# Track –
# Next Generation Data Center Fabrics:
# Do New Technologies meet Operator Needs?

NANOG 76
10 June 2019

# Agenda

- *Advancing Ethernet for Data Center Networks*
  - Roger Marks, EthAirNet Associates
- *Towards Hyperscale High Performance Computing with RDMA*
  - Omar Cardona, Microsoft
- *The IEEE P802.1Qcz Project on Congestion Isolation*
  - Paul Congdon, Tallac Networks
- *Discussion: Identifying Problems and Inefficiencies in Current Data Center Operations*

# Advancing Ethernet for Data Center Networks

Roger Marks

roger@ethair.net
+1 802 capable

Prepared: 3 June 2019

# Disclaimer

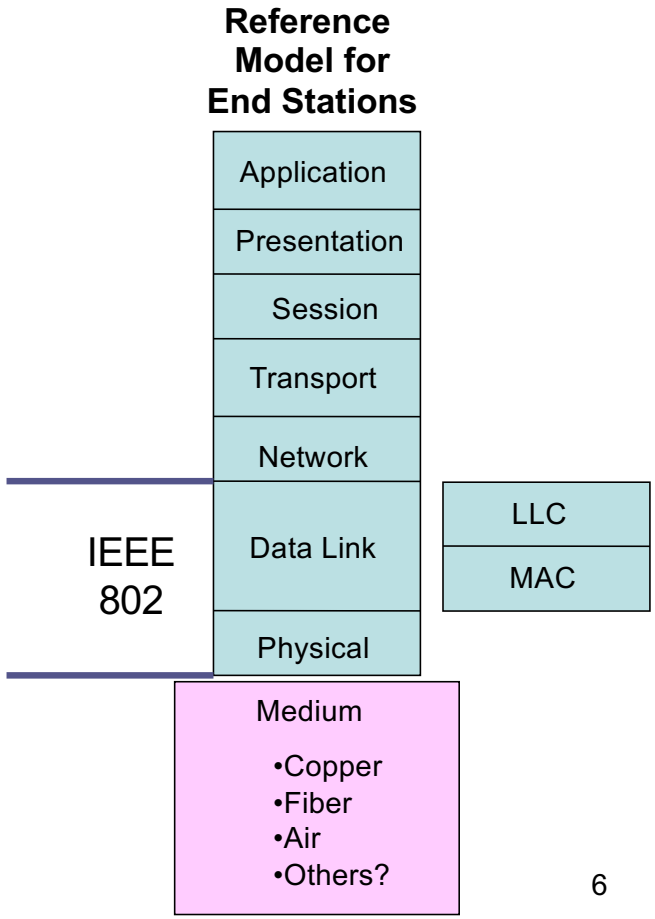
- All speakers presenting information on IEEE standards speak as individuals, and their views should be considered the personal views of that individual rather than the formal position, explanation, or interpretation of the IEEE.

# IEEE 802: The LAN/MAN Standard Committee (LMSC)

- IEEE 802:
  - LAN/MAN Standards Committee (LMSC)
- Develops standards, for
  - Local Area Networks (LAN),
  - Metropolitan Area Networks (MAN),
  - Regional Area Networks (RAN),
  - Personal Area Networks (PAN),
  - Wireless Specialty Networks, etc.
- In operation since March 1980

# Lower-Layer Focus

- IEEE 802 standards emphasize the functionality of the lowest two layers of the OSI reference model, and the higher layers as they relate to network management
  - physical layer (PHY, Layer 1)
  - data link layer (DLL, Layer 2)
- IEEE 802 divides DLL into:
  - Medium Access Control (MAC)
    - Multiple specifications
  - Common logical link control (LLC)
- See details in IEEE Std 802
  - "IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture"






6

# 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
- Chair: Roger Marks
- Open to all participants; no membership

# Nendica Motivation and Goals

- "The goal of this activity is to assess… emerging requirements for IEEE 802 wireless and higher-layer communication infrastructures, identify commonalities, gaps, and trends not currently addressed by IEEE 802 standards and projects, and facilitate building industry consensus towards proposals to initiate new standards development efforts."

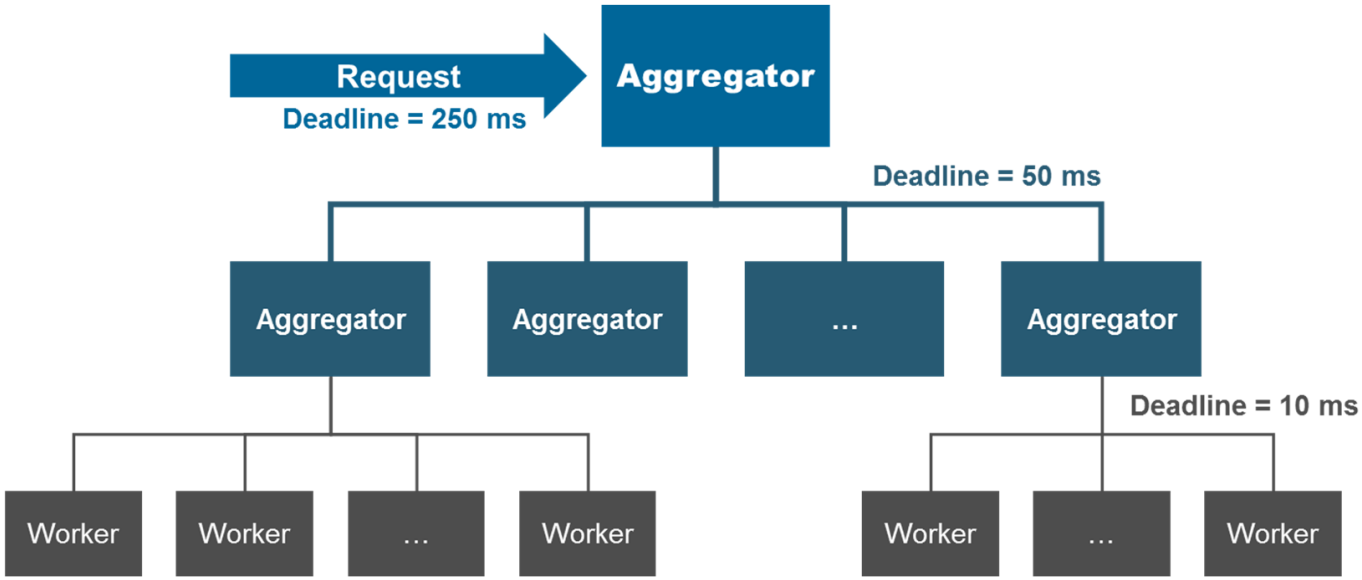
# Nendica Report: *The Lossless Network for Data Centers*

- 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 may be revised.

# Use Cases: *The Lossless Network for Data Centers*

- The scale of Data Center Networks continues to grow
  - Online Data Intensive (OLDI) Services
  - AI Deep Learning and Model Training
  - Cloud High-Performance Computing
  - Financial Trading
  - Distributed Storage
    - Non-Volatile Memory Express (NVMe) over Fabrics
- Data Center => High Performance Computer
- New requirements for lossless Ethernet fabric

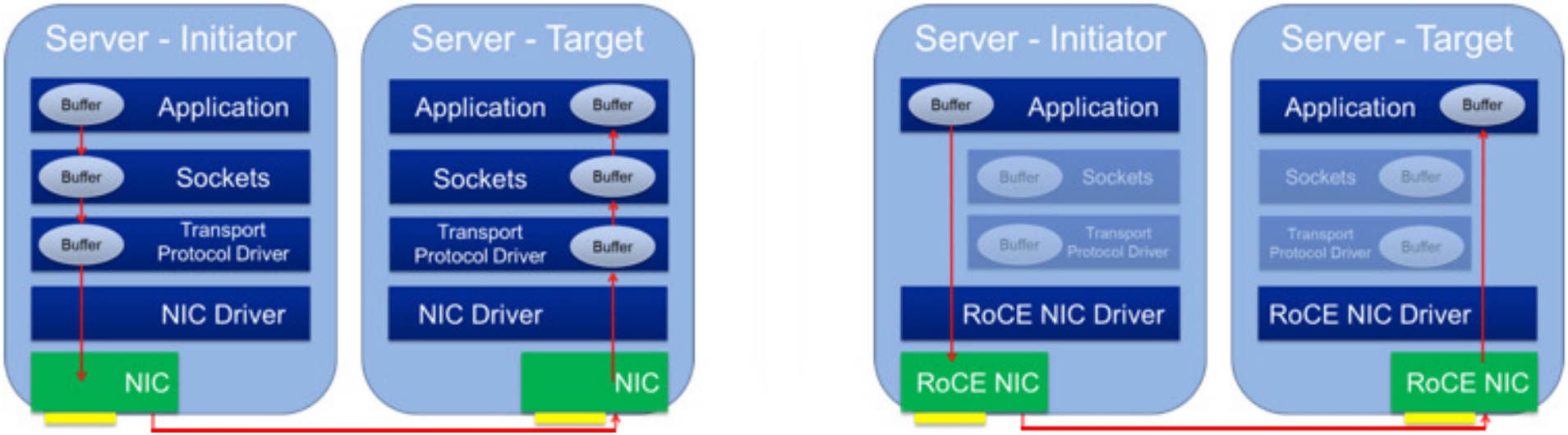
# Data Center Applications are distributed and latency-sensitive




- Tend toward congestion; e.g. due to incast
- Packet loss leads to retransmission, more congestion, more delay

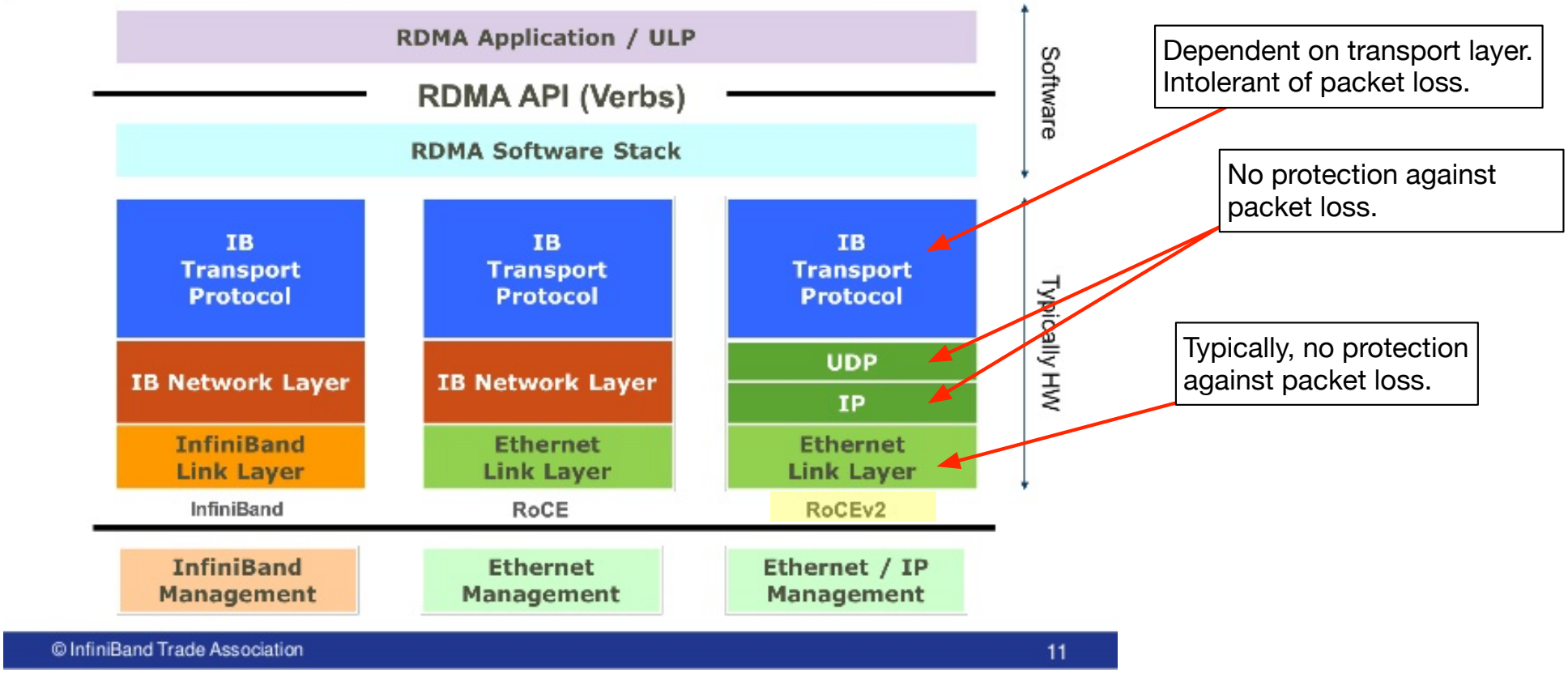
Source: IEEE 802 Nendica Report: The Lossless Network for Data Centers [2]

# Remote Direct Memory Access (RDMA) in the Data Center




Source: InfiniBand Trade Association <http://www.roceinitiative.org/roce-introduction/>

# RoCE: RDMA over Converged Ethernet



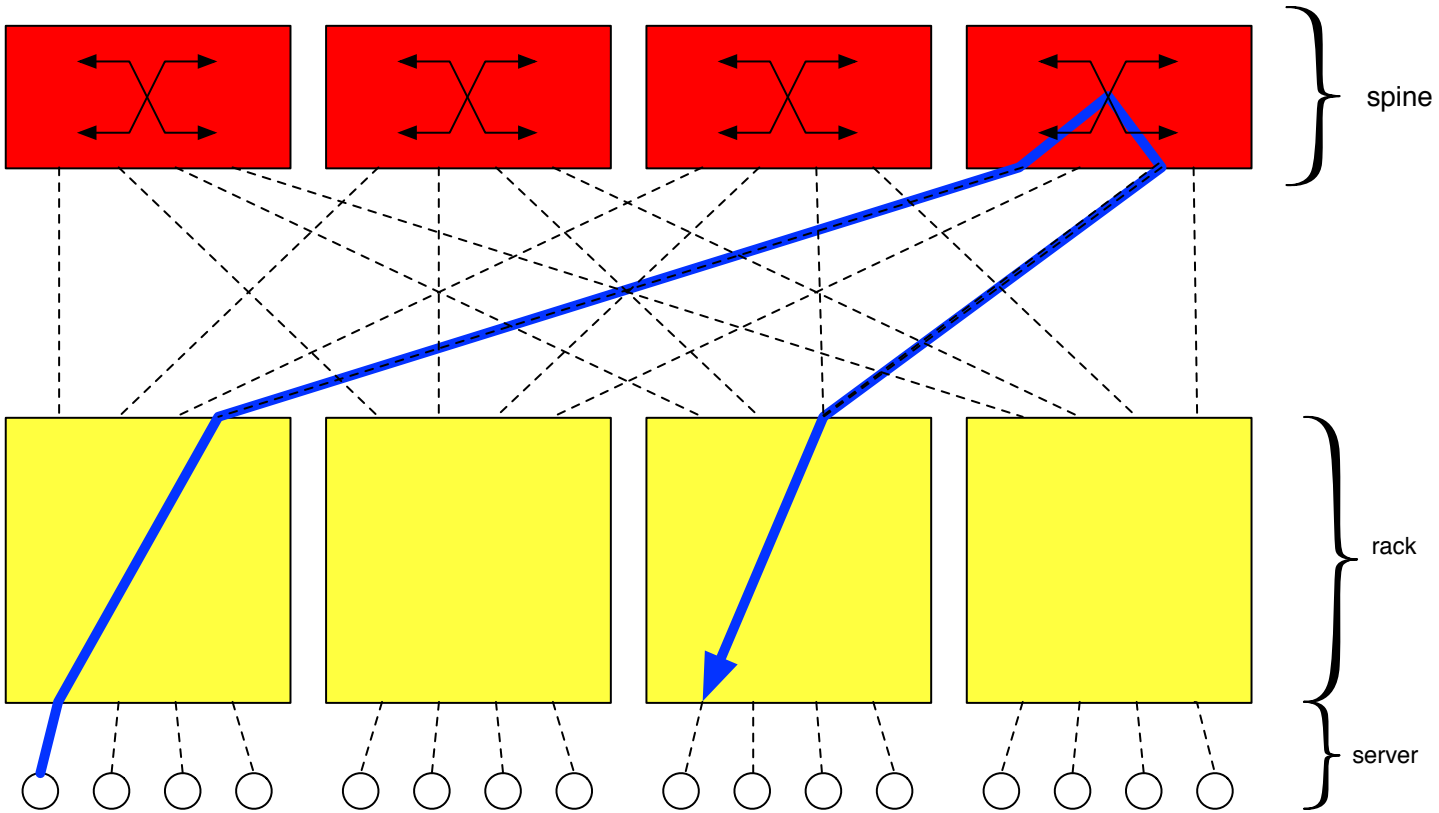
# Lossless Ethernet is a Foundation of the New Data Center

- Data Centers are becoming High Performance Computers
- Most Data Center networking is at Layer 3/4, but it all rides on Layer 1/2
- Ethernet is preferred option for Layer 1/2
  - Ethernet infrastructure can support multiple upper layers (e.g. TCP and RoCEv2) simultaneously
- Most Data Center networking requires low latency
- Some Data Center networking requires a lossless Ethernet fabric

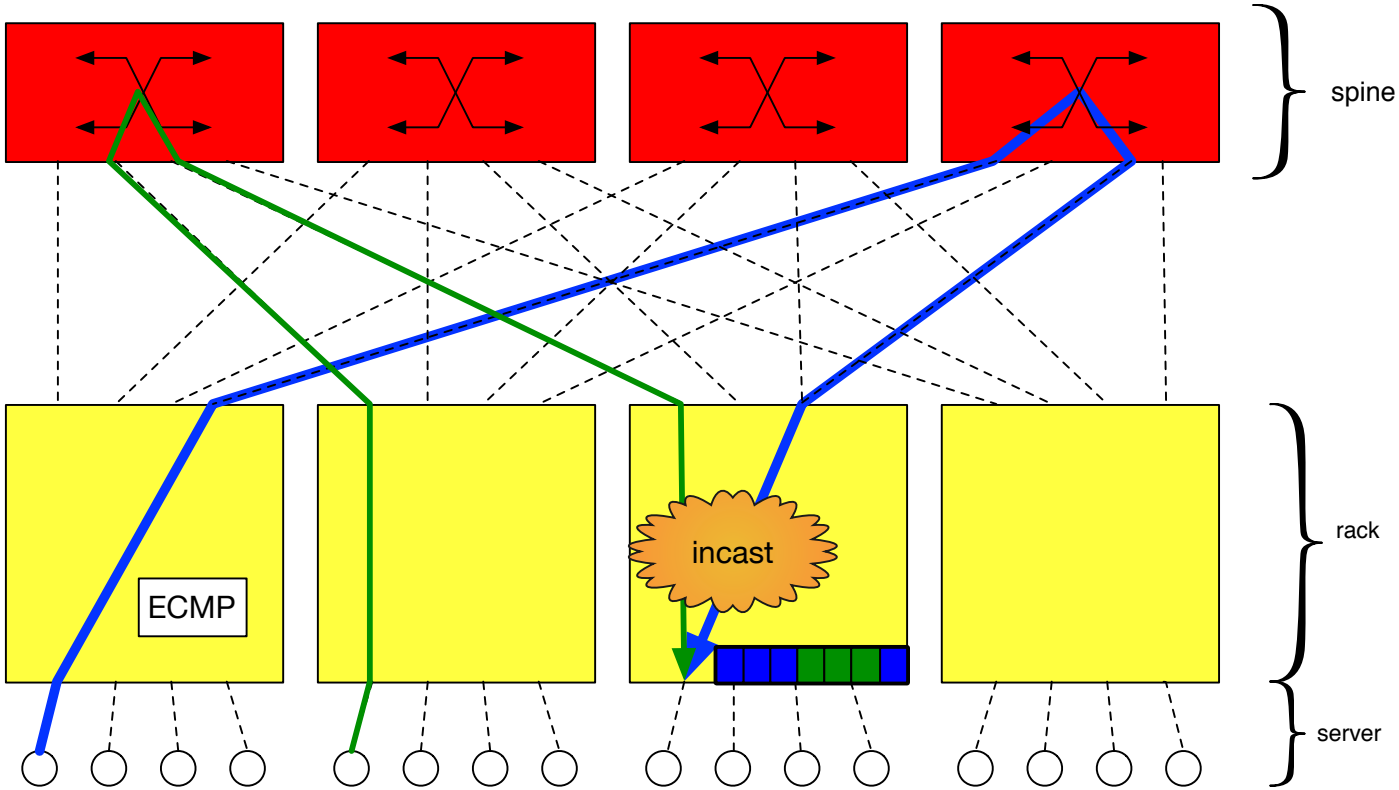
# IEEE 802 Developed Data Center Bridging

- In the IEEE 802.1 Working Group, the Data Center Bridging Task Group developed many standards to enhance Layer 2 support for the data center
  - Beginning around 2006
- Key technologies include:
  - Priority-based Flow Control (PFC)
  - Congestion Notification
  - Enhanced Transmission Selection (ETS)

# Folded-Clos Network:
# Many Paths from Server to Server

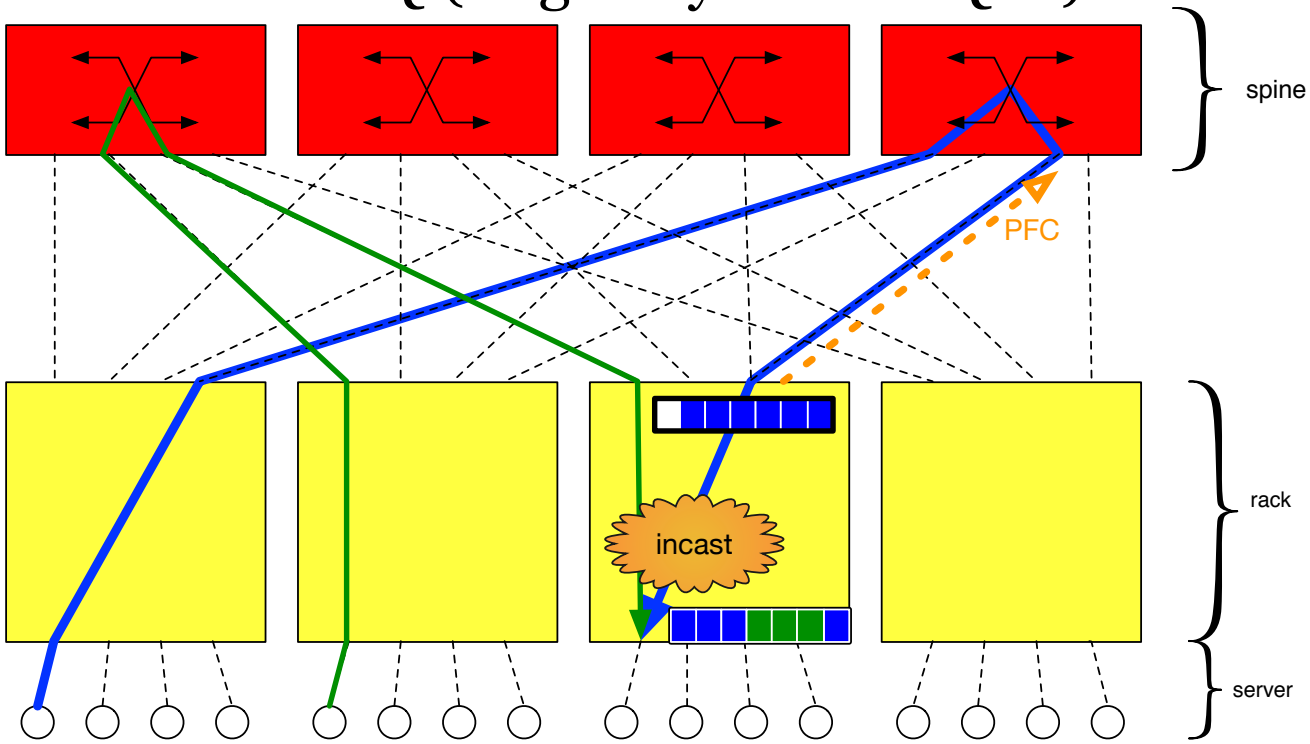


# Incast fills output queue (note: ECMP cannot help)

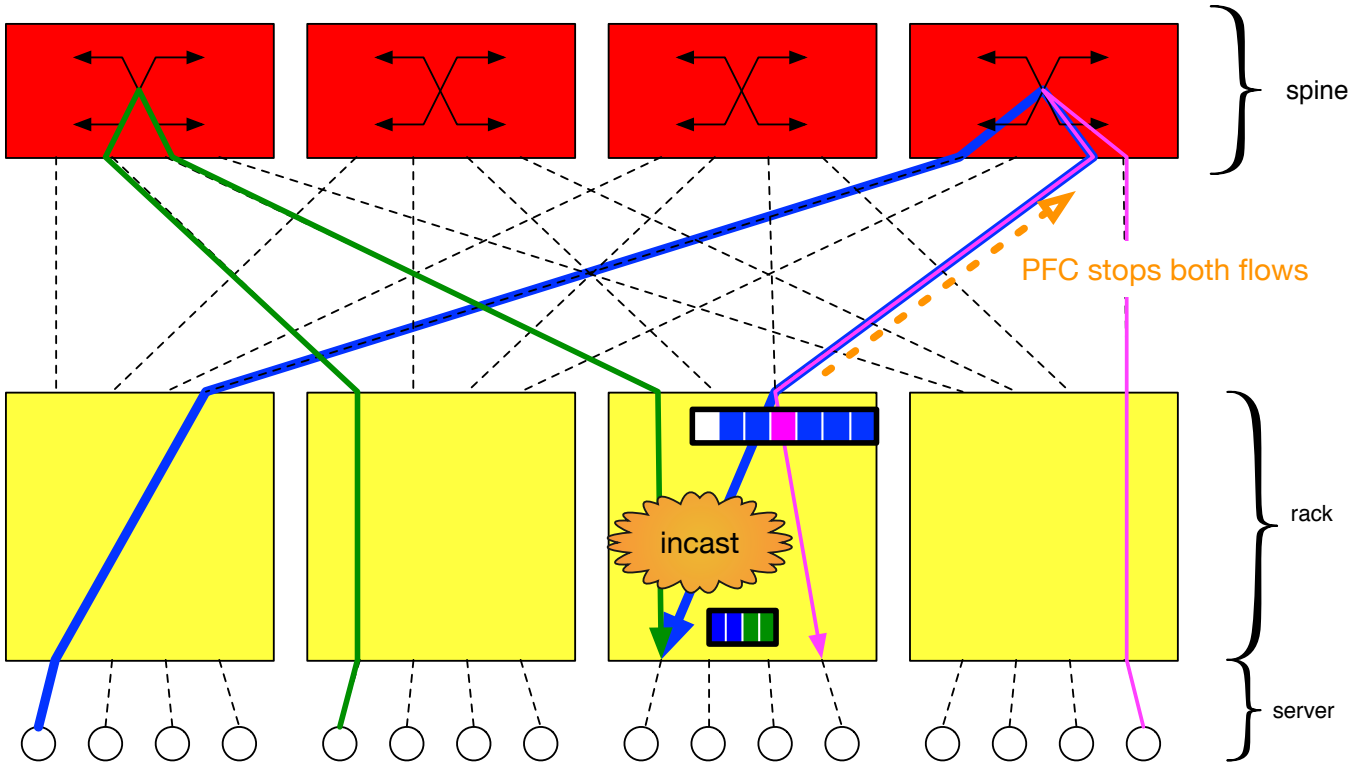


# Priority flow control (PFC)

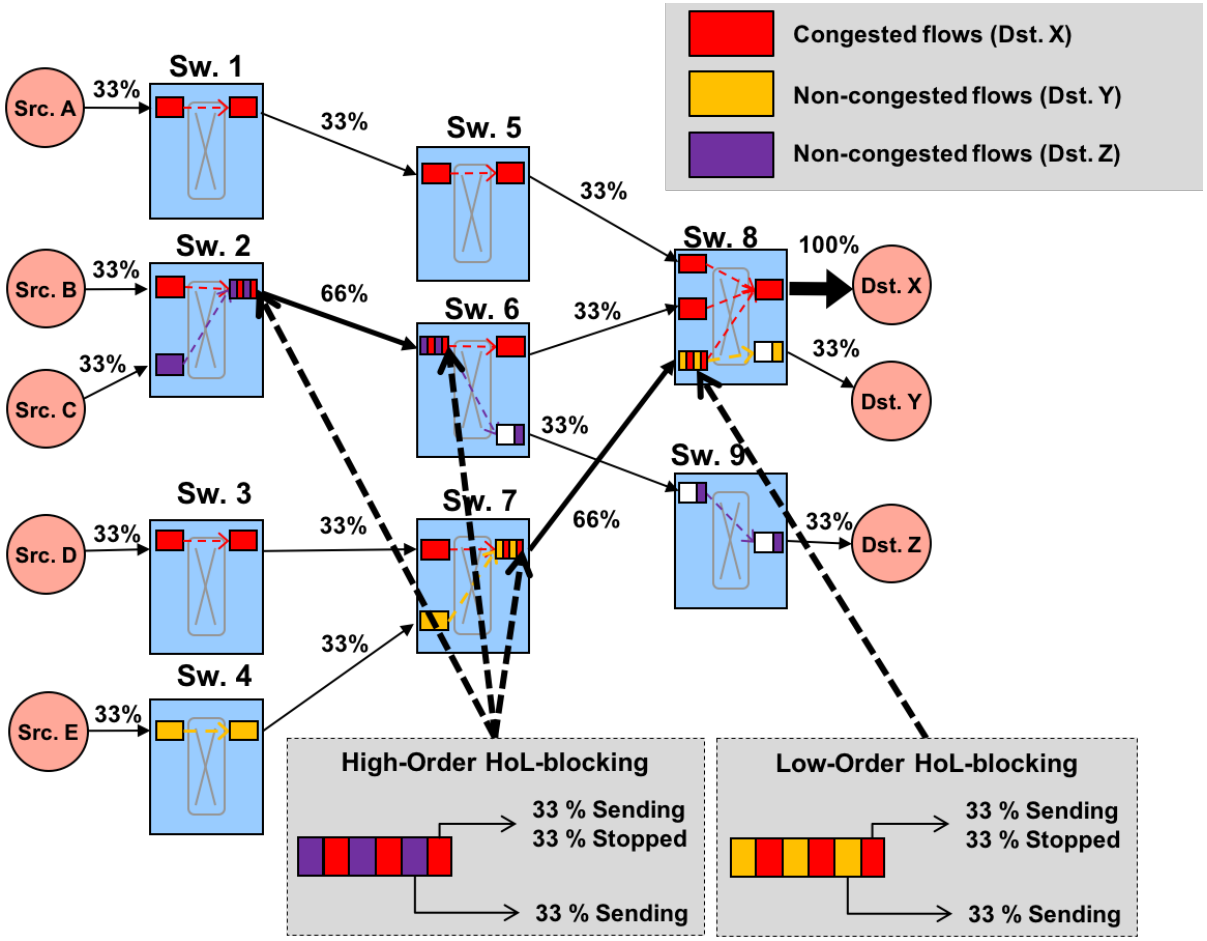
- Output backup fills ingress queue
- PFC can be used to pause input per QoS class
- IEEE 802.1Q (originally in 802.1Qbb)



# PFC pauses all flows of the class including “victim” flows

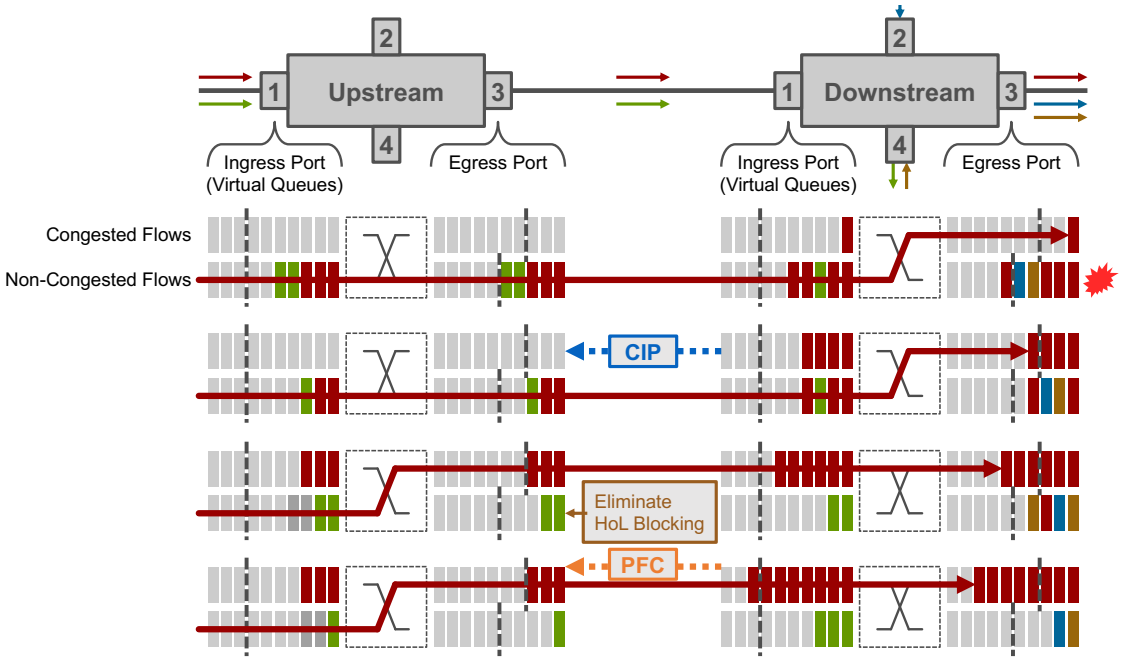


# Traffic-class blocking analysis




Source: IEEE 802.1-19-0012-00 [4]

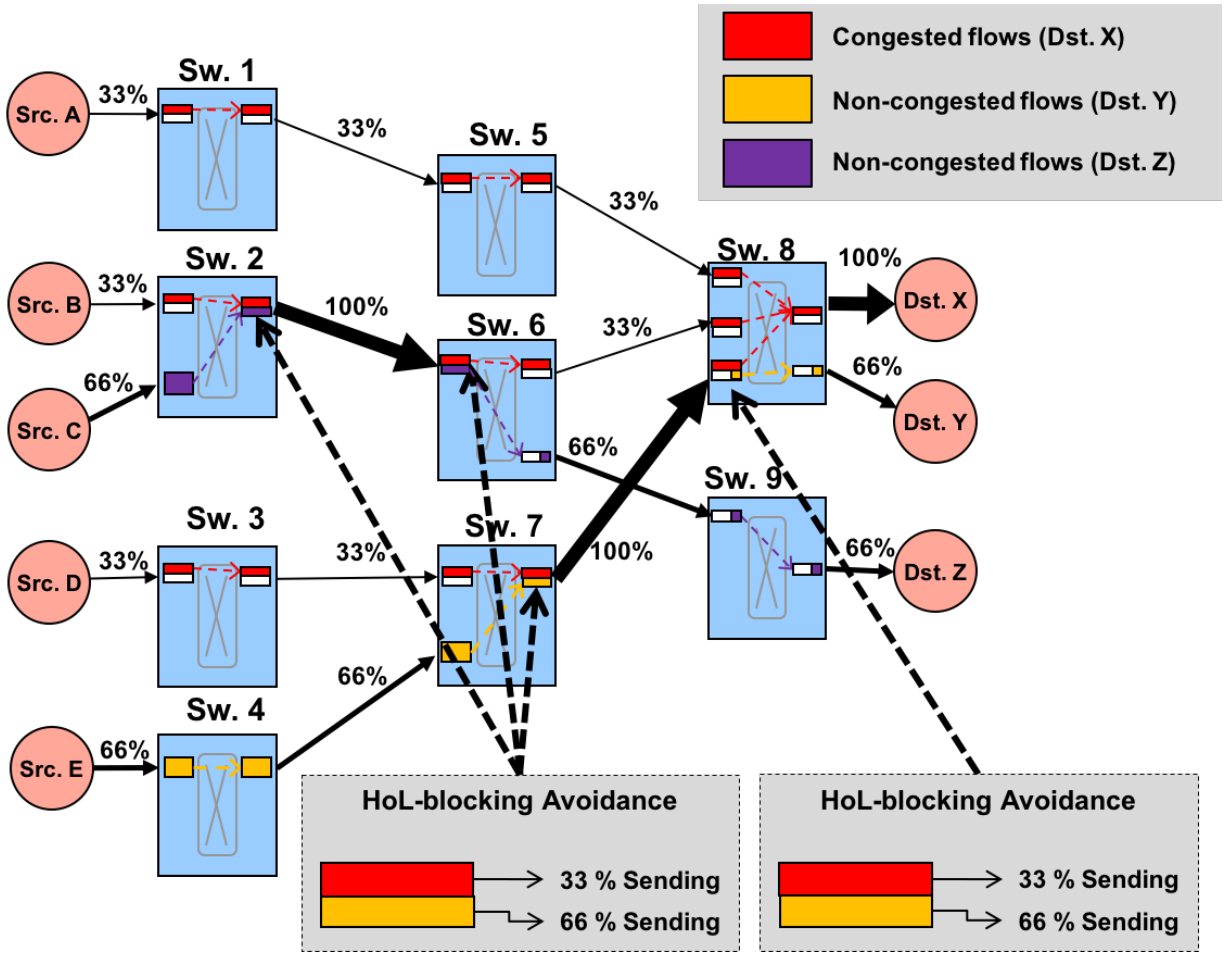
# Congested-flow Isolation (see IEEE Project P802.1Qcz [5])




1. Identify the flow causing congestion and isolate locally
2. Signal to neighbor when congested queue fills
3. Upstream isolates the flow too, eliminating head-of-line blocking
4. If congested queue continues to fill, invoke PFC for lossless

Source: IEEE 802 Nendica Report: The Lossless Network for Data Centers [2]

# Congested-flow Isolation Analysis




Source: IEEE 802.1-19-0012-00 [4]

# Bibliography


1) IEEE 802 Orientation for New Participants
   - https://mentor.ieee.org/802-ec/dcn/18/ec-18-0117-02.pdf
2) IEEE 802 "Network Enhancements for the Next Decade" Industry Connections Activity (Nendica)
   - https://1.ieee802.org/802-nendica
3) IEEE 802 Nendica Report: "The Lossless Network for Data Centers" (18 August 2018)
   - https://mentor.ieee.org/802.1/dcn/18/1-18-0042-00.pdf
4) Paul Congdon, "The Lossless Network in the Data Center," IEEE 802.1-17-0007-01, 7 November 2017
   - https://mentor.ieee.org/802.1/dcn/17/1-17-0007-01.pdf
5) Pedro Javier Garcia, Jesus Escudero-Sahuquillo, Francisco J. Quiles, and Jose Duato, "Congestion Management for Ethernet-based Lossless DataCenter Networks," IEEE 802.1-19-0012-00, 4 February 2012
   - https://mentor.ieee.org/802.1/dcn/19/1-19-0012-00.pdf
6) IEEE P802.1Qcz Project: "Congestion Isolation"
   - https://1.ieee802.org/tsn/802-1qcz

# Going forward

- IEEE 802 Nendica Report: “The Lossless Network for Data Centers” (18 August 2018) is published but open to further comment.
  - Comments are welcome from all
- Could open an activity to revise the report, addressing new issues.
  - Proposal [5] may be discussed in future teleconference.
- Report could help identify issues in need of further research or unified action.
- Nendica could instigate standardization of key topics
  - Mainly in IEEE 802; perhaps also in e.g. IETF

# Nendica Participation

- Nendica is open to all participants: please join in!
  - no membership requirements
  - Comment by Internet or contribute documents
  - Call in to teleconferences
  - Attend meetings
- Nendica web site
  - https://1.ieee802.org/802-nendica
  - Details of in-person and teleconference meetings
- Feel free to contact the Chair (see first slide)