

Overview:

# IEEE 802 Nendica Report on The Lossless Network for Data Centers

Roger Marks (Huawei)

roger@ethair.net  
+1 802 capable

10 November 2018

# 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.

# Nendica

- Nendica: IEEE 802 “Network Enhancements for the Next Decade” Industry Connections Activity
  - An IEEE Industry Connections Activity
- Organized under the IEEE 802.1 Working Group
- Chartered March 2017 - March 2019
  - may be extended
- Chair (until March 2018): Glenn Parsons
- Chair (from March 2018): Roger Marks

# IEEE Industry Connections Activity

- Under IEEE-SA, but not standardization.
- “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, but they do not themselves develop IEEE Standards.”
- IEEE 802.3 manages another Industry Connections Activity (“New Ethernet Applications”).

# 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.
- Encouraged topics include enhancements of IEEE 802 communication networks and vertical networks as well as enhanced cooperative functionality among existing IEEE standards in support of network integration.
- Findings related to existing IEEE 802 standards and projects are forwarded to the responsible working groups for further considerations.”

# Nendica Work Items

- The Lossless Network for Data Centers
  - published Nendica Report, 2018-08-17
    - IEEE 802.1-18-0042-00
    - [Circulated to IETF New Work during development]
  - Published report invites further comments
  - Stimulated new standardization project IEEE P802.1Qcz (Congestion Isolation)
- Flexible Factory IOT
  - Draft report 802.1-18-0025-06
  - Significant focus on wireless
  - Comment resolution underway

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

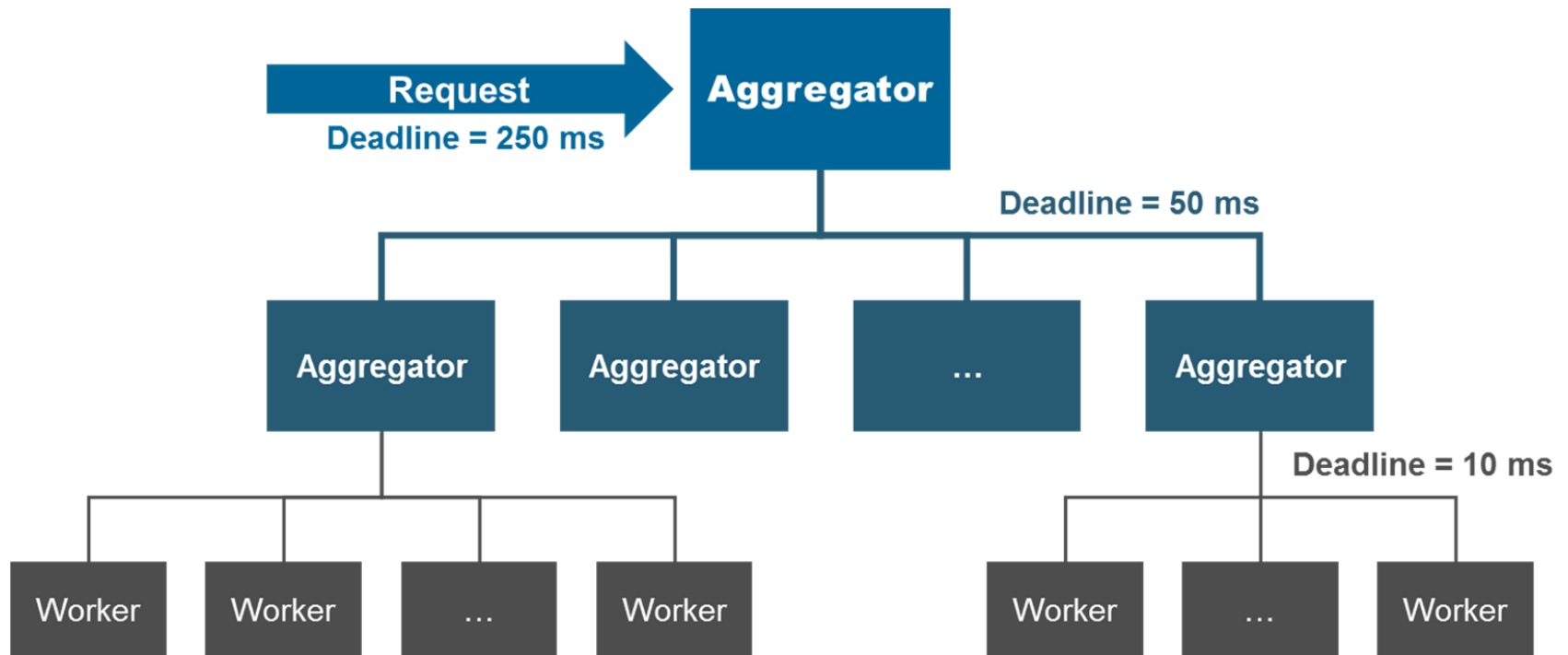
- Paul Congdon, Editor
- Key messages regarding the data center :
  - Packet loss leads to large delays.
  - Congestion leads to packet loss.
  - Conventional methods are problematic.
  - Even in a Layer 3 network, we can take action at Layer 2 to reduce congestion and thereby loss.
  - The paper is 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*

- Online Data Intensive (OLDI) Services
  - Deep Learning and Model Training
  - Non-Volatile Memory Express (NVMe) over Fabrics
  - Cloudification of the Central Office
- 
- Overall theme is dependence of parallel computation on the network

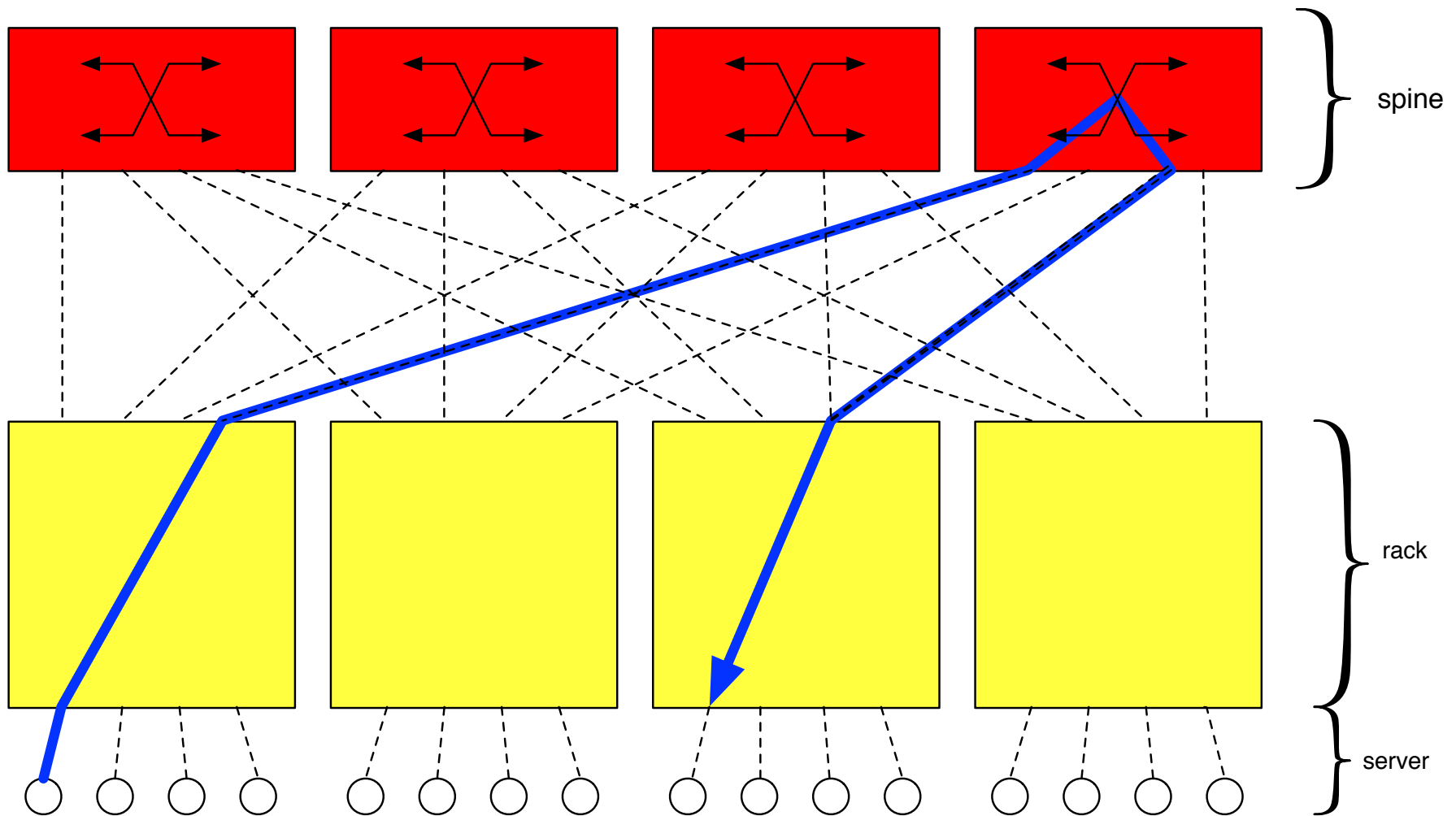


# Data Center Applications are distributed and latency-sensitive

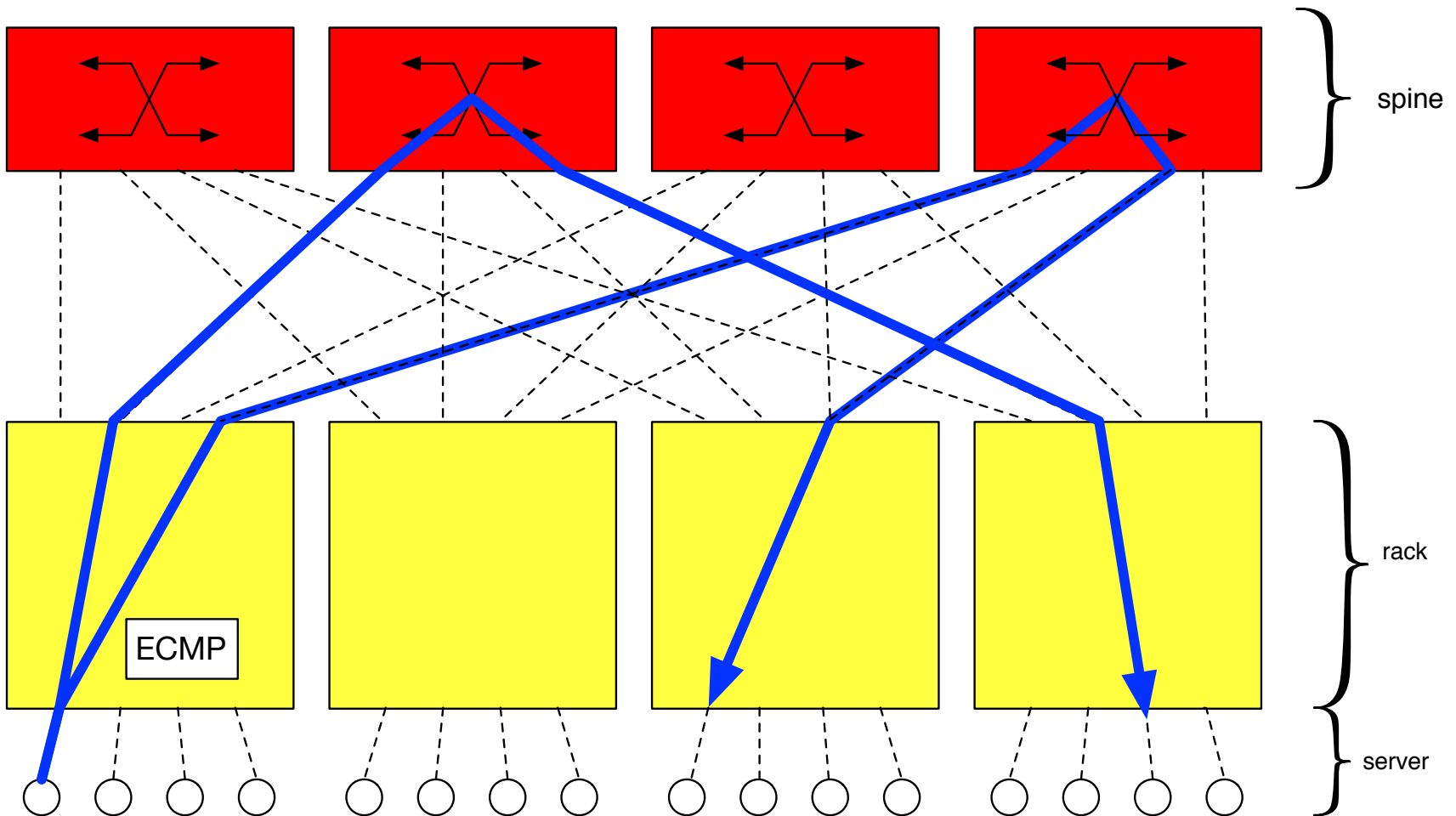


- Tend toward congestion; e.g. due to incast
- Packet loss leads to retransmission, more congestion, more delay

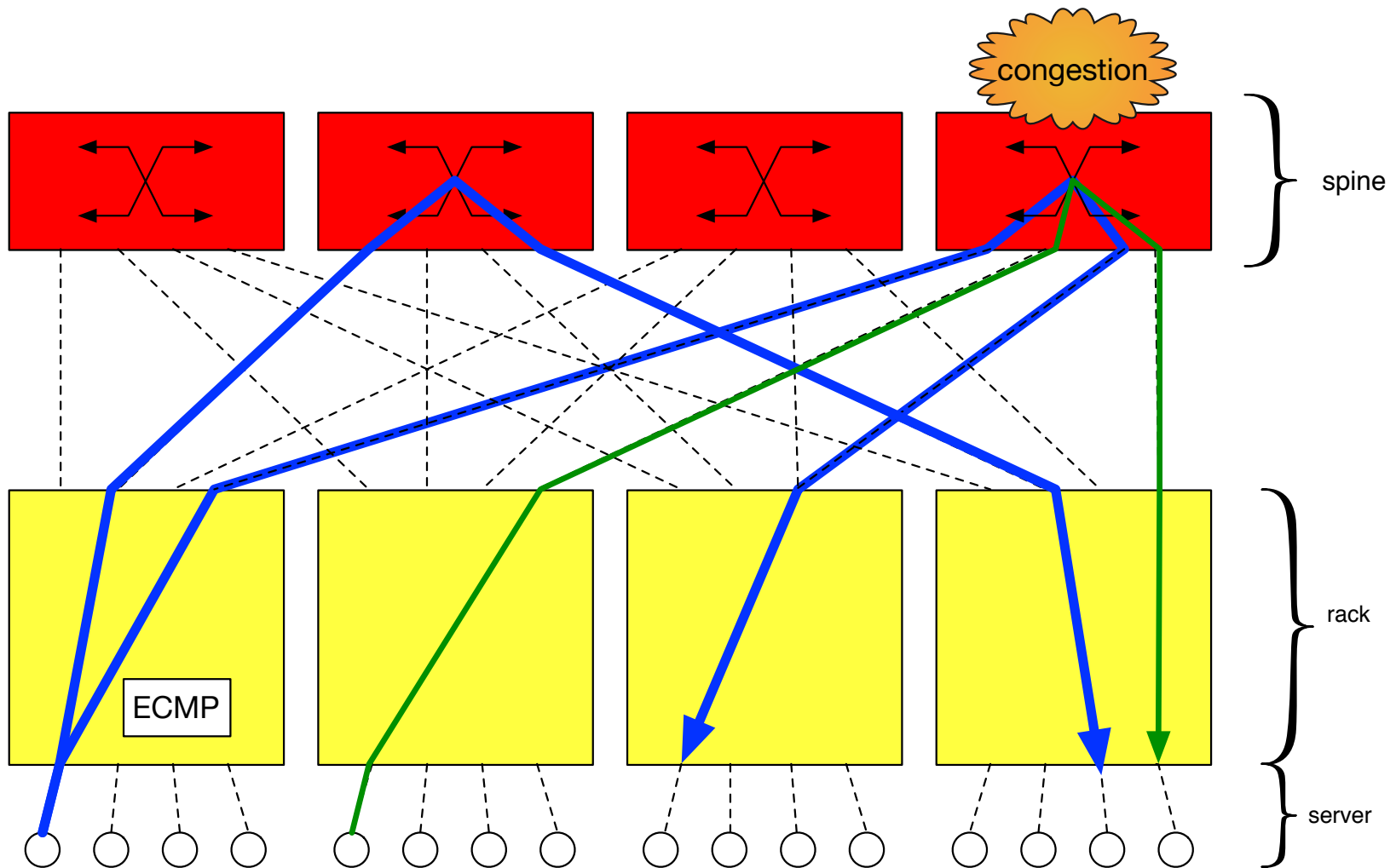
# Folded-Clos Network: Many Paths from Server to Server



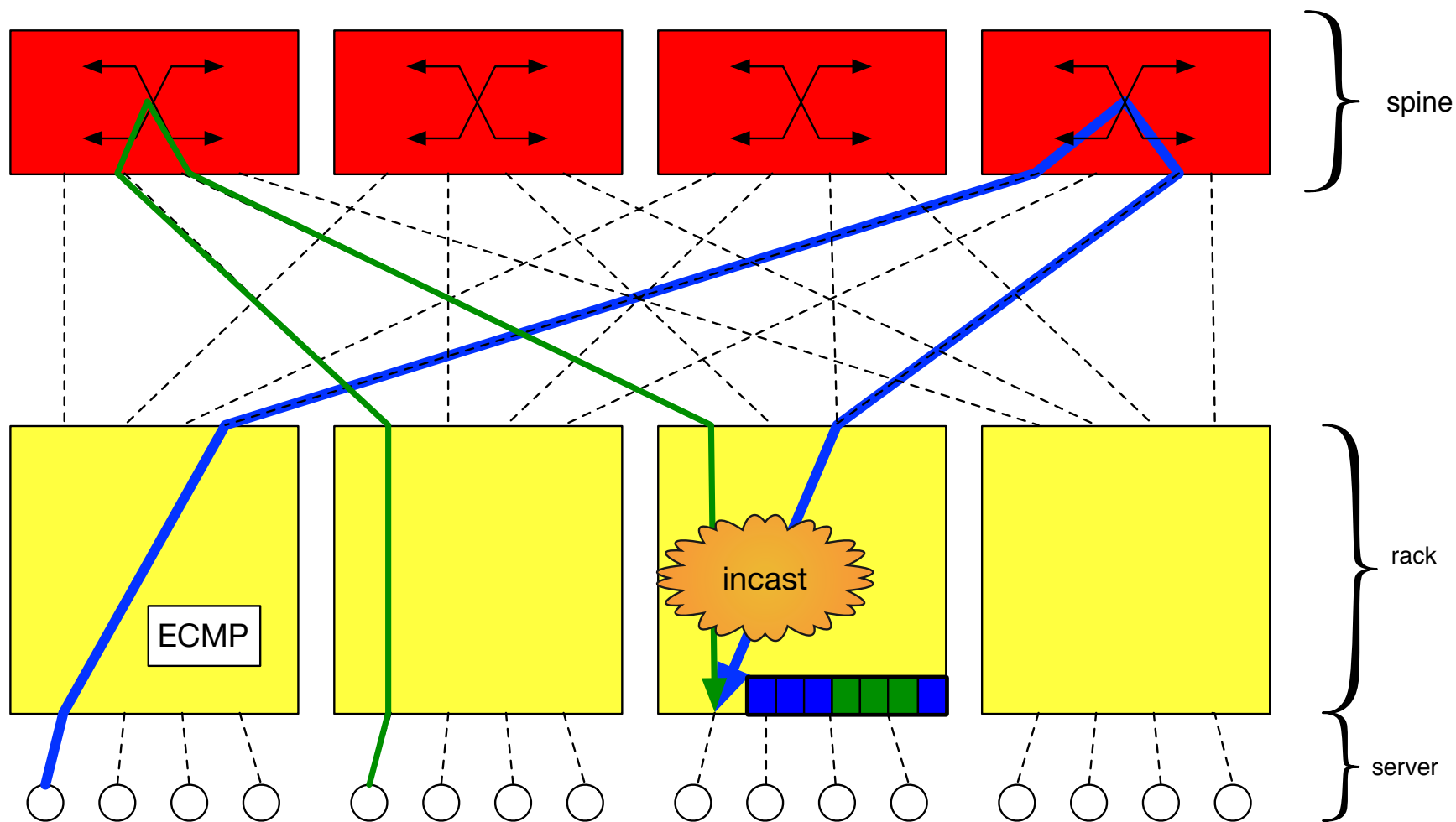
# Equal-Cost Multi-Path (ECMP): Path assigned per flow (~random)



# ECMP may still lead to congestion; e.g. large flows may collide

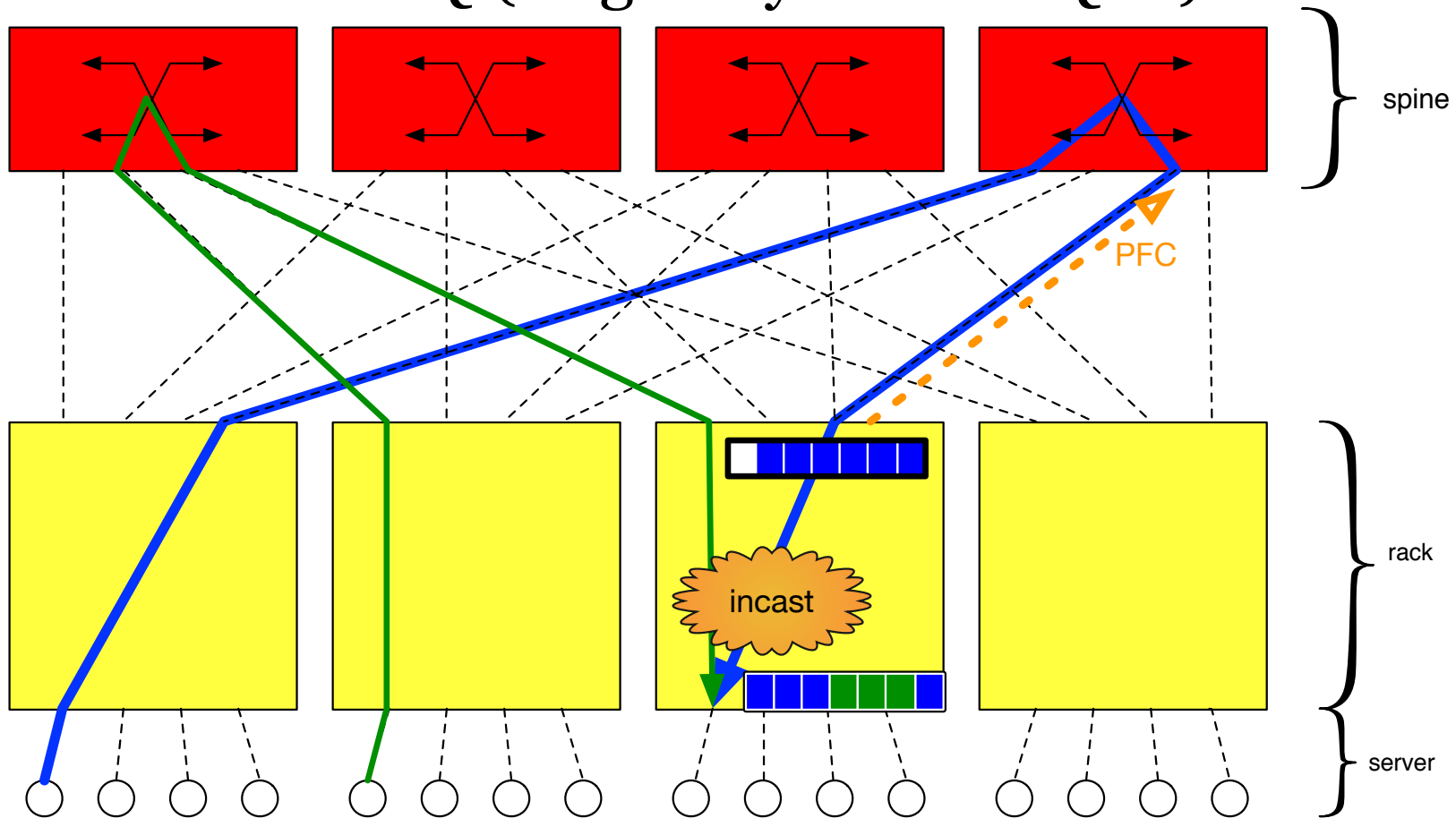


# Incast fills output queue (note: ECMP cannot help)

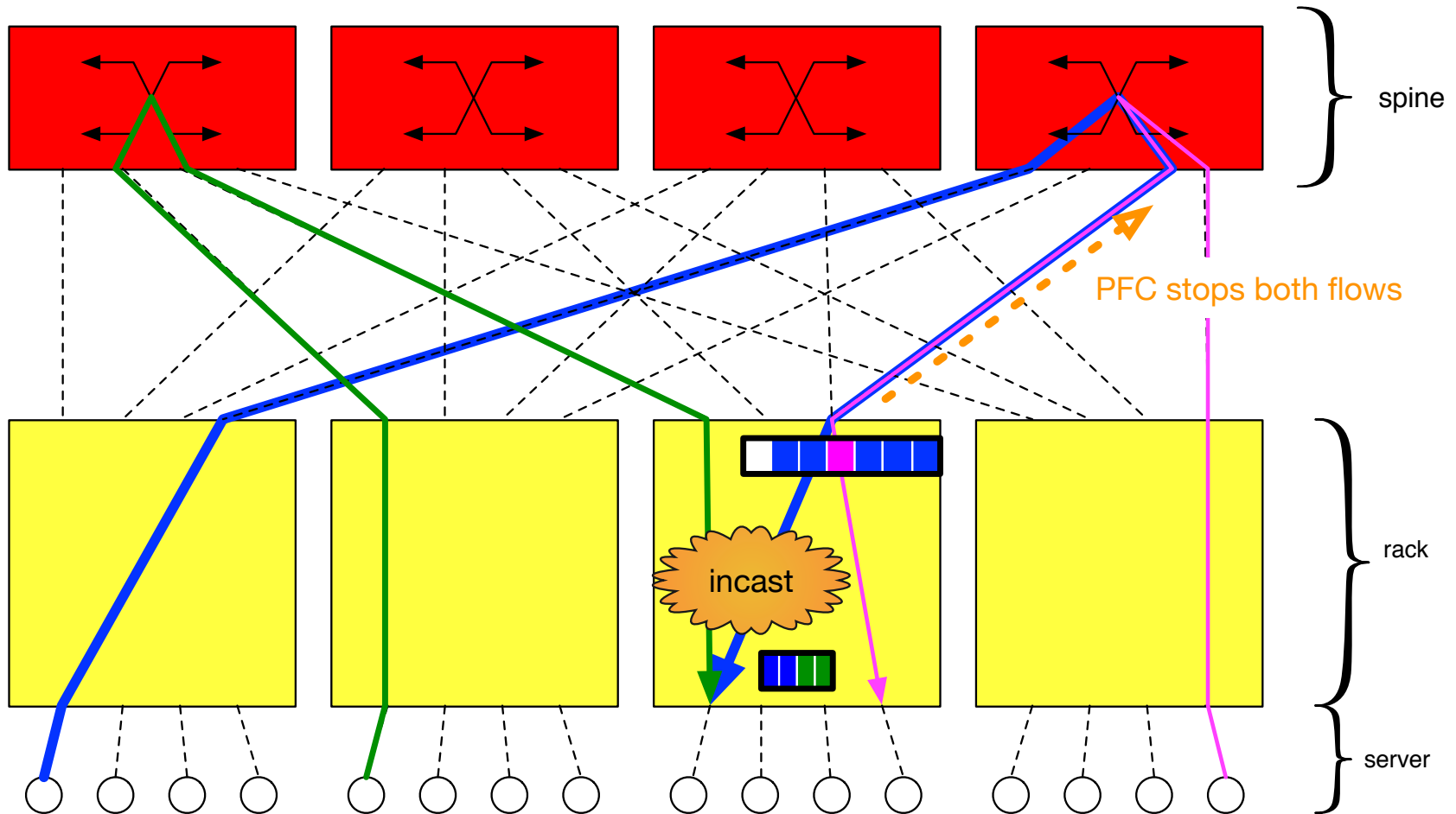


# Priority flow control (PFC)

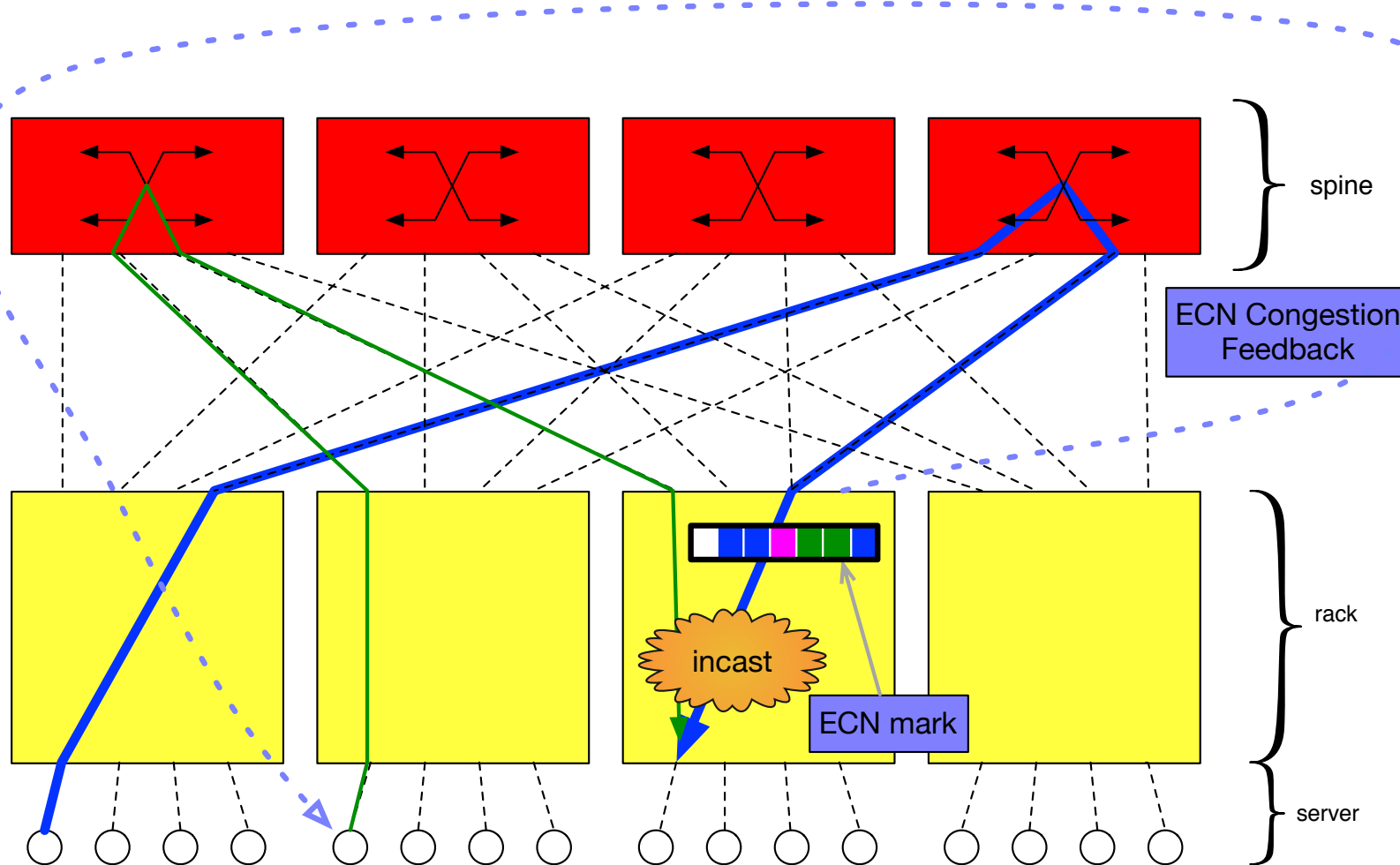
- Output backup fills ingress queue
- PFC can be used to pause input per QoS class
- IEEE 802.1Q (originally in 802.1Qbb)



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

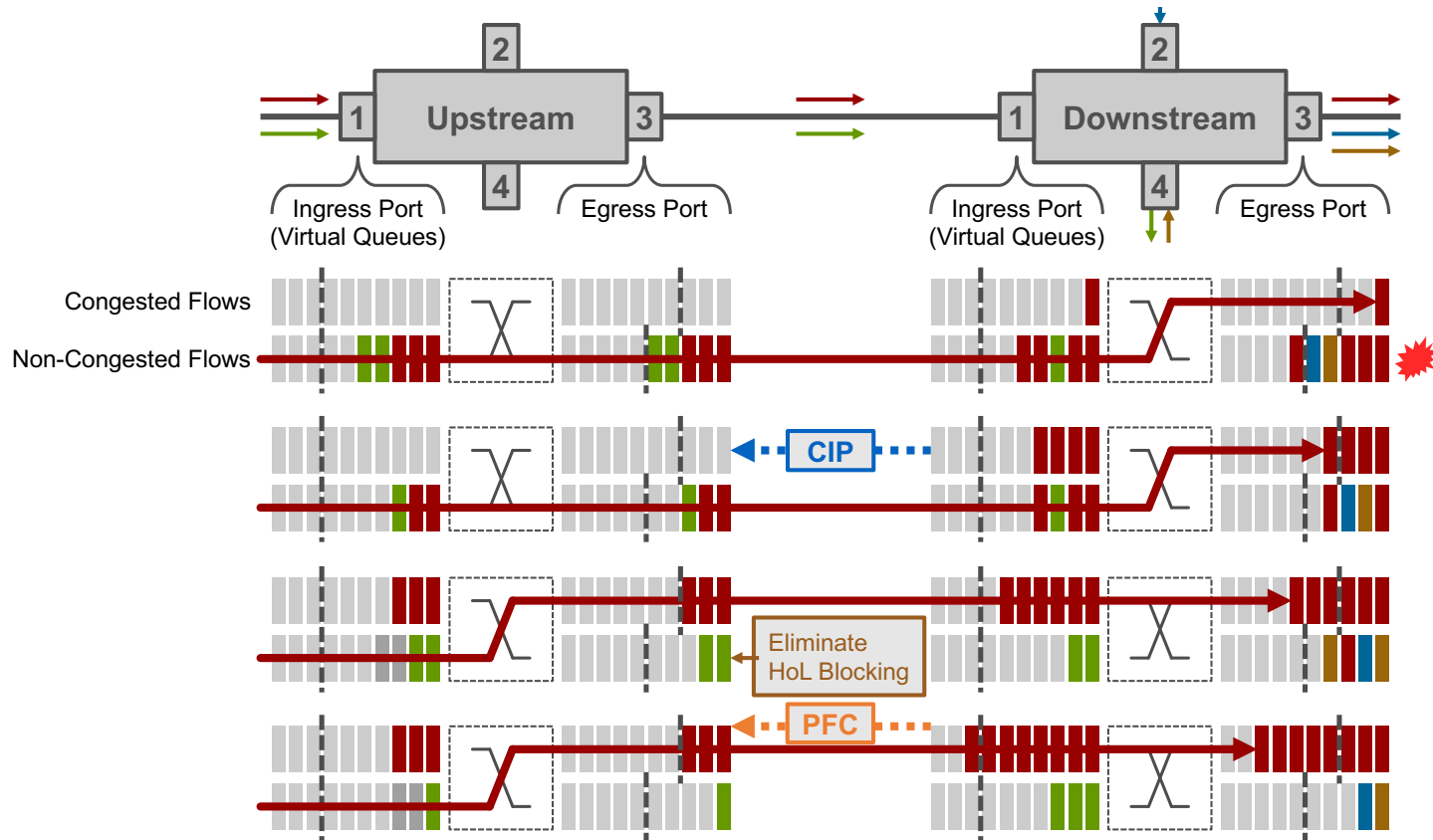


# Explicit Congestion Notification (ECN) pauses flows at source





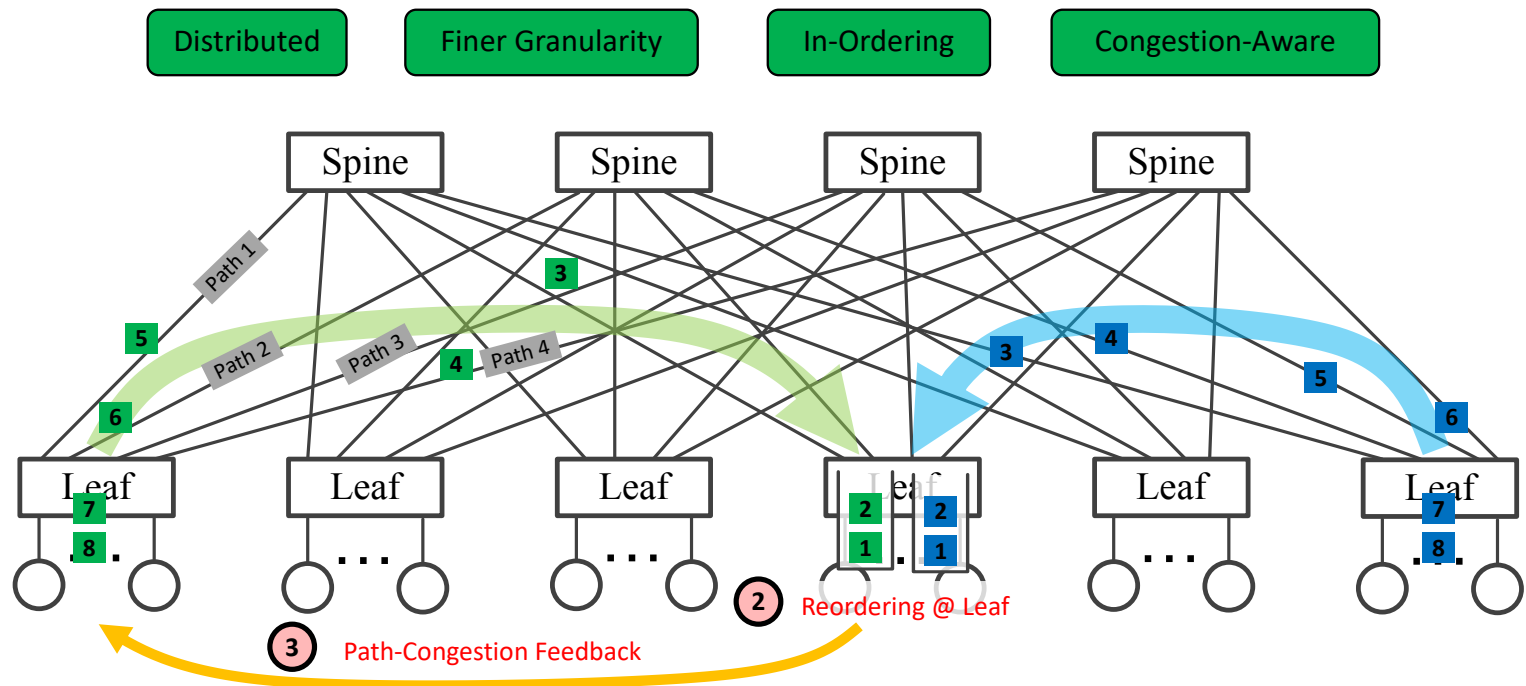
# Dynamic Virtual Lanes (DVL)



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

# Load-Aware Packet Spraying (LPS)

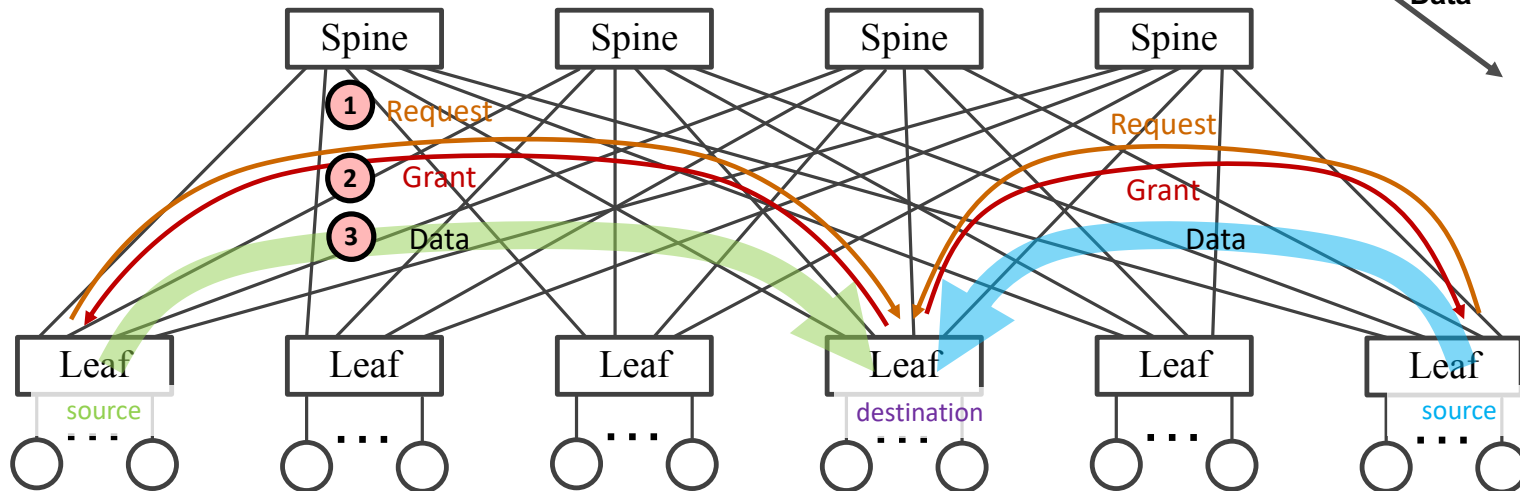
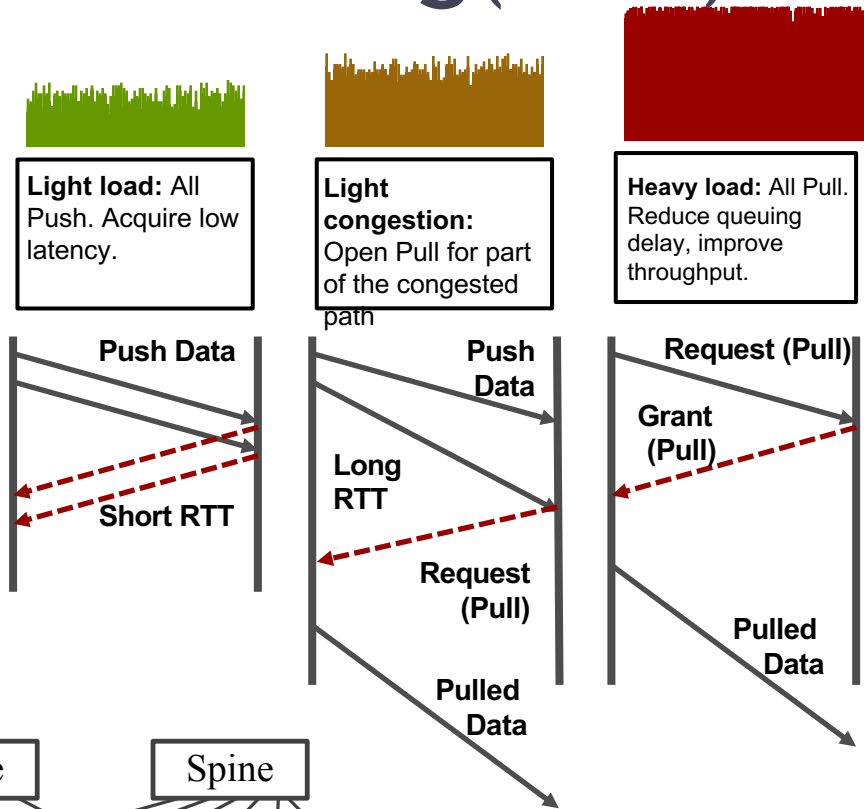
LPS = Packet Spraying + Endpoint Reordering + Load-Aware



# Push & Pull Hybrid Scheduling (PPH)

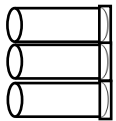
Congestion aware edge switch scheduling

- Push when load is light
- Pull when load is high



# Key Issues: Nendica Report on Lossless Network for Data Centers

## Congestion Cause




Priority-based Flow Control is coarse. Victim flows paused due to congested flows

**Isolate Congestion**




## Mitigation



Allow time for end-to-end congestion control. Move congested flows out of the way. Eliminate victim blocking.


## Innovation

**Dynamic Virtual Lane**




Unbalanced load sharing. Multiple elephant flows congest and block mice flows..

**Spread the Load**



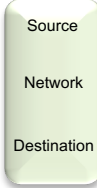
Load-balance flows at higher granularity. Use congestion awareness to avoid collisions

**Load-aware Packet Spraying**



Unscheduled incast without awareness of network resources leads to packet loss.

**Schedule Appropriately**



Schedule using integrated information from source, network, and destination.

**Push & Pull Hybrid Scheduling**

# Bibliography

- IEEE 802 “Network Enhancements for the Next Decade” Industry Connections Activity (Nendica)
  - <https://1.ieee802.org/802-nendica/>
- 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>
- 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>

# Next Steps

- IEEE 802 Nendica Report: “The Lossless Network for Data Centers” (18 August 2018) is published but open to further comment.
- Would a useful revision document point to complementary directions in 802 and IETF?
- Is it time to open a revision activity?