IEEE 802.1 Data Center Network Standard

Enabling Data Center Fabric Convergence onto Ethernet

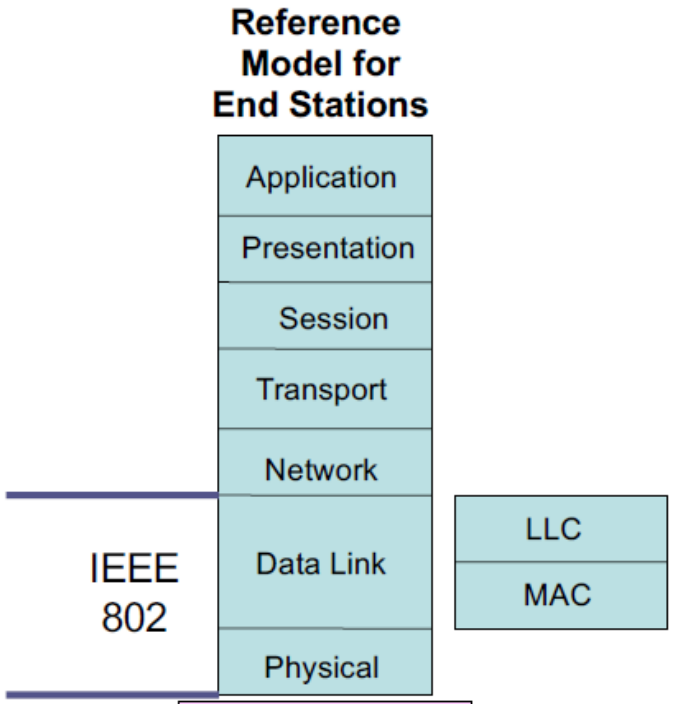
September 2023

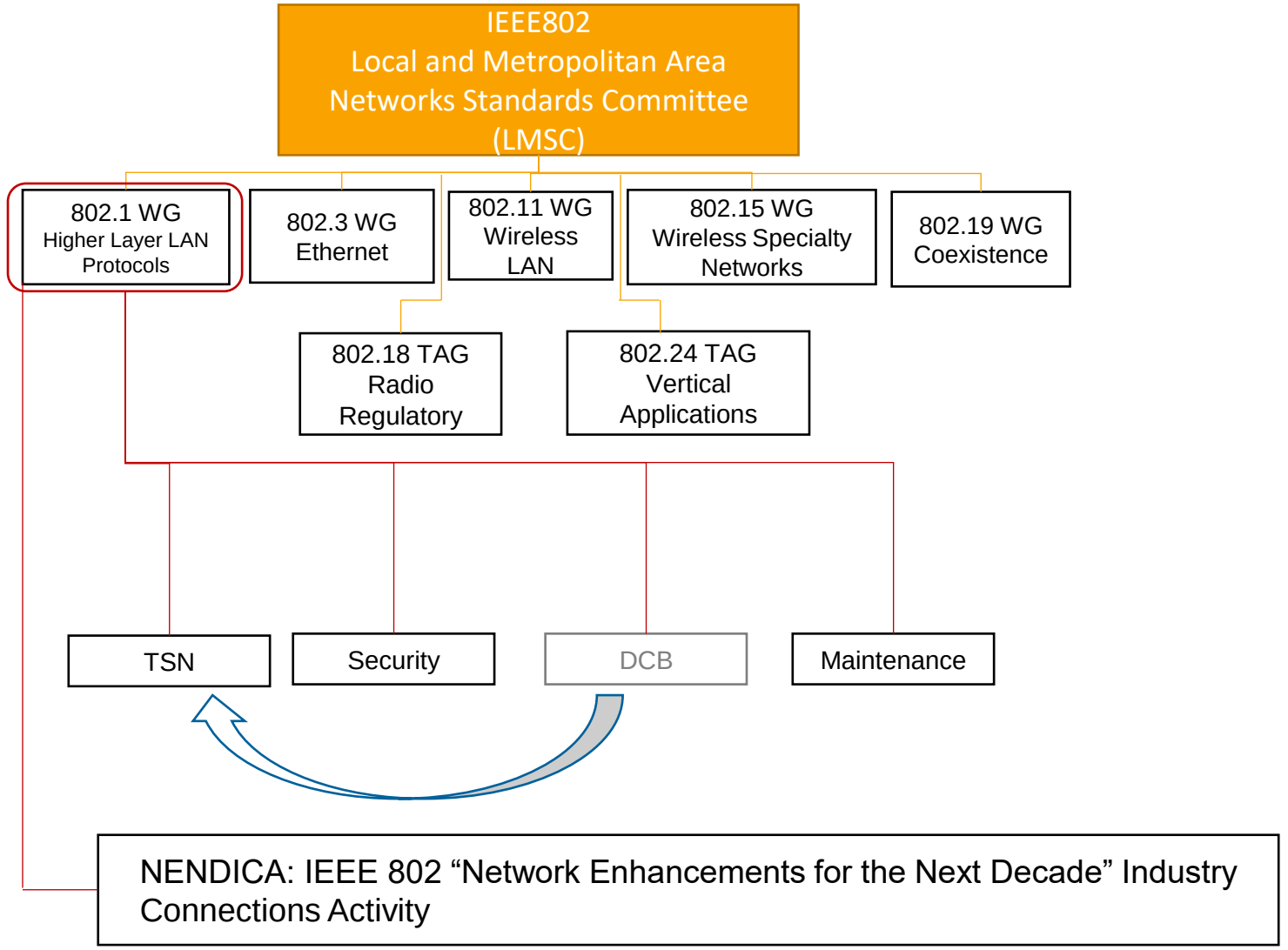
Presenter: Lily Yunping Lv, IEEE 802.1 Working Group Editor

IEEE COMPUTER SOCIETY

IEEE

IEEE 802.1 is Higher Layer LAN Protocols Working Group

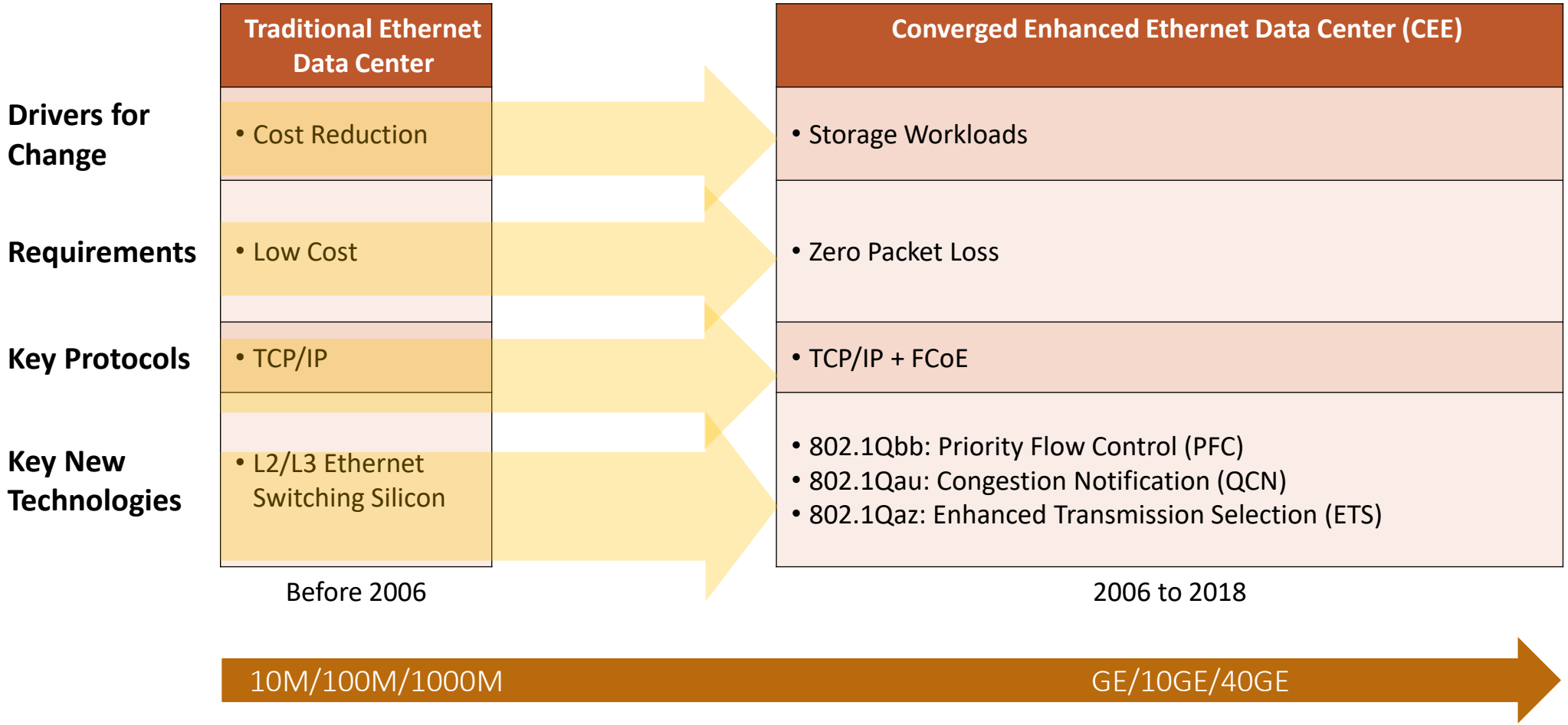




IEEE COMPUTER SOCIETY

IEEE

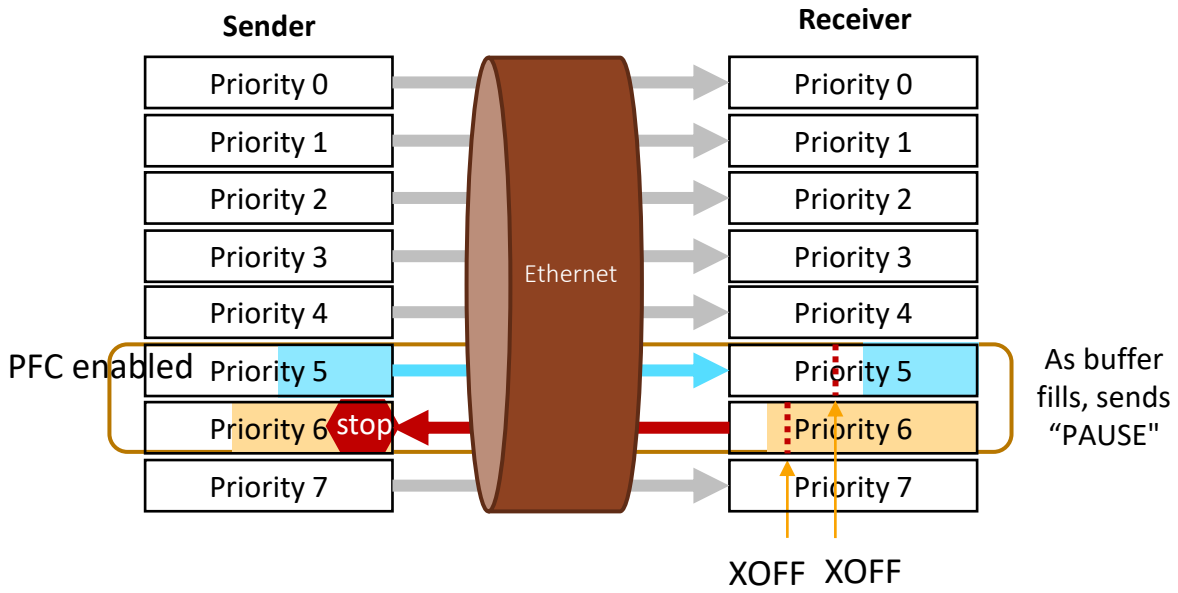
IEEE 802.1 Developed Standards for DC Fabric Since 2006



PFC

PFC

- IEEE 802.1Q defines 8 priorities
- Priority-based Flow Control ‘pauses’ individual priorities, while other priorities continue
  - “PFC allows link flow control to be performed on a per-priority basis. In particular, PFC is used to inhibit transmission of data frames associated with one or more priorities for a specified period of time. PFC can be enabled for some priorities on the link and disabled for others.” (Std 802.1Q-2018)



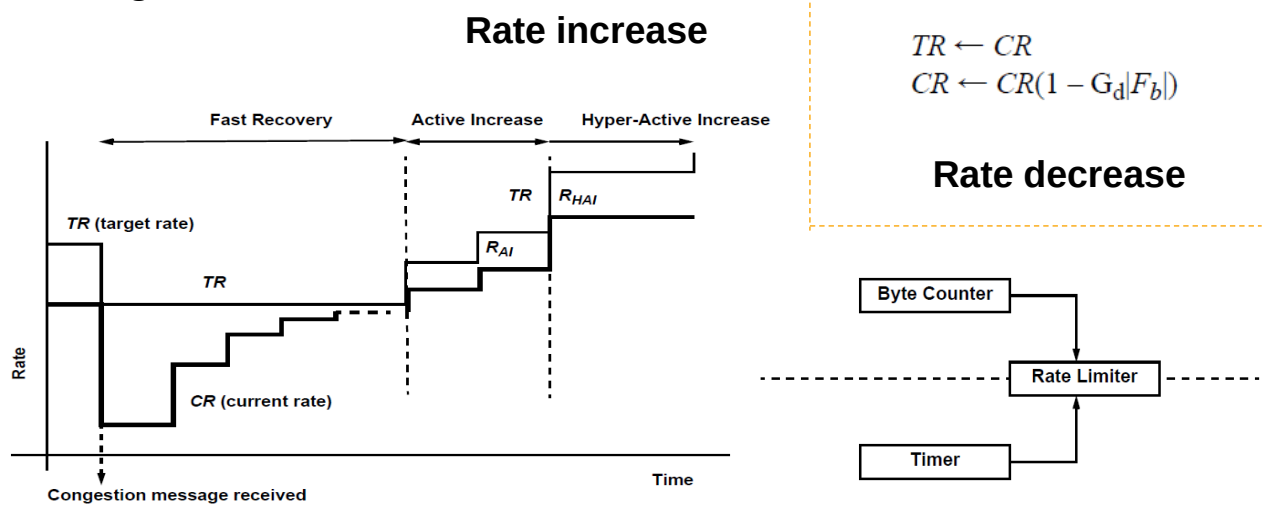
Example:

- XOFF threshold which invokes PFC is set on each PFC enabled priority
- Priority 6 reaches threshold XOFF
- PFC pause frame is triggered and sent upstream
- Upstream priority 6 transmission is stopped

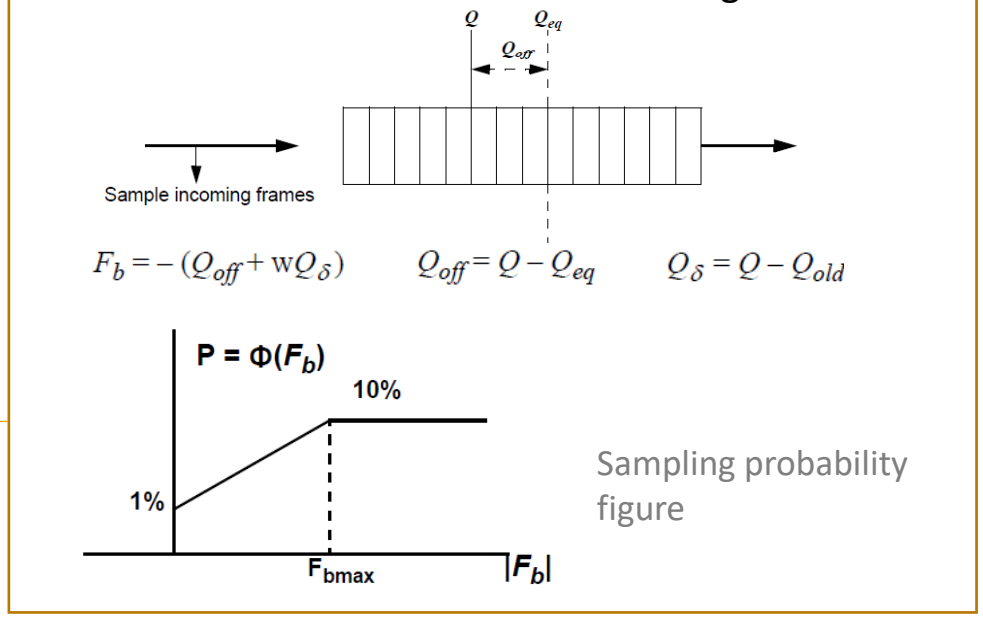
QCN

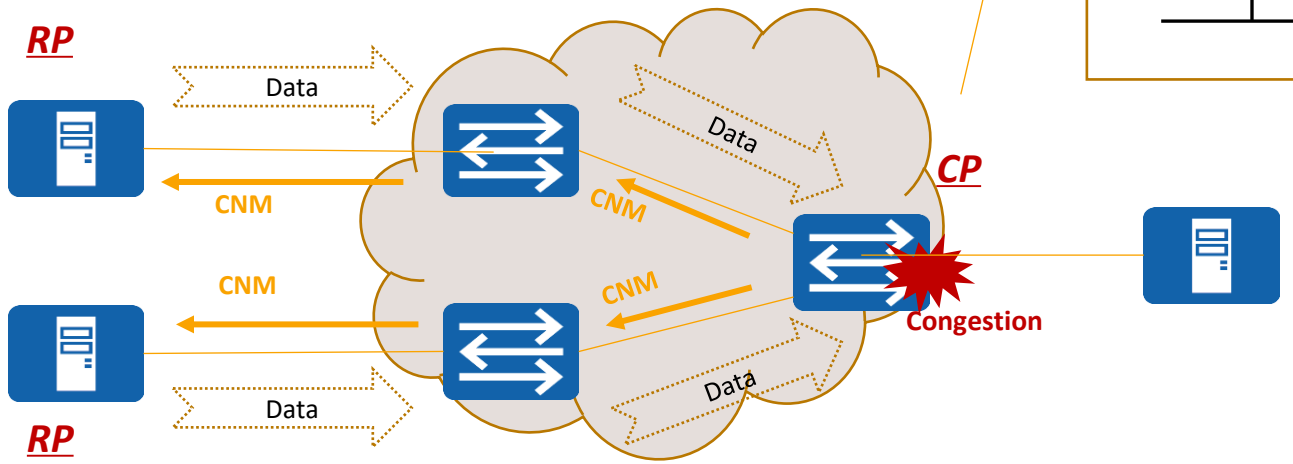
QCN When congestion point (CP) detects congestion according to CP algorithm, it generates congestion notification message (CNM) to reaction point (RP), causing reaction point to adjust flow rate.

RP Algorithm



CP Algorithm

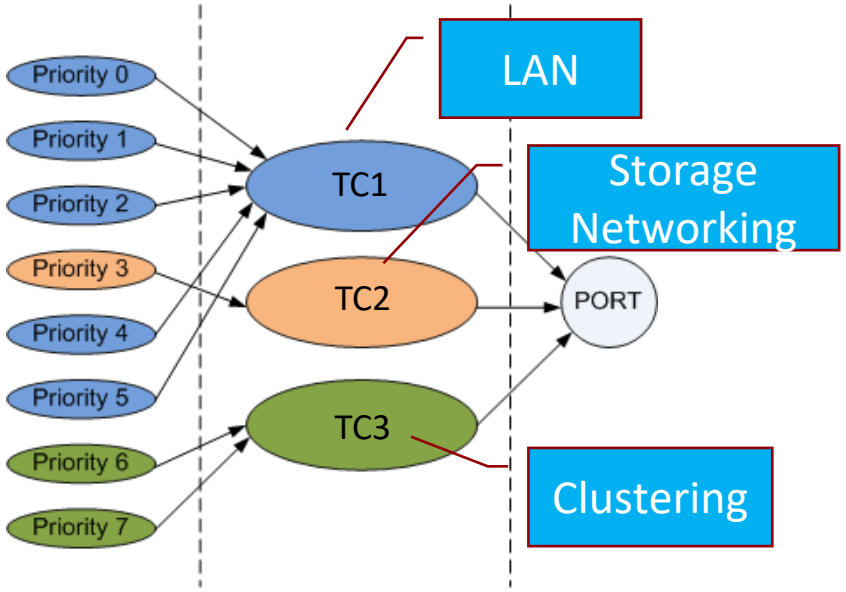


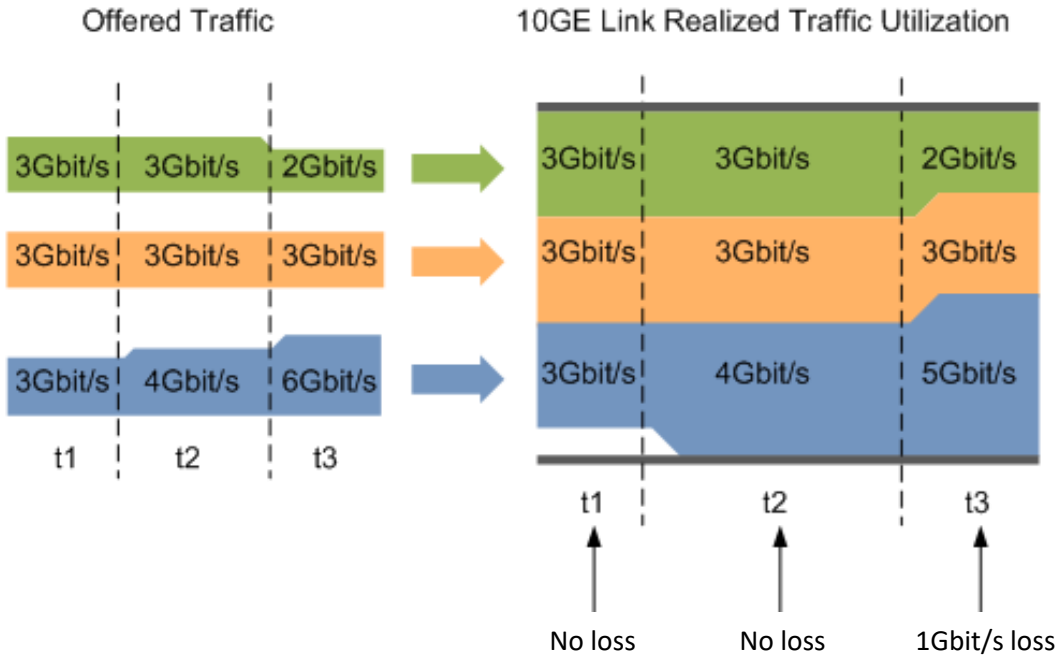


ETS

ETS

ETS allows a uniform management of dynamic bandwidth allocation among traffic classes on a single network. When the offered load in a traffic class doesn't use its allocated bandwidth, other traffic classes are allowed to use the available bandwidth.





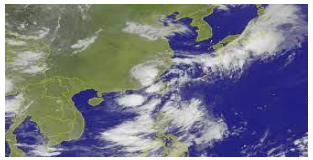
It’s Time For CEE Evolvement (1/3)

Driven by new emerging applications, IDC predicts the Global Data will grow from 45 Zettabytes in 2019 to 175 Zettabytes by 2025

Manufacturing

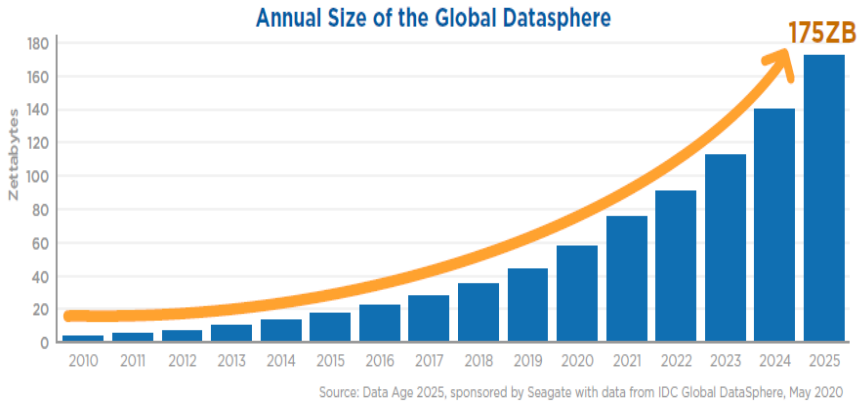
Autopilot

Weather forecast



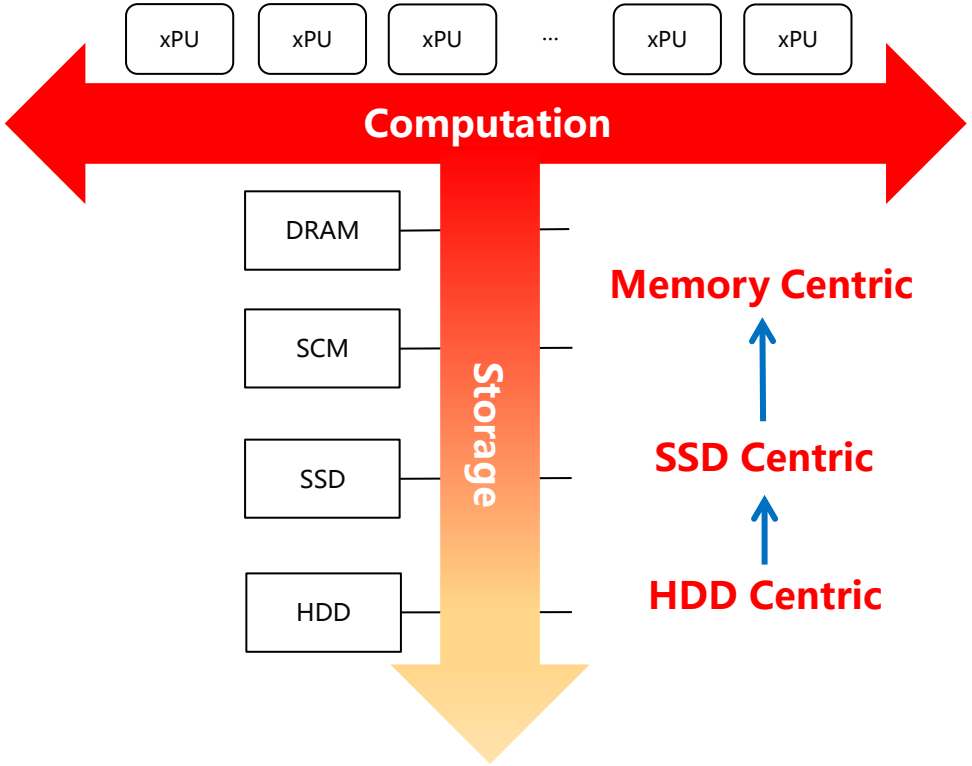
Biomedicine

.....



1B → 100B → 1000B (AI Model Size)

100PFlop/s → EFlop/s → 10EFlop/s (HPC Perf.)



IEEE COMPUTER SOCIETY

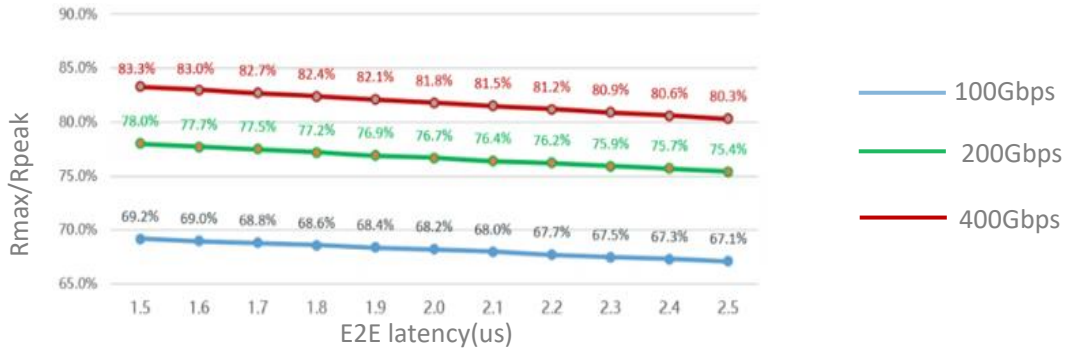
IEEE

It’s Time For CEE Evolvement (2/3)

AI: Bandwidth hungry; Burst of low number, large flows requires network load balancing

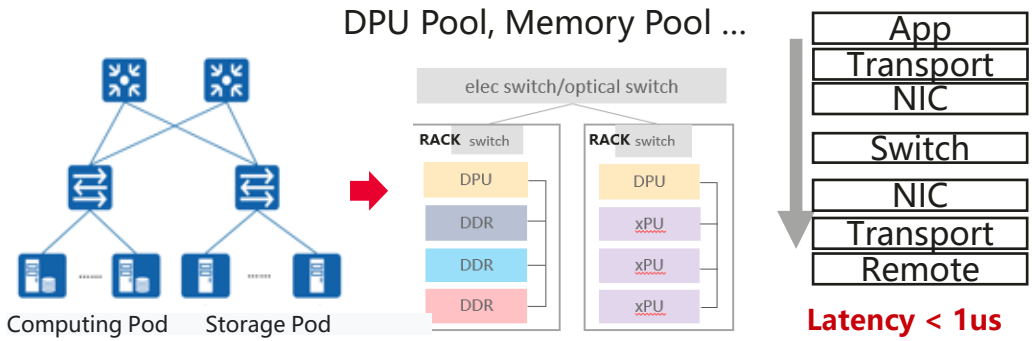
Parallel Mode	Characteristic	BW Requirement
TP	• Intra node communication (AllReduce) • 100s GB level traffic	★ ★ ★ ★ ★
PP	• Inter nodes communication (send/recv) • 100s MB ~ GB level traffic	★ ★ ★
DP	• Inter nodes communication (AllReduce) • GB level traffic	★ ★ ★ ★
MOE	• Inter nodes communication (AlltoAll/AllReduce) • GB level traffic	★ ★ ★ ★

HPC: Bandwidth and latency are key to HPL performance. Reducing 1us increases 2%~3% performance



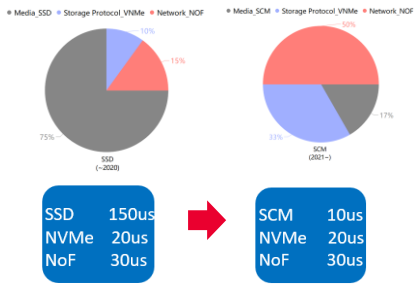
HPL: High Performance Linpack, a classic HPC benchmark used in TOP500
Rmax: Maximal LINPACK performance

Cloud: resource utilization ratio becomes crucial, resource pooling requires sub-us latency and determinism.



Latency < 1us

Storage: scale: 10K → 100K; IOPS: 1M → 10M; latency: ~150us → ~10us



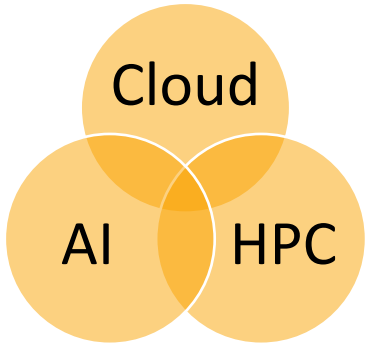
Proportion of network Latency: **15%**

Proportion of network Latency: **50%**

Network Scale: **10K**
IOPS: **1.5M**

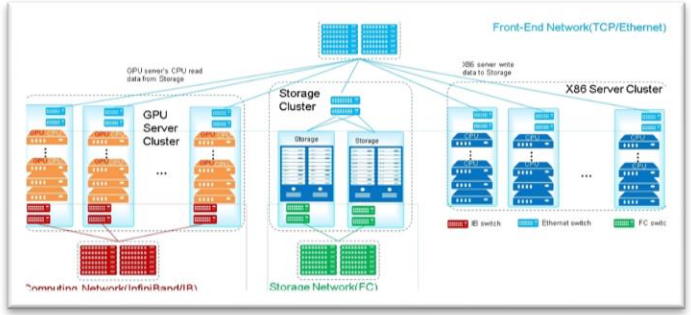
Network Scale: **100K**
IOPS: **6M**

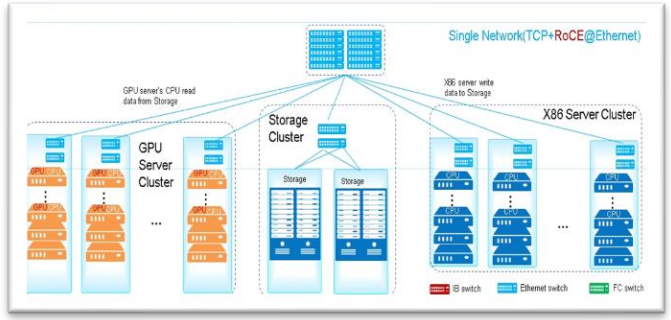
It’s Time For CEE Evolvement (3/3)



AI, HPC and cloud computing are converging

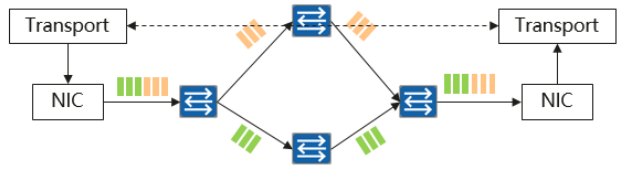
It is time to consider CEE evolvement supporting computing systems.

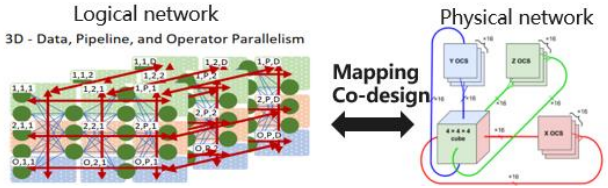




Large Scale **High Throughput** **Low Latency**

- ✓ Fully utilize underlying network resource to achieve High Throughput & Low Latency.
- ✓ Minimize network resources for computing tasks while minimizing completion time.





IEEE 802.1 DCN Standards Continue to Evolve

	Traditional Ethernet Data Center	Converged Enhanced Ethernet Data Center (CEE)	Hyper Converged Ethernet Data Center （HCE）
Drivers for Change	• Cost Reduction	• Storage Workloads	• Computing Workloads (AI & HPC & Cloud & Storage)
Requirements	• Low Cost	• Zero Packet Loss	• Large Scale • Low Latency • High Throughput
Key Protocols	• TCP/IP	• TCP/IP + FCoE	• TCP/IP + NVMe-oF + RoCE
Key New Technologies	• L2/L3 Ethernet Switching Silicon	• Priority Flow Control (PFC) • Enhanced Transmission Selection (ETS) • Quantized Congestion Notification (QCN)	• Nendica reports • 802.1Qcz: Congestion Isolation (CI) • P802.1Qdt: PFC Enhancement • P802.1Qdw: Source Flow Control (SFC) • ……
	Before 2008	2008 to 2018	After 2018

10M/100M/1000M GE/10GE/40GE 100GE/400GE/800G

Nendica Reports

IEEE 802 Nendica published two reports on data center fabrics.

- 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)

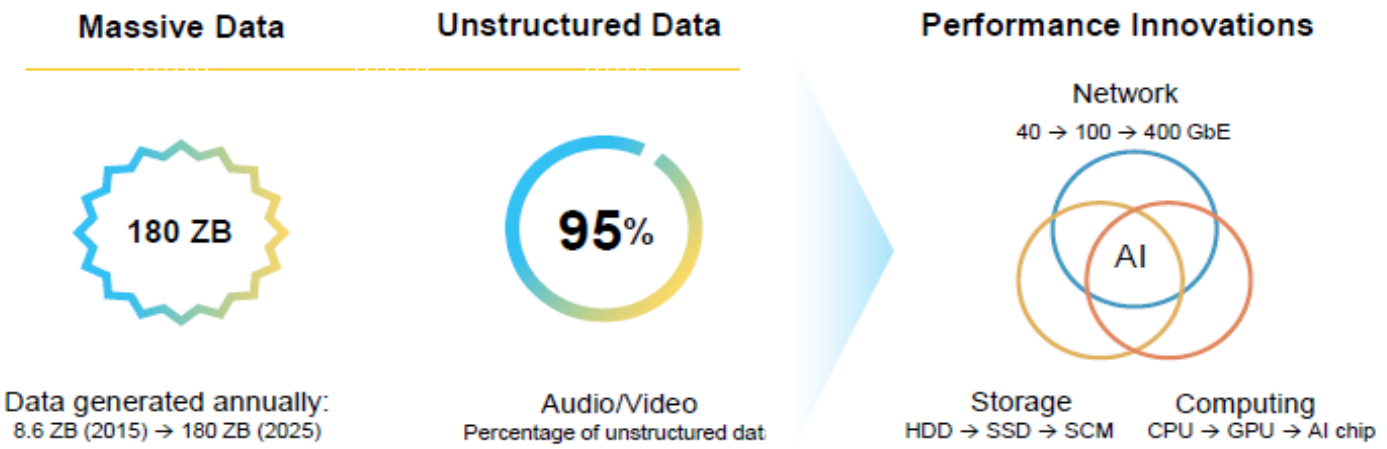
Nendica reports discuss **future opportunities** in data center Fabric and has incubated **802.1 standard** projects: 802.1Qcz, P802.1Qdt, P802.1Qdw.

Nendica Reports

Use cases

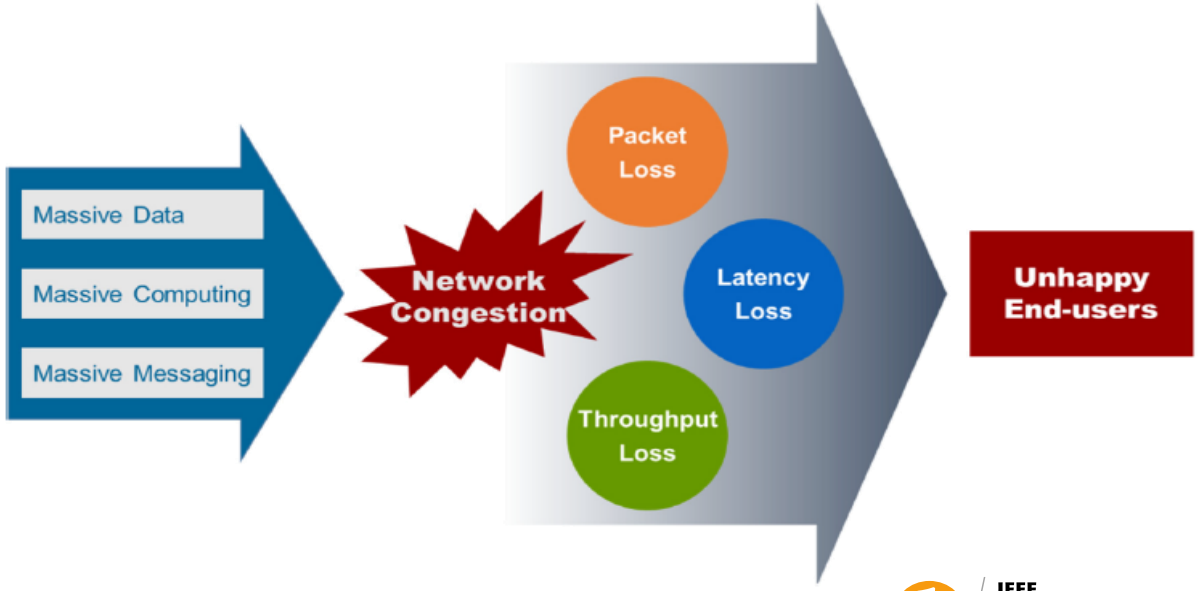
Massive amounts of data and computing

- Online Data Intensive (OLDI) Services
- Deep learning
- NVMe over Fabrics
- Cloudification of the Central Office



Challenges

Congestion Creates the Problems



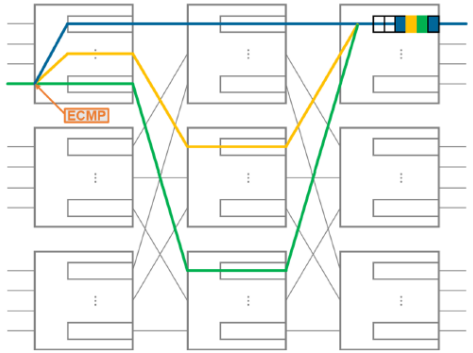
IEEE COMPUTER SOCIETY

IEEE

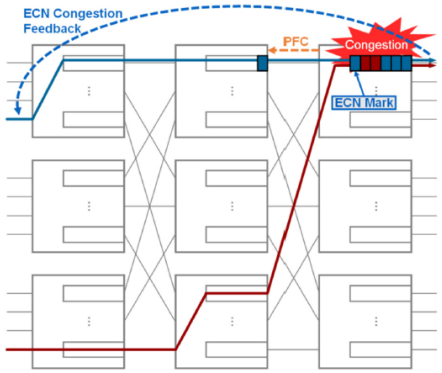
Nendica Reports

Future Technologies

Congestion management

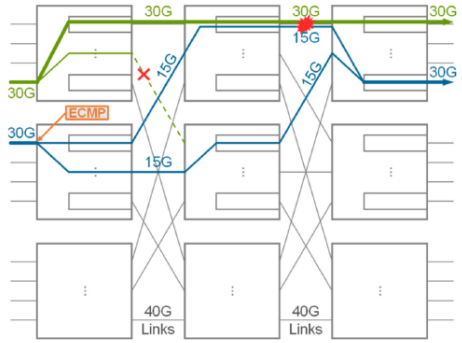


ECMP

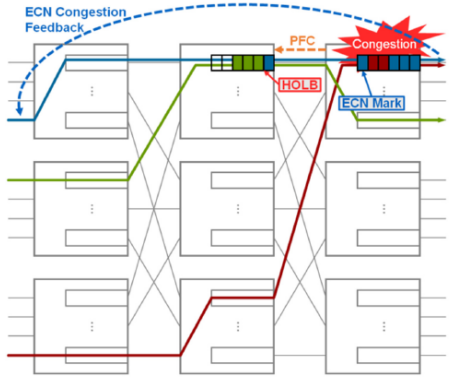


PFC+ECN

Issues



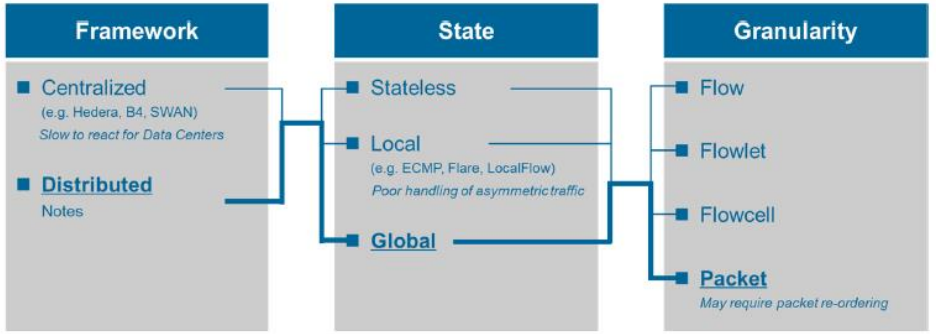
ECMP collisions



PFC drawbacks (HOLB, deadlock)
ECN control loop delays
Configuration complexity (PFC headroom, ECN threshold)

Innovations

- Load Balancing Design Space (**key for AI training**)



- Virtual input queuing
- **Dynamic virtual lanes -> 802.1Qcz**
- PFC deadlock prevention using topology recognition

- Push and Pull Hybrid Scheduling
- **Improving Congestion Notification -> P802.1Qdw**

- **Buffer optimization to reduce the complexity of PFC headroom configuration -> P802.1Qdt**
- Intelligent ECN threshold optimization

IEEE COMPUTER SOCIETY

IEEE

Congestion Isolation

Standard status：

- Initiated in 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”
- Published in August 2023 -- https://standards.ieee.org/ieee/802.1Qcz/

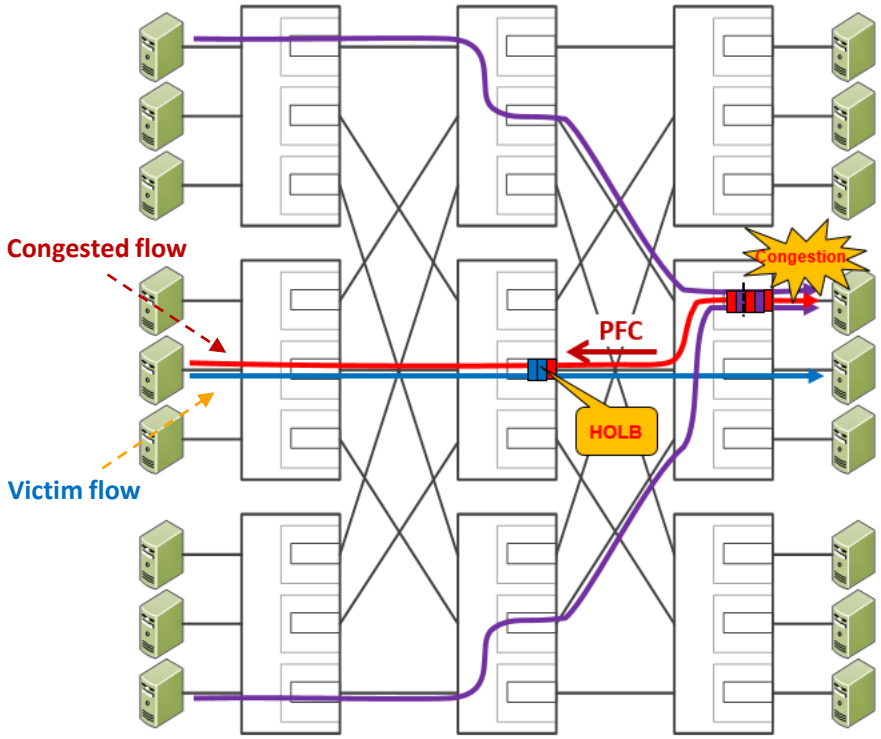
Project scope & goal:

- Amendment to IEEE 802.1Q to support the isolation of congested data flows within data center, supporting lossless and low-loss environments
- Work in conjunction with higher-layer end-to-end congestion control (ECN, BBR, etc)
- Reduce the frequency of relying on PFC for a lossless environment

Find detail information (including scope, history etc.) on website:
P802.1Qcz – Congestion Isolation | (ieee802.org)

Congestion Isolation

Problem to solve： Issue with PFC head-of-line blocking

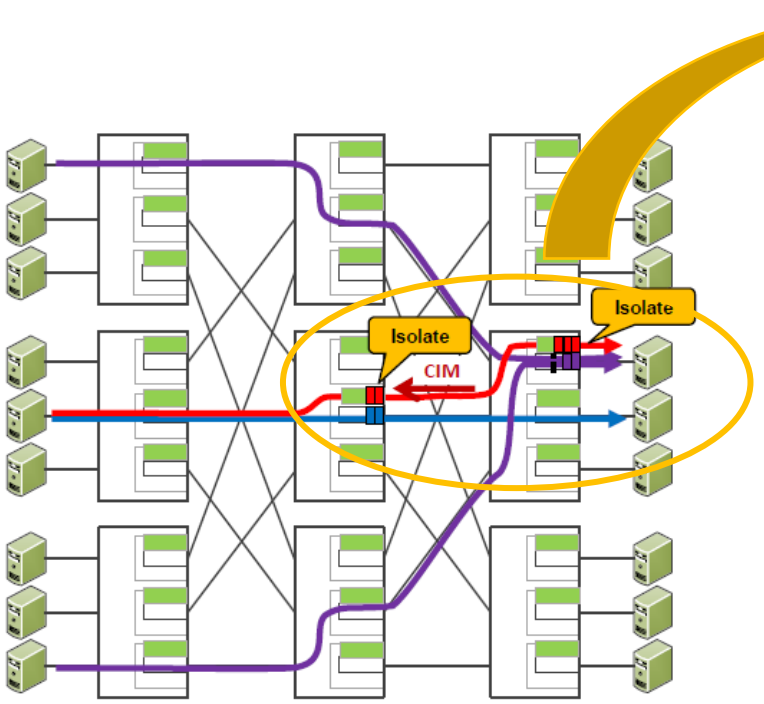


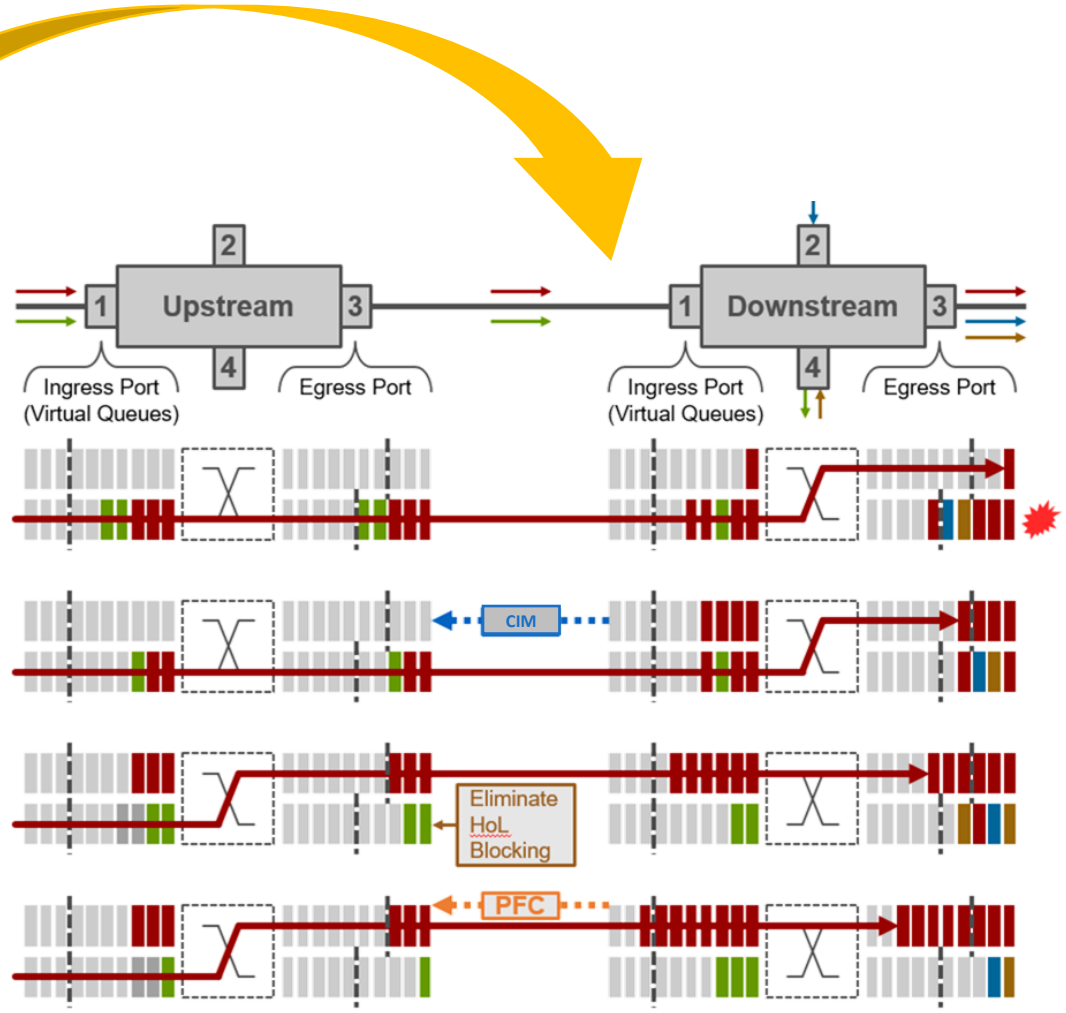
As the Ethernet data center network scales in size, speed and number of concurrent flows, the current environment creates head-of-line blocking for flows sharing the same queue.

- Increase jitter and latency of victim flow, especially harmful when victim flow is mice flow
- Increase frequency of congestion spreading which causes more victim flows and potential PFC deadlock.

Congestion Isolation

Key technology : Isolation on congested flows to mitigate HOLB





Congested Flow

Victim Flow

Congested Queue

Normal Queue

CI Procedure:

1. Identify the flow causing congestion and isolate locally

2. Signal to neighbor when congested queue fills

3. Upstream isolates the flow too, eliminating HOLB.

4. If congested queue continues to fill, invoke PFC for lossless

PFC Enhancement

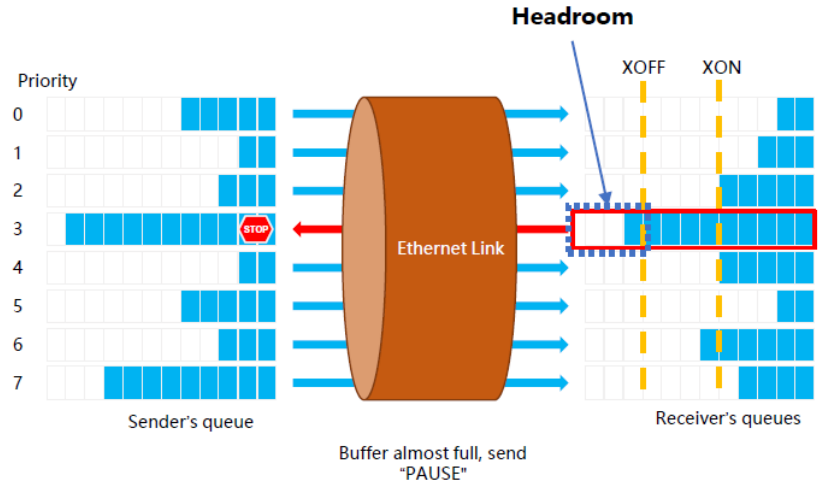
Standard status:

- (modified) PAR approved in June 2023 -- 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”
- Task group discussions, draft0.2 is ready

Project scope & goal :

- Amendment to IEEE 802.1Q, emphasizing on the requirements for low latency and lossless transmission in large-scale and geographically dispersed data centers.
- Automatic calculation for *PFC headroom*
- Protection on PFC frames by MACsec

PFC headroom

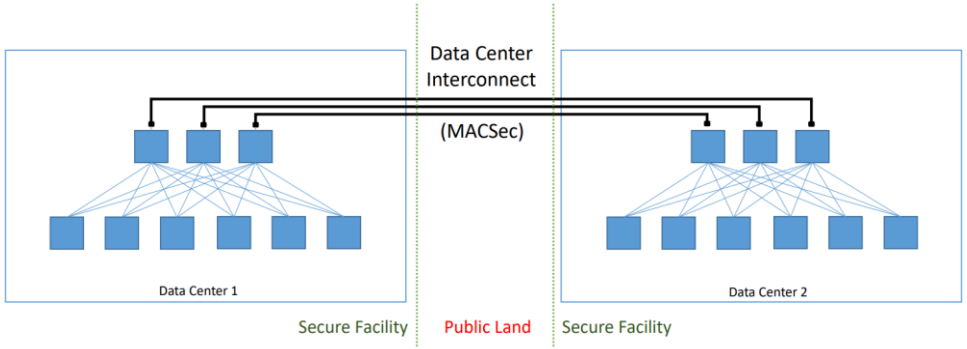


- There is a time delay between PFC invocation on sender and pause action on receiver.
- This PFC delay requires the PFC sender reserve buffer to absorb in-flight packets.
- The reserved buffer is called ‘headroom’.

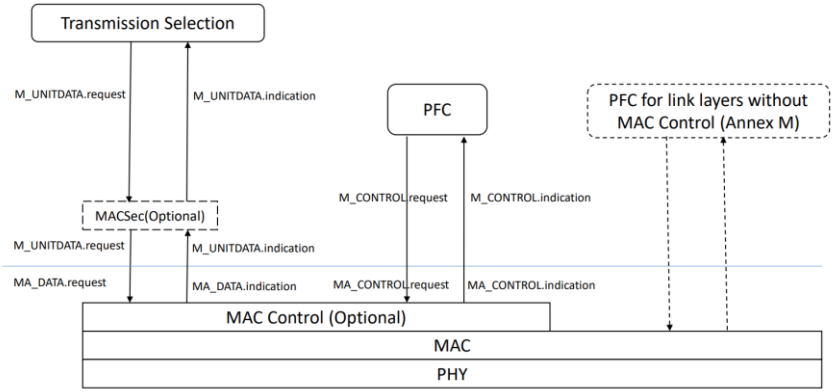
PFC Enhancement

Problem1 to solve： Issue with PFC interoperation with MACSec

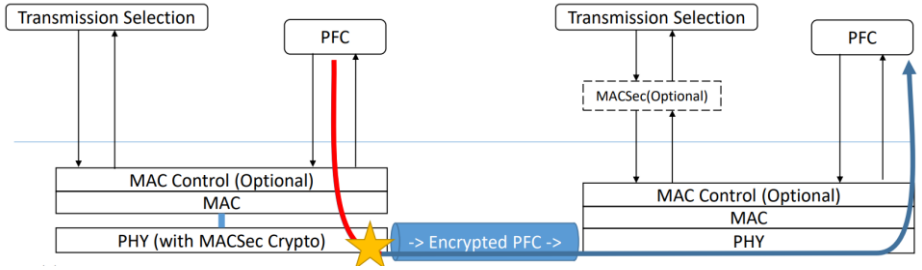
Use Case To Consider MACsec protection on PFC



- 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 don’t support PFC frames encryption.

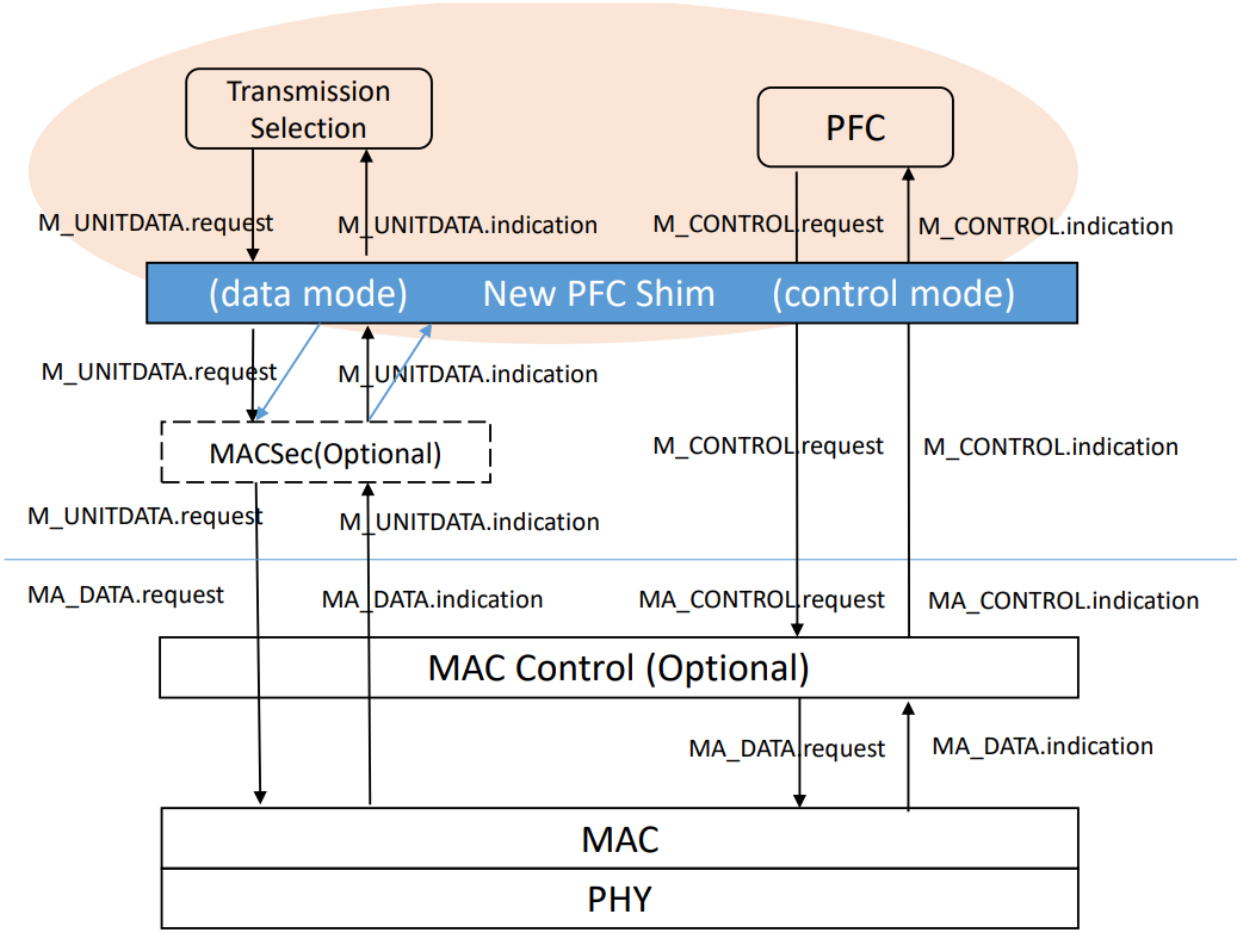


Implementations in the field have interoperability issues.

- 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

PFC Enhancement

Key technology : New PFC shim enabling protected PFC frames



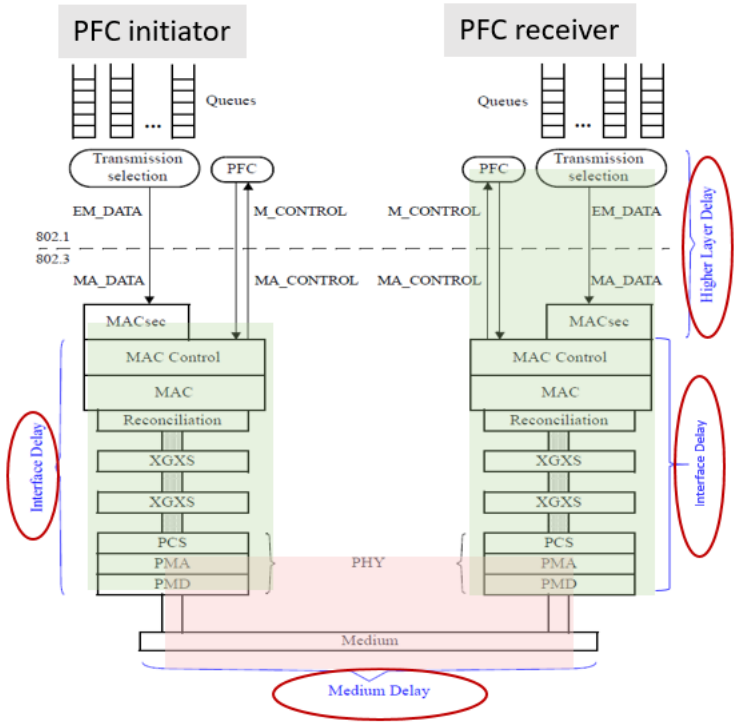
- Minimal (or no) impact to current PFC implementations
- Shim passes through existing MAC Control interface in ‘control mode’ (with no delay)
- Shim configured to generate and consume PFC frames if ‘data mode' is desired
- Internal delay calculation depends on Shim configuration

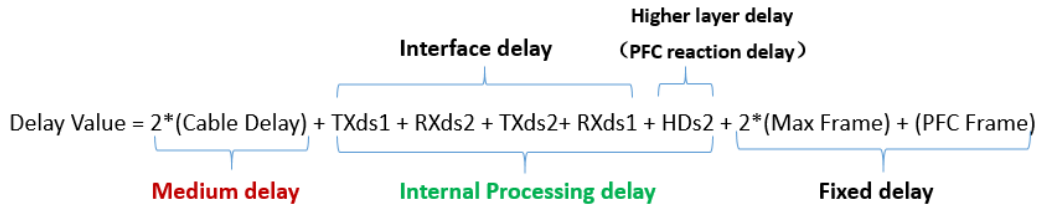
NOTE: MACsec delay is bounded and can be small and fixed.

PFC Enhancement

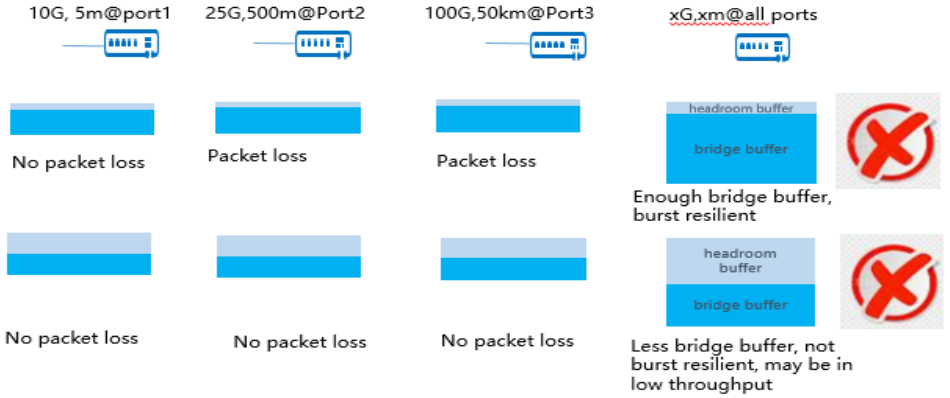
Problem2 to solve： Issue with PFC headroom configuration

PFC headroom calculation （worst case considered）





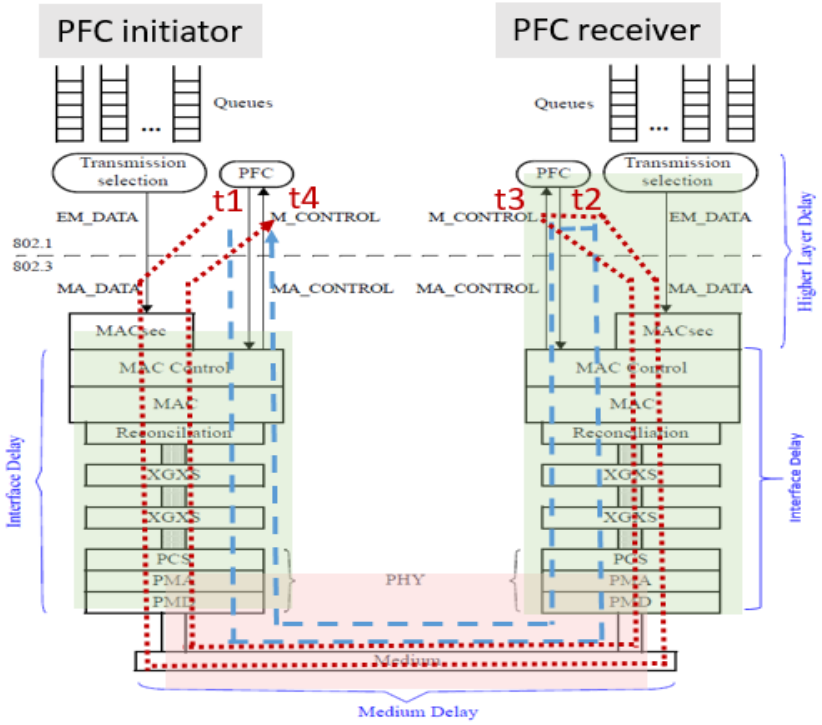
Current manual PFC headroom reservation in network management is inefficient



- 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 resource, and do not work in certain circumstances (e.g. long distance data center interconnection)

PFC Enhancement

Key technology: point-to-point roundtrip measurement enabling adaptive PFC headroom calculation



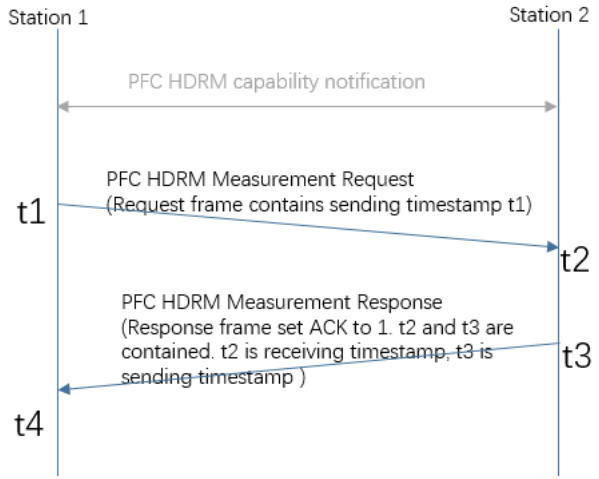
Delay value = t2-t1+ t4-t3

- t1: RX queue is above threshold and invokes signal to PFC module.
- t2: TX queue receives signal from PFC module and stops transmission.
- t3: last packet is sent after TX queue is stopped
- t4: last packet is received by RX queue

Non PTP-based mode:

Specify a new request-response measurement to measure the “round trip delay”

- Consider piggybacking mechanism in order to complete the measurement faster especially in request lost case



PTP-based mode:

Reuse PTP protocol but define separate mechanism to get peer node internal processing delay

- Reuse PTP protocol (Pdelay procedure) to measure cable delay
- Develop additional procedure (DCBX enhancement) to request peer node internal processing delay

Source Flow Control

Standard status：

- Initiated in September 2021, PAR approved in September 2022 -- IEEE 802.1 agreed to develop a Project Authorization Request (PAR) and Criteria for Standards Development (CSD) to amend IEEE 802.1Q with “Standards Development (CSD) to amend IEEE 802.1Q with “Source Flow Control”
- Individual contributions

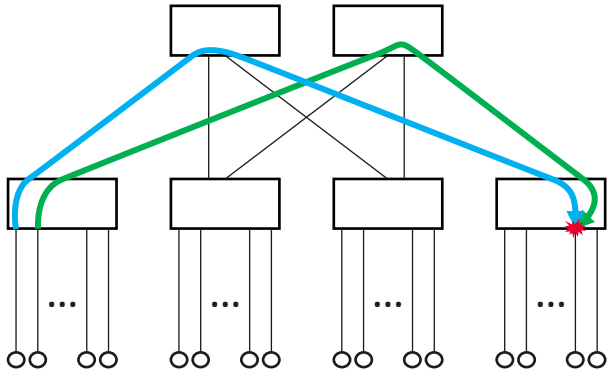
Project scope & goal:

- Amendment to IEEE 802.1Q, reducing latency in large scale data centers when congestion control is less effective.
- Remote invocation of flow control at the source of transmission
- Work in conjunction with other congestion control, such as DCQCN, DCTCP, Congestion Isolation
- Support L3 environments
- Enable early deployment without Server upgrades via Source

Source Flow Control

Problem to solve： Issues with in-cast congestion management

In-cast congestion

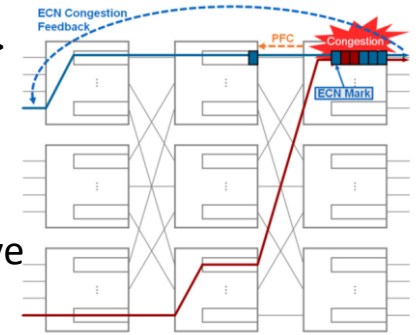


- Many-to-one traffic pattern
- Mostly at the last-hop
- Governs tail latency

End-to-end congestion control, e.g DCQCN

Issues:

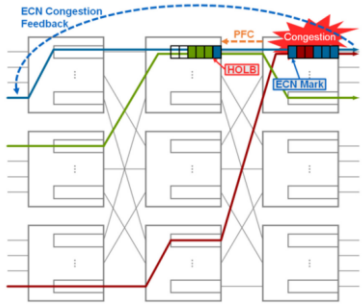
- Faster link speed-> shorter RTTs to finish a message-> need sub-RTT reaction
- Relies on forward signal, packets carrying the signals delayed by the congestion
- May need multiple RTTs to flatten down the rate curve (e.g CNP cuts rate by half）



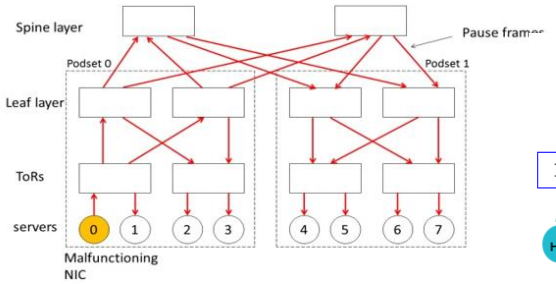
Hop-by-hop flow control, e.g PFC

Issues:

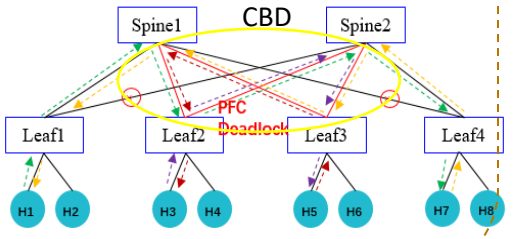
- HOLB



- PFC storm



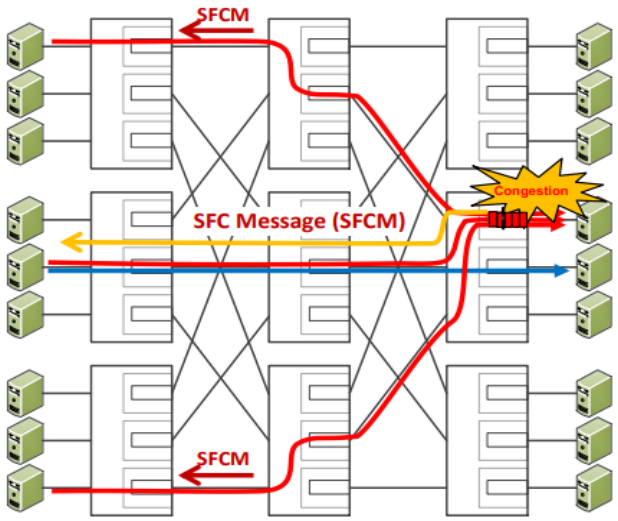
- PFC deadlock



Source Flow Control

Key technology : Invoking 'PFC' on source side directly when congestion detected

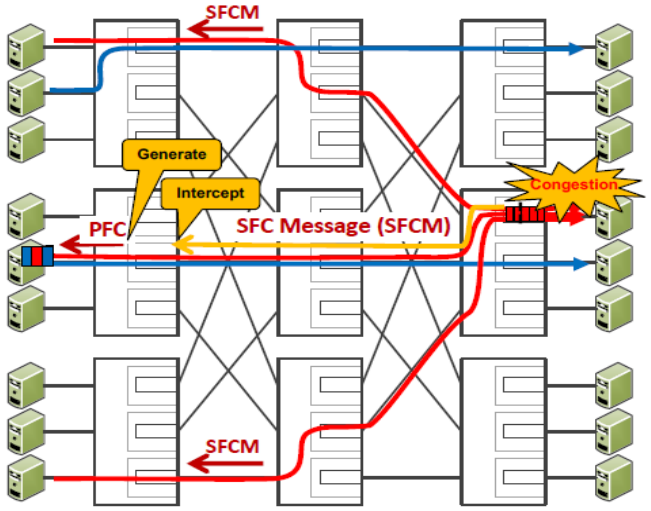
Source flow control



- Can be combined with Congestion Isolation
- If congestion persists, Edge-to-Source signaling using L3 message
- Signal from switch directly to traffic source: can allow for per-flow pausing

For early deployment

Source flow control (TOR proxy)



- ToR intercepts SFCM at egress port connected to non-supporting host using an egress stream_filter matching SFCM UDP port number
- ToR generates traditional PFC frame from SFCM

Proxy mode is a fast way to deploy the feature in the field.

- Uses PFC which is widely deployed in datacenter network.
- No changes on hosts, only changes on TOR switches.
- Simple interaction between switches, no new protocol between switch and host.

Basic information discussed in SFCM： source IP, dst UDP port, pause duration, DSCP

Summary

- Explosion of computing power growth ignites another wave of network technologies innovation in data center.
- IEEE802.1 already started research and standardization work to meet the emerging AI/HPC/cloud requirements.

- It is expected to see more work happening in IEEE802.1 and coordination between IEEE802.1 and other SDOs， to enable hyper converged ethernet supporting today’s high performance applications.

Thank You !

IEEE COMPUTER SOCIETY

IEEE