne-sfc-q-changes.pdf - 01/27/2022
            - https://www.ieee802.org/1/files/public/docs2022/new-bottorff-sfc-0322-v6.pdf - 02/24/2022
        - IETF Awareness
            - Topic raised at IEEE 802 / IETF Coordination call – 10/25/2021
            - https://datatracker.ietf.org/meeting/112/materials/slides-112-iccrg-source-priority-flow-control-in-data-centers-00 - 11/08/2021
            - Upcoming at the IETF-113 – HotRFC session – 03/20/2022, Scheduled side-meeting discussion – 03/23/2022
    - March 2022 – IEEE 802.1 agreed to develop a Project Authorization Request (PAR) and Criteria for Standards Development (CSD) to amend IEEE 802.1Q with “Source Flow Control”
- Project Status
    - April 2020 – Initial public discussion
    - September 2021 – Introduced to IEEE 802
    - March 2022 – Approval to develop PAR and CSD
    - May 2022 – Prepare PAR & CSD for pre-circulation in July
- So what is Source Flow Control?

Source Flow Control

P802.1Qcz - Congestion Isolation



Source Flow Control (w/ ToR Proxy)



Implementation details

- Congesting flows are isolated locally first
- As queues continue to congest, CIM is generated and sent to upstream bridge/router
- CIM can be L2 or L3 message to support L3 networks (common deployment model).

Details

- Can be combined with Congestion Isolation
- Edge-to-Source signaling using L3 message
- Like an L3 version of 802.1Qau (L3-QCN), but no Reaction Point (RP) rate controller defined – this is Flow Control
- Optional source Top-of-Rack switch involvement

Status: New project proposal

Source Flow Control - Goals

- Work in conjunction with other congestion control, such as DCQCN, DCTCP, Congestion Isolation
- Reduce latency in large scale data centers when congestion control is less effective.
  - In heavy in-cast congestion (large number of flows), ECN/CNP adjustment does not help in controlling queue length or reducing flow rate.
  - In transient congestion, end to end congestion control does not provide fast enough control loop.
  - Provide sub-RTT reaction time
- Provide the benefits of PFC at the source, while avoiding the negatives of PFC (congestion spreading, head-of-line blocking, PFC storms, and deadlocks)
- Provide a simpler solution than Qau (no Reaction Point (RP) just Flow-Control) and support L3 environments
- Enable early deployment without Server upgrades via Source ToR Proxy
- Carry flow information for more intelligent decisions at the source.

SFC Design Team

Team

Paul Bottorff (HPE), Paul Congdon (Huawei), Claudio Desanti (Dell) , Uri Ezur (Intel), JK Lee (Intel), Lily Lv (Huawei)

Current List of Topics

1. UDP port number for SFCM
2. How to secure SFCM
3. Contents of SFCM
4. Identifying the source priority/traffic-class to pause
5. Operation in overlay networks (VxLAN, Geneve)
6. Calculation of pause interval
7. SFCM suppression
8. Multicast considerations
9. Source ToR intercept of SFCM packets
10. Consideration of DCBX enhancements

Design Team and Participation

- P802.1Qcz is at the finish line!
- P802.1Qdt is relatively straight forward and in the early stages of drafting a specification.
- SFC is just beginning with approval to craft a new project. A standards related technical design team exists with multiple vendors involved.
- Other technologies, from PHY to Transport, are of interest for consideration in traditional standards organizations or elsewhere.

There is a strong desire to see Ethernet as the leader in a high performance, low-latency, low-loss, high reliability fabric/interconnect for HPC/AI and modern workloads

A New Ethernet for the Data Center