DCN 1-20-0044-00-ICne

Scaling the Intelligent Lossless Data Center with PFC Deadlock Prevention

PAUL CONGDON, CTO TALLAC NETWORKS

IN COOPERATION WITH IEEE 802 NENDICA AND HUAWEI TECHNOLOGIES, LTD.

JULY 16TH, 2020

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."

Background: Nendica

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

IEEE Industry Connections Activities “provide an efficient environment for building consensus and developing many different types of shared results. Such activities may complement, supplement, or be precursors of IEEE Standards projects”

Organized under the IEEE 802.1 Working Group

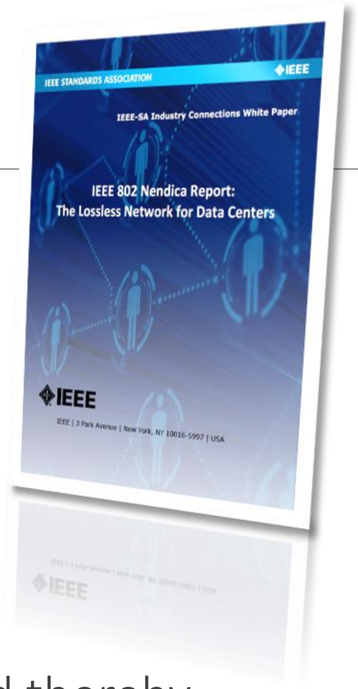
Chartered through March 2021

Open to all participants; no membership

https://1.ieee802.org/802-nendica/

Nendica Report: August 2018

The Lossless Network for Data Centers



Paul Congdon, Editor

Key messages regarding the data center :

- Packet loss leads to large delays.
- Congestion leads to packet loss.
- Conventional methods are problematic.

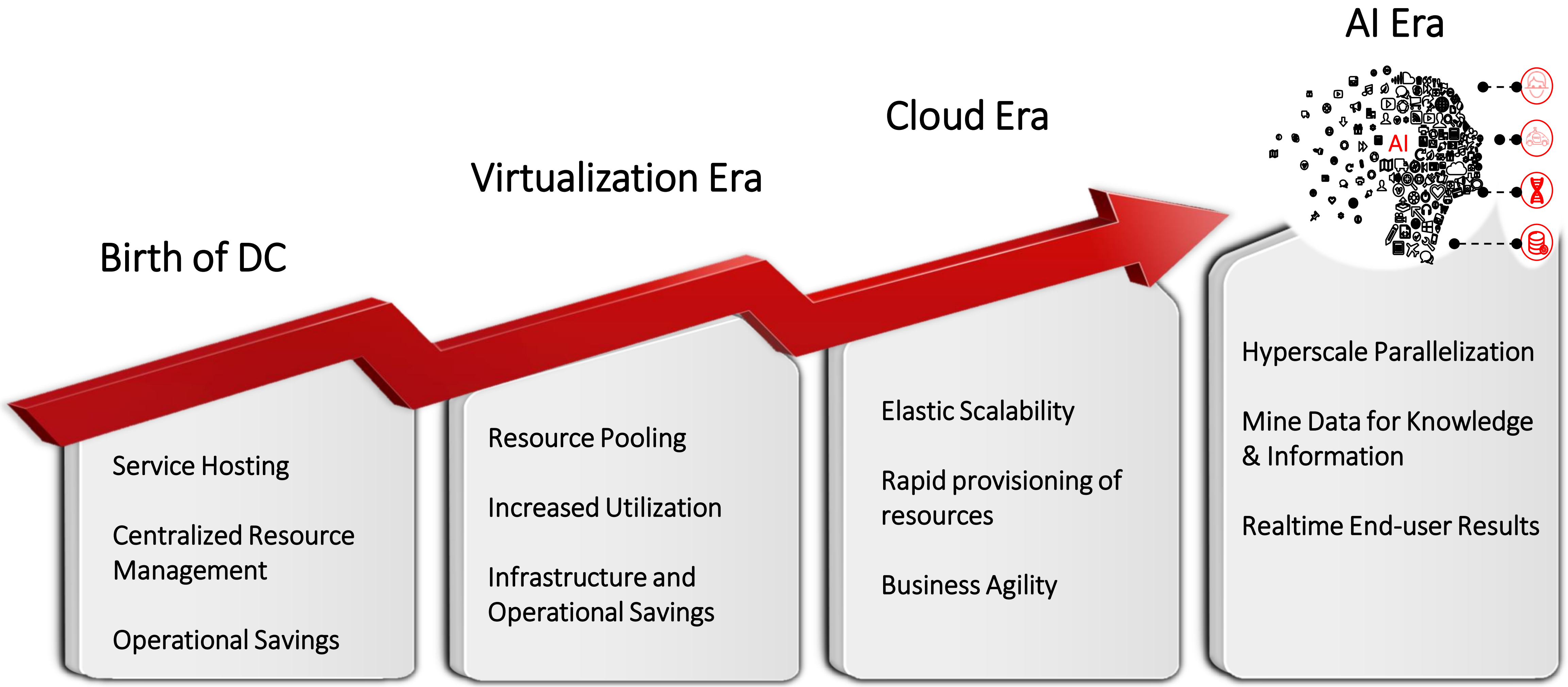
A Layer 3 network uses Layer 2 transport; action at Layer 2 can reduce congestion and thereby loss.

The paper is not specifying a “lossless” network but describing a few prospective methods to progress towards a lossless data center network in the future.

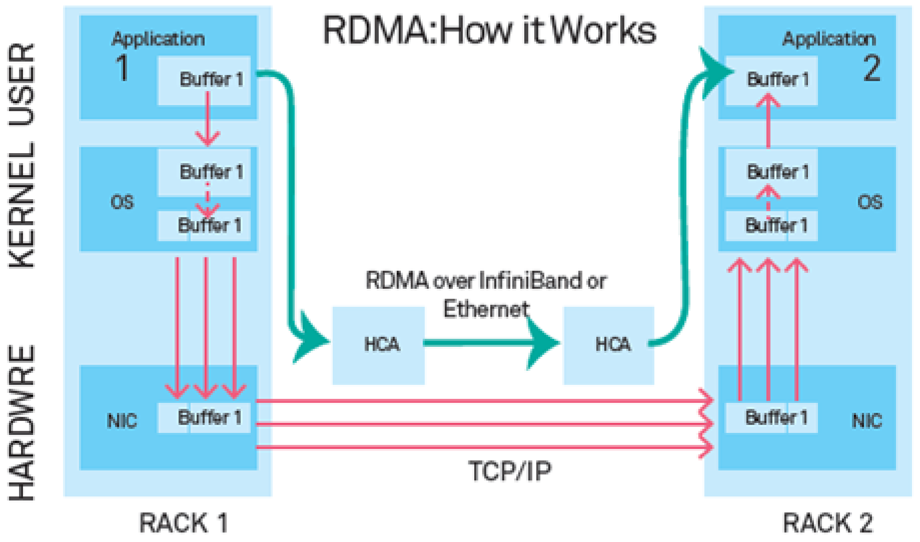
The report is open to comment and currently being revised.

https://ieeexplore.ieee.org/document/8462819

Data centers stepping into the AI era



RDMA is an essential protocol for the AI era



TCP disadvantages

- Slow startup and low throughput
- Three copy operations, resulting in a long latency
- CPU consumed by traffic: 1 Hz per bit

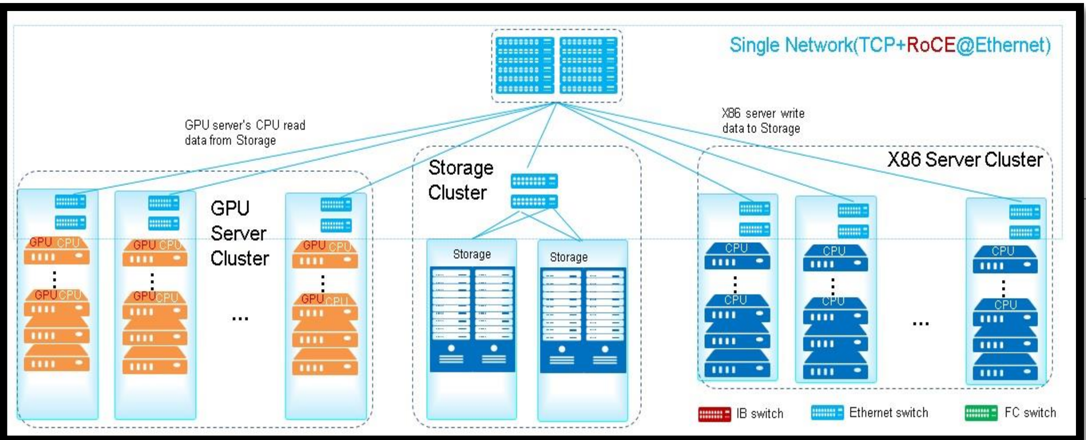
RDMA advantages

- Fast startup, maximizing the bandwidth usage
- One copy operation, effectively reducing the kernel latency
- Zero CPU resources consumed upon network adapter uninstallation

RDMA advantages are more significant as link bandwidth increases (e.g. 400 Gbps)

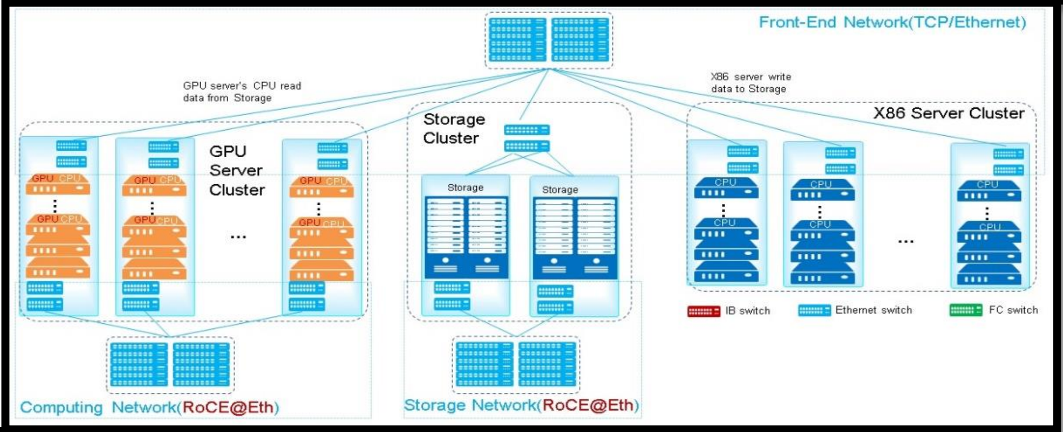
- Traditionally deployed in custom, closed and expensive Infiniband networks
- Adapted to Ethernet networks for better scale, lower cost and manageability.
- Network innovation is preparing RDMA for wide scale use

Hyperscaling HPC/RDMA



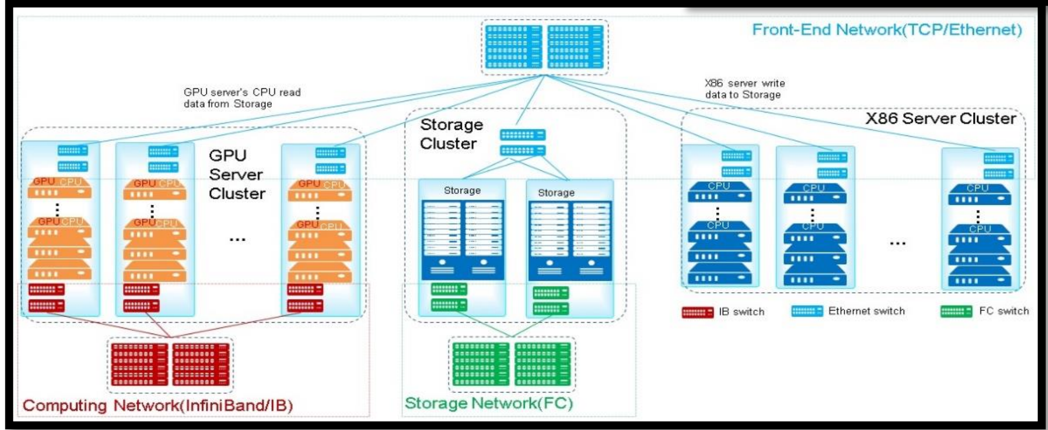
Separate Networks

- ✓ Multiple O&M
- ✓ Multiple domains of expertise
- ✓ Different hardware, different lifecycles, multiple HBAs, NICs



The Single Network

- ✓ Reduce **30%+** calculation time, increase **20%+** storage throughput, improve application performance, reduce overall wiring complexity
- ✓ Reduce TCO by **30%,** one technology, one network, multiple services, and unified O&M



Innovations have allowed Ethernet performance to be equivalent to Infiniband and Fibre Channel: Ethernet can replace Infiniband and Fibre Channel

Next Generation DCN Needs

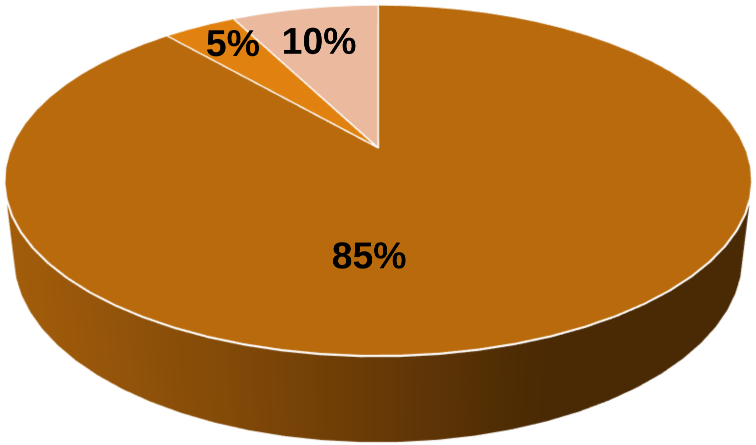
	Traditional Ethernet Data Center	Converged Enhanced Ethernet Data Center (CEE)	Next Generation Ethernet Data Center
Drivers for Change	• Cost Reduction	• Storage Workloads	• Information Workloads (AI & OLDI)
Requirements	• Low Cost	• Zero Packet Loss	• Large Scale • Low Latency • High Throughput
Key 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)	• Scalable Solutions for no Packet Loss • Solutions for Latency Reduction • Load Balancing for High Throughput • Solutions for Congestion Elimination
	Before 2008	2008 to 2020	After 2020

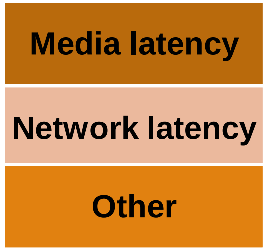
10M/100M/1000M — GE/10GE/40GE — 25GE/100GE/400GE

A Challenge for NextGen Network Storage

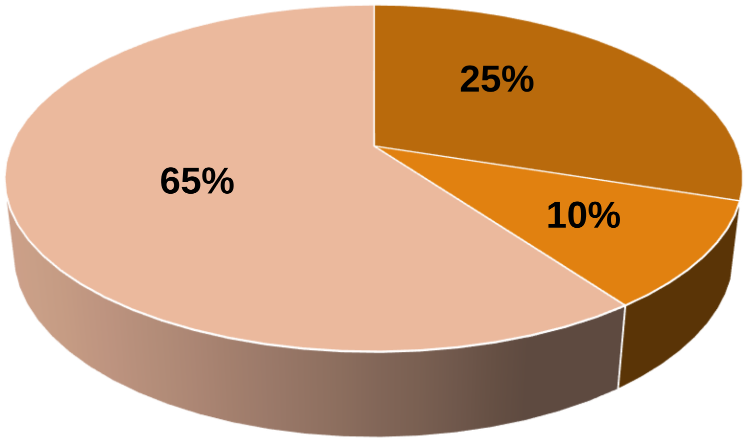
Network Storage Evolution

Traditional NAS





NVMe-oF



Network latency in HDD scenario: negligible

Current latency: > 300 μs

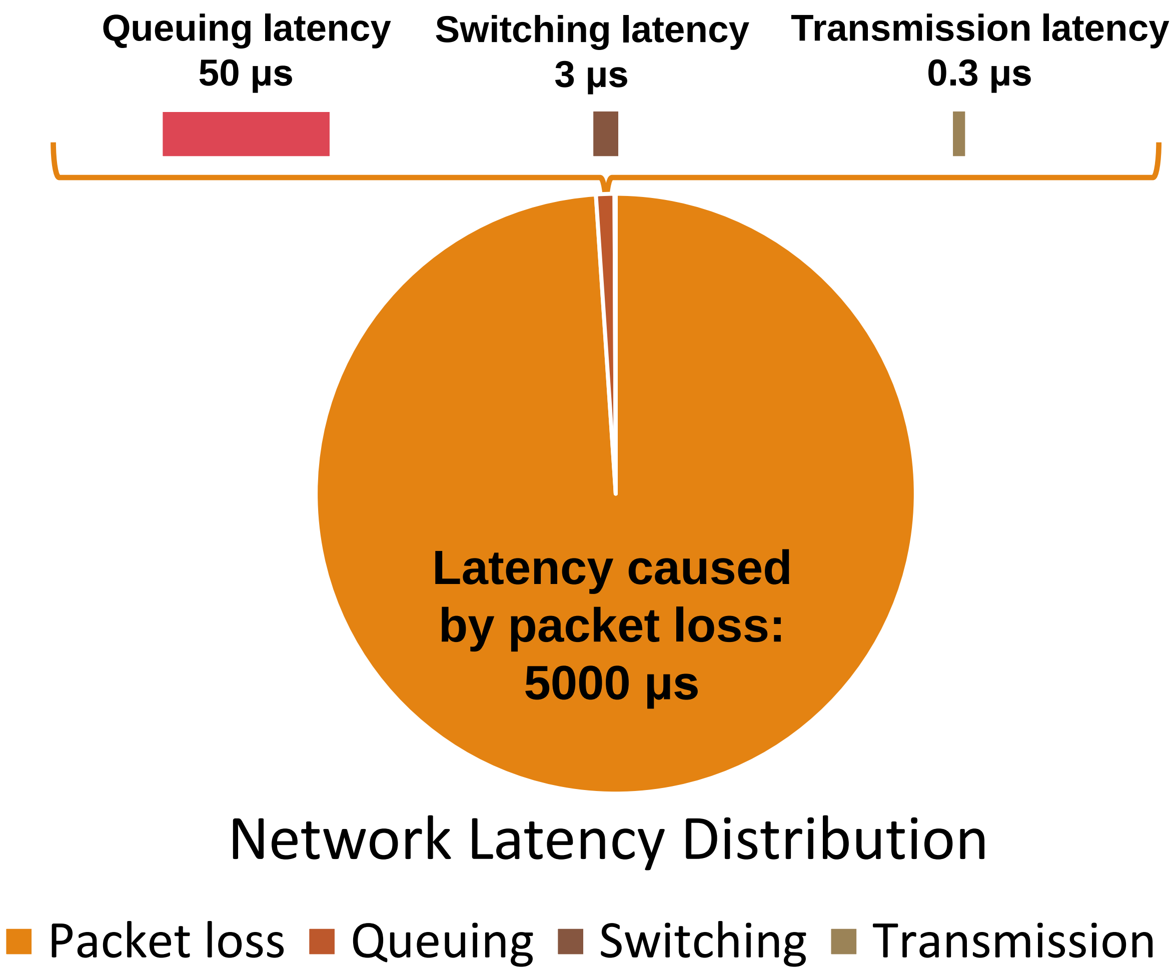
Network latency in SSD scenario: bottleneck

Latency target: < 50 μs

Latency is a limiting factor to improving storage IOPS

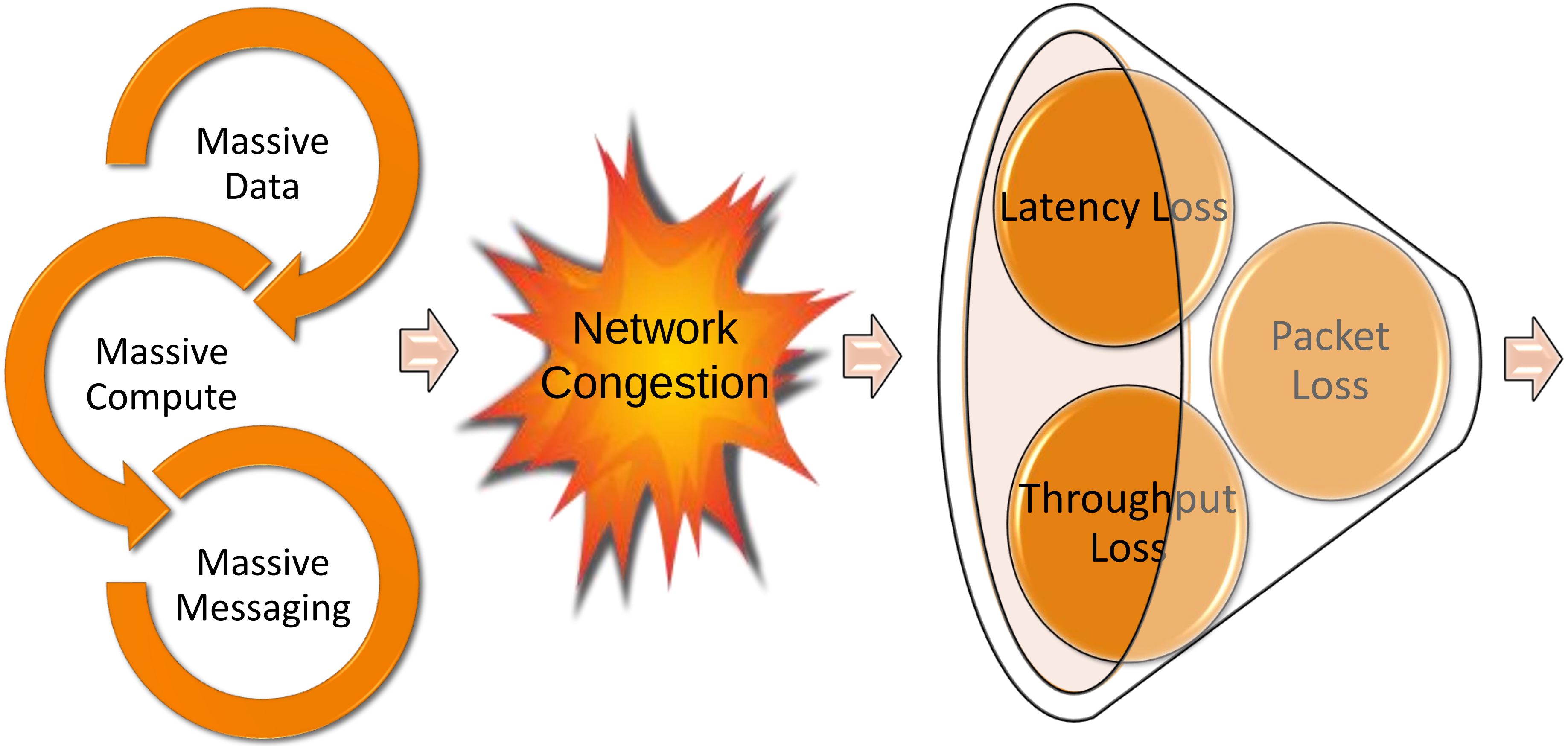
The Key to Reducing Network Latency:

Focus on Dynamic Latency



Dynamic Latency = Queuing Delay + Packet Loss Delay

Congestion is the problem



Scaling HPC/RDMA can lead to Congestion which Leads to Loss which Leads to Unhappy End-users

Mitigating Congestion in the Ethernet DCN

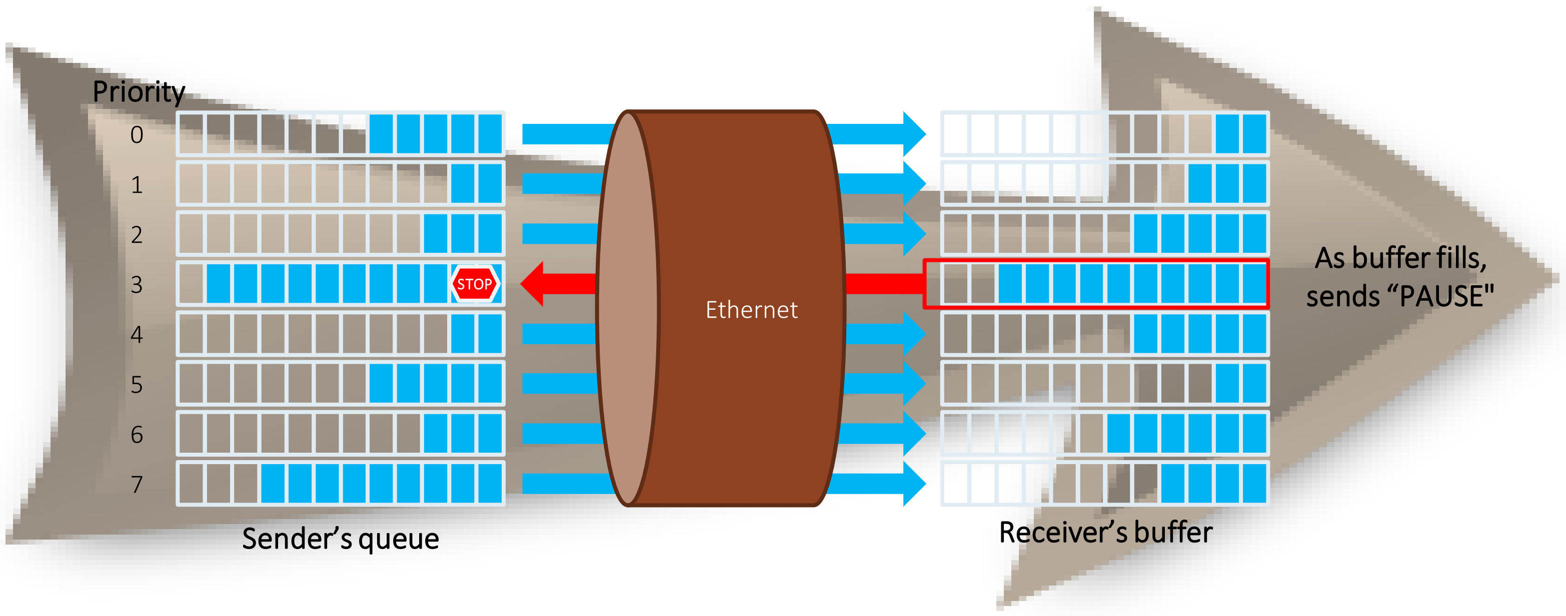
- Historical perspective (partial list)
  - 802.3x – Pause (1997)
  - 802.1Qau – Congestion Notification (2010)
  - 802.1Qaz – Enhanced Transmission Selection (2011)
  - 802.1Qbb – Priority-based Flow Control (2011)
  - RFC 2309 - Recommendations on Queue Management and Congestion Avoidance in the Internet (1998)
  - RFC 3168 - The Addition of Explicit Congestion Notification (ECN) to IP (2001)
  - RFC 5562 - Adding ECN Capability to TCP's SYN/ACK Packets (2009)
  - RFC 7141 - Byte and Packet Congestion Notification (2014)

IEEE Data Center Bridging (DCB) Task Group

- Recent solutions (partial list)
  - RoCEv2 - RDMA over Converged Ethernet v2 (2014)
  - DCQCN - Data Center Quantized Congestion Notification (2015)
  - RFC 8257 - Data Center TCP (DCTCP): TCP Congestion Control for Data Centers (2017)
  - 802.1Qcz – Congestion Isolation (expected in 2021)

NOTE: Many approaches reduce loss, but to eliminate loss, PFC is required…

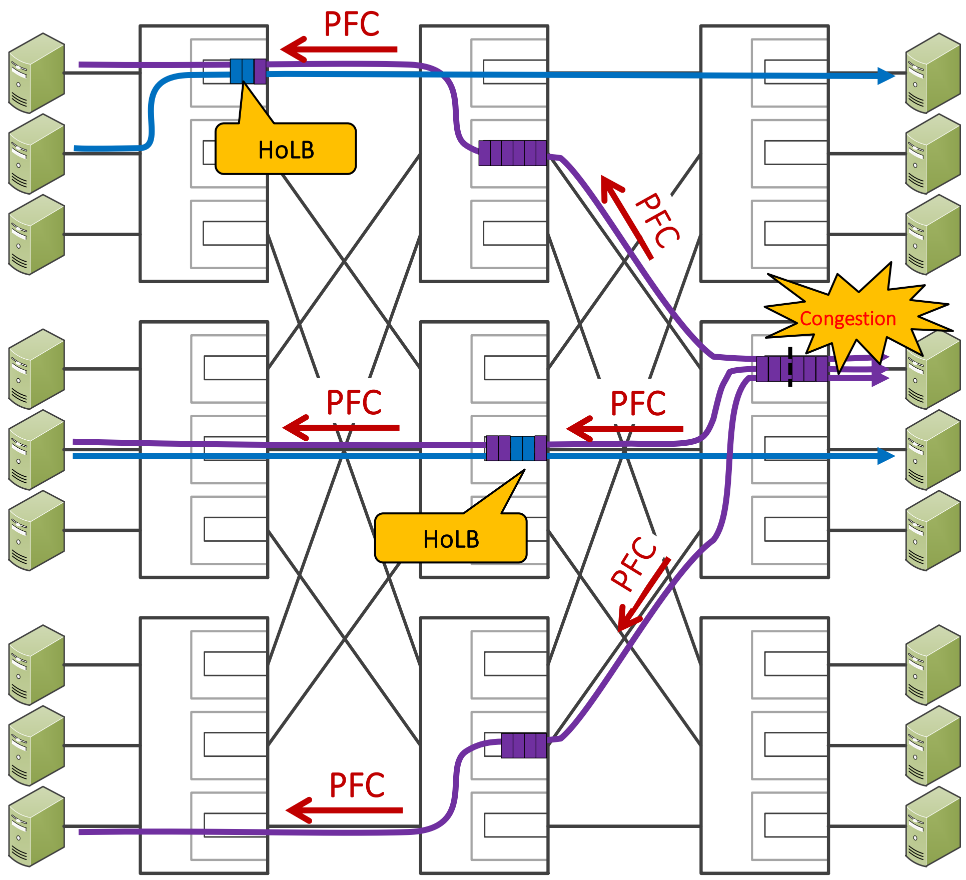
Priority base Flow Control (PFC)



- IEEE 802.1Q Defines 8 Traffic Classes (aka Queues)
- Priority-based Flow Control 'pauses' individual traffic classes, while other classes continue
- Necessary for a 'lossless' environment
- Motivated to allow Ethernet to used in HPC/RDMA networks

The dark side of PFC

802.1Qbb - Priority-based Flow Control



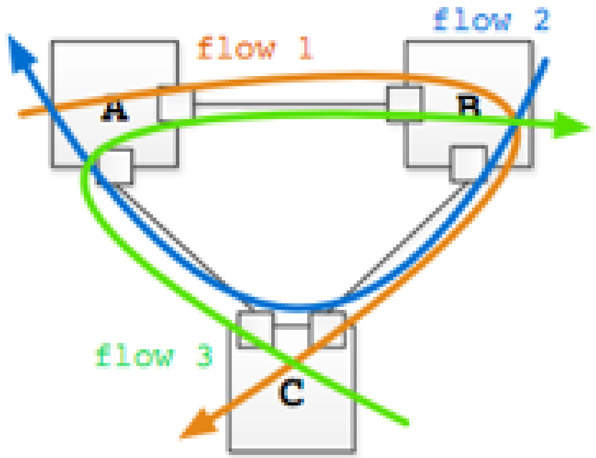
Concerns with over-use

- Hard to configure lossless environment
- Head-of-Line blocking (HoLB)
- Congestion spreading
- Buffer Bloat, increasing latency
- Increased jitter reducing throughput
- **Deadlocks!**

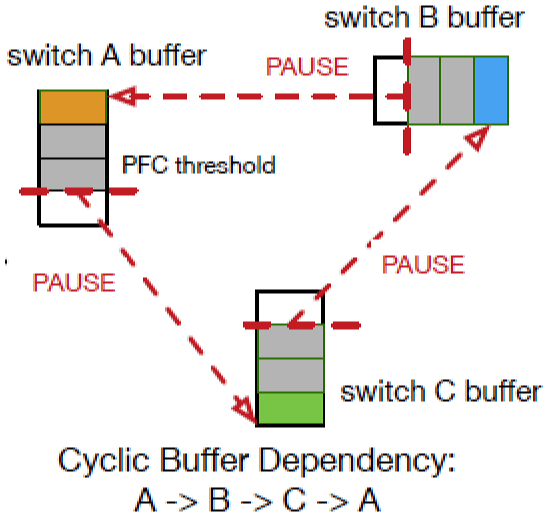
How do PFC deadlocks form?

- Cyclic Buffer Dependency (CBD) is a necessary condition for deadlock formation
- Flow loop is a necessary condition for CBD

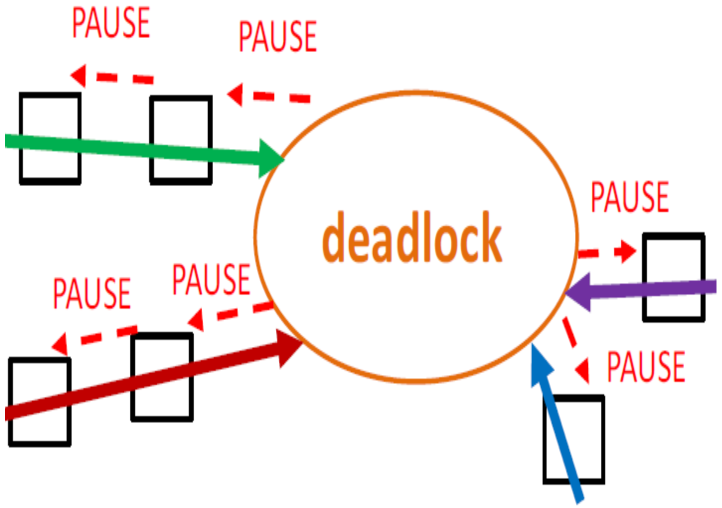
Flows loops create buffer dependencies



Cyclic buffer dependency

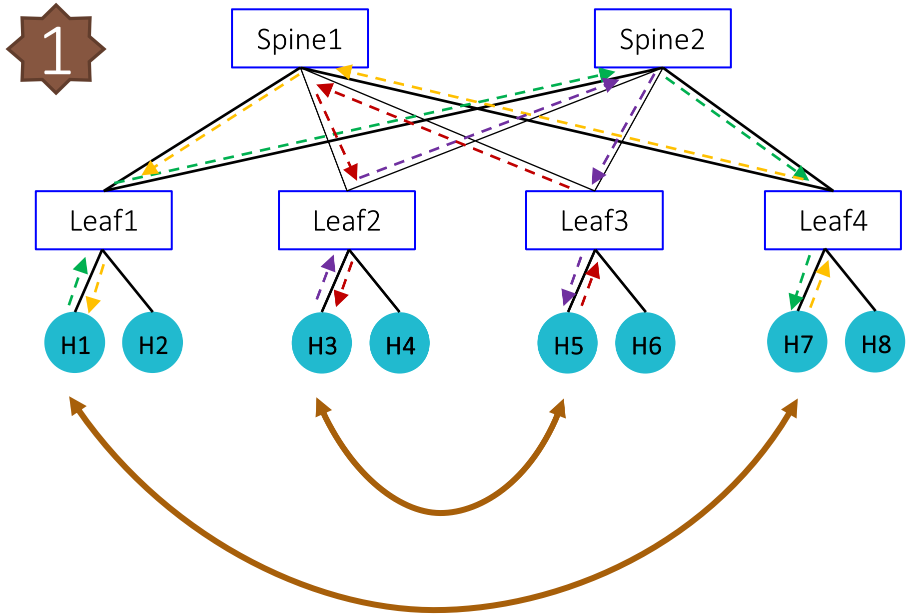


PFC Deadlock



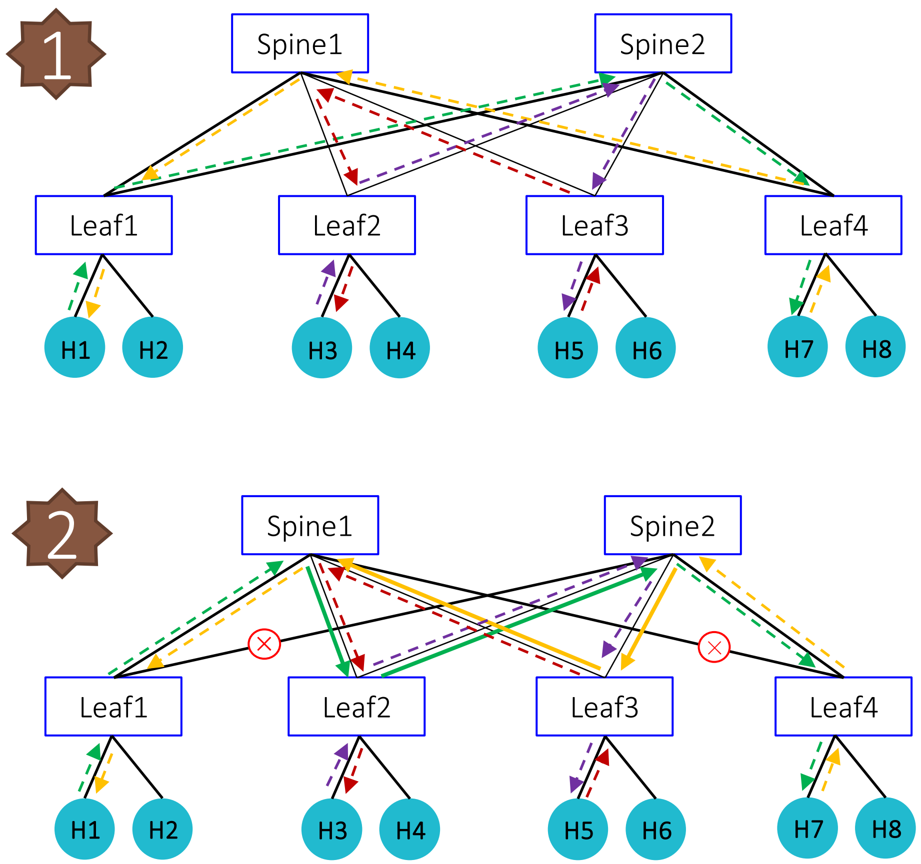
Hu, Shuihai, et al. "Tagger: Practical PFC Deadlock Prevention in Data Center Networks." *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*. ACM, 2017.

Example of PFC Deadlock



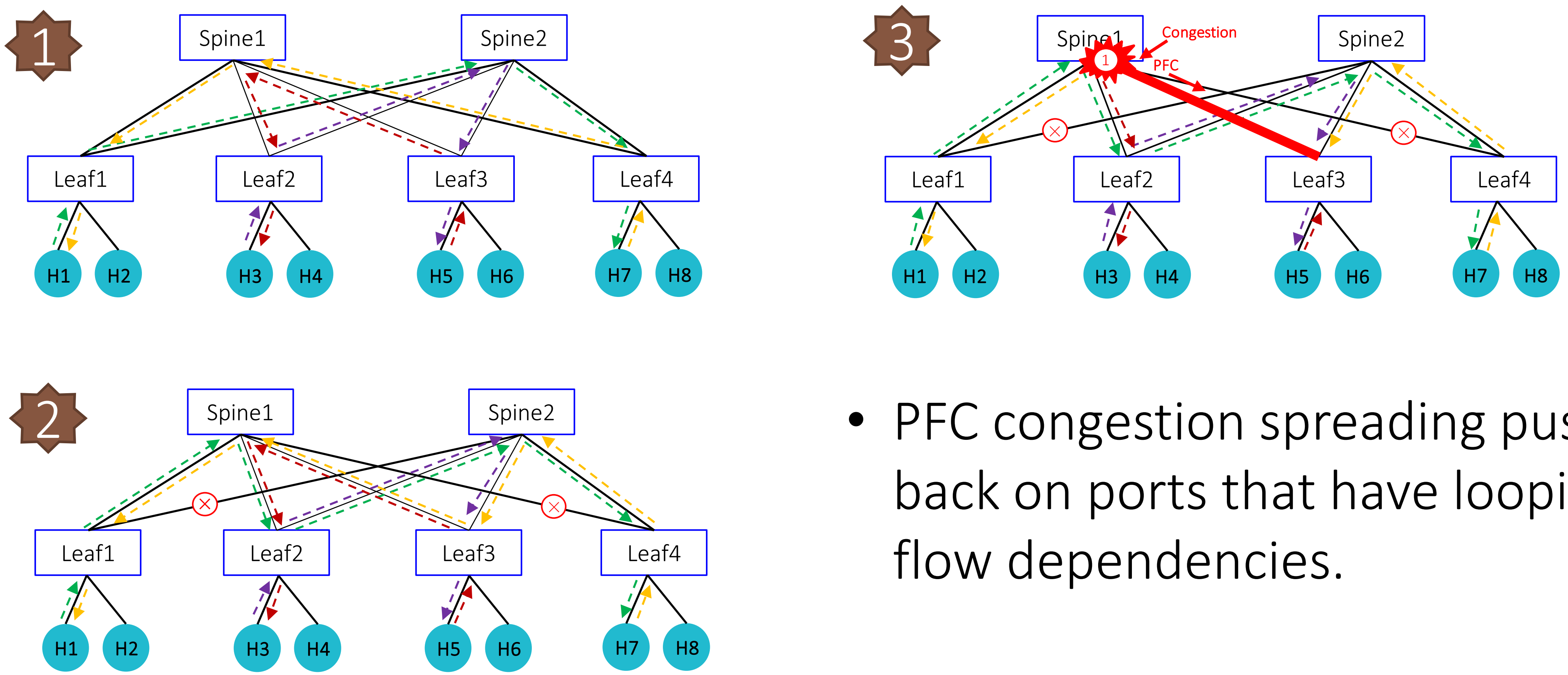
- ECMP load balanced flows across the Clos network
- Flows traverse ‘up’ from leaves to spines and ‘down’ from spines to leaves

Example of PFC Deadlock



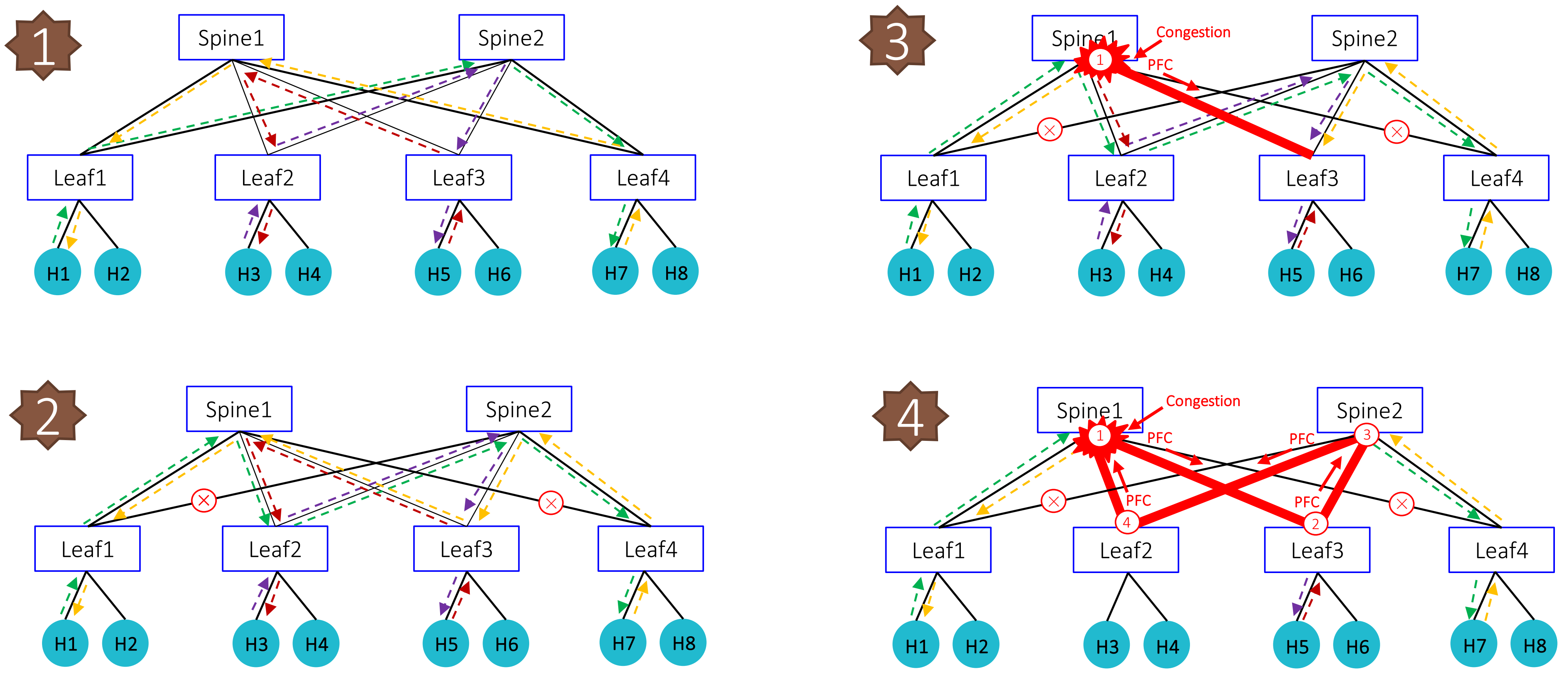
- Link or node failures cause ECMP traffic to be re-distributed, increasing the probability of congestion points
- Flows at leaves may now traverse from ‘uplink’ to ‘uplink’

Example of PFC Deadlock



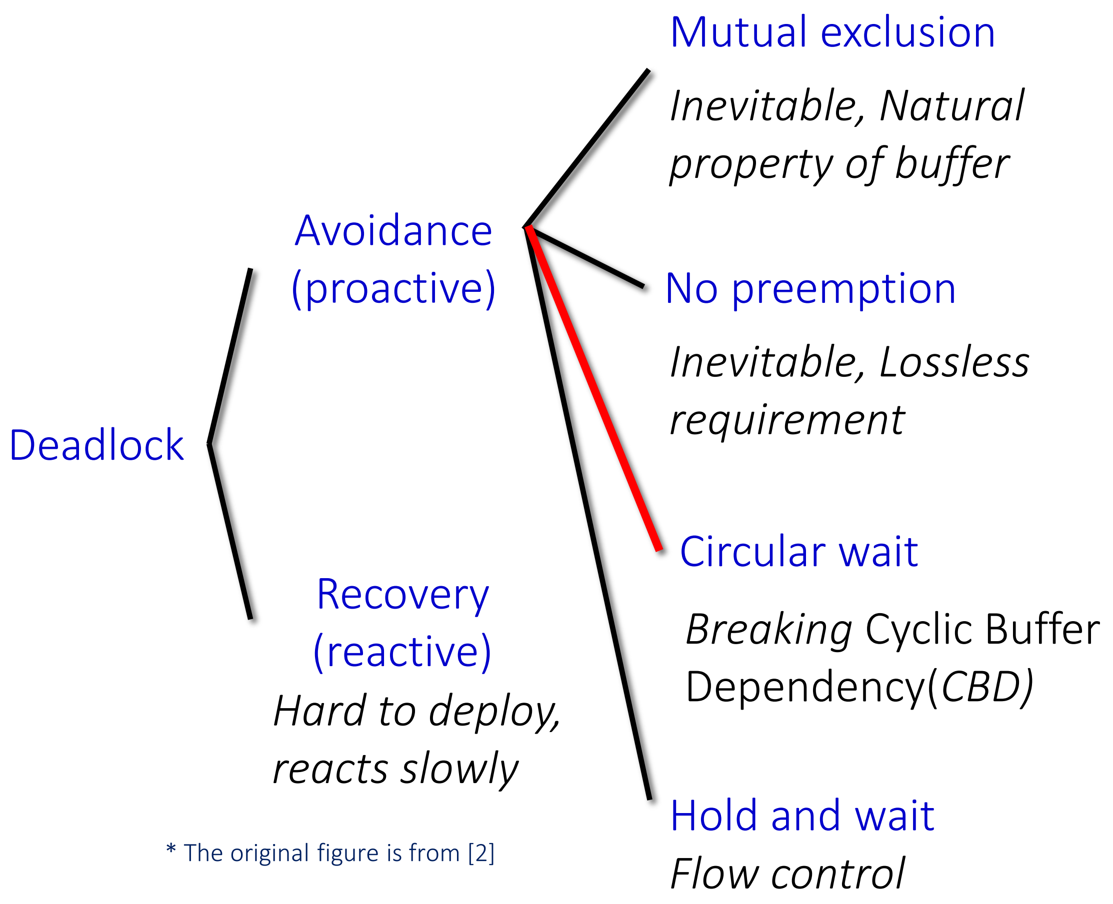
- PFC congestion spreading pushes back on ports that have looping flow dependencies.

Example of PFC Deadlock



Avoiding Deadlocks

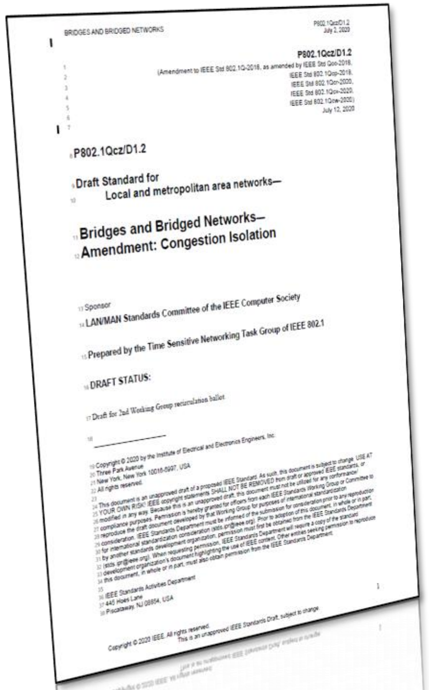
- There are four necessary conditions for deadlock occurrence[1]. To prevent deadlocks, we must ensure that at least one of these conditions never holds [2].
- Years of research and many approaches, often related to deadlock free routing.
- The Ethernet legacy is that simple and scalable solutions prevail.



* The original figure is from [2]

[1] Abraham Silberschatz, Peter Baer Galvin, and Greg Gagne. 2014. Operating system concepts essentials. John Wiley & Sons, Inc.
[2] Qian, Kun, et al. "Gentle flow control: avoiding deadlock in lossless networks." *Proceedings of the ACM Special Interest Group on Data Communication*. ACM, 2019.

P802.1Qcz – Congestion Isolation


BRIDGES AND BRIDGED NETWORKS

P802.1Qcz/D1.2
July 2, 2020

P802.1Qcz/D1.2
(Amendment to IEEE Std 802.1Q-2018, as amended by IEEE Std Qcc-2018,
IEEE Std 802.1Qcp-2018,
IEEE Std 802.1Qcr-2020,
IEEE Std 802.1Qcx-2020,
IEEE Std 802.1Qcw-2020)
July 12, 2020

P802.1Qcz/D1.2

Draft Standard for
Local and metropolitan area networks—

Bridges and Bridged Networks—
Amendment: Congestion Isolation

Sponsor
LAN/MAN Standards Committee of the IEEE Computer Society

Prepared by the Time Sensitive Networking Task Group of IEEE 802.1

DRAFT STATUS:

Draft for 2nd Working Group recirculation ballot

Copyright © 2020 by the Institute of Electrical and Electronics Engineers, Inc.
Three Park Avenue
New York, New York 10016-5997, USA
All rights reserved.

IEEE Standards Activities Department
445 Hoes Lane
Piscataway, NJ 08854, USA

Copyright © 2020 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

Project Background

- Initiated in November 2017
- Amendment to IEEE 802.1Q-2018 to Support the Isolation of Congested data flows within ***Data Center Environments***, such as high-performance computing, AI/RDMA fabrics, and distributed storage networks.
- Motivation discussed in Nendica report of “802 Network Enhancements For the Next Decade”

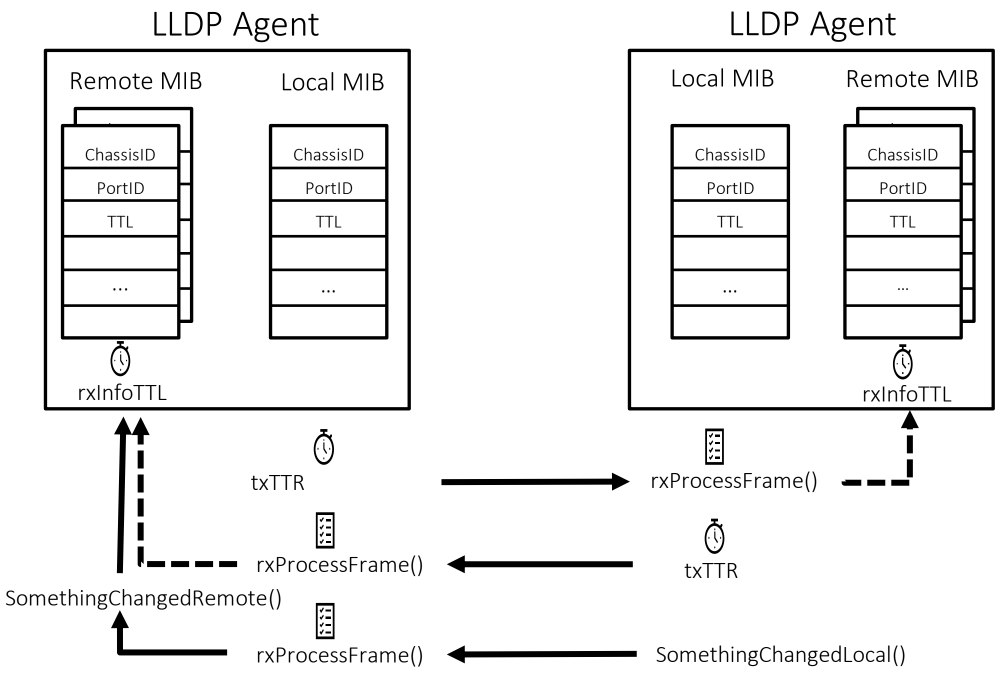
- Two key technologies:
    - Congestion Isolation
    - Topology Recognition (via LLDP)

Project Status

July 2020 – Completing Working Group Ballots
Early 2021 - Anticipated publication

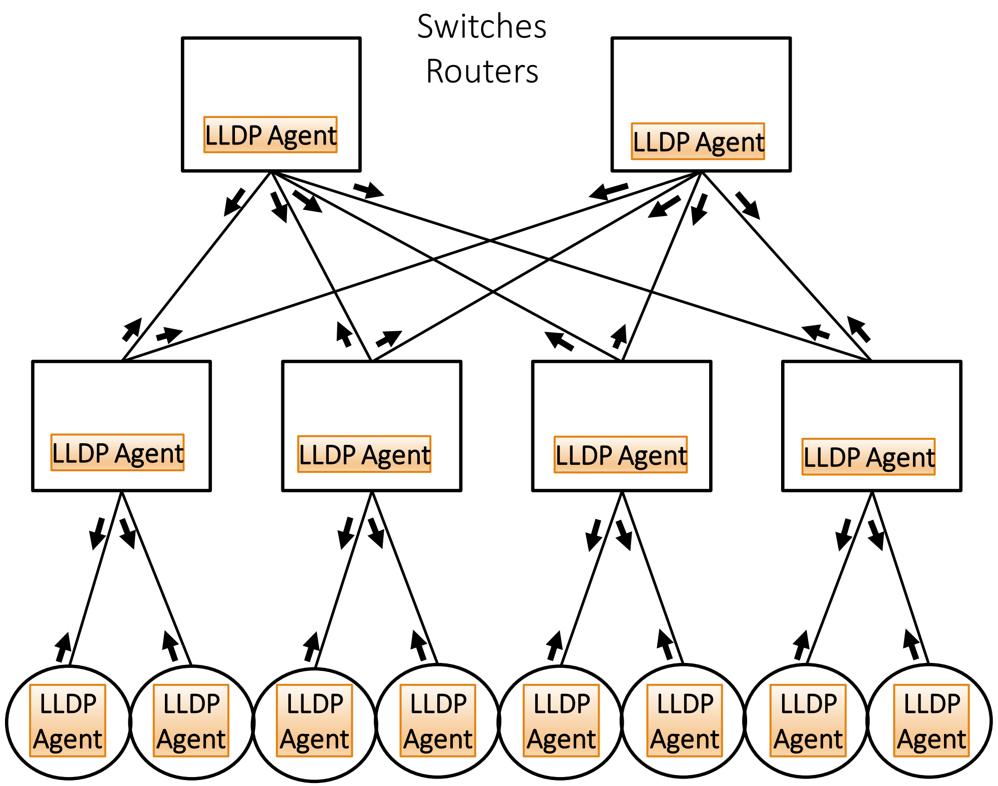
What is LLDP and how does it work?

LLDP Agent Communication



Information is packed into Type-Length-Value (TLV) objects

Network Wide Discovery



Topology Recognition via LLDP

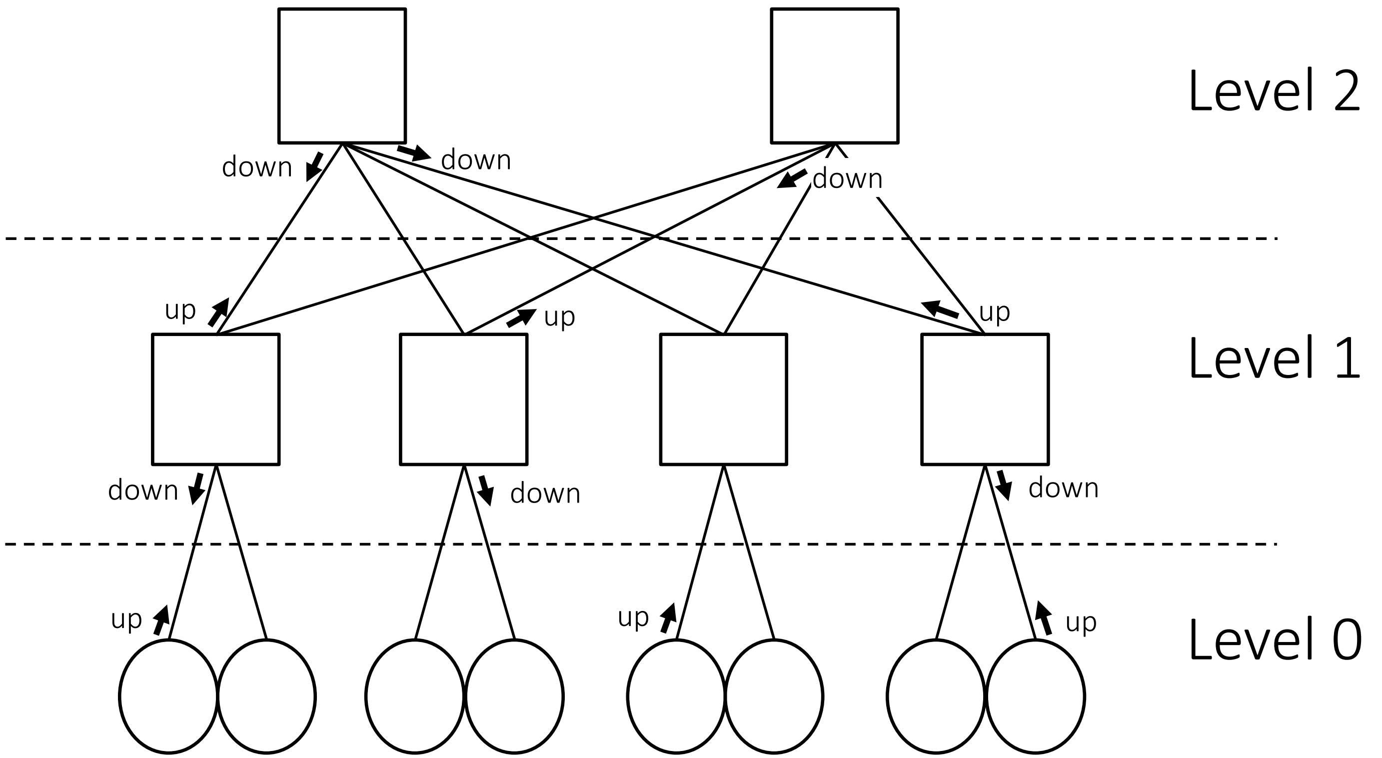
Through the exchange of LLDP TLVs automatically determine:

1. Topology level of devices in network
    - 0 = End-station or server edge
    - 1 = Leaf
    - n+1 = Spine
2. Port orientation for each link
    - Uplink
    - Downlink
    - Crosslink

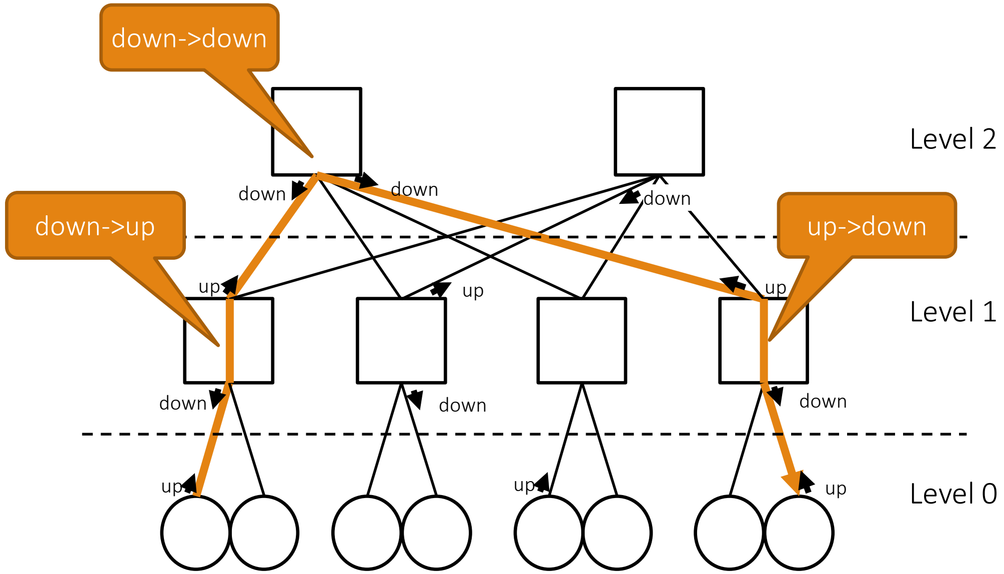
HINT: Servers are always at level 0 with uplinks.

Useful for:

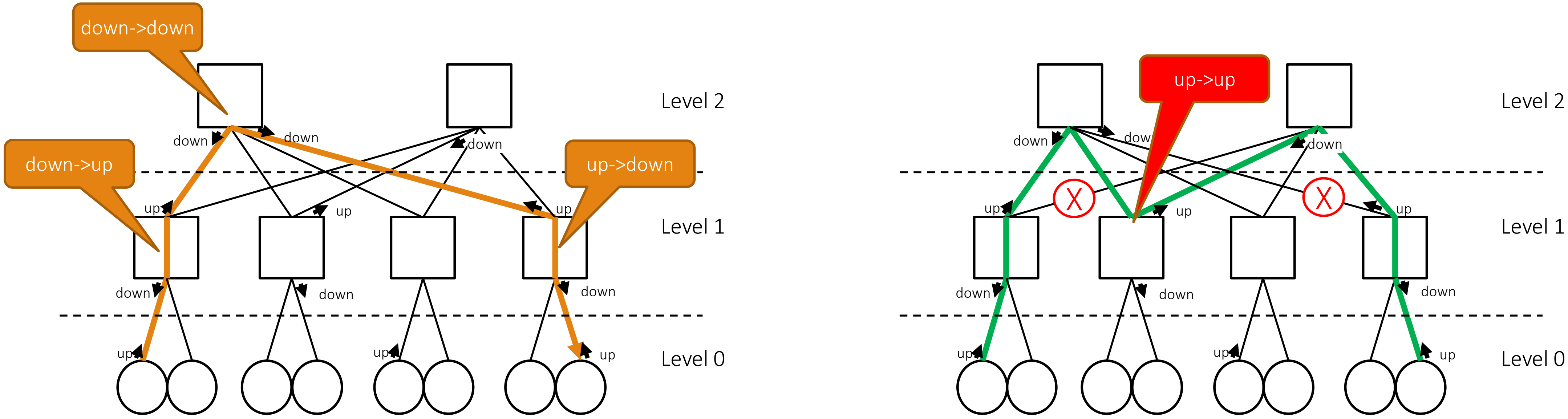
- PFC deadlock prevention
- Resetting a changed DSCP or PCP
- Detecting incast vs in-network congestion



Topology Aware Forwarding Perspective

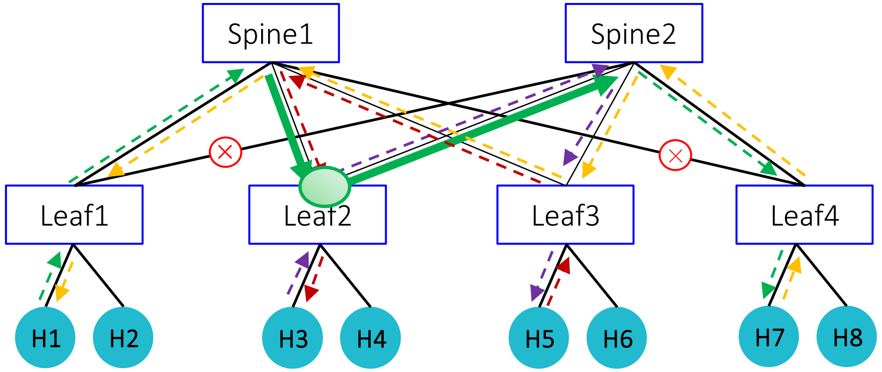


Topology Aware Forwarding Perspective

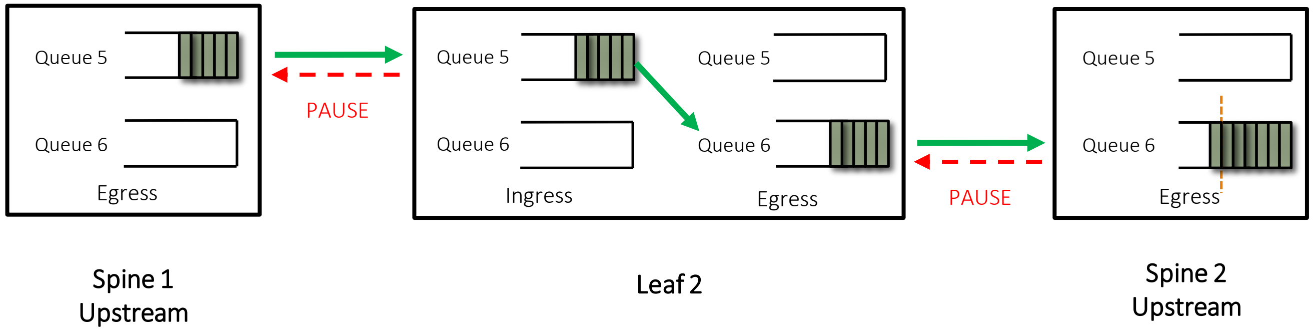


Deadlock free mechanism (Proactive)

- Identify a CBD breaking point and prevent PFC deadlock
- Consideration:
  - Although the traffic in a CLOS network has no loops, topology changes due to failure may cause rerouting which may form a CBD.
  - Determine if rerouted traffic creates a CBD by knowing topology level and port orientation.
  - Eliminate CBD by deploying independent resources for dependent flows (i.e. use a different priority queue).



- Recognize down-up reroute.
- Identify the CBD breaking point



- Example Queues 5&6 are lossless queues (Enable PFC)
- Leaf 2 judges the flow and enqueue to Queue6, modify the DSCP
- If PFC is triggered, it will be on separate queues.

Summary

- The lossless data center in the era of AI needs to scale to meet future demands
- Priority-based flow control is necessary for a lossless network, but creates issues such as **Deadlock**
- New standards are underway to enable simple and scalable solutions to PFC Deadlock
- All of this is part of the IEEE 802 Network Enhancements for the Next Decade Industry Connections Activity (NENDICA)
- Participation in NENDICA is free, open and welcomed to all.
- https://1.ieee802.org/802-nendica

Thank You!

PAUL.CONGDON@TALLAC.COM