Towards a New Ethernet for High-Performance Data Centers

Activities and Enhancements in IEEE 802.1

Paul Congdon

17-OCT-2022

Towards a New Ethernet for High-Performance Data Centers

Activities and Enhancements in IEEE 802.1

17-OCT-2022

Paul Congdon

NANOG™

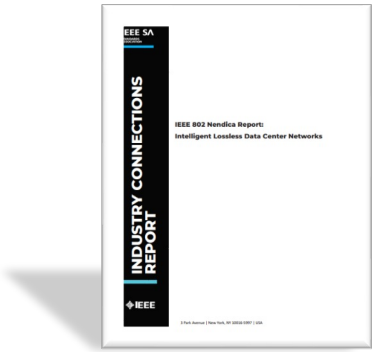
Disclaimer

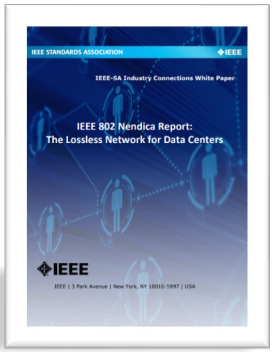
- This presentation should be considered as the personal view of the presenter not as a formal position, explanation, or interpretation of IEEE.

- 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.”

IEEE 802 Nendica Reports

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





- 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)

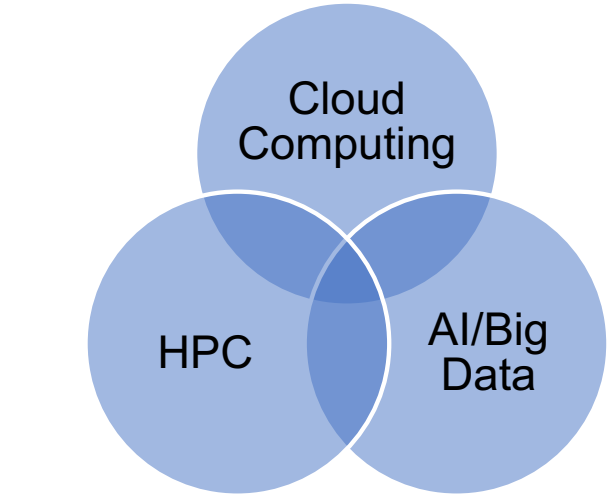
Change in the ‘Data-Centric’ Era

Our daily lives are changing from the combination of mobility, cloud computing, high performance computing and AI/Big Data.

Their cause and effect has brought an explosive growth of data.

The exponential growth of data is forcing the evolution of computing systems to 'data-centric' computing systems.

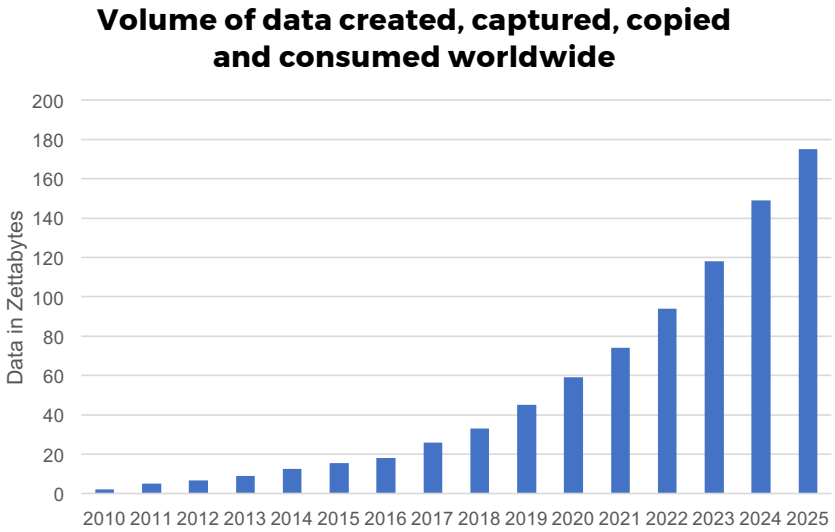
- High performance is essential for storing, moving and processing data.
- Computing systems require breakthroughs in processing power, storage performance and network connectivity.



NANOG™

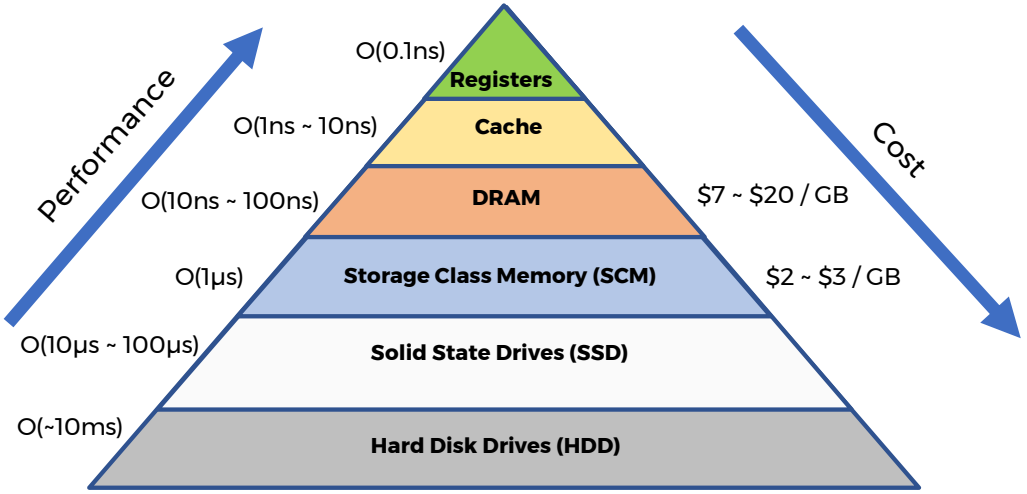
‘Data-Centric’ is enabled by Data

Big Data



IDC predicts that the Global Data will grow from 45 Zettabytes in 2019 to 175 Zettabytes by 2025

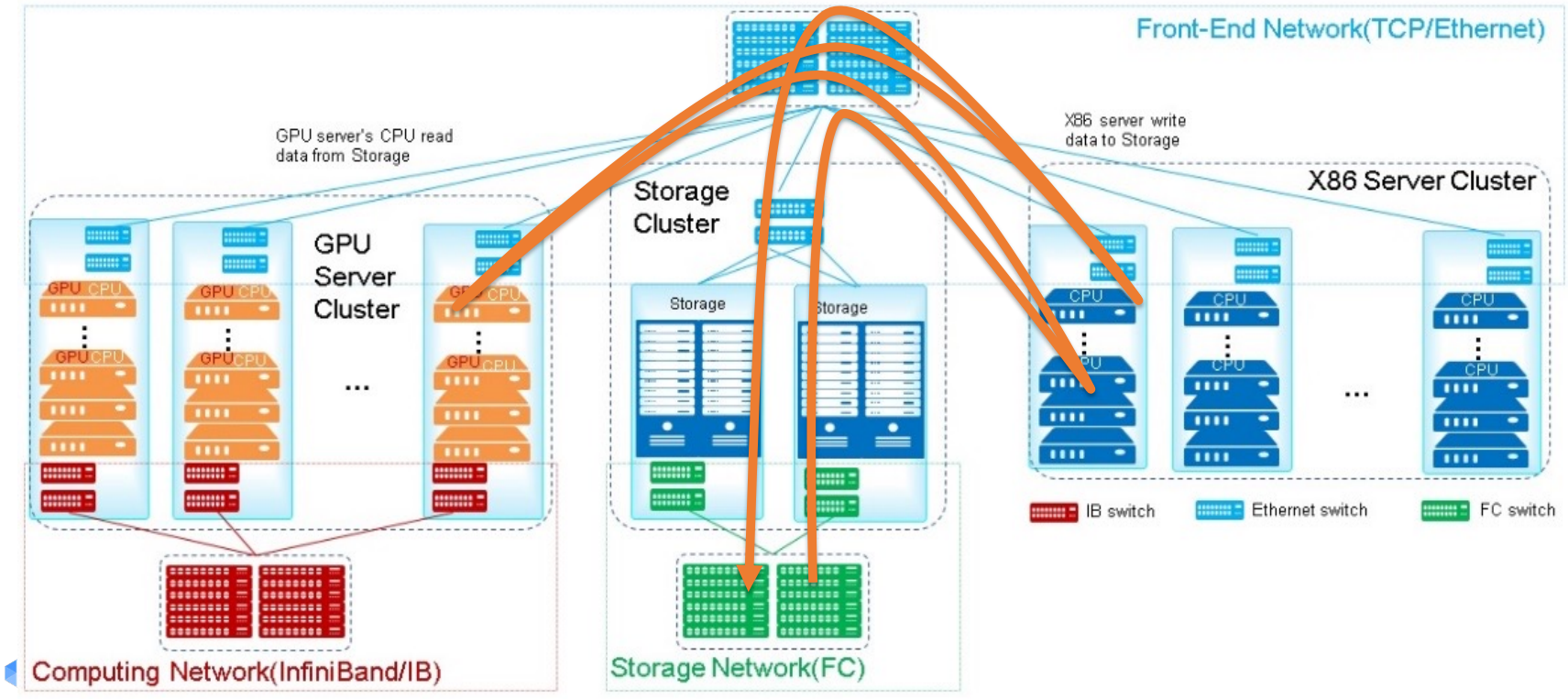
Fast Data



Persistent storage latencies are approaching memory latencies with the latest Storage Class Memory (SCM) technology. Accessed over the network using NVMeoF

Traditional approaches can’t keep up

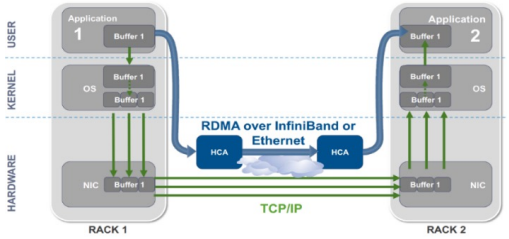
Moving the data to the compute doesn’t scale or perform

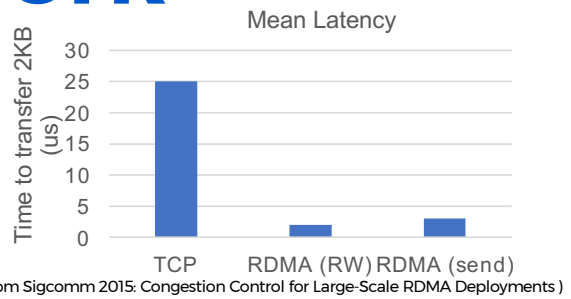


New Tech stresses the Network

RDMA

Reduce memory bandwidth, CPU cycles and read/write time with efficiency and zero copy at end-points. **E2E latency** is mainly contributed by network.

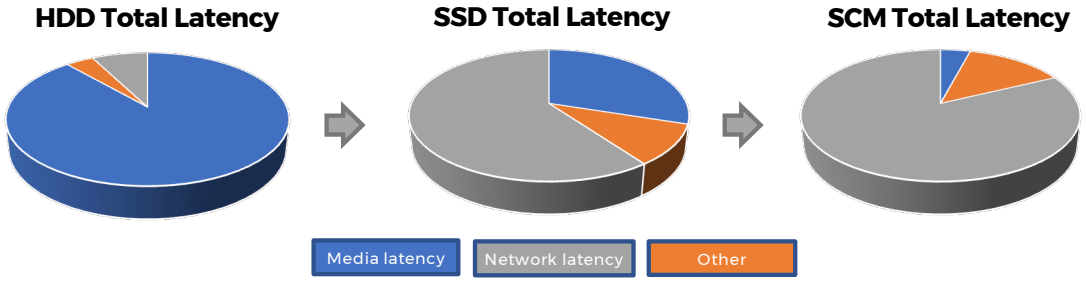




(Data from Sigcomm 2015: Congestion Control for Large-Scale RDMA Deployments)

NVMe

Faster storage media has higher IOPS and lower latency. **Network latency** become the biggest portion in the total latency.

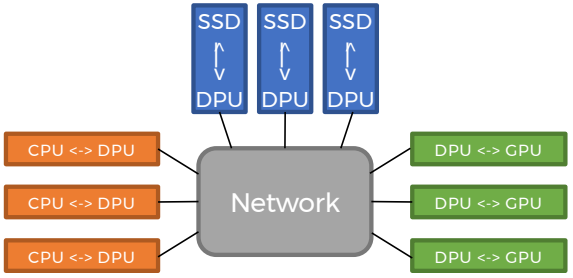


Distributed HW with Parallel SW

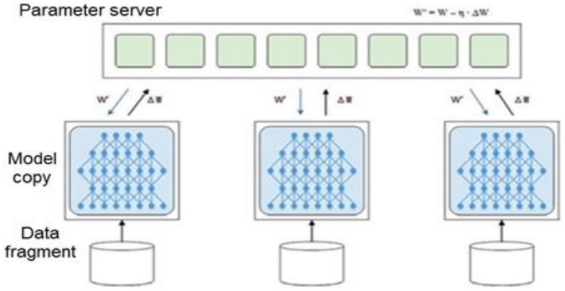
Distributed infrastructure using Data Processing Units (DPUs/SmartNICs) and parallel software architectures increase communication and require data transmission with **lower tail latency**.

NANOG™

Distributed Infrastructure with DPUs



Parallel AI Software



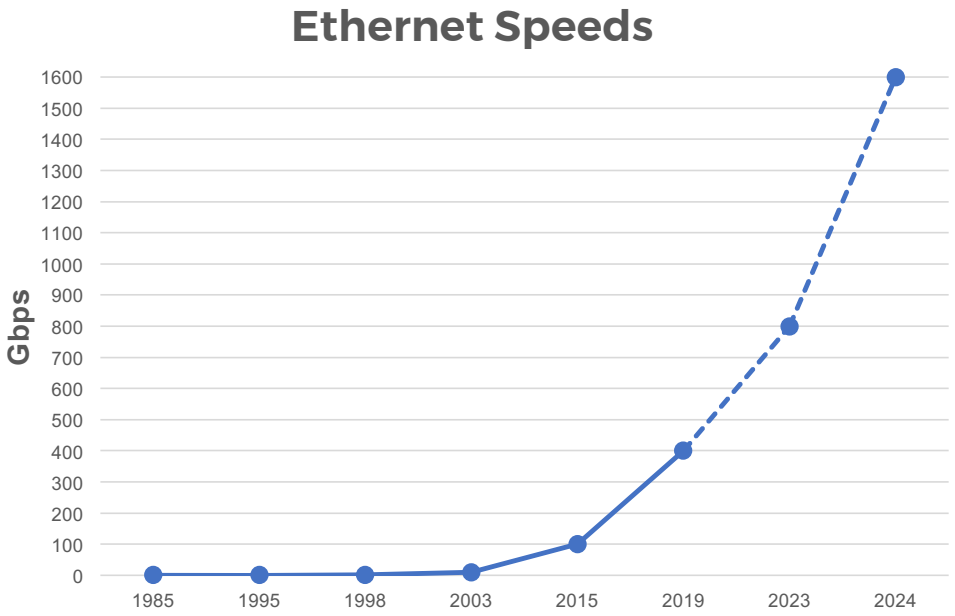
Ethernet for the 'Data-Centric' Network Fabric

Ethernet is **ubiquitous technology**.

- Cost-effective solution
- Relatively easy to deploy and manage
- Leading technology development

Ethernet provides **large bandwidth** connectivity

- up to 400 Gbps, 100G for single lane
- towards 800 Gbps, 200G for single lane

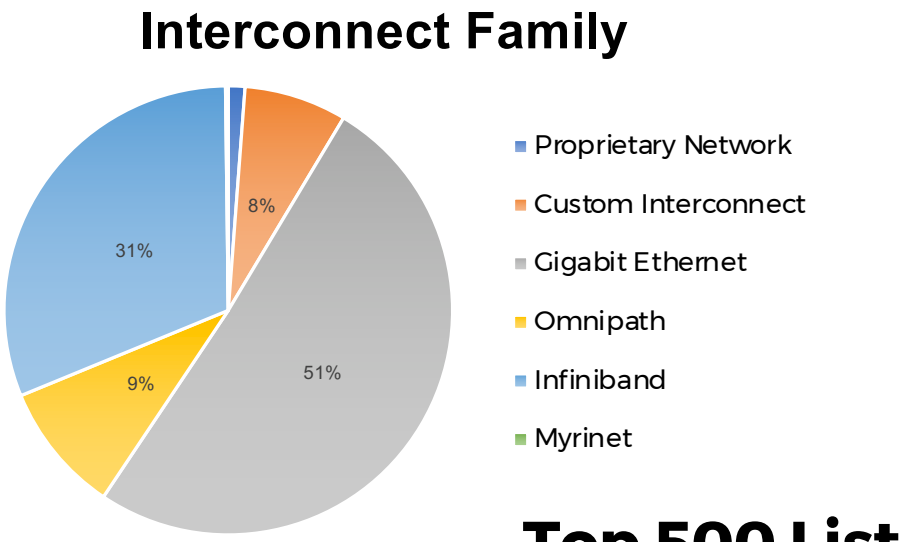


NANOG™

Ethernet is the popular choice

While dedicated protocols are designed for computing and storage networks, like IB for HPC, FC for storage, **people see the advantages of Ethernet**.

- In **TOP500**, which is a list of the world's 500 most powerful computer systems, 51% supercomputers use high performance Ethernet fabrics, more percentage than IB and other proprietary technologies.



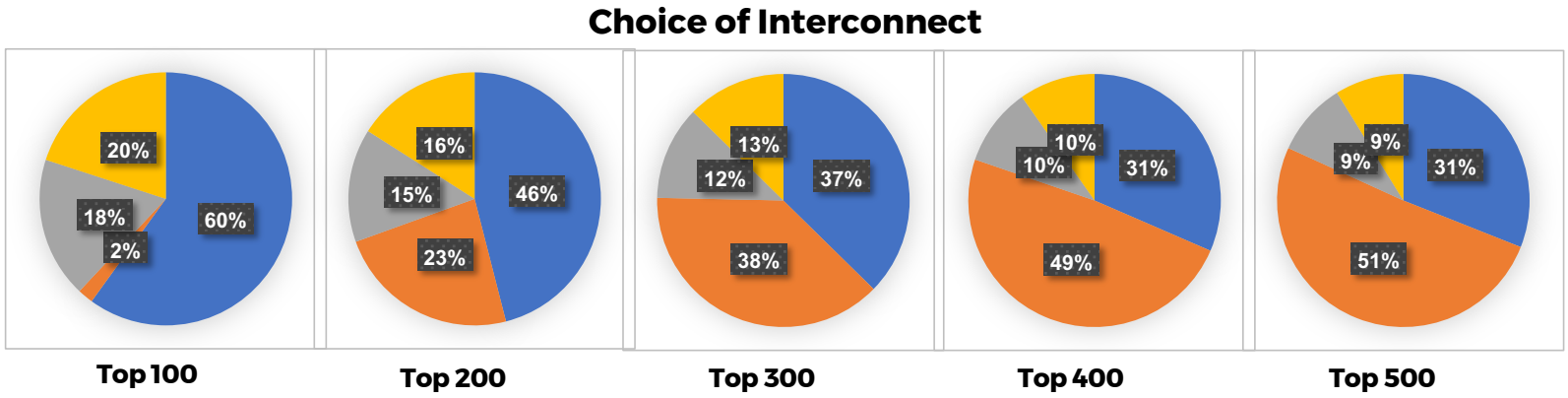
Top 500 List

However, Improvements are needed

In the Top500, the Ethernet advantage is most likely due to price-performance.

There is still a performance gap.

- In Top500 performance metric, the fastest Ethernet system is 100 times slower than #1.
- In Top100 Ethernet interconnect is far behind Infiniband.



Perceived Advantages of Other Fabrics

Some claims for why Infiniband is superior

- Guaranteed delivery at the HW level. Credit based flow control
- Hardware based re-transmission. Higher throughput
- Better lossless congestion management
- Cut through design with late packet invalidation
- 16 Virtual Lanes vs 8 Traffic Classes in Ethernet
- Preselected failover paths and switches for instant recovery
- Lower cost IB switch chips due to technical differences

Some known proprietary tweaks to Ethernet for HPC

- Per-flow, credit basis congestion control
- Reduced minimum frame size (< 64B) with local addressing
- Auto-negotiation between Standard Ethernet and HPC Ethernet features
- Low-latency FEC, link-level retry to tolerate transient errors

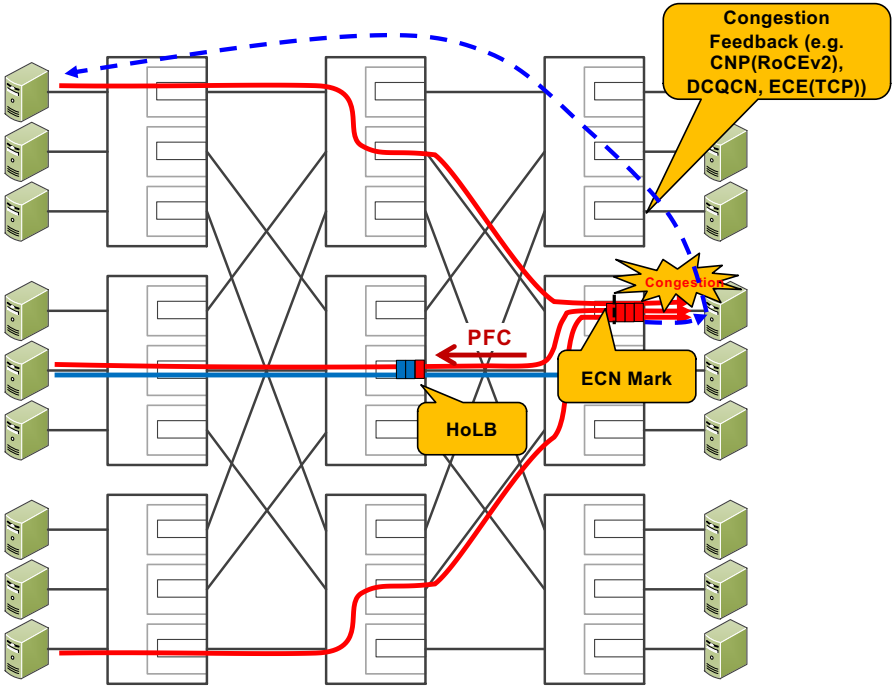
IEEE 802.1 DCB Made Progress

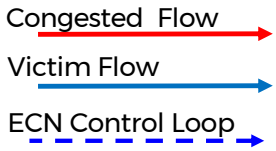
Standard	Description	Contribution
802.1Qau-2010	Congestion Notification	Layer-2 end-to-end congestion control
802.1Qaz-2011	Enhanced Transmission Selection	Bandwidth sharing between traffic classes
802.1Qbb-2011	Priority-based Flow Control	Lossless traffic classes
802.1Qcz-2023	Congestion Isolation	Avoid head-of-line blocking

However, a lot more is needed…

- Avoid/mitigate incast congestion
- Improve congestion detection and signaling
- Further reduce network latency
- Automate and/or simplify configuration
- Define attributes for proactive analytical response to congestion

Still the DCN state-of-the-art environment

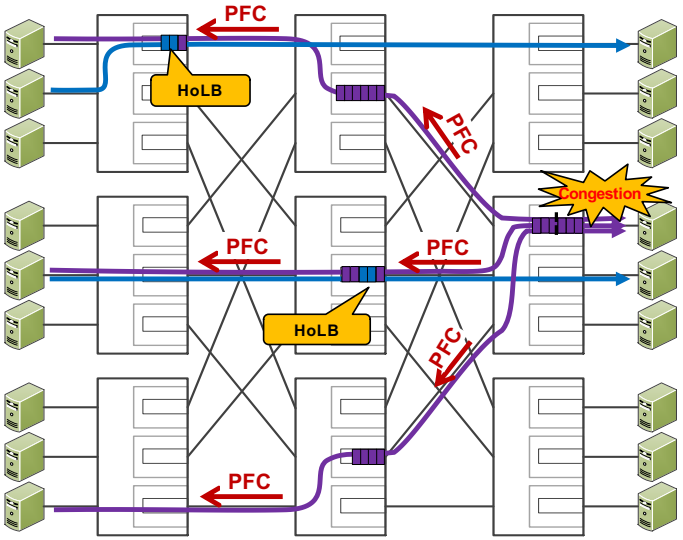




- DCNs are primarily L3 CLOS networks
- ECN is used for end-to-end congestion control
- Congestion feedback can be protocol and application specific – including new proprietary transports
- PFC still used as a last resort to ensure lossless environments – perhaps just at the edge.
- Traffic classes for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

Existing 802.1 CM Tools

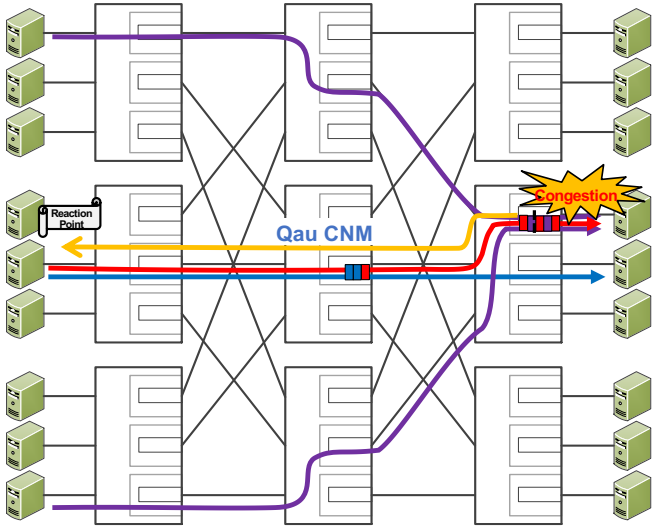
802.1Qbb - Priority-based Flow Control



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

Three New Initiatives of Interest

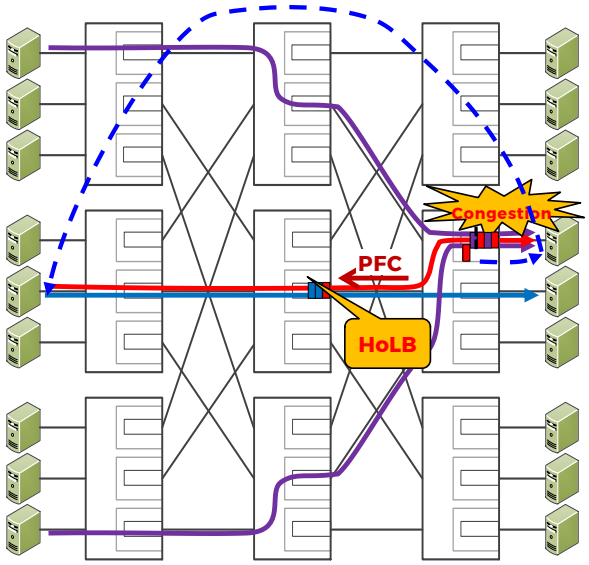
Motivated to enable low-latency, low-loss, high-reliability Ethernet-based Data Center Networks supporting RDMA and AI/HPC workloads.

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

These are all ‘amendments’ to IEEE Std 802.1Q

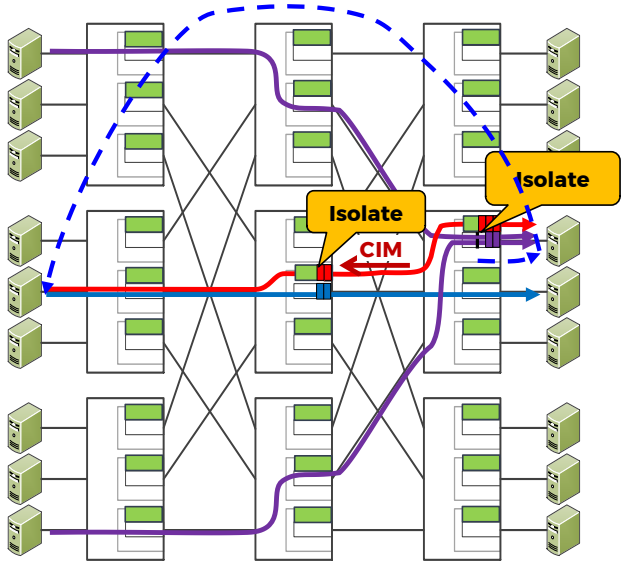
P802.1Qcz - Congestion Isolation

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



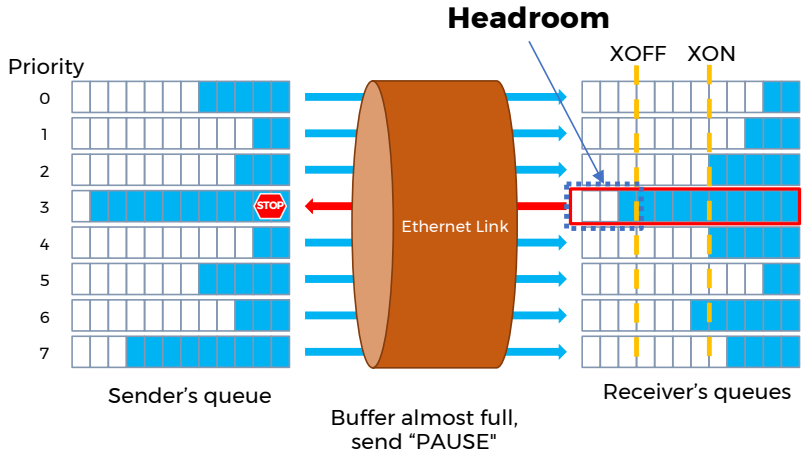
1. Move congesting flows to a separate queue and signal your upstream neighbor
2. Upstream neighbor moves congesting flows to separate queue

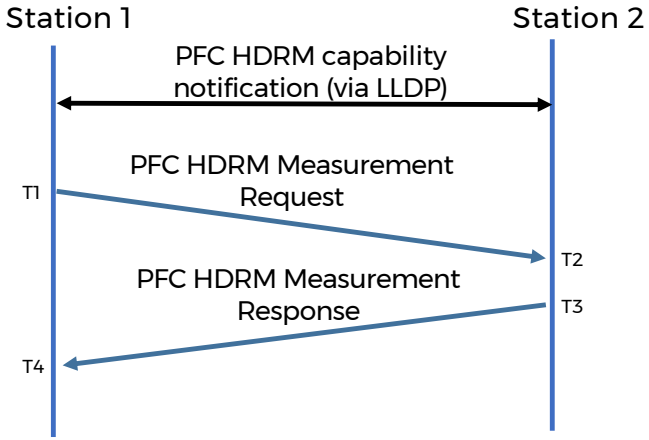
Congestion Isolation - Goals

- 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

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



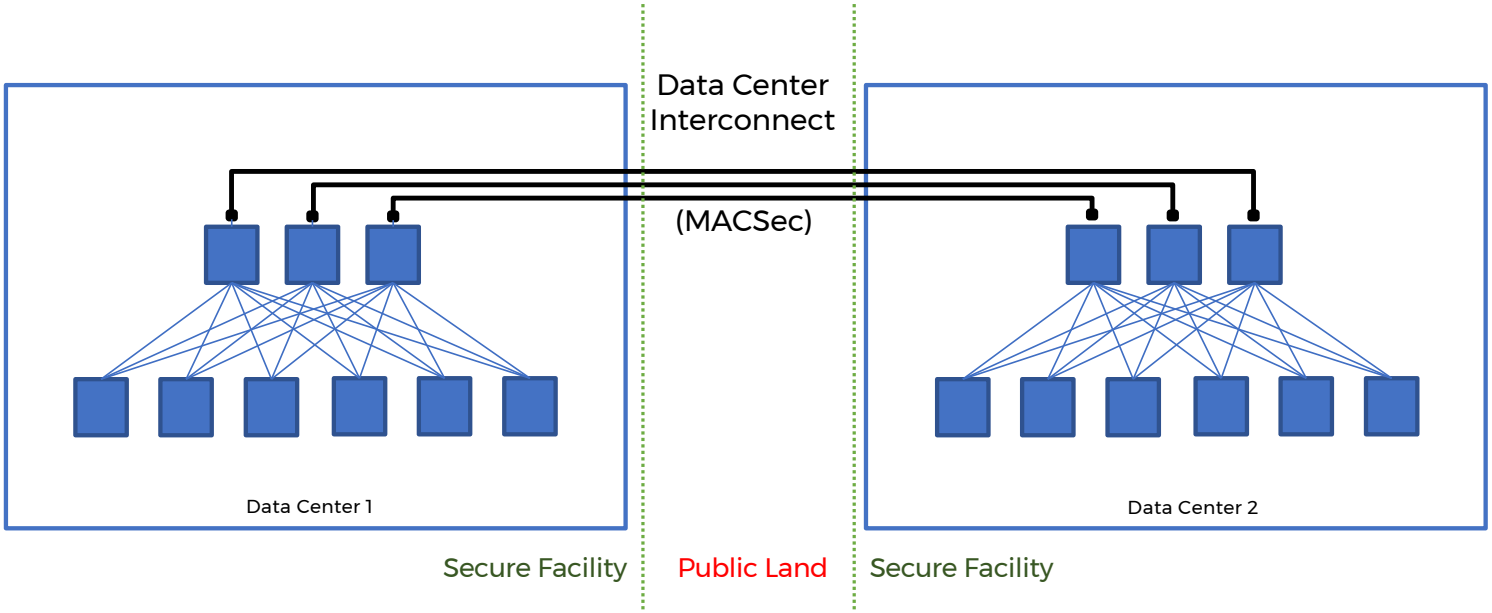


Headroom needed = (Port speed * (T4-T1-(T3-T2)) + 2*(Max Frame) + (PFC Frame)) * Alpha

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

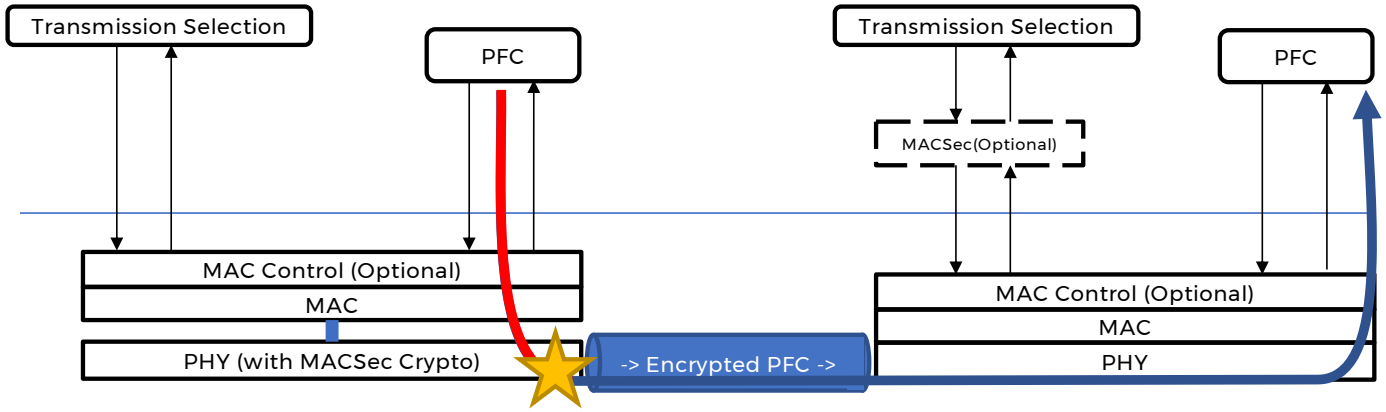


See: https://youtu.be/CJP1rJnPVG8?t=712

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

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 IEEE Std 802.1AE
  - These early implementations do not run key agreement protocol and may suffer outages

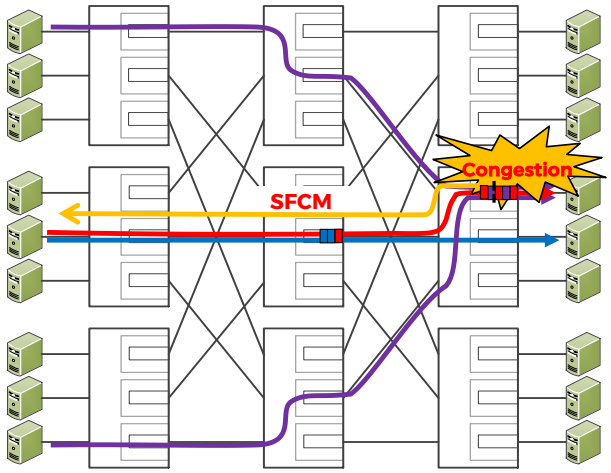


PFC Enhancements - Goals

- 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

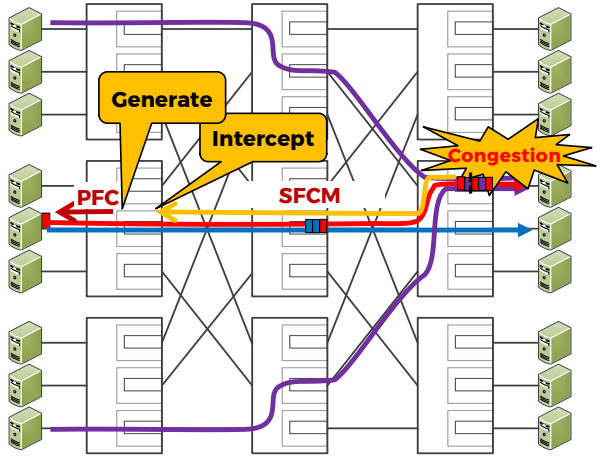
P802.1Qdw - Source Flow Control

Source Flow Control



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

Source Flow Control (w/ ToR Proxy)



- Optional source Top-of-Rack switch involvement
- Support SFC un-aware servers
- Intercept Edge-to-Source signaling and convert to PFC
- Simplifies deployment and migration to new functionality

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.

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.
- P802.1Qdw (SFC) is just beginning. A standards related technical design team exists with multiple vendors (Broadcom, Dell, Intel, HPE, Huawei) 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

Additional Concepts Considered

The current initiatives are small steps – For more advanced ideas see:

- 2021-06-22 - IEEE 802 Nendica Report: Intelligent Lossless Data Center Networks (ISBN: 978-1-5044-7741-3)

1. Hybrid Transport Protocols
2. Supplemental Congestion Notifications
3. PFC Deadlock Free Mechanisms
4. Dynamic ECN Threshold Adjustments
5. AI Models Built From Network Telemetry

Potential Future Nanog Presentations?

Thank you

17-OCT-2022

NANOG™