Lossless Data Center Networks: Opportunities for NANOG Engagement with IEEE 802 Nendica

Roger Marks
Chair, Nendica

roger@ethair.net
+1 802 capable

5 February 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.

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
- Open to all participants; no membership

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.”

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
  - Paul Congdon, Editor
  - published Nendica Report, 2018-08-17
    - IEEE 802.1-18-0042-00
  - Published report invites further comments
  - Stimulated new standardization project IEEE P802.1Qcz (Congestion Isolation)
- Flexible Factory IOT
  - Nader Zein, Editor
  - Draft report 802.1-18-0025-06
  - Significant focus on wireless
  - Comment resolution underway

Nendica Report: *The Lossless Network for Data Centers*

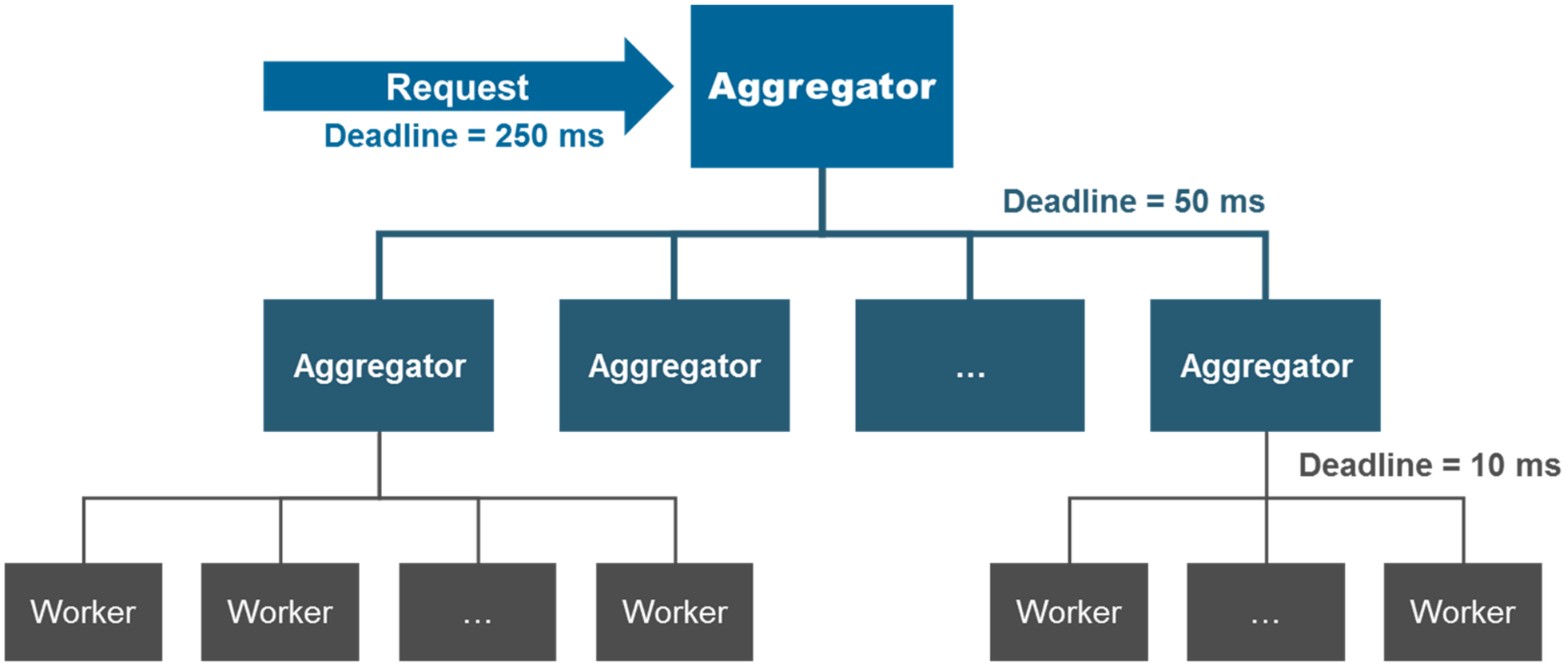
- Paul Congdon, Editor
- Key messages regarding the data center :
  - Packet loss leads to large delays.
  - Congestion leads to packet loss.
  - Conventional methods are problematic.
  - Even in a Layer 3 network, we can take action at Layer 2 to reduce congestion and thereby loss.
  - The paper is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

- An overall theme of these use cases is the dependence of parallel computation on the network.

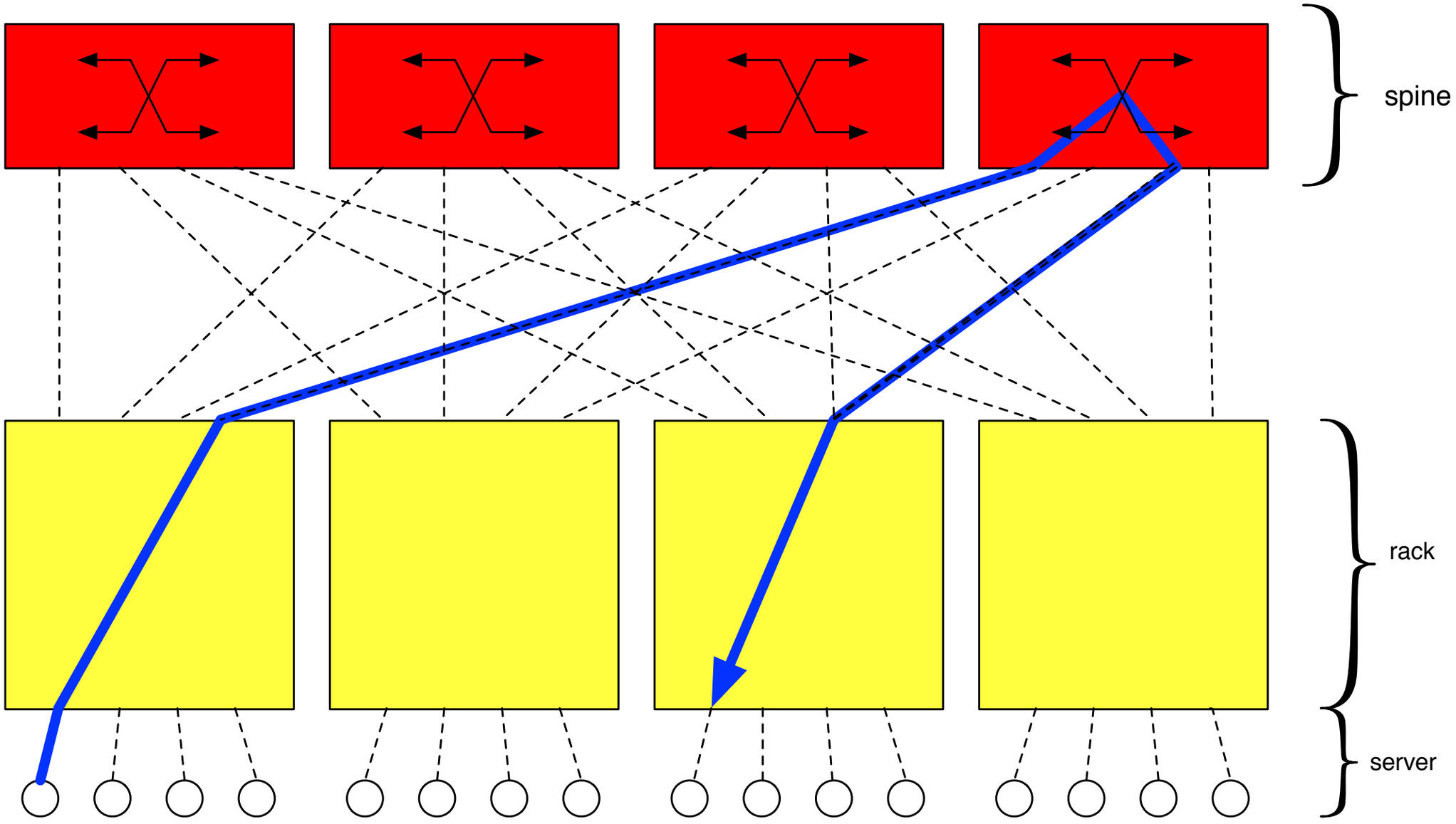
Data Center Applications are distributed and latency-sensitive



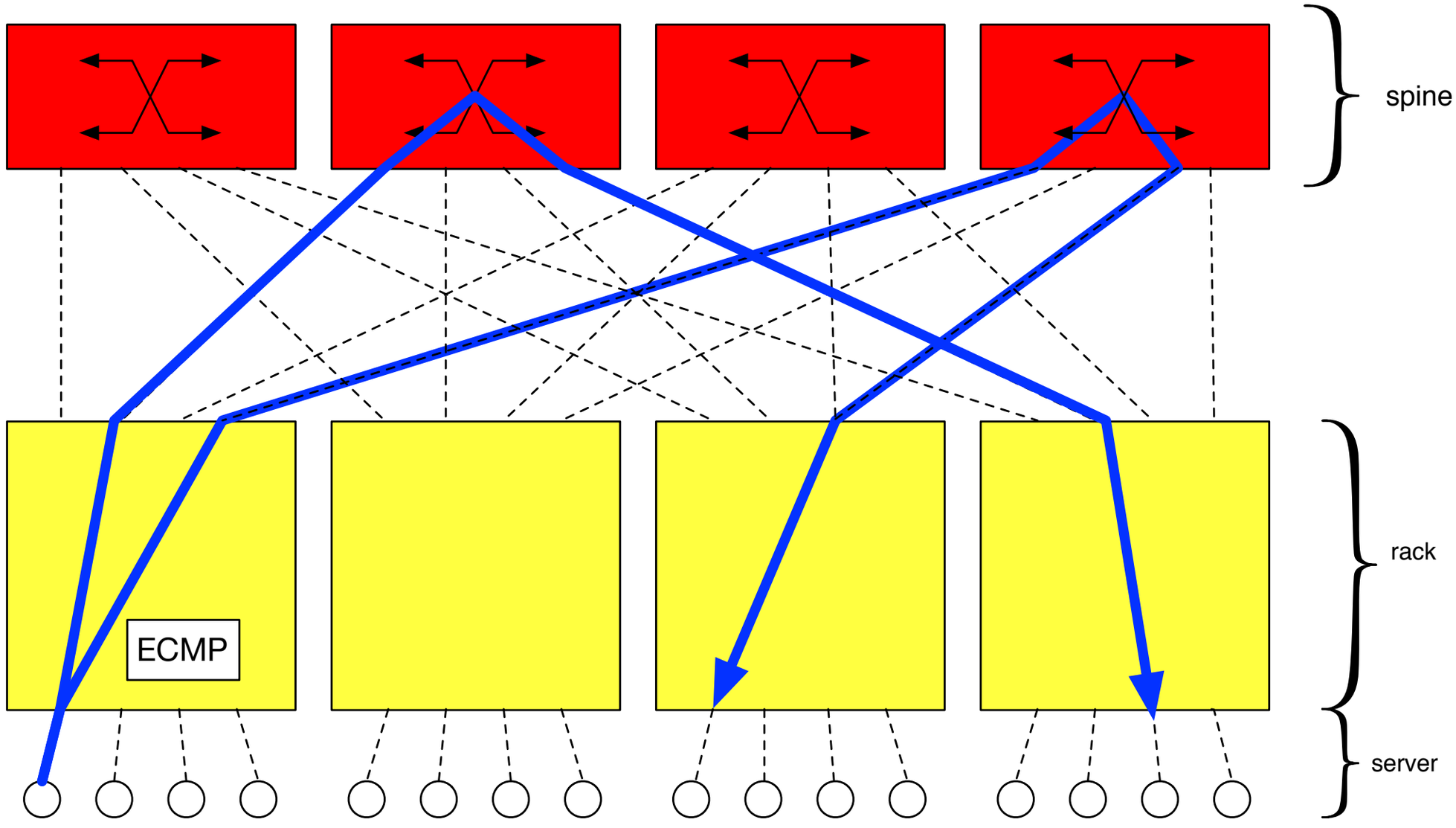
- Tend toward congestion; e.g. due to incast
- Packet loss leads to retransmission, more congestion, more delay

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

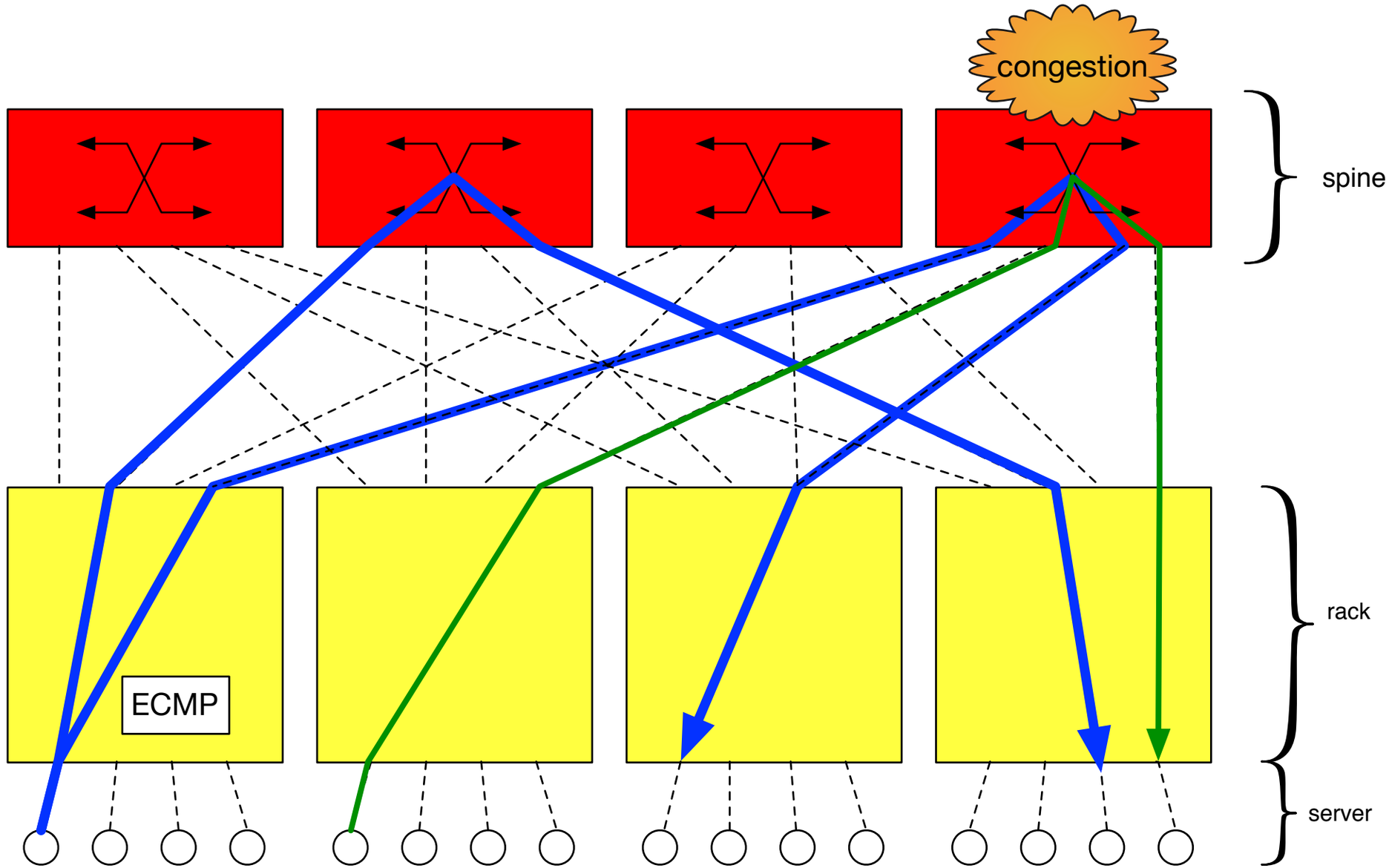
Folded-Clos Network:
Many Paths from Server to Server



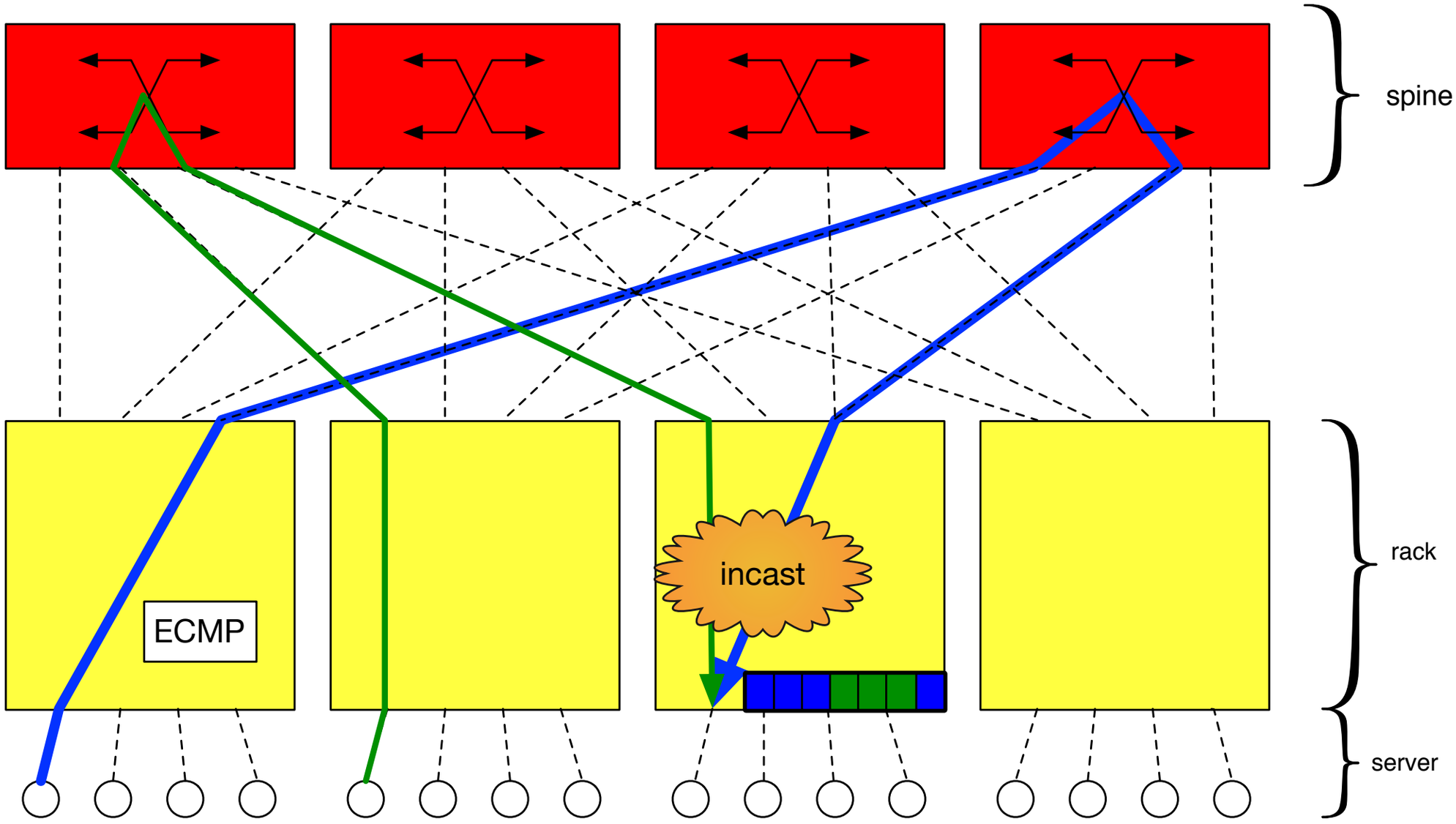
Equal-Cost Multi-Path (ECMP): Path assigned per flow (~random)



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

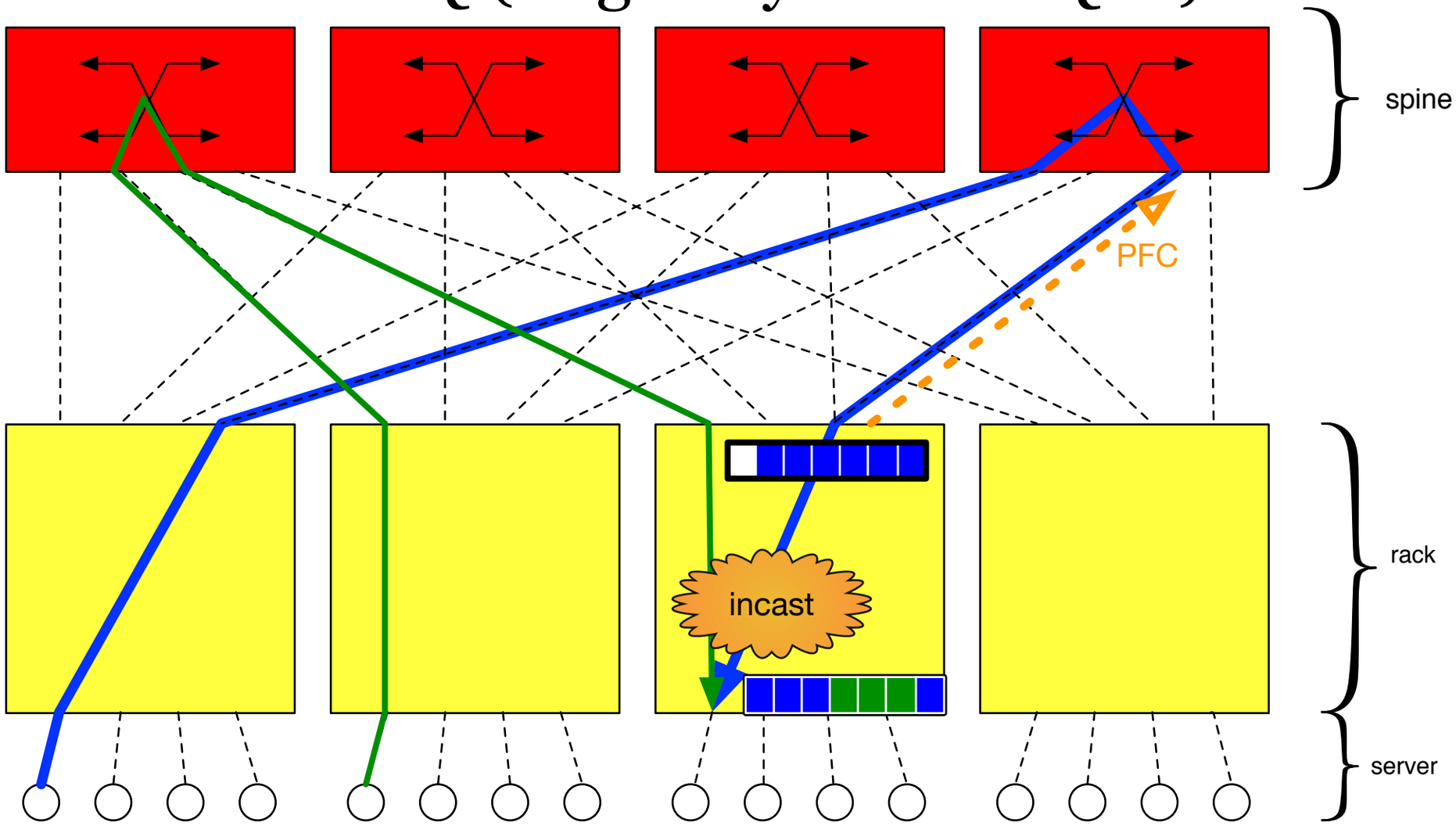


Incast fills output queue (note: ECMP cannot help)

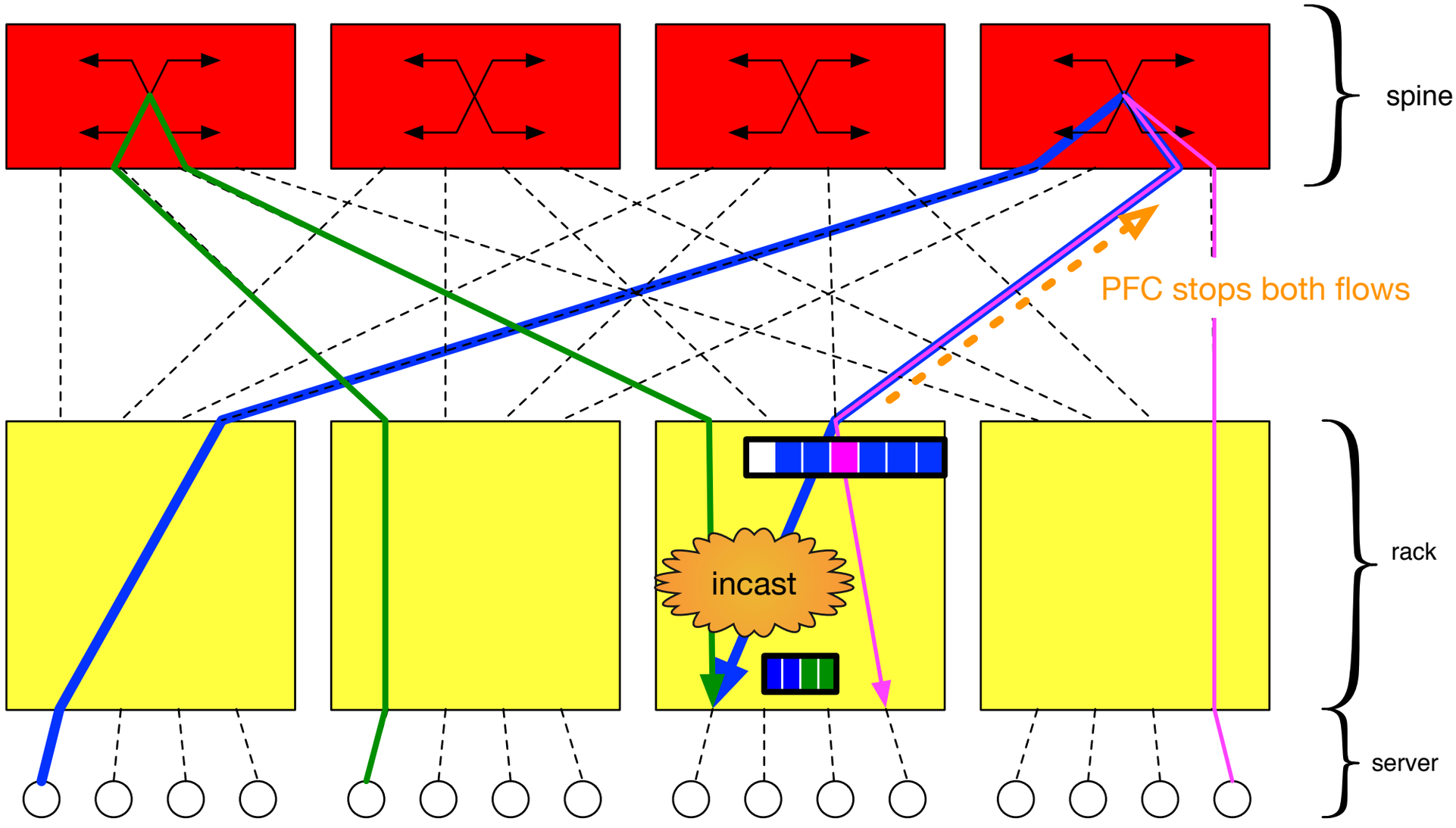


Priority flow control (PFC)

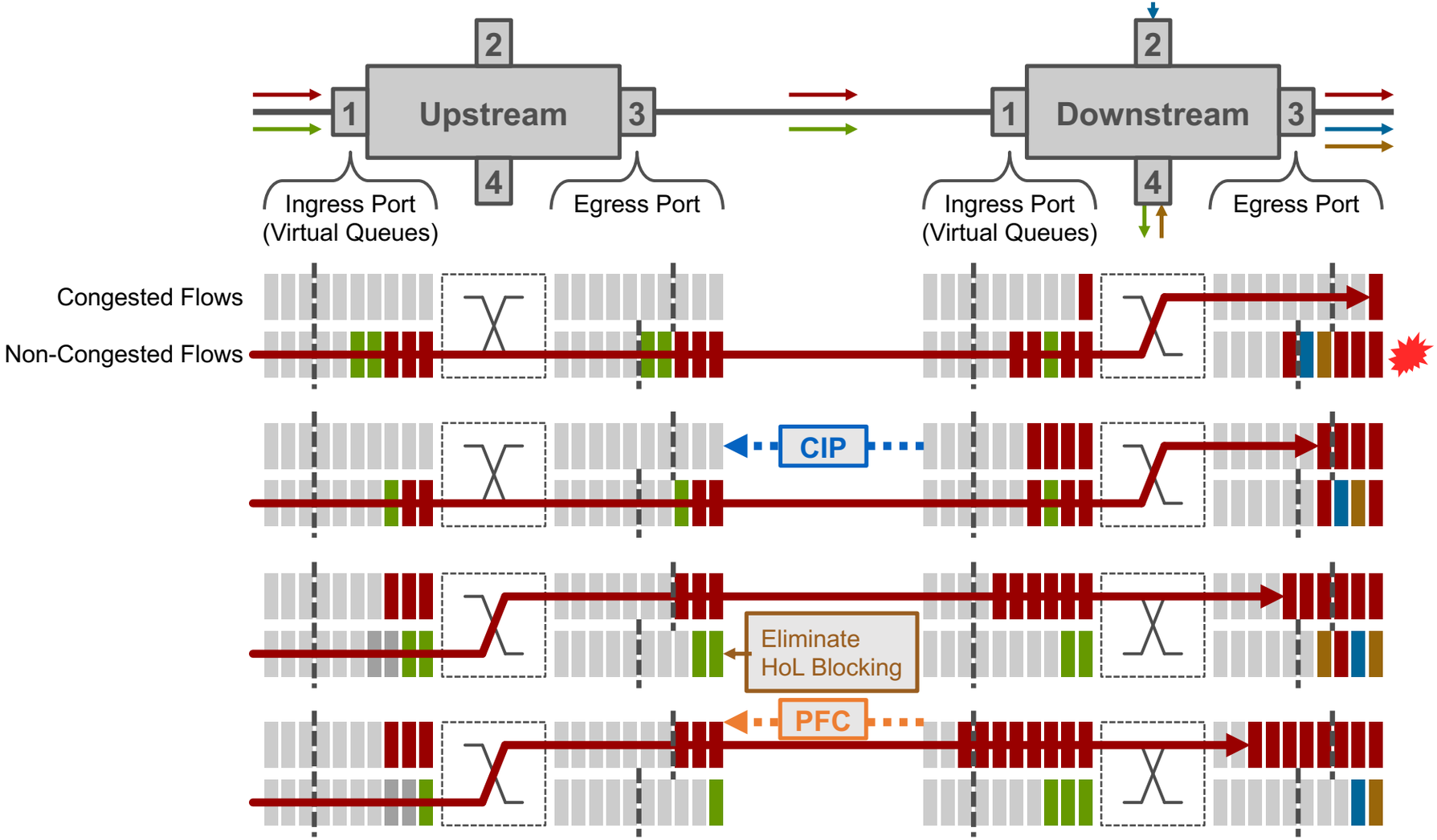
- Output backup fills ingress queue
- PFC can be used to pause input per QoS class
- IEEE 802.1Q (originally in 802.1Qbb)



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



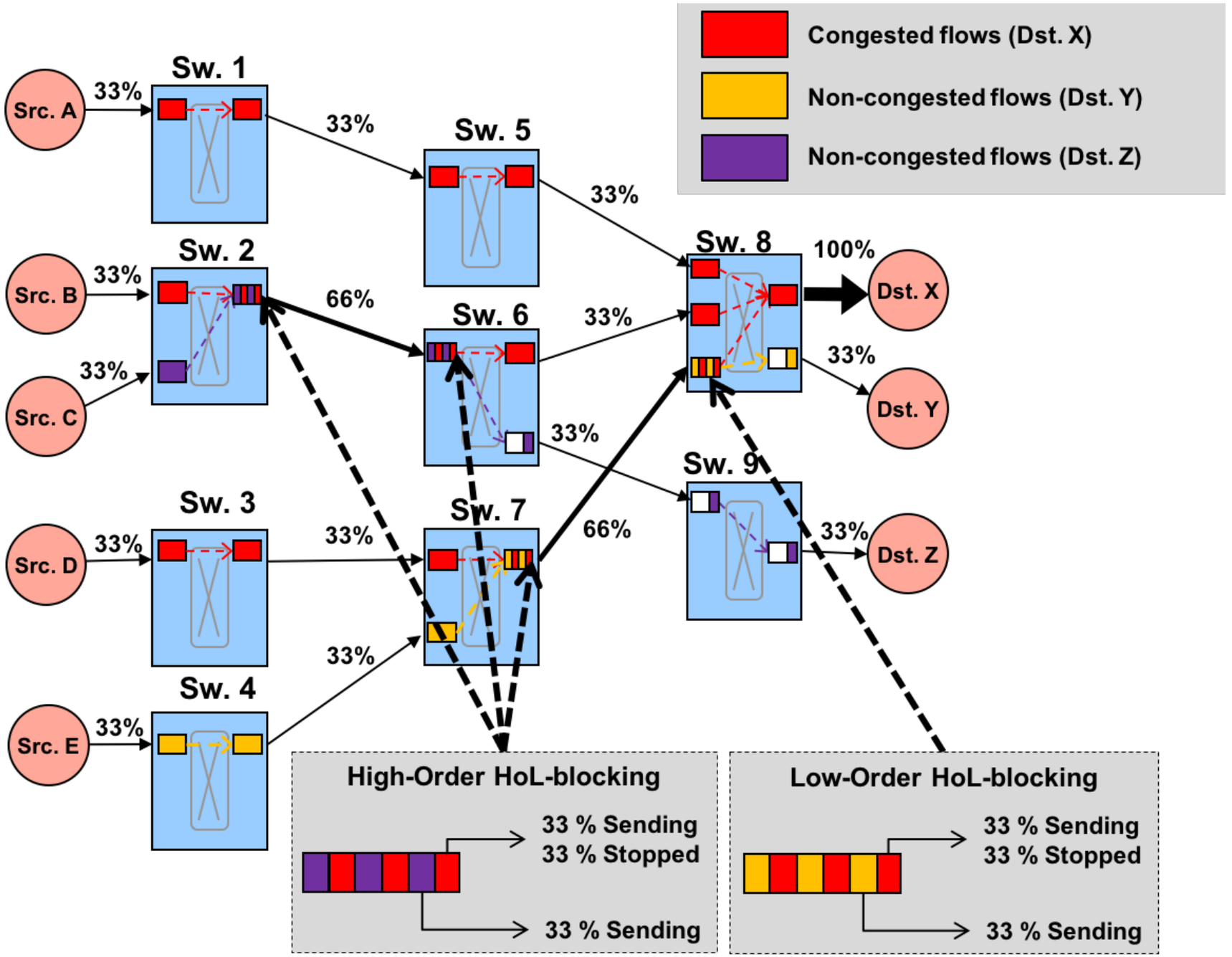
Congested-flow Isolation (see IEEE Project P802.1Qcz [5])



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

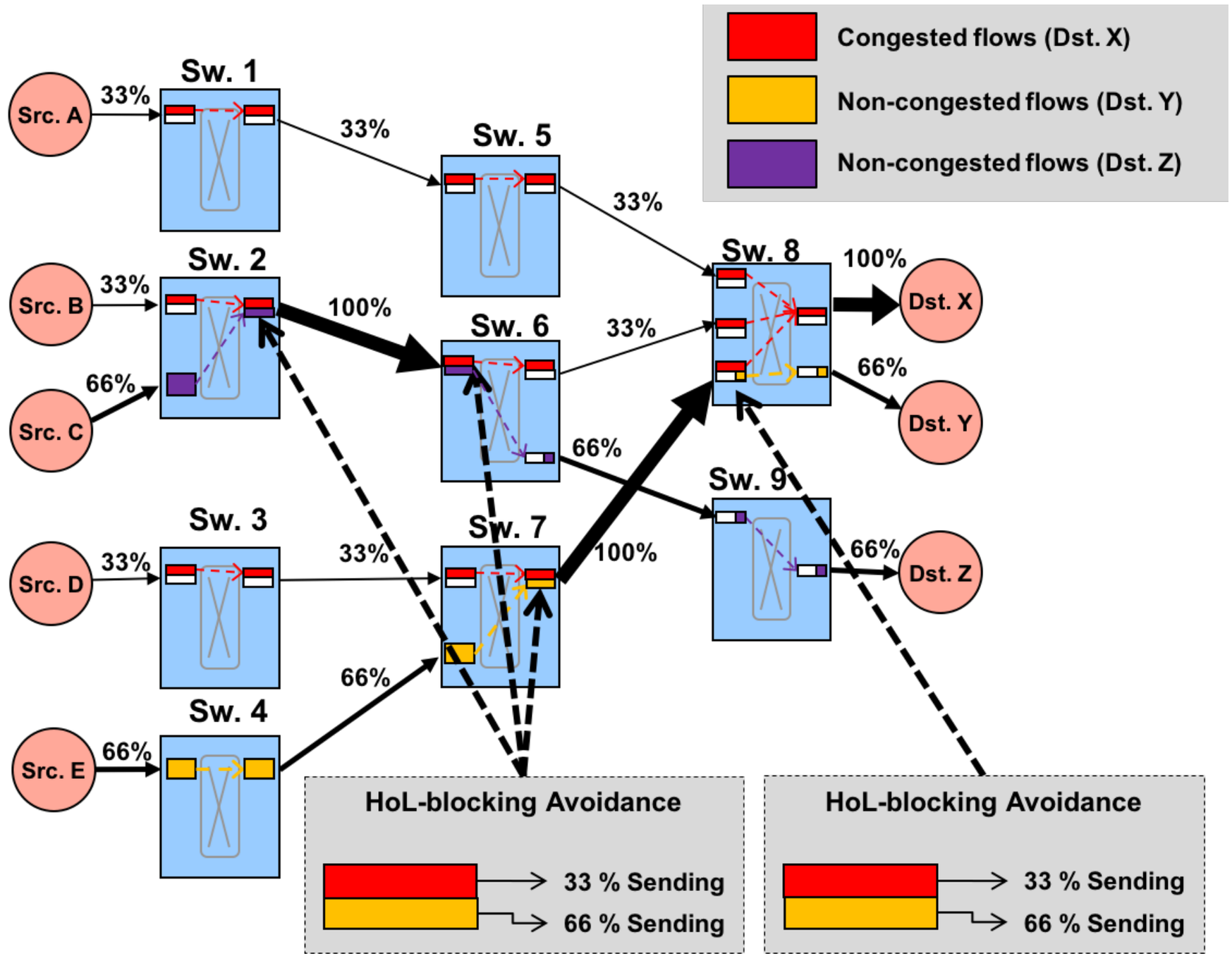
Source: IEEE 802 Nendica Report: The Lossless Network for Data Centers [2]

Traffic-class blocking analysis



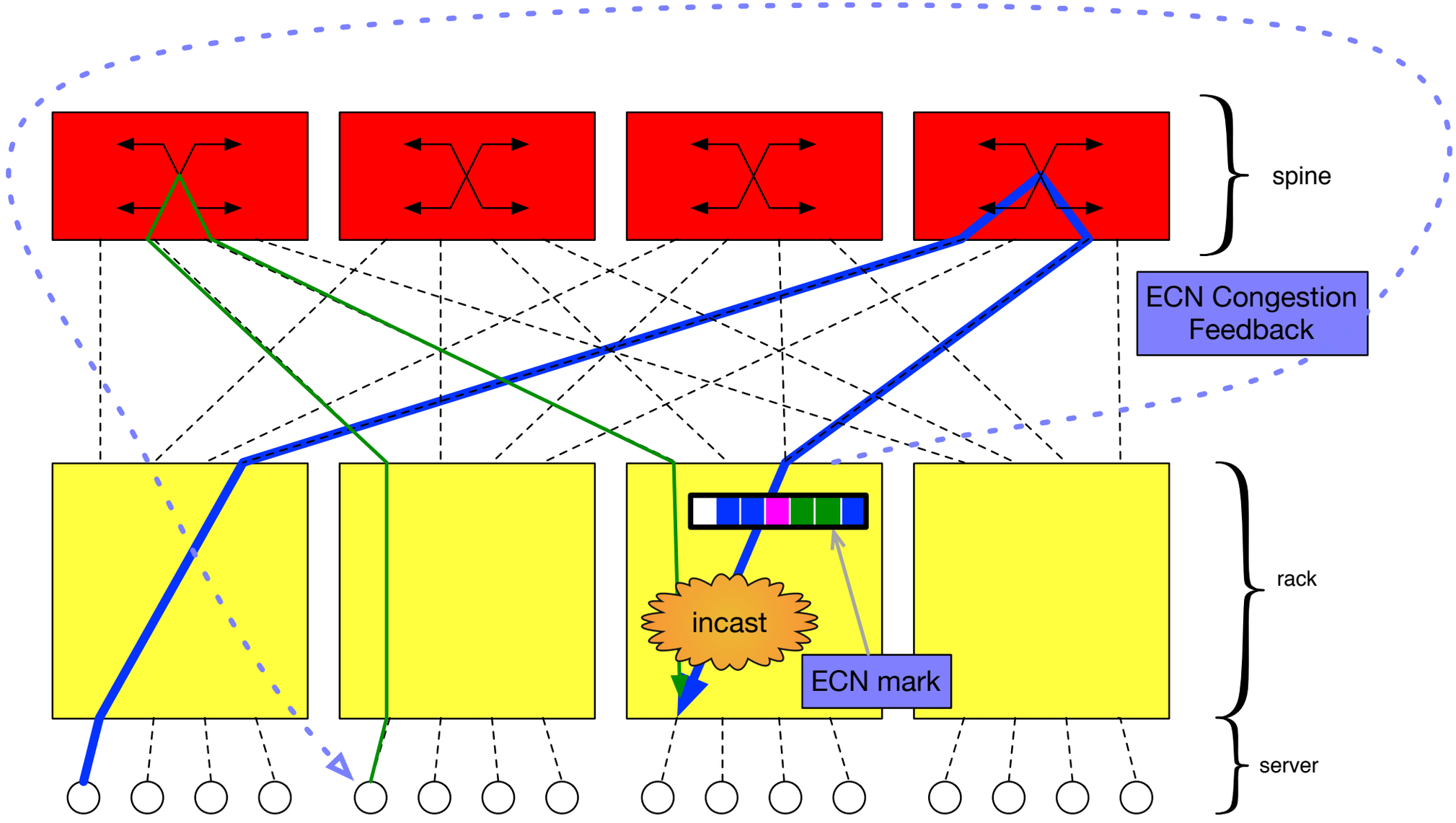
Source: IEEE 802.1-19-0012-00 [4]

Congested-flow Isolation Analysis



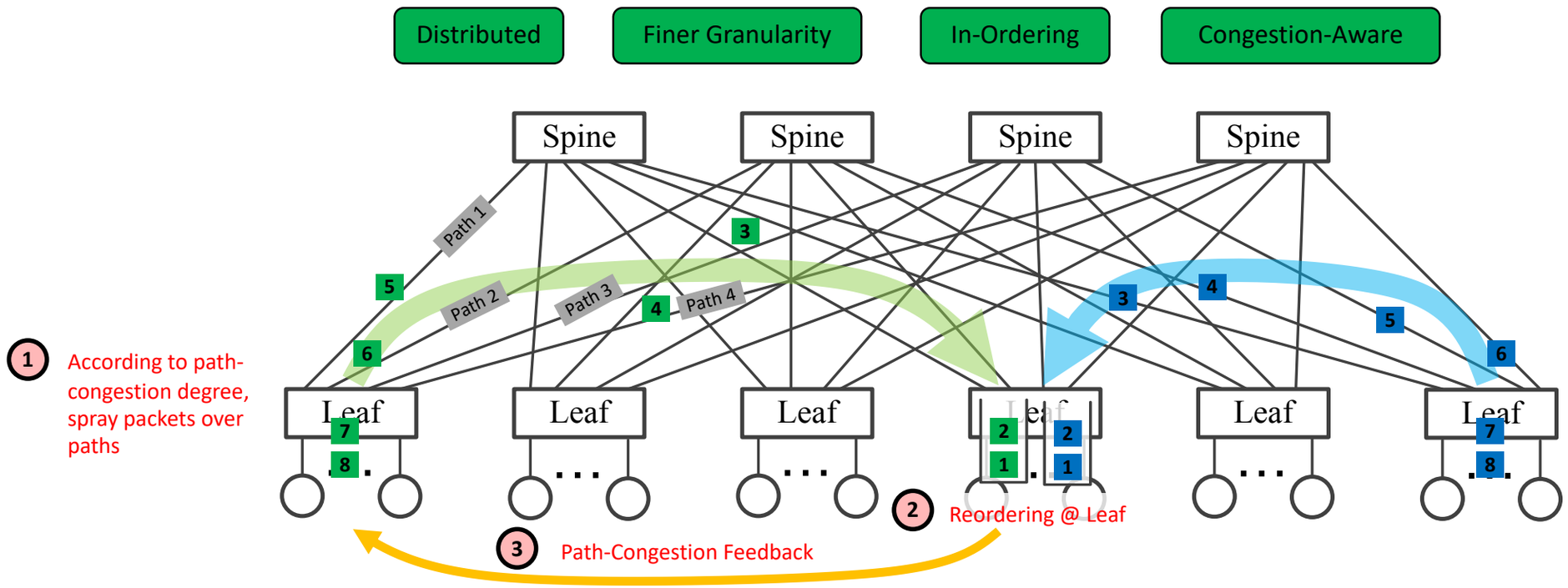
Source: IEEE 802.1-19-0012-00 [4]

Explicit Congestion Notification (ECN) pauses flows at source



Load-Aware Packet Spraying (LPS)

LPS = Packet Spraying + Endpoint Reordering + Load-Aware

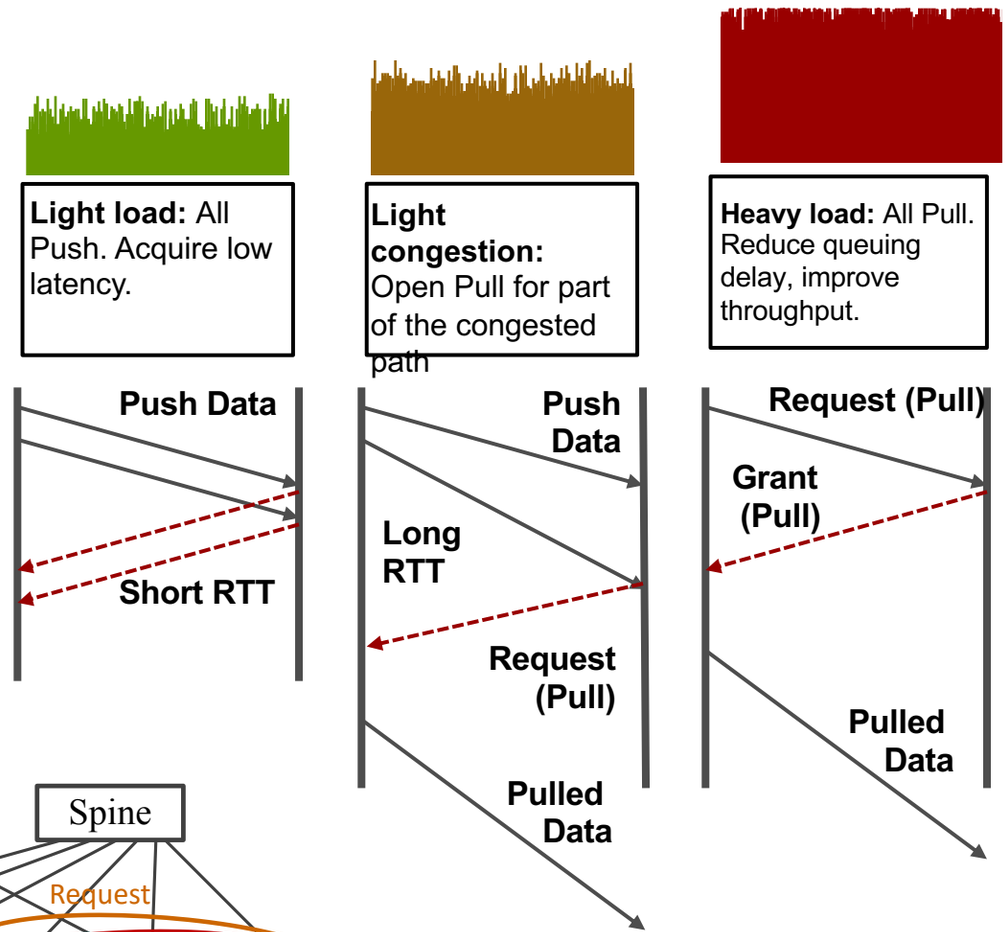


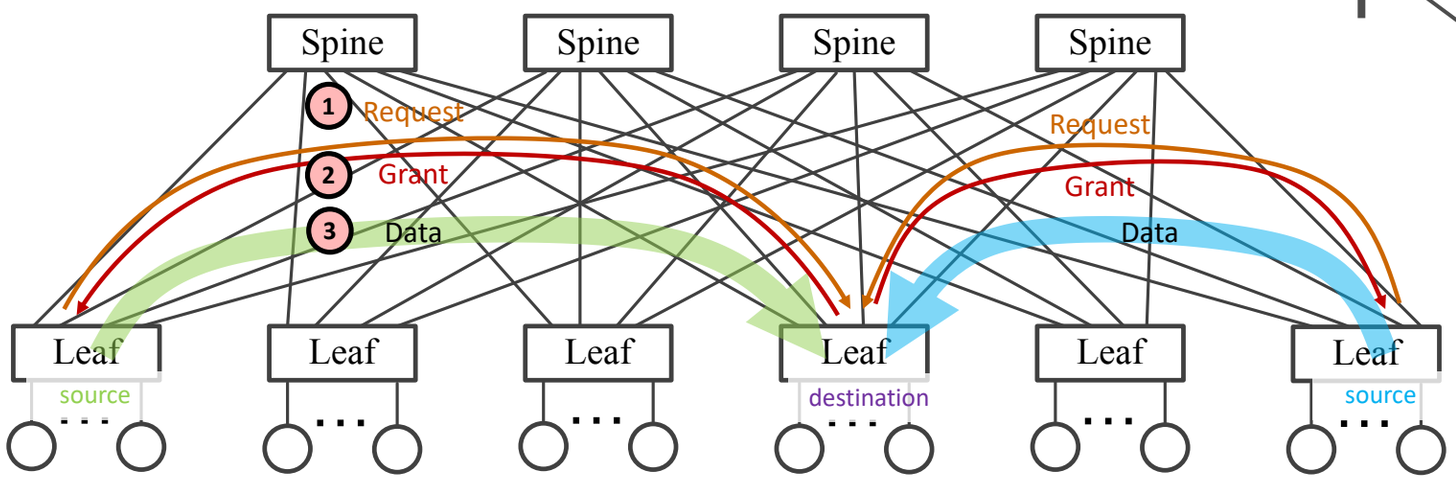
Source: IEEE 802.1-17-0007-01 [3]

Push & Pull Hybrid Scheduling(PPH)

Congestion aware edge switch scheduling

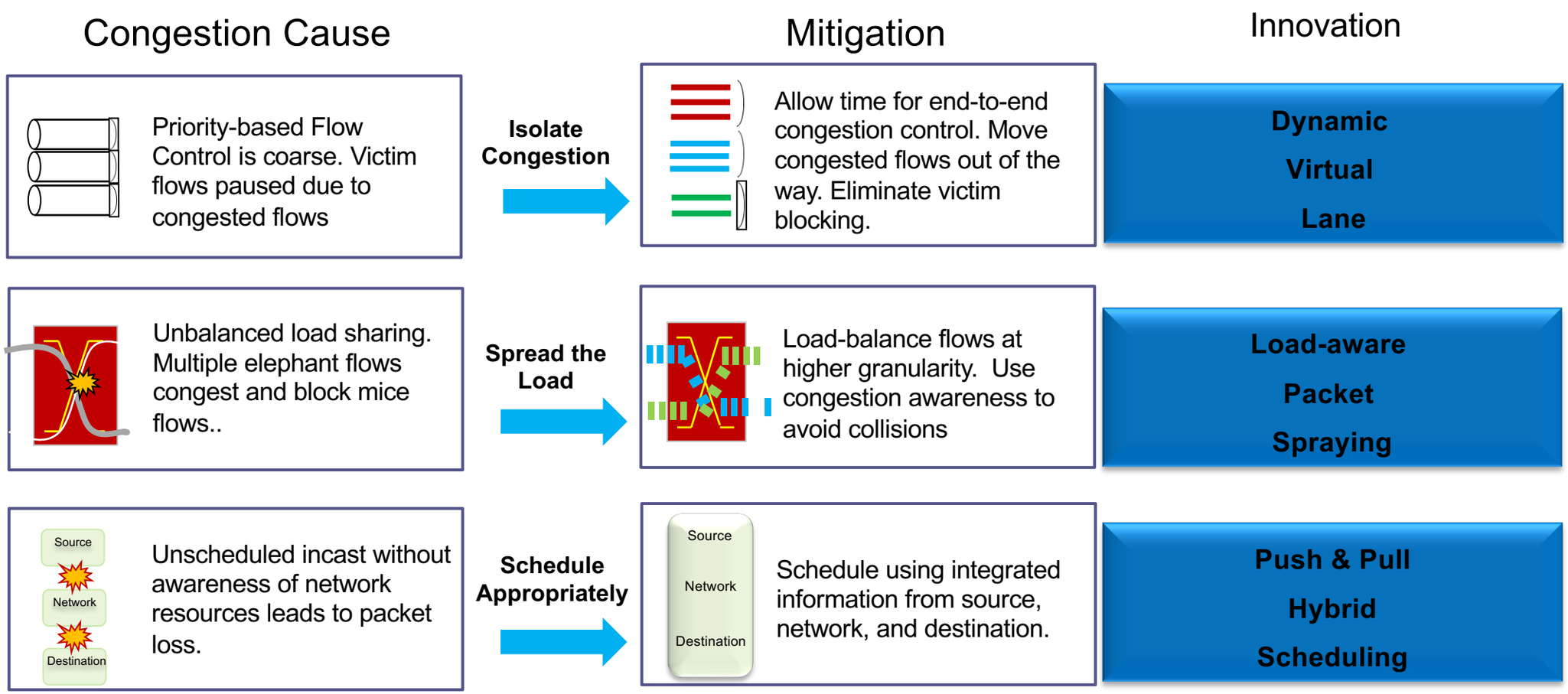
- Push when load is light
- Pull when load is high





Source: IEEE 802.1-17-0007-01 [3]

Key Issues: Nendica Report on Lossless Network for Data Centers



Extracted from IEEE 802.1-17-0007-01 [3]

Bibliography

1) IEEE 802 “Network Enhancements for the Next Decade” Industry Connections Activity (Nendica)
   - https://1.ieee802.org/802-nendica
2) 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
3) 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
4) 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
5) 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/NANOG Cooperation?

- Nendica is open to all participants; no membership
  - e.g. teleconference participation; comment process
- IEEE 802 Nendica Report: “The Lossless Network for Data Centers”.
  - Comments are welcome from NANOG participants
- An activity to revise the report could address issues important to advance NANOG goals.
- Might be useful to convene Nendica meetings in conjunction with NANOG meetings:
  - NANOG 76 (Washington, 10-12 June 2019)
  - NANOG 77 (Austin, 28-30 Oct 2019)

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)
  - or the Work Item Editors

Possible next steps

- Teleconferences targeted at identifying issues regarding Data Center networks
  - NANOG participants welcome
- Opening of activity to revise “The Lossless Network for Data Centers” report
- Decision to convene Nendica meetings in conjunction with NANOG 76
- Detailed presentation, open to NANOG participants, regarding the new P802.1Qcz project on Congestion Isolation [5]