IEEE 802 Industry Connections Report

The Next Generation Lossless Network

in the Data Center

BrightTalk, Data Center Transformation 3.0, January 2019

Paul Congdon, PhD

Disclaimer

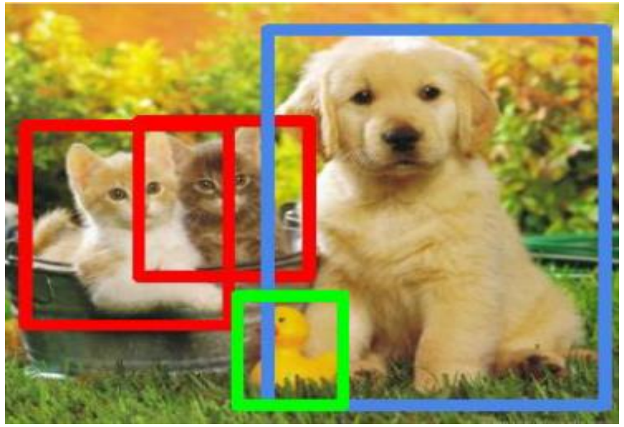
- All speakers presenting information on IEEE standards speak as individuals, and their views should be considered the personal views of that individual rather than the formal position, explanation, or interpretation of the IEEE.

Acknowledgements

- The initial technical contribution and sponsorship for this work was provided by Huawei Technologies Co., Ltd.
- This presentation summaries work from the IEEE 802 Network Enhancements for the Next Decade Industry Connections Activity (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
  - https://1.ieee802.org/802-nendica/
  - Report Freely Available at: https://ieeexplore.ieee.org/document/8462819

Our Digital Lives are driving Innovation in the DC

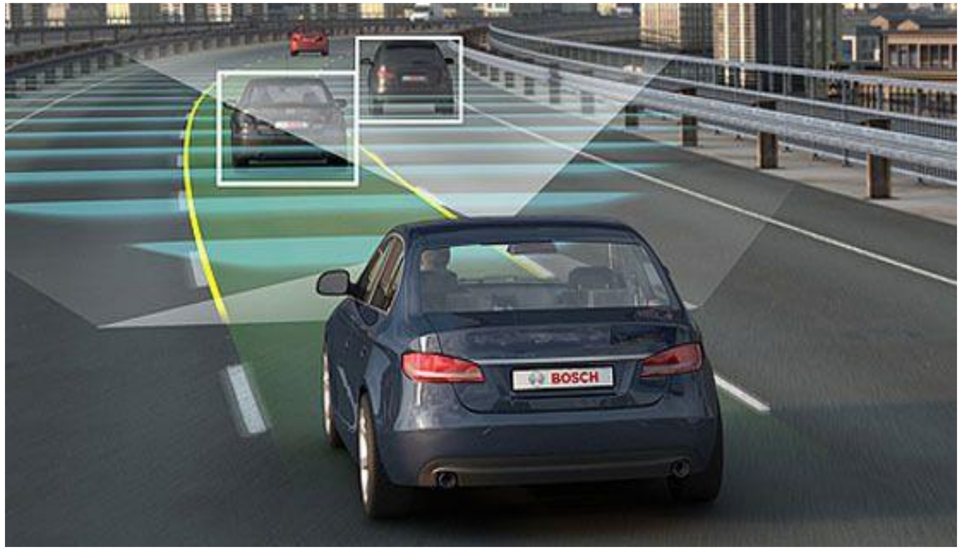
Interactive
Speech
Recognition



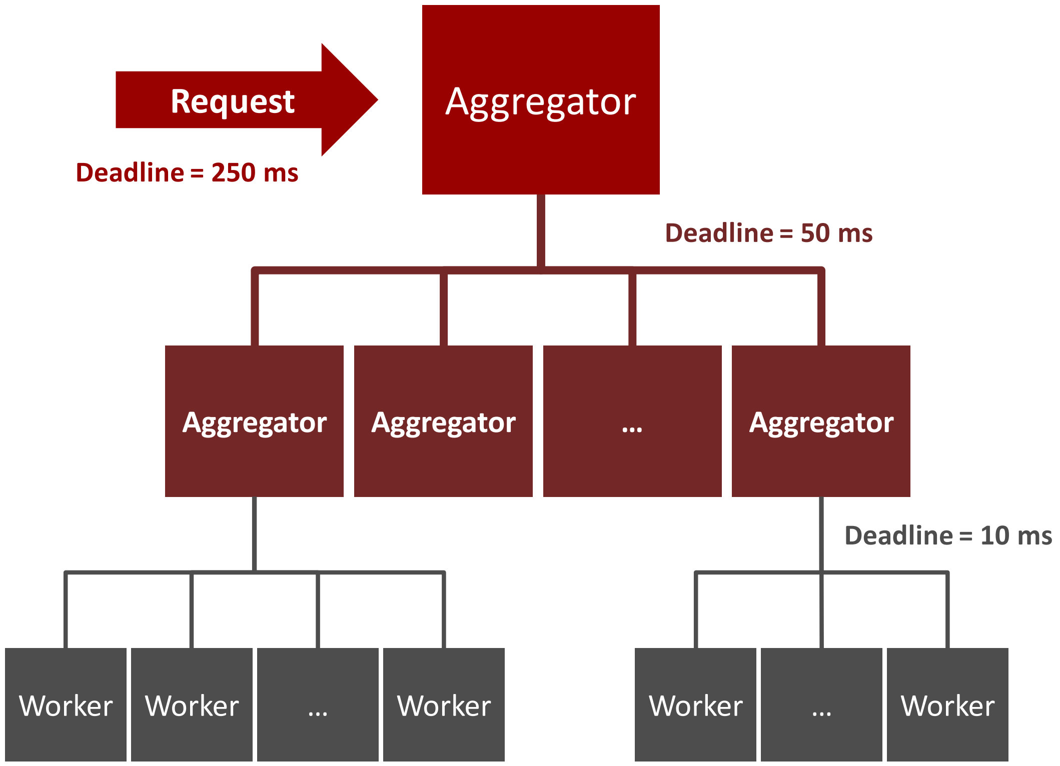
Interactive
Image
Recognition

Human / Machine
Interaction

Autonomous
Driving

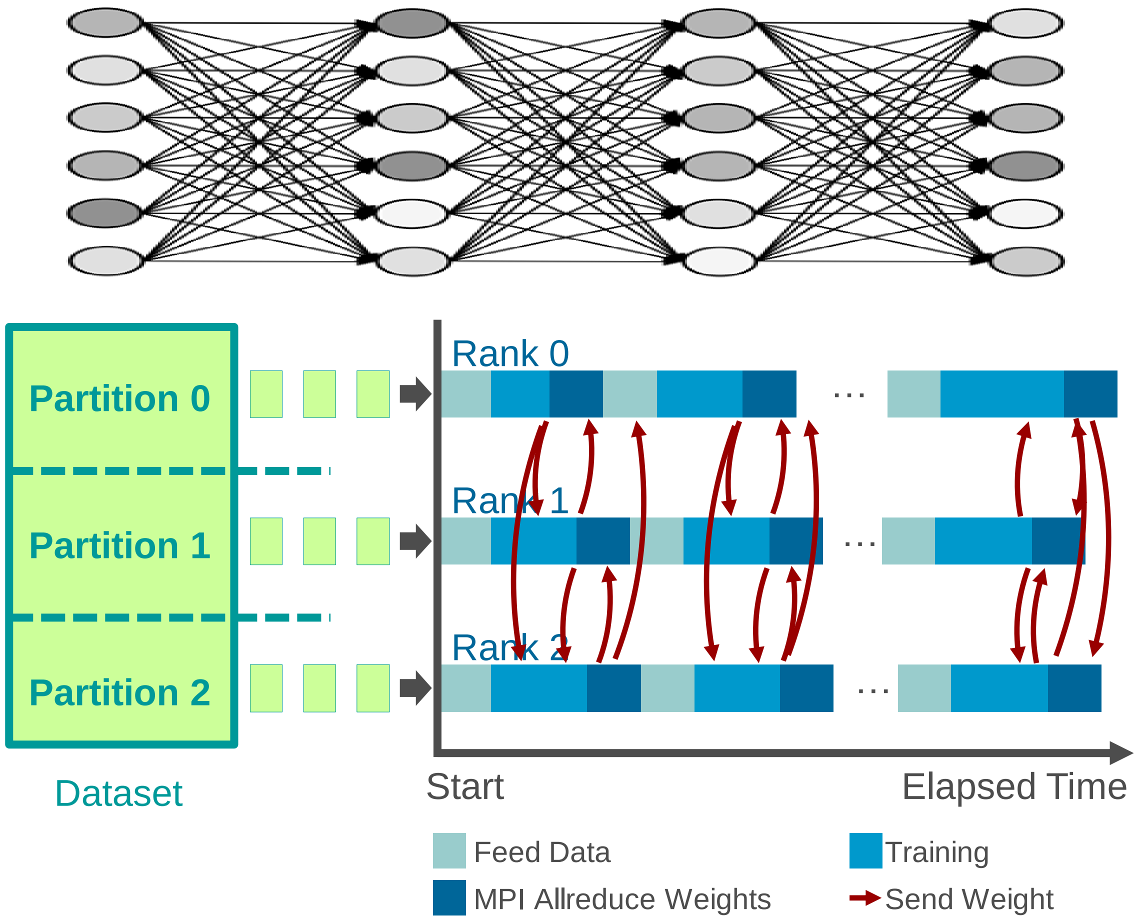


Critical Use Case – Online Data Intensive Services (OLDI)

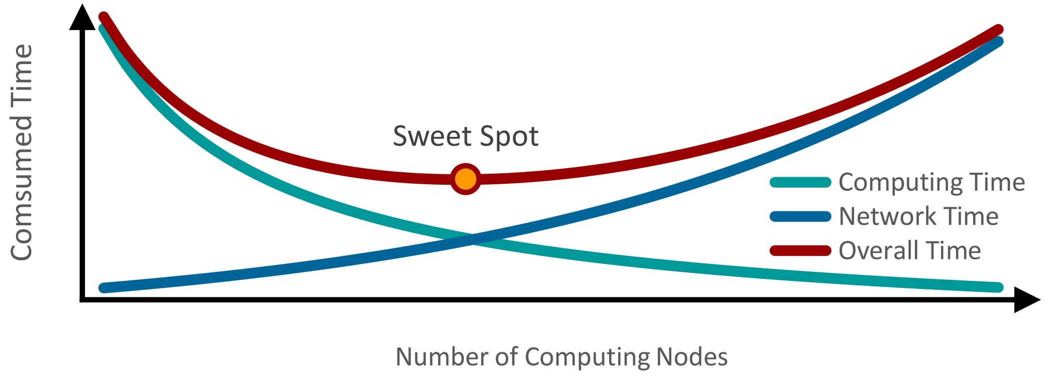


- OLDI applications have real-time deadlines and run in parallel on 1000s of servers.
- Incast is a naturally occurring phenomenon.
- Tail latency reduces the quality of the results

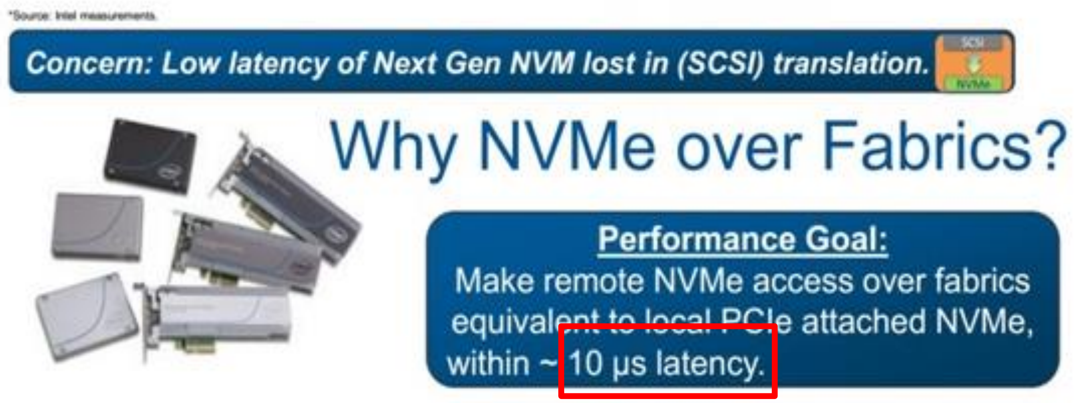
Critical Use Case – Deep Learning

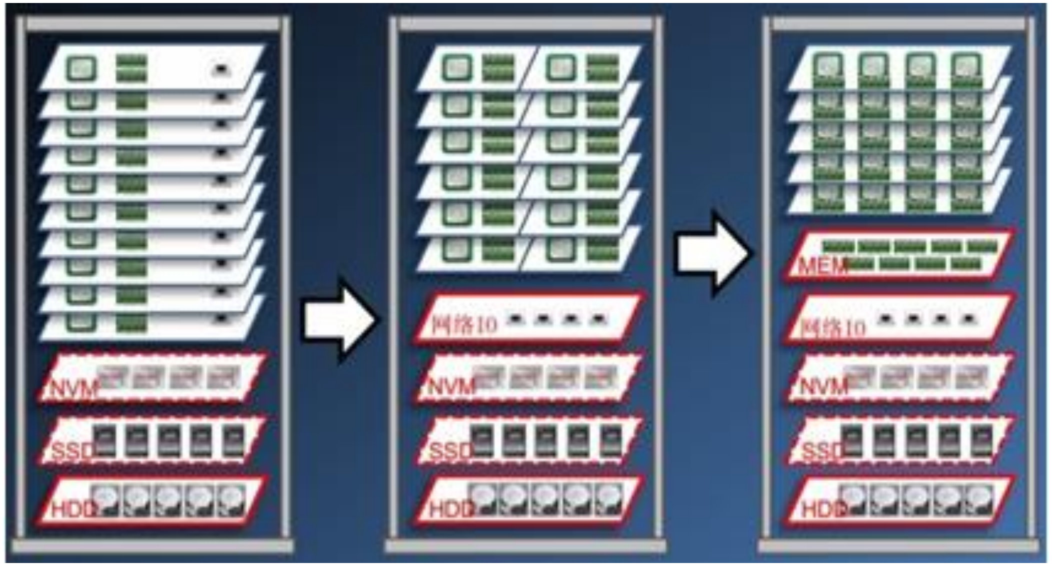


- Massively parallel HPC applications, such AI training, are dependent on low latency and high throughput network.
- Billions of parameters.
- Scale out is limited by network performance.



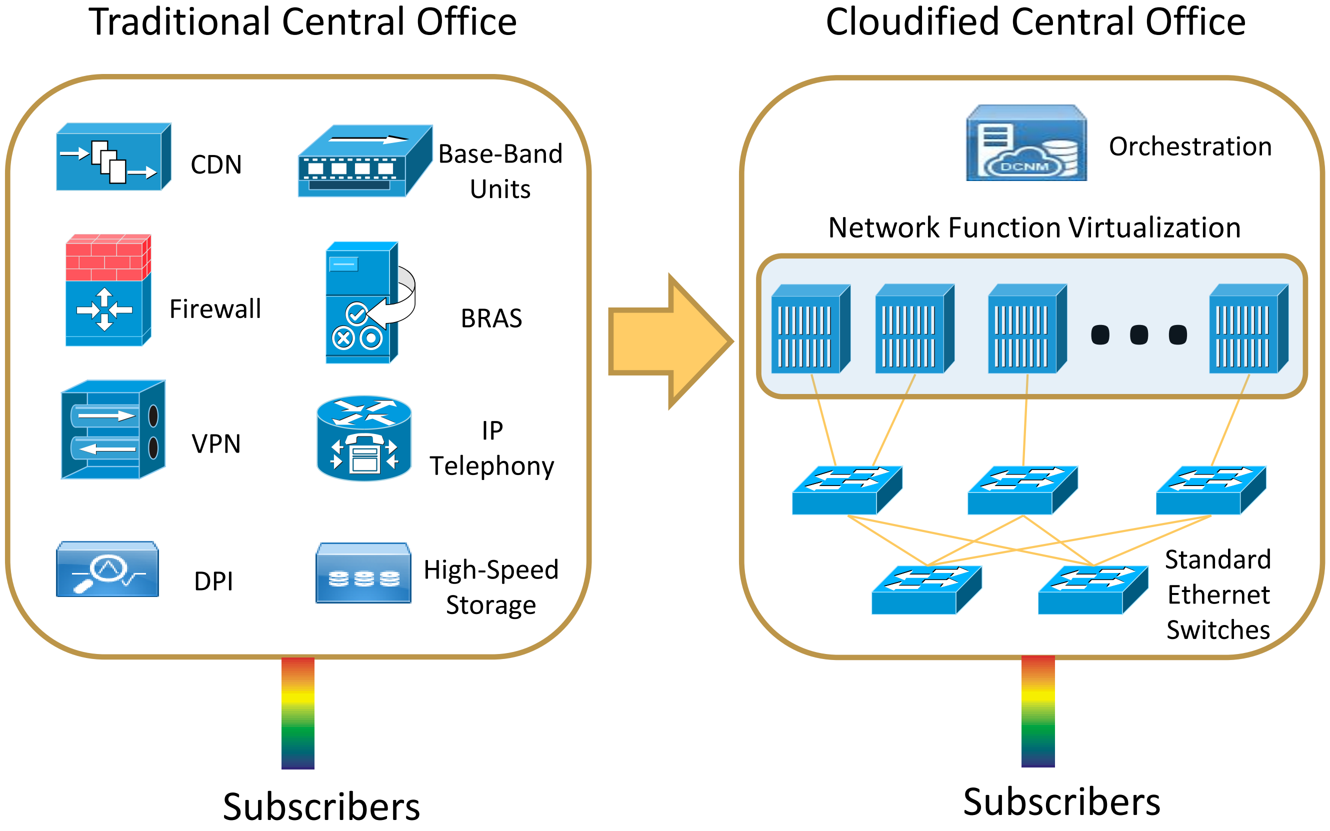
Critical Use Case – NVMe Over Fabrics





- Disaggregated resource pooling, such as NVMe over Fabrics, use RDMA and run over converged network infrastructure.
- Low latency and lossless are critical.
- Ease of deployment and cloud scale are important success factors.

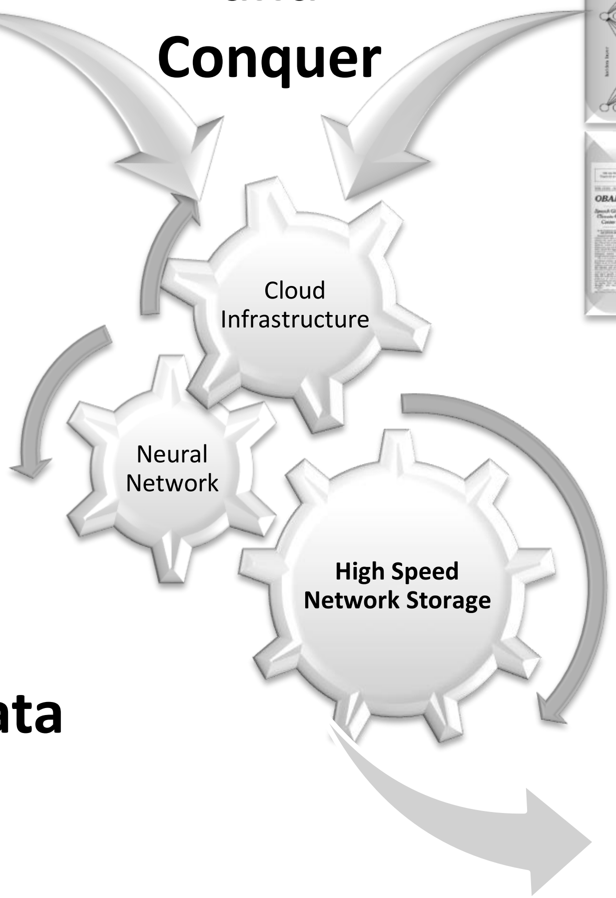
Critical Use Case – Cloudification of the Central Office

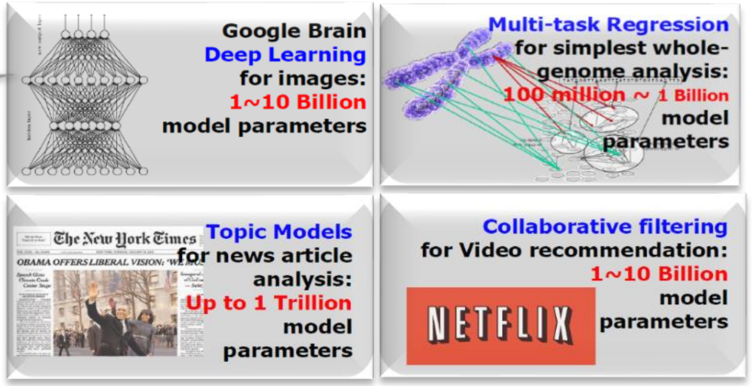


- Massive growth in Mobile and Internet traffic is driving Infrastructure investment
- To meet performance requirements of traditional purpose built equipment, SDN and NFV must run on low-latency, low-loss, scalable and highly available network infrastructure

We are dealing with massive amounts of data and computing

Divide and Conquer



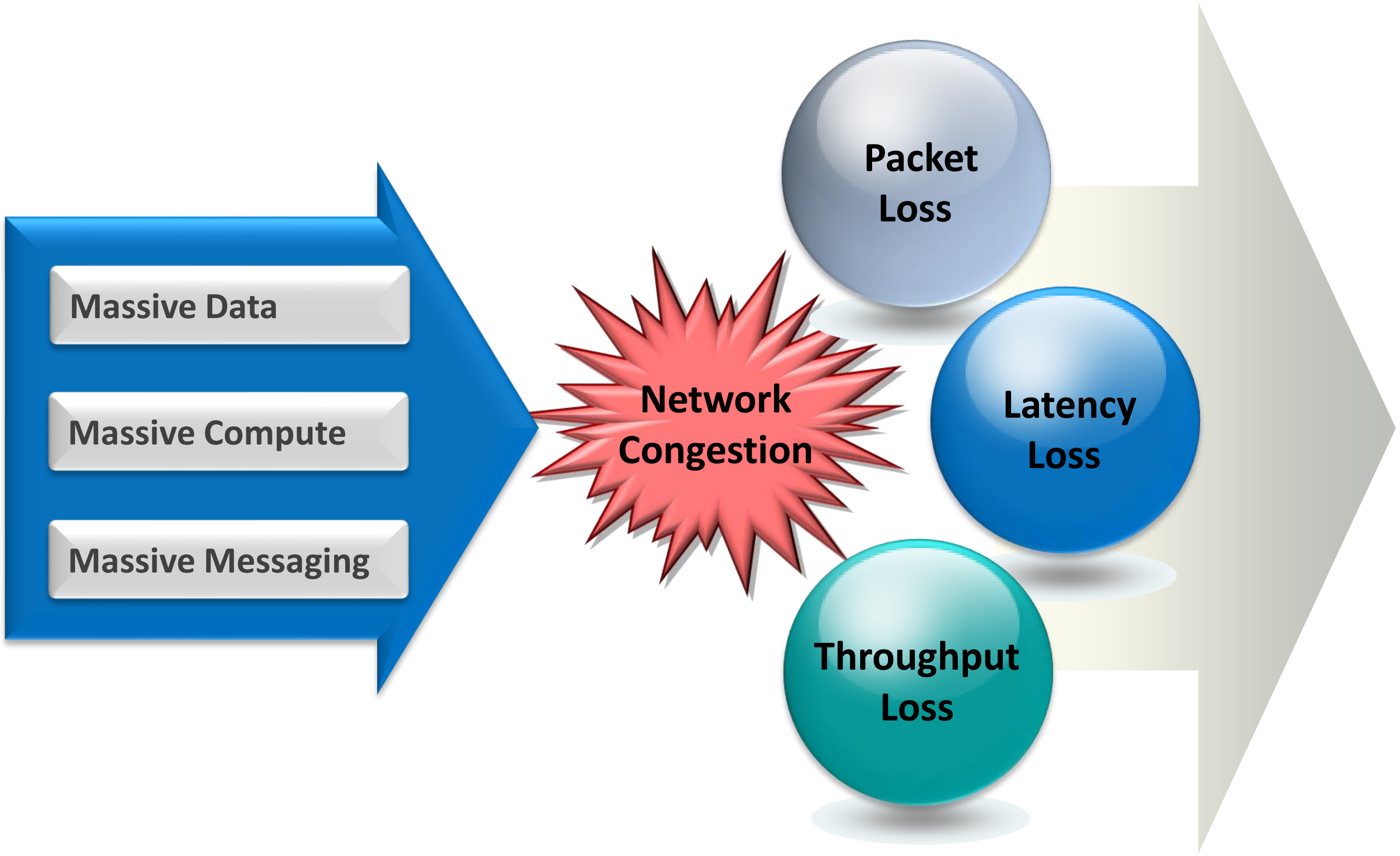


Requirements:

- **Fast-scalable storage**
- **Parallel applications and data**
- **Cloud-ified Infrastructure**

Real-time Natural Human/Machine Response

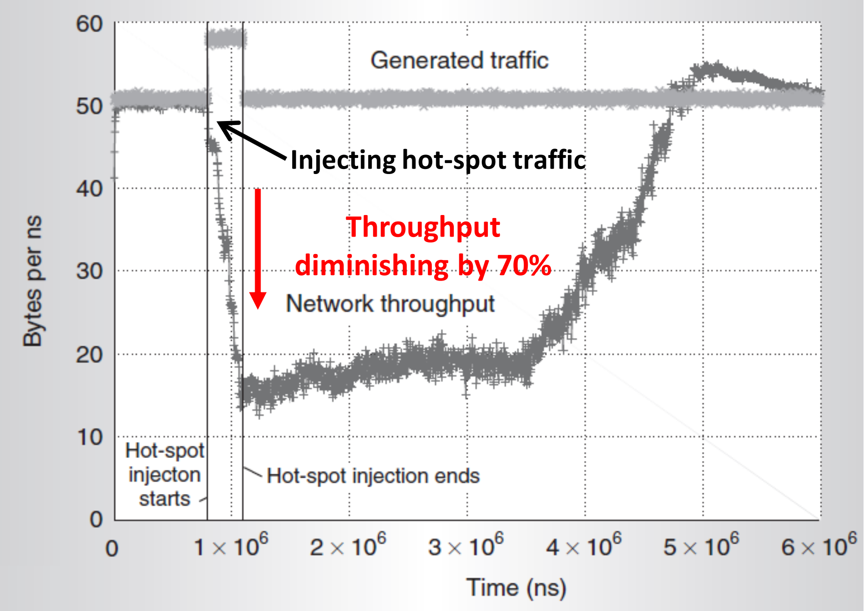
Congestion Creates the Problems



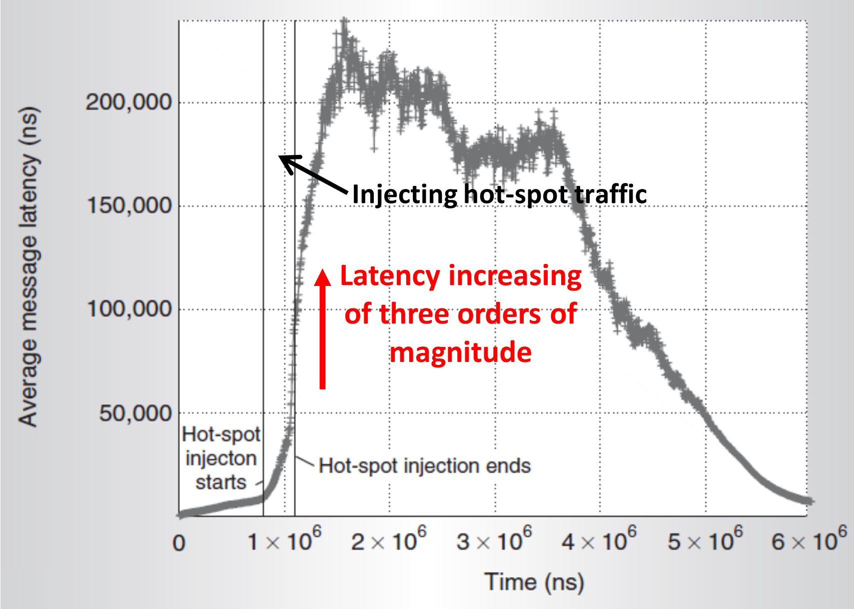
Parallelism can create congestion which leads to loss making end-user unhappy

The Impact of Congestion in Lossless Network

- The impact of congestion on network performance can be very serious.
- As shown in paper (Pedro J. Garcia et al, IEEE Micro 2006)[1]:



Network Throughput and Generated Traffic



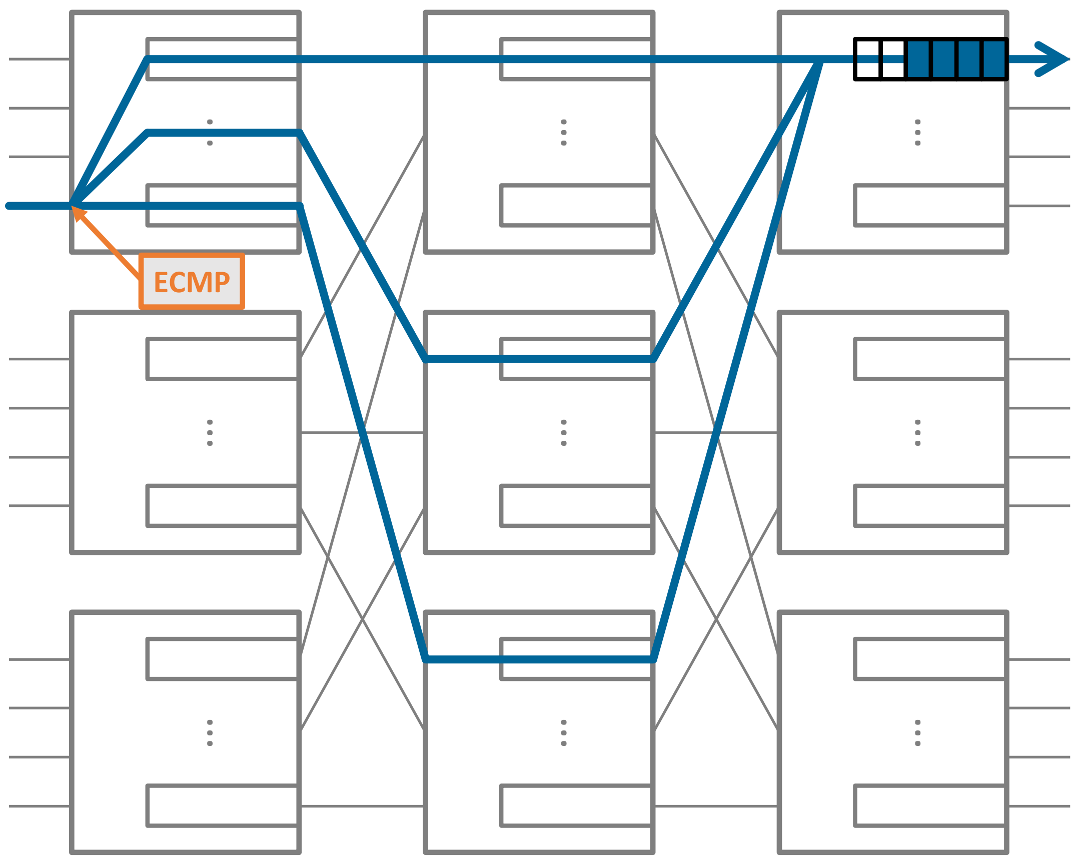
Average Packet Latency

Network Performance Degrades Dramatically after Congestion Appears

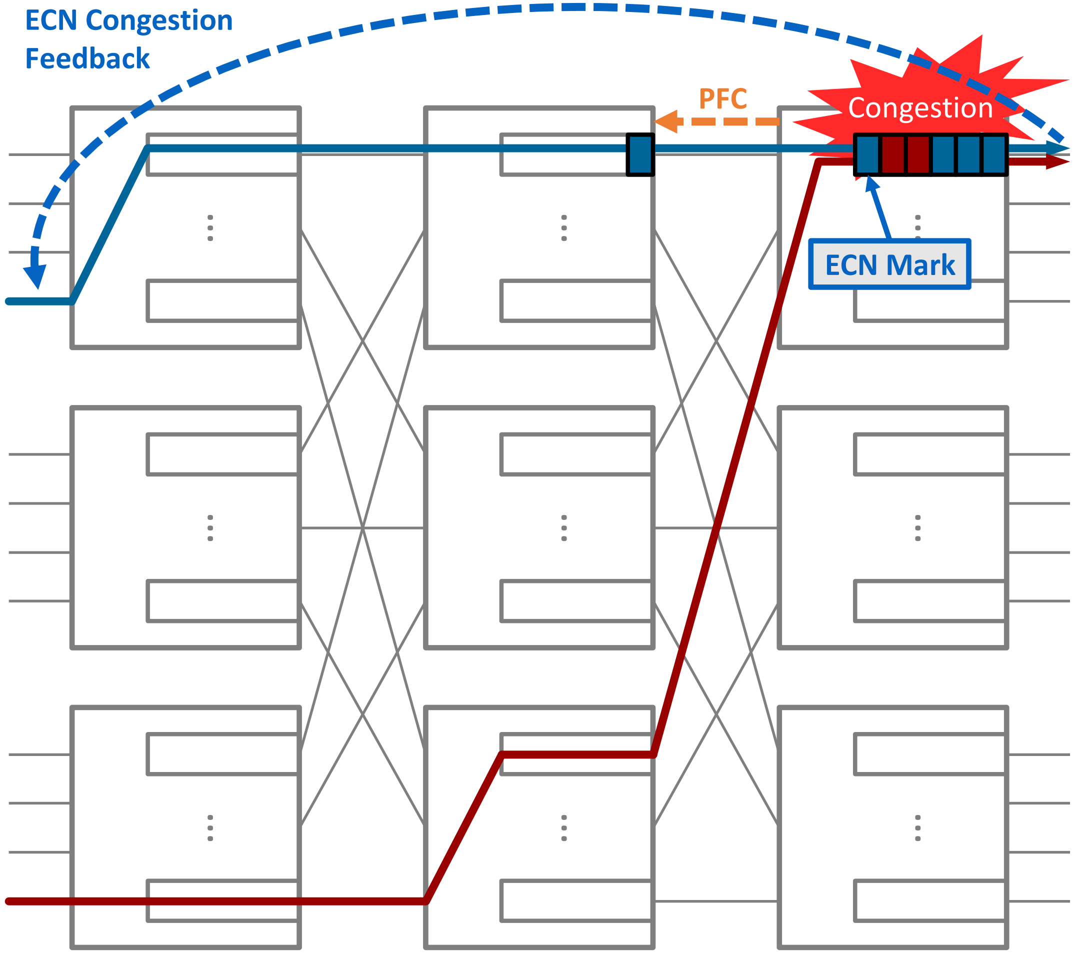
[1] Garcia, Pedro Javier, et al. "Efficient, scalable congestion management for interconnection networks." *IEEE Micro* 26.5 (2006): 52-66.

Dealing with Congestion today

ECMP – Equal Cost MultiPath Routing

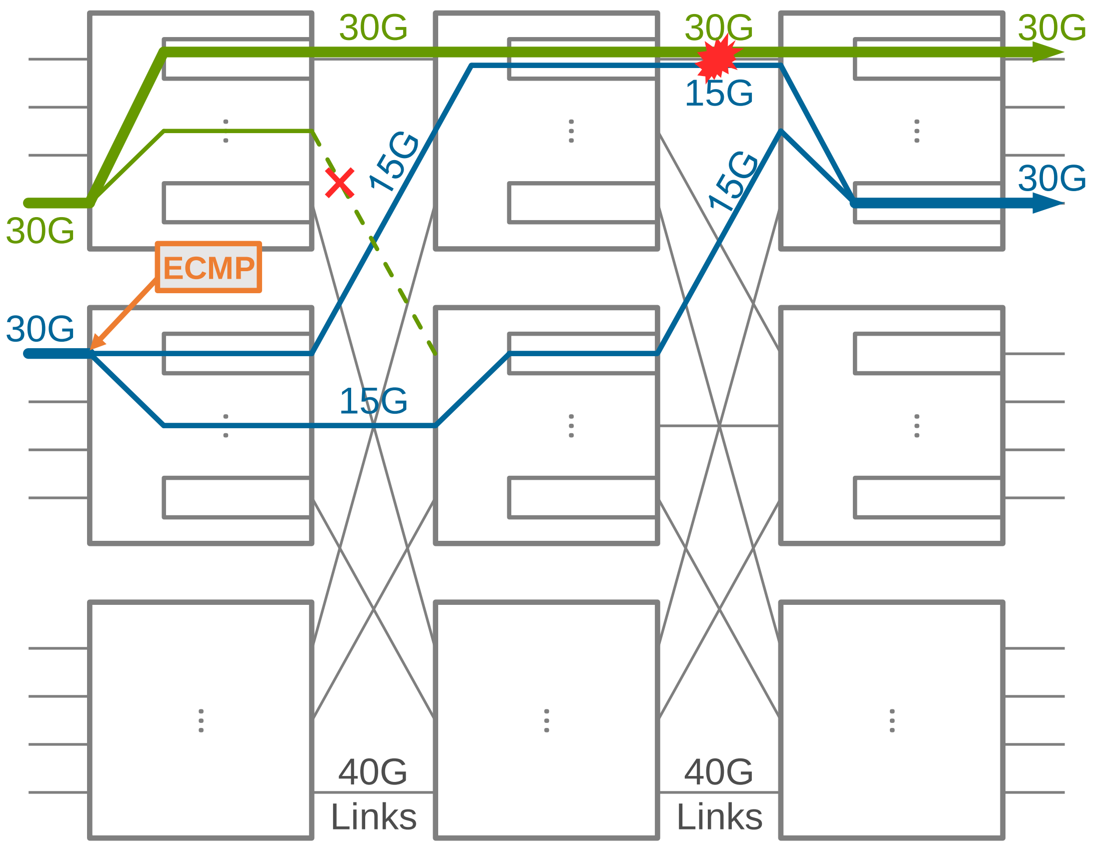


Explicit Congestion Notification (ECN) +
Priority-based Flow Control (PFC)

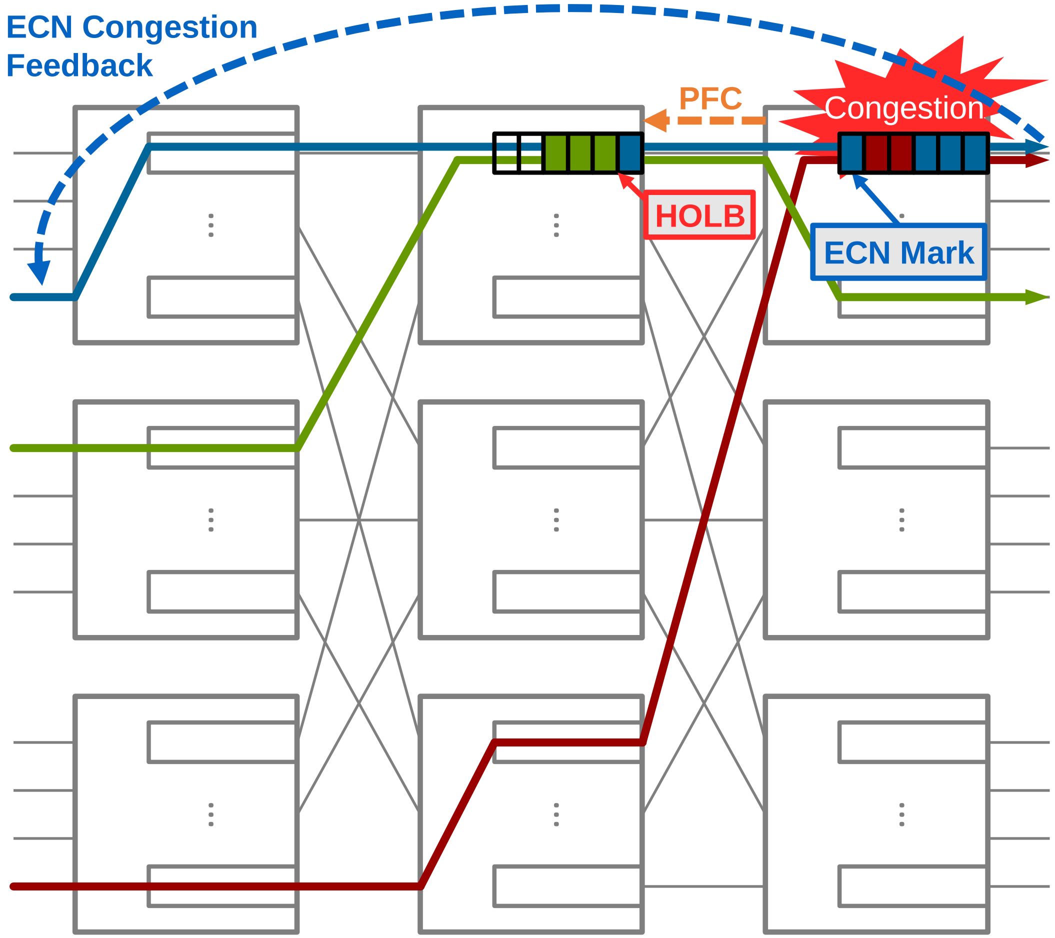


Ongoing challenges with congestion

ECMP Collisions



ECN Control Loop Delay
Head-of-line Blocking



Potential New Lossless Technologies for the Data Center

Goal = No Loss

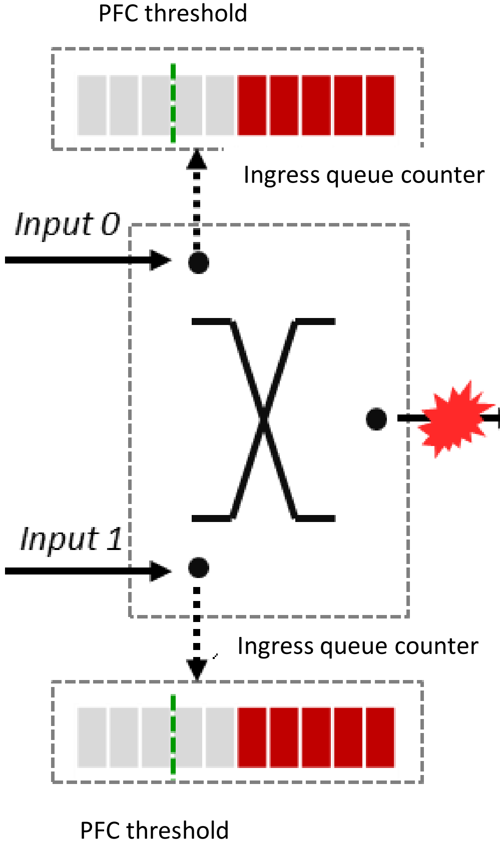
- No Packet Loss
- No Latency Loss
- No Throughput Loss

Solutions

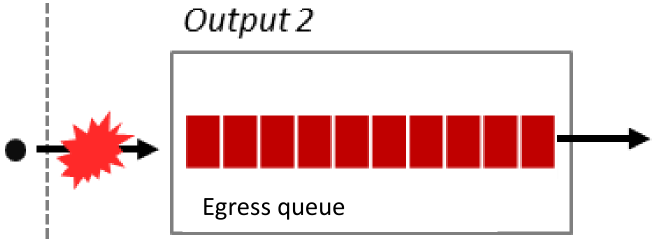
- Virtual Input Queuing - VIQ
- Dynamic Virtual Lanes - DVL
- Load-Aware Packet Spraying - LPS
- Push & Pull Hybrid Scheduling - PPH

VIQ (Virtual Input Queues)： Resolve Internal Packet Loss

Incast Congestion leading to internal packet loss

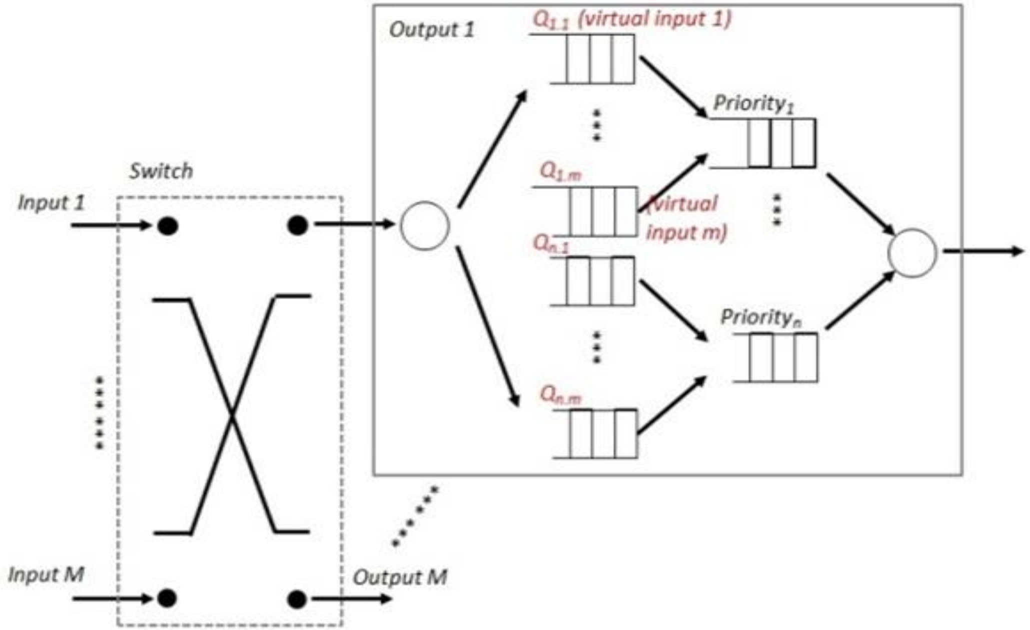


1. During incast scenario, ingress queue counter doesn’t exceed the PFC threshold, so will not send PFC Pause frame to upstream. Packet will always come in from ingress port.



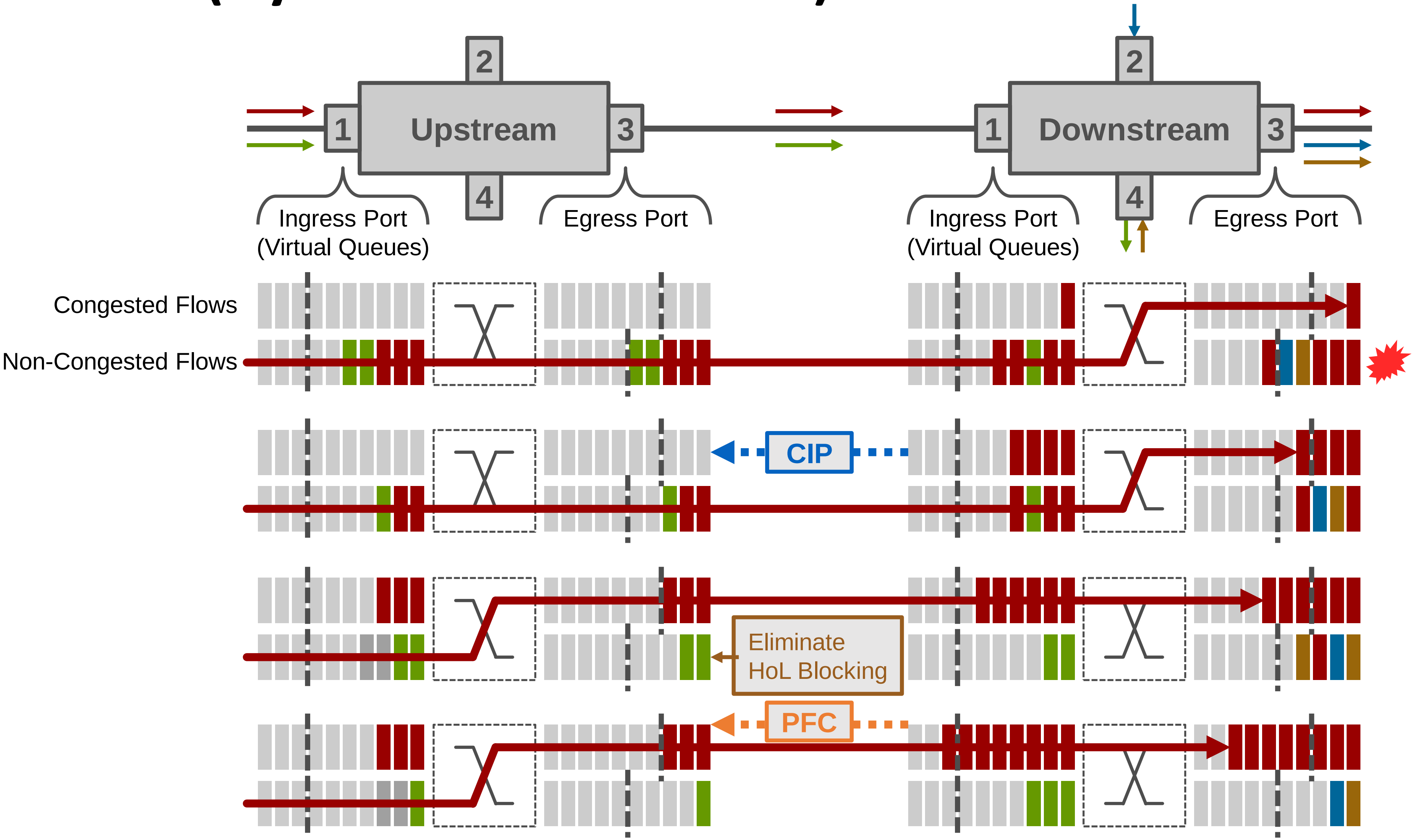
2. But the physical egress queue has backlog because of convergence effect. Packet loss occurs without egress-ingress coordination.

Coordinated egress-ingress queuing



VIQ could be looked as: that on out port, assign a dedicated queue for every in port. Memory changes from sharing to virtually monopolized according to in ports. So that every in port could get fair scheduling. The tail latency of business could be controlled effectively.

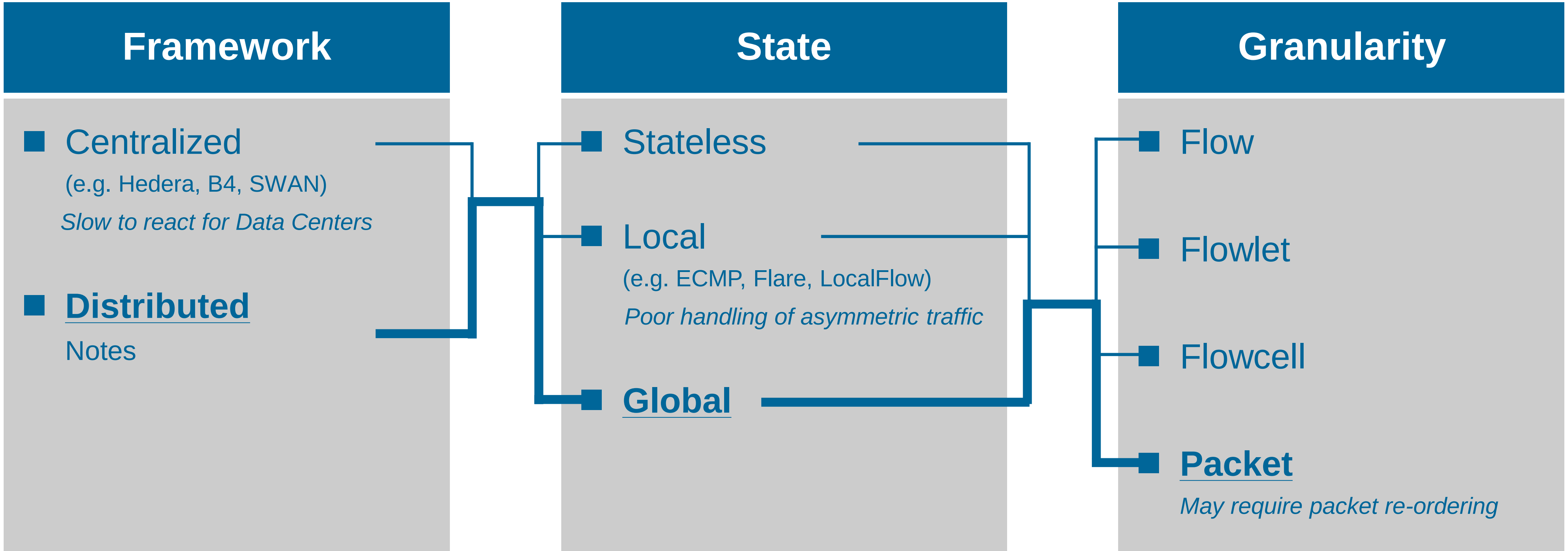
DVL (Dynamic Virtual Lanes)



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

LPS (Load-Aware Packet Spraying)

Load Balancing Design Space



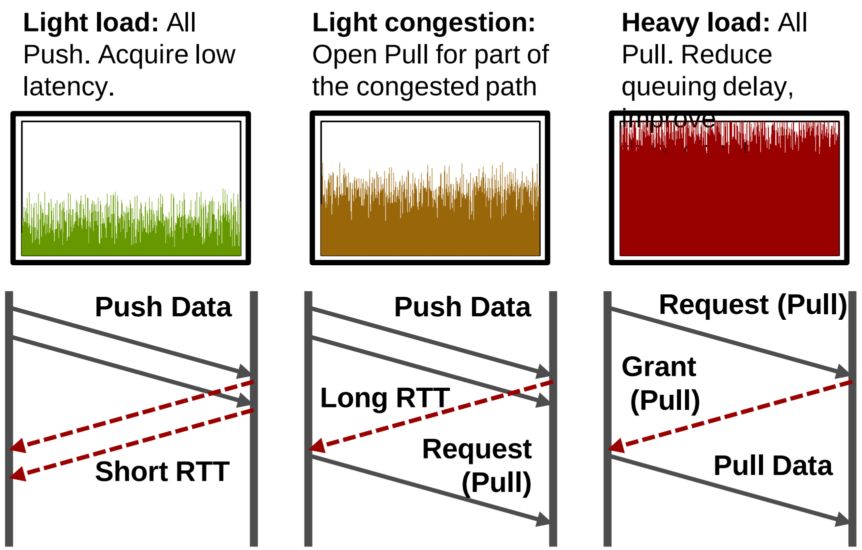
LPS = Packet Spraying + Endpoint Reordering + Load-Aware

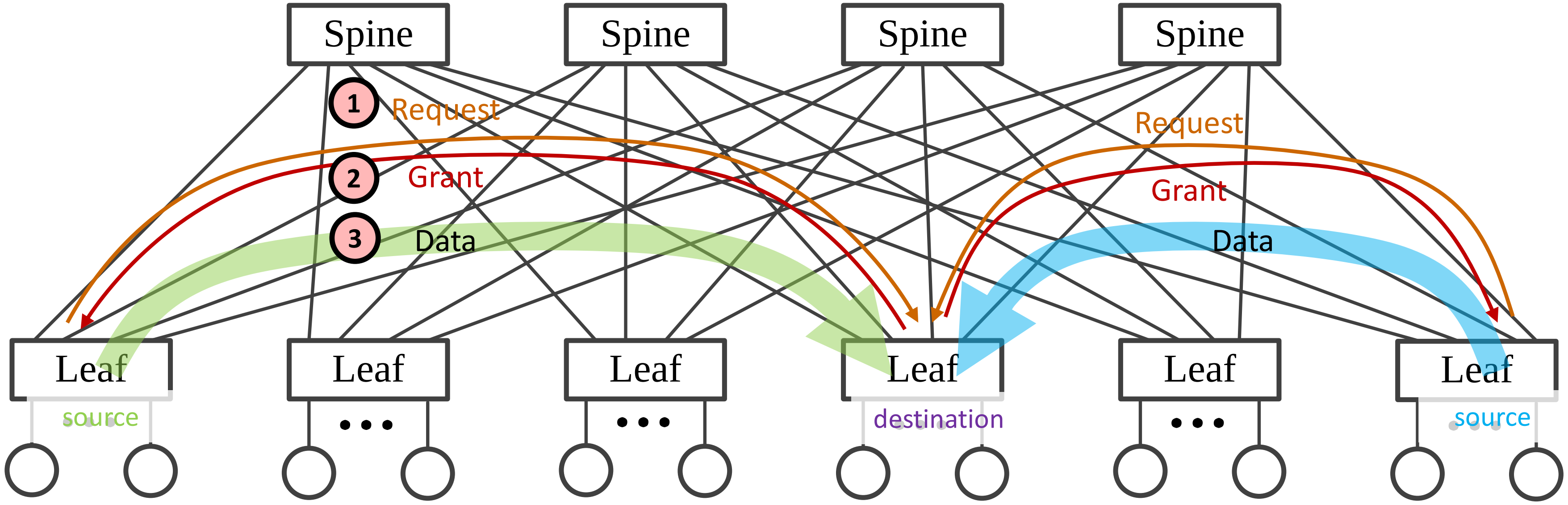
PPH (Push & Pull Hybrid Scheduling)

PPH = Congestion aware traffic scheduling

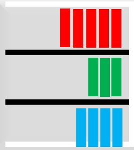
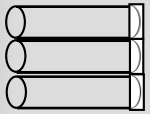
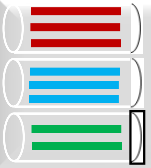
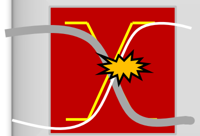
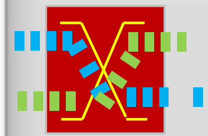
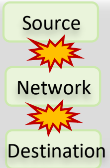
Push when load is light

Pull when load is high





Innovation for the Lossless Network

Congestion Impact		Mitigating Congestion	Innovation
 Ingress thresholds unrelated to egress buffer availability. Incast causes internal packet loss.	**Coordinated Resources**	 Coordinate egress availability with ingress demand. Avoid internal switch packet loss	**Virtual Input Queues**
 Priority-based Flow Control (Coarse grain). Victim flows hurt by the congested flows	**Isolate Congestion**	 Allow time for end-to-end congestion control. Move congested flows out of the way. Eliminate head-of-line blocking.	**Dynamic Virtual Lane**
 Unbalanced load sharing. Elephant flow collisions block mice flows.	**Spread the Load**	 Load-balance flows at higher granularity. Use congestion awareness to avoid collisions	**Load-aware Packet Spraying**
 Unscheduled and network resource unaware many-to-one communication leads to incast packet loss	**Schedule Appropriately**	 Scheduling decision integrated the information from source, network and destination.	**Push & Pull Hybrid Scheduling**

Thank You