Towards Hyperscale High Performance Computing with RDMA

NANOG 76

2019.06.12

Omar Cardona - Microsoft

Outline

- Drivers for making RDMA and HPC a critical part of modern cloud networks
- Trends and directions for network storage (SCM) and CPUs (GPUs, custom for HPC)
- RDMA fundamentals and Fabric impacts
- What are some problems with today's solution that keep it from scaling?
  - Go-back N makes packet loss a huge penalty
  - Configuring a lossless network is a challenge
  - PFC and HoL Blocking problems
  - Delays in end-to-end control loop
  - Mixing flows with different congestion controllers (TCP & RoCE)
  - Traffic Class separation
  - Persistent memory need for RDMA
- Discussion

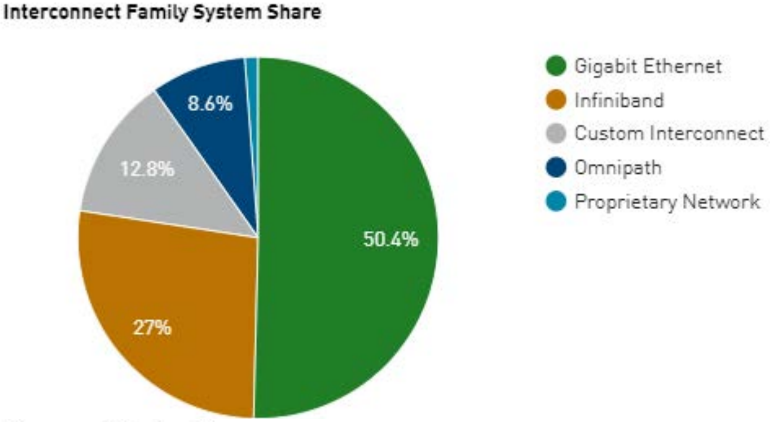
Current HPC/RDMA networks

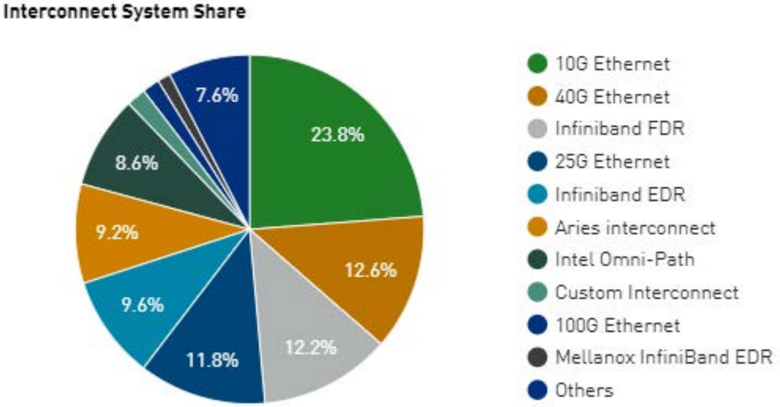
“Future datacenters of all kinds will be built like high performance computers”

-Nvidia CEO, Jensen Huang

- Traditional HPC runs over custom lossless technologies
  - Infiniband - with L2 credit-based Flow Control
- Increasingly also runs over IP infrastructure
  - iWARP - RDMA over HW Offload TCP
  - RoCEv2 – Infiniband Transport over Converged Ethernet
- Benefits applicable via integration in:
  - *artificial intelligence*
  - *machine learning*
  - *data analytics*
  - *data science workloads*

TOP 500 The List.

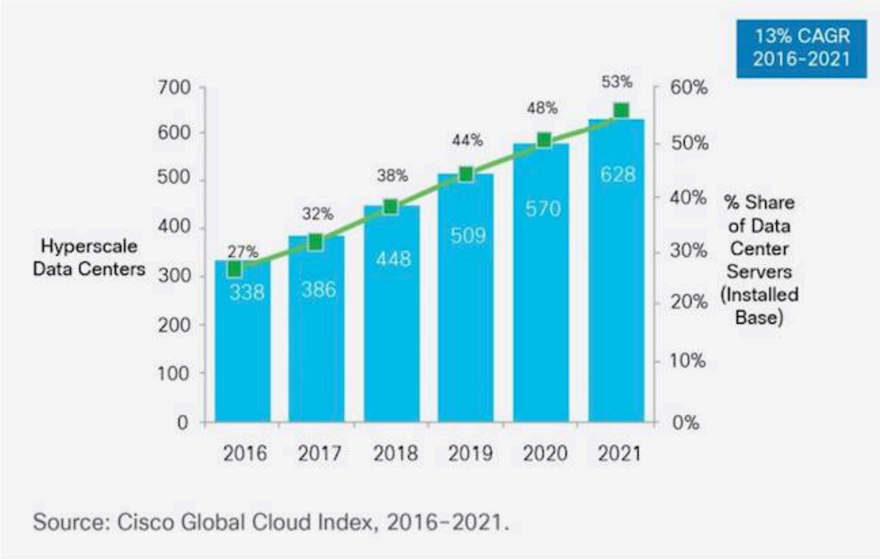




What does it mean to be Hyperscale

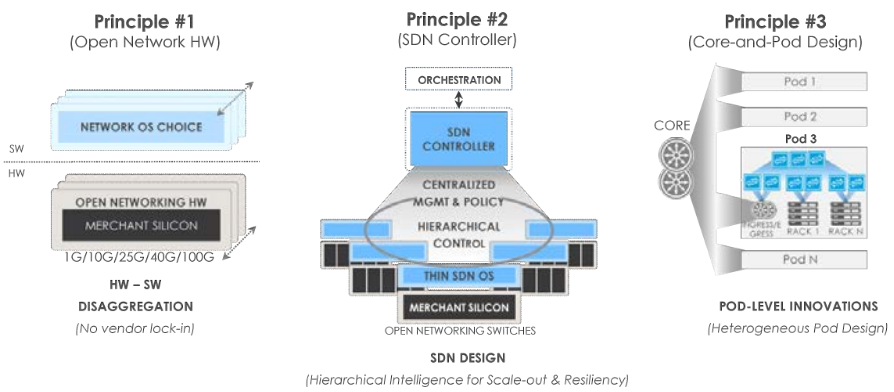
Architecture's ability to scale with increasing demand.

- Common scale infrastructure
- Dynamic and automated provisioning
- Diverse workload mix
- Service Level Agreements
    - Consistency, Low-latency, high-throughput

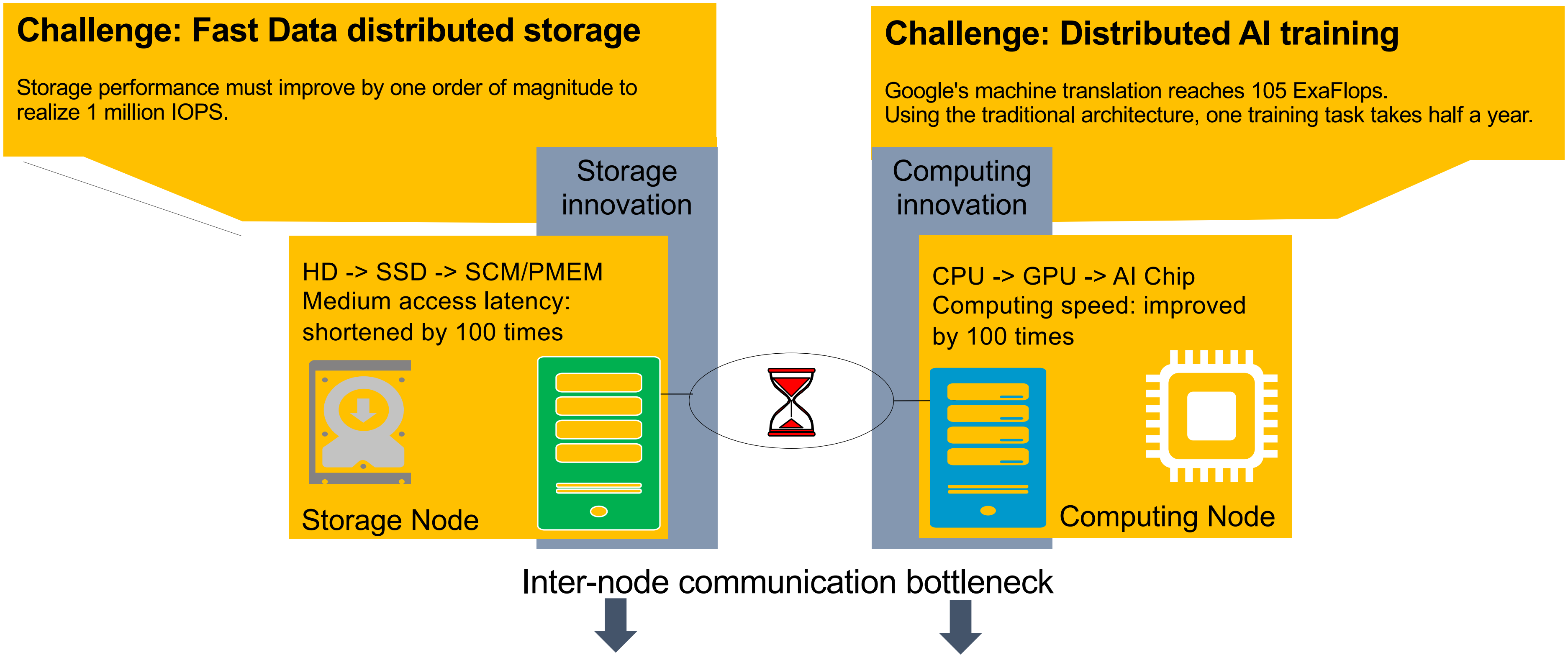


Source: Cisco Global Cloud Index, 2016-2021.

HYPERSCALE DESIGN PRINCIPLES



Storage and CPU create network pressure



Challenge: Fast Data distributed storage

Storage performance must improve by one order of magnitude to realize 1 million IOPS.

Challenge: Distributed AI training

Google's machine translation reaches 105 ExaFlops.
Using the traditional architecture, one training task takes half a year.

Storage innovation

HD -> SSD -> SCM/PMEM
Medium access latency: shortened by 100 times

Storage Node

Computing innovation

CPU -> GPU -> AI Chip
Computing speed: improved by 100 times

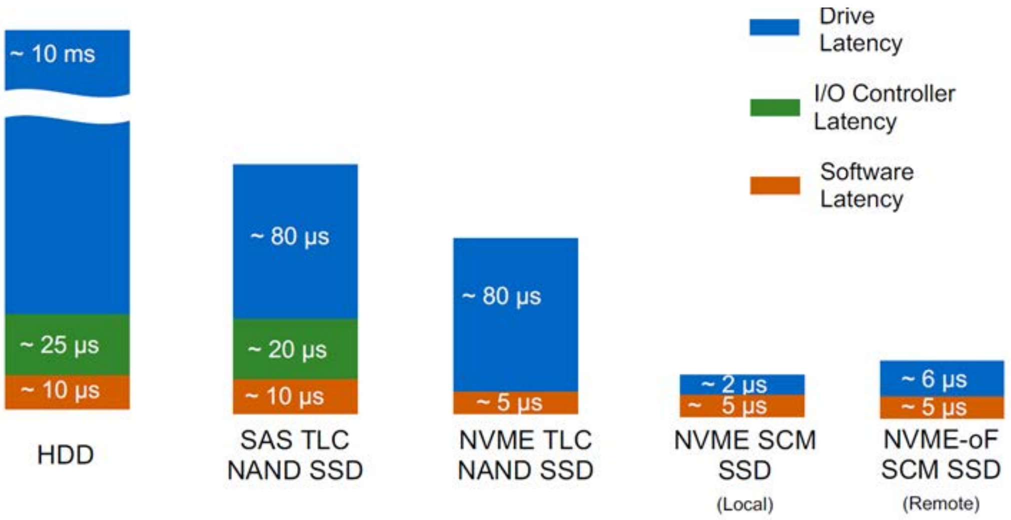
Computing Node

Inter-node communication bottleneck

Communication LAT **>50%** of the total storage access LAT.

Wait time for GPU communication **>50%** of the job completion time (JCT).

Remote Storage Class Memory (SCM)



SCM @ usec access

Requires <= usec network access

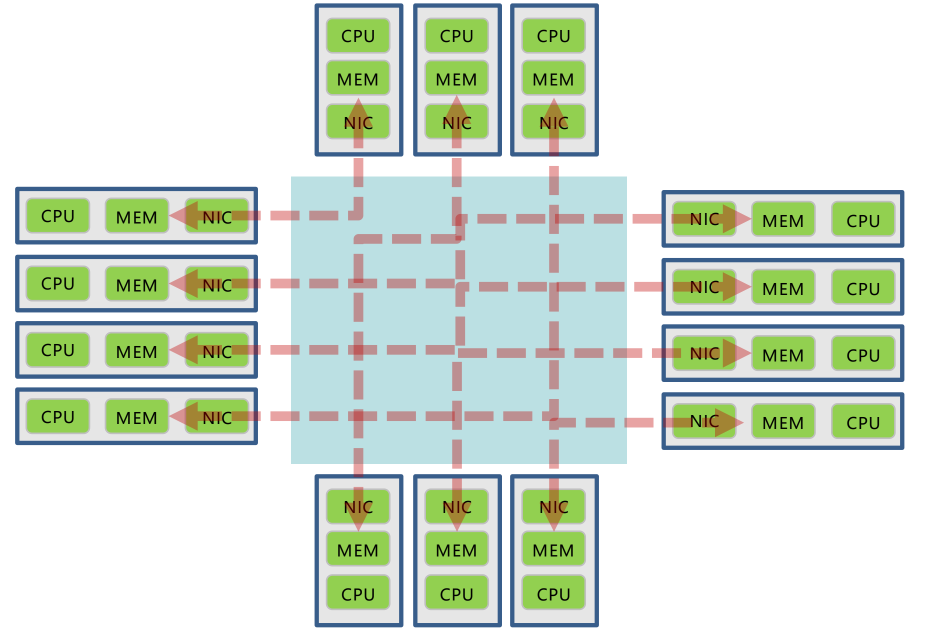
Network Options:

1. iWARP
2. RoCE
3. DPDK or similar

~~Host mediated access…~~

Packet loss stalls HPC Applications

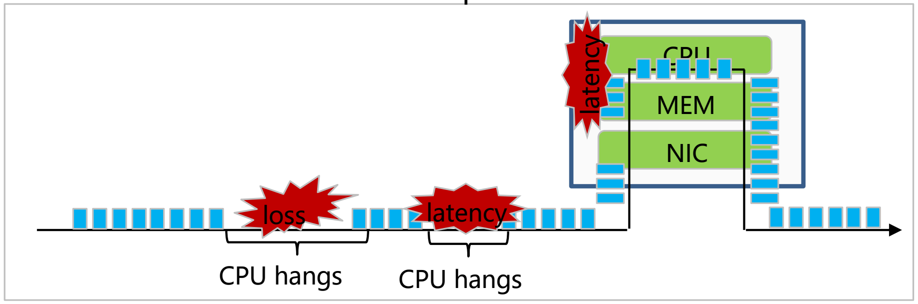
The key application of computing cloud: distributed HPC



- HPC requires network data copy
- Network LAT should ideally to memory access LAT
- RDMA provides lowest LAT solution

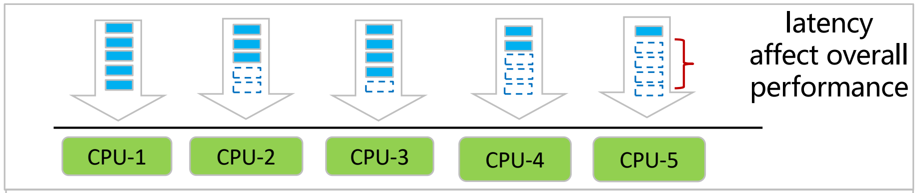
Single processor perspective:

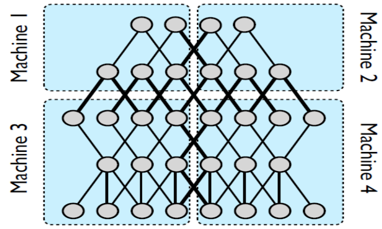
network loss -> workload interruption



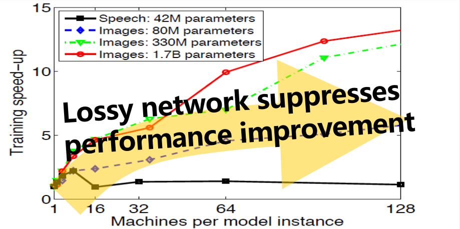
Multi-processors perspective:

Under the synchronous parallel compute model, the slowest data arrival drags down the overall performance





Google AlphaGo parallel compute model

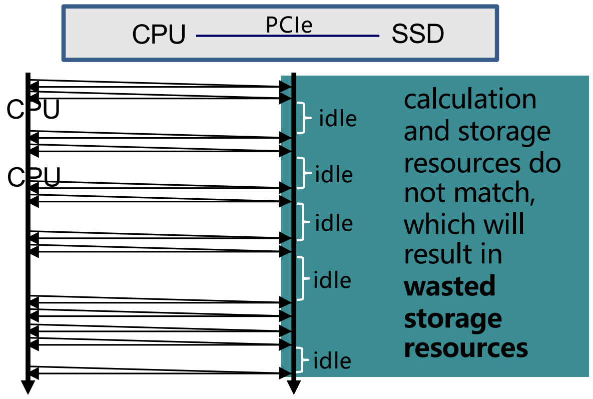


Limited by network capability, simply increasing processors does not lead to a linear increase in overall performance.

Packet loss stalls storage

Direct Storage

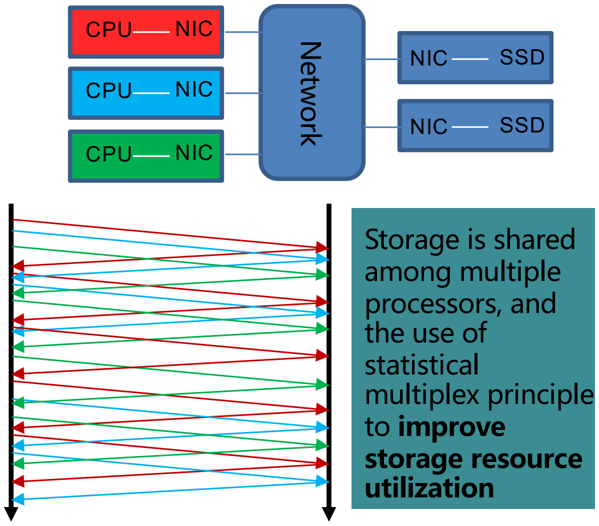
Inefficiency in storage resources



calculation and storage resources do not match, which will result in **wasted storage resources**

Network Storage

Improves resource utilization

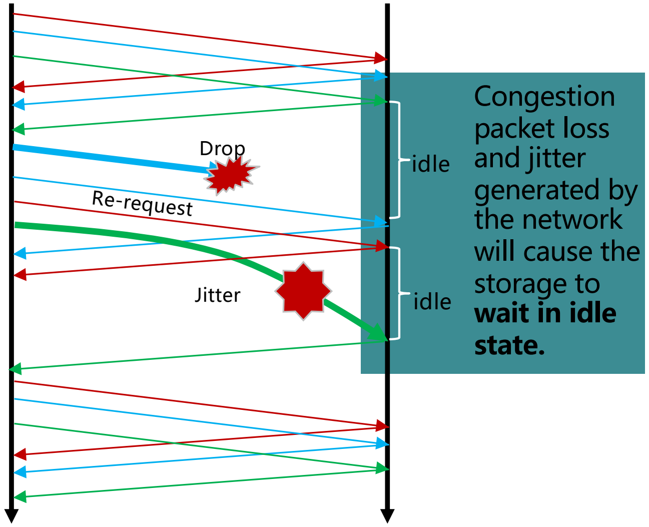


Storage is shared among multiple processors, and the use of statistical multiplex principle to **improve storage resource utilization**

- Storage cloud utilizes the principle of statistical multiplexing to improve storage resource utilization

Lossy Network

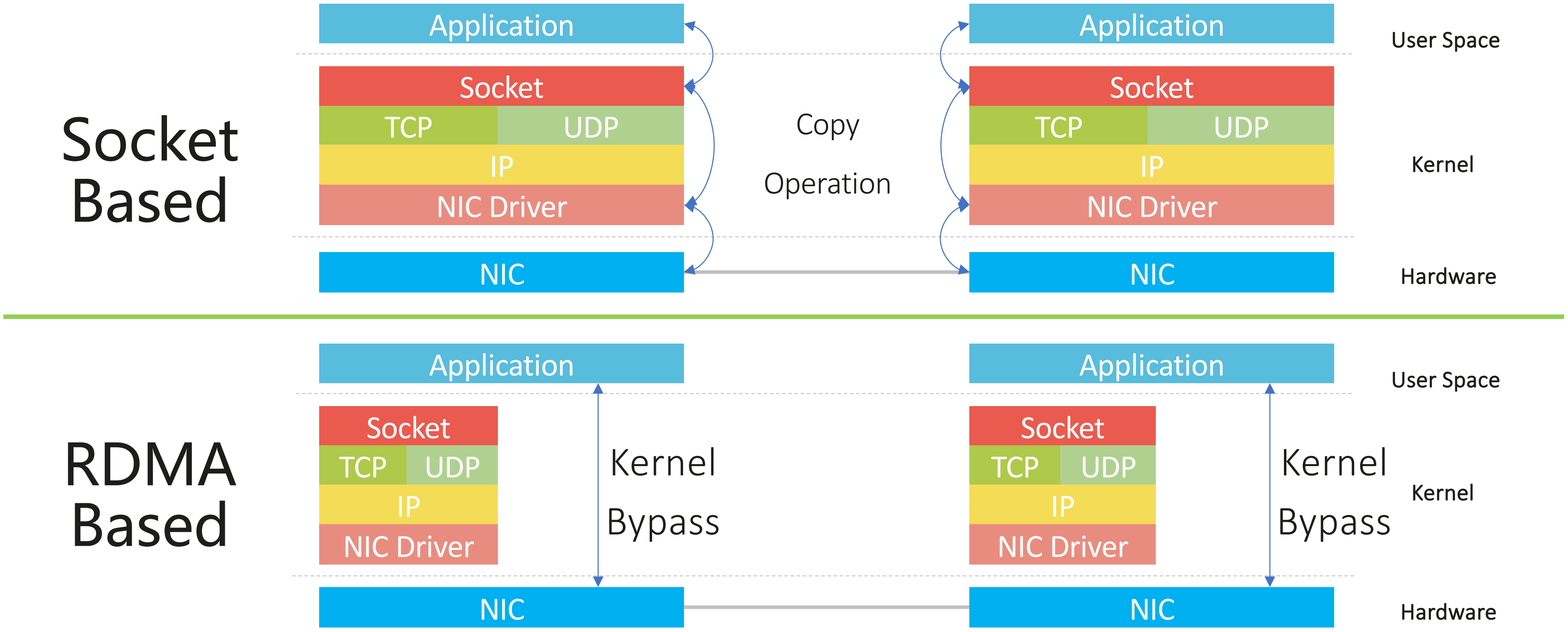
Impacts storage performance



Congestion packet loss and jitter generated by the network will cause the storage to **wait in idle state.**

- Storage cloud based on lossy network will cause access delay due to network congestion, packet loss, and jitter, which seriously affects the effect of storage cloud.

RDMA vs Traditional Messaging



Socket Based

Application | Socket | TCP | UDP | IP | NIC Driver | NIC

Copy Operation

Application | Socket | TCP | UDP | IP | NIC Driver | NIC

User Space

Kernel

Hardware

RDMA Based

Application | Socket | TCP | UDP | IP | NIC Driver | NIC

Kernel Bypass

Application | Socket | TCP | UDP | IP | NIC Driver | NIC

Kernel Bypass

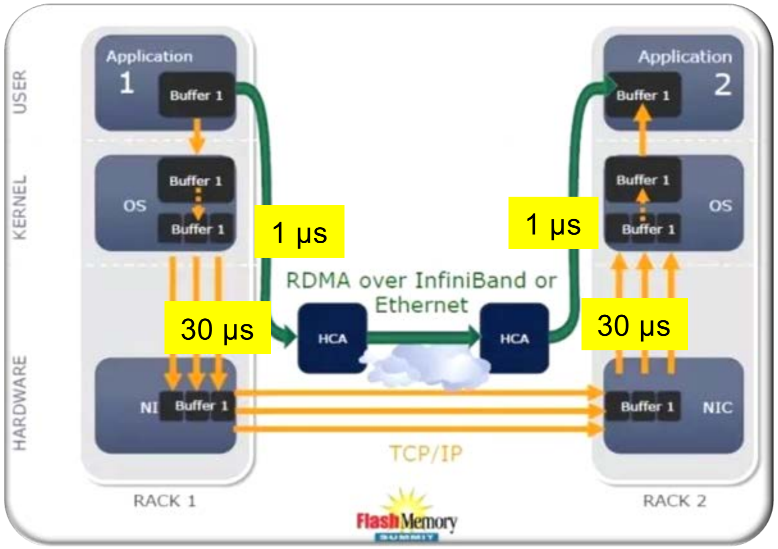
User Space

Kernel

Hardware

RDMA eliminates: Context Switch, Intermediate Data Copies, and Protocol Processing

RDMA is an essential protocol for the AI era



TCP disadvantages

- Three copy operations, resulting in a long latency
- CPU consumed by traffic: 1 Hz per bit

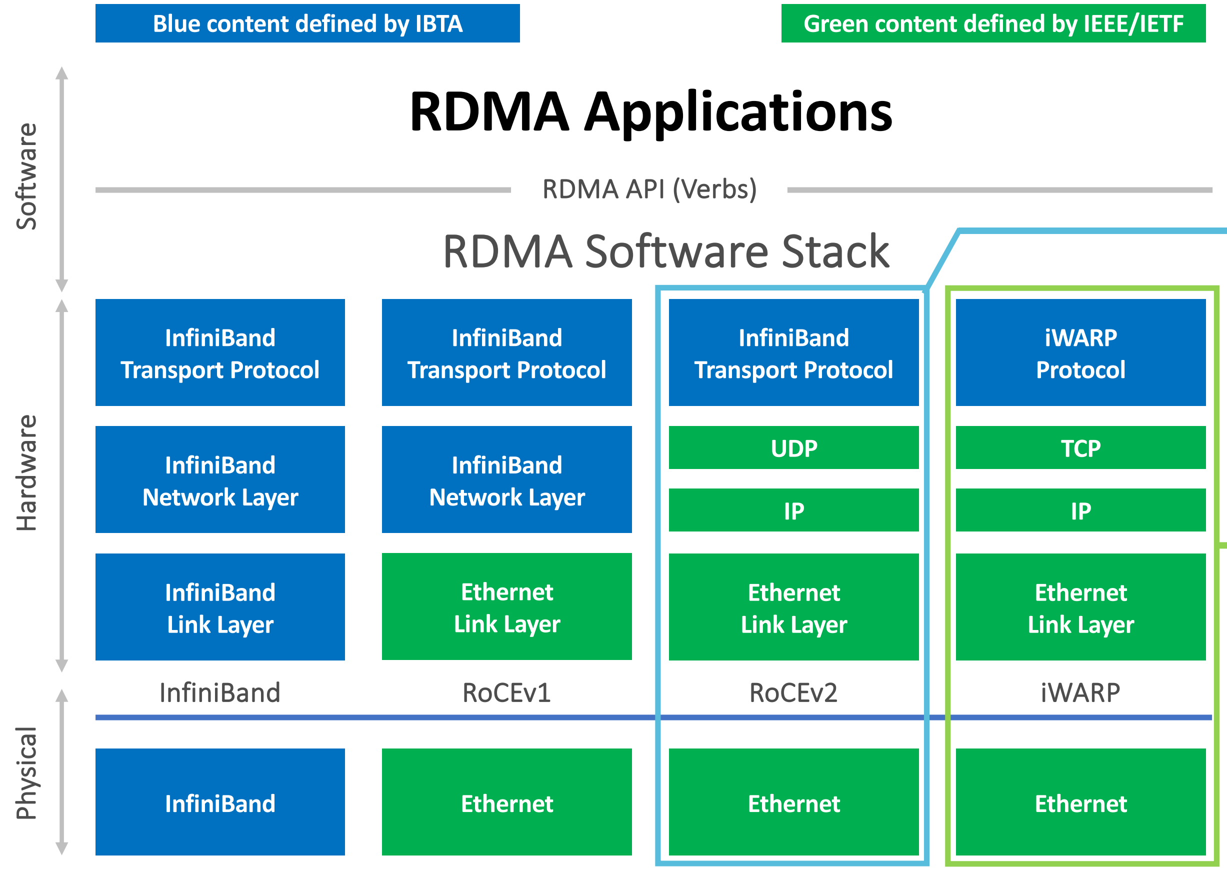
RDMA advantages

- Fast startup, maximizing the bandwidth usage
- One copy operation (DMA), effectively reducing the kernel latency
- Minimal <5% CPU resources consumed for Kernel transfers.
- ~0 kernel CPU usage for Userspace RDMA

- 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

RDMA advantages are most effective in reduced latency and minimal cycles/byte costs

RoCEv2 and iWARP are RDMA over Ethernet



Blue content defined by IBTA

Green content defined by IEEE/IETF

Software

RDMA Applications

RDMA API (Verbs)

RDMA Software Stack

	InfiniBand	RoCEv1	RoCEv2	iWARP
Hardware	InfiniBand Transport Protocol	InfiniBand Transport Protocol	InfiniBand Transport Protocol	iWARP Protocol
	InfiniBand Network Layer	InfiniBand Network Layer	UDP IP	TCP IP
	InfiniBand Link Layer	Ethernet Link Layer	Ethernet Link Layer	Ethernet Link Layer
Physical	InfiniBand	Ethernet	Ethernet	Ethernet

RoCEv2 vs InfiniBand:

- Eliminates IB fabric req
- Minimal Congestion Control Capability

RoCEv2 vs RoCEv1:

- Support L3 Forwarding
- Requires Lossless L2 - DCB:PFC/ETS + ECN

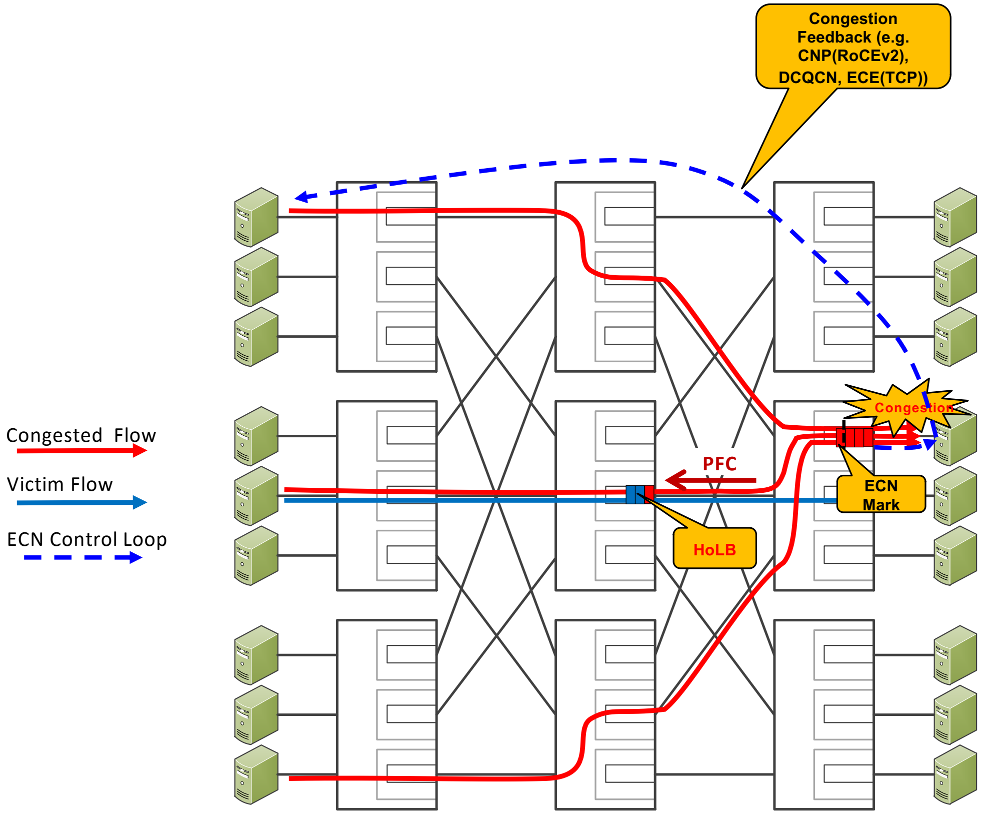
iWARP vs RoCEv2:

- L4 based Congestion Control
- Uses standard TCPIP CC and SAK
- No L2 lossless req – DCB:ETS optional

Marketing spins of RoCEv2 vs iWARP:

- ~~Easy~~
- ~~Cost-effective~~
- ~~Better Performance~~

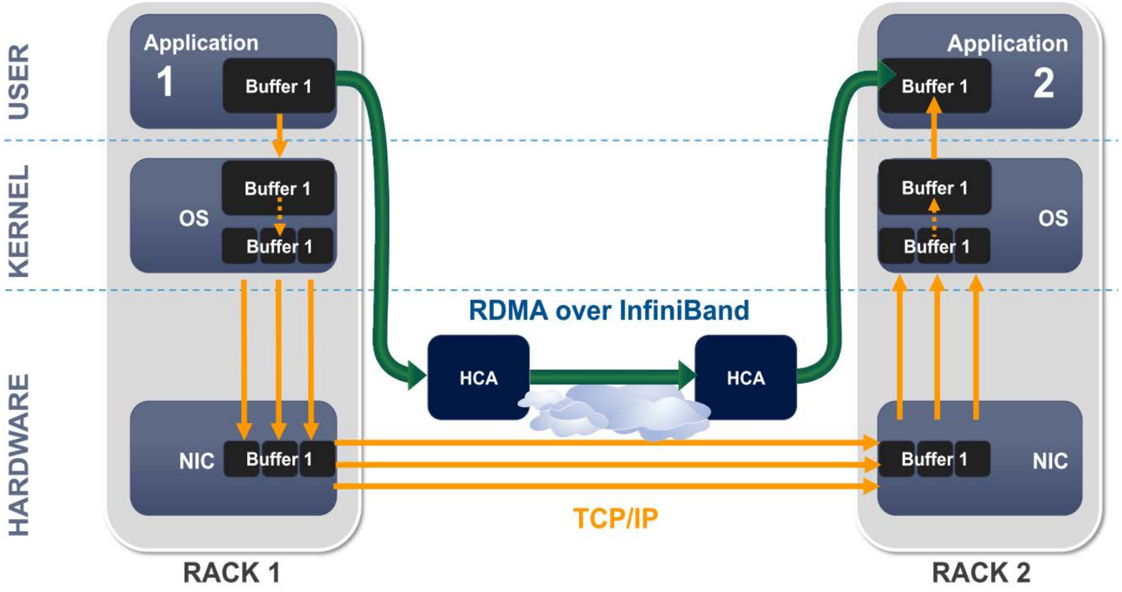
Basic RoCEv2 Network: DCB:ETS / PFC + ECN

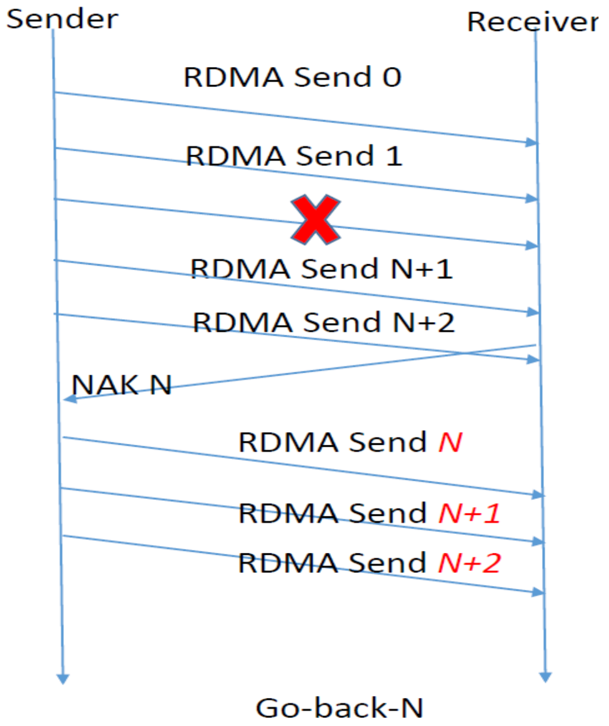


- **ECN** - Explicit Congestion Notification
  - End-to-end congestion control
  - **CNP** - Congestion Notification Packet
    - Feedback at connection Granularity
    - Source quench @ Source Queue Pair
- **PFC** – Priority Flow Control
  - Last resort to ensure lossless environment
  - May cause L2 congestion spreading if improperly configured
- **ETS** - Enhanced Transmission Selection
  - 802.1p COS, 8 classes
  - Traffic Class Egress BW reservation

RoCE Congestion Control

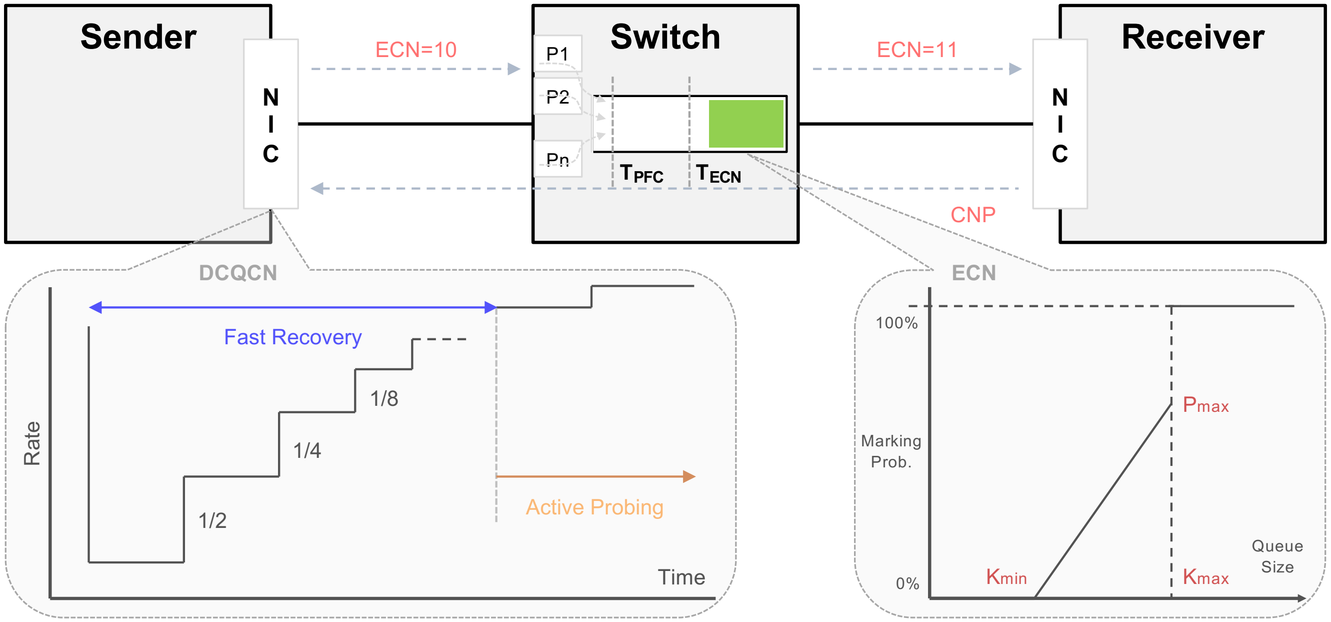
- No slow start to sample initial load
- No Selective Acknowledgement for retransmits
- Uses Go-Back-N batch retransmits

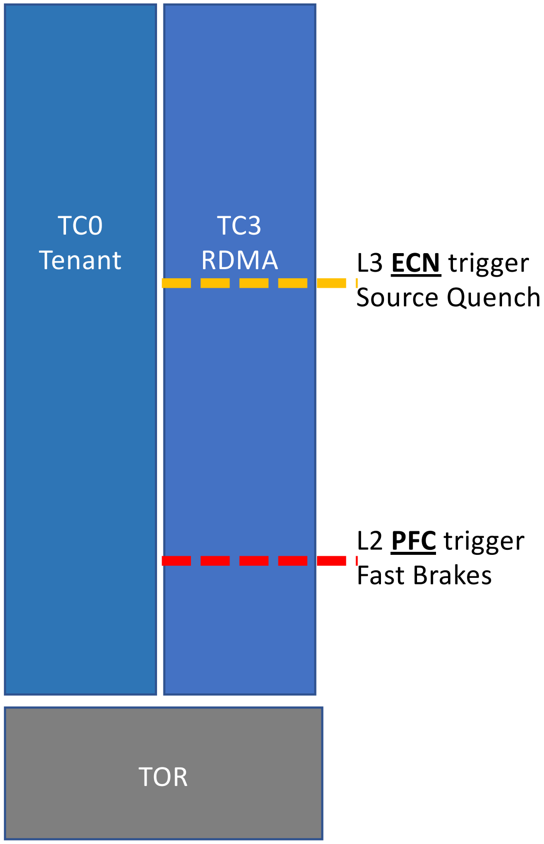




- Go-back-n window may cause fabric LIVELOCK if loss within window
- iWARP uses standard TCP Selective Acknowledge + Granular Fast-Retransmits

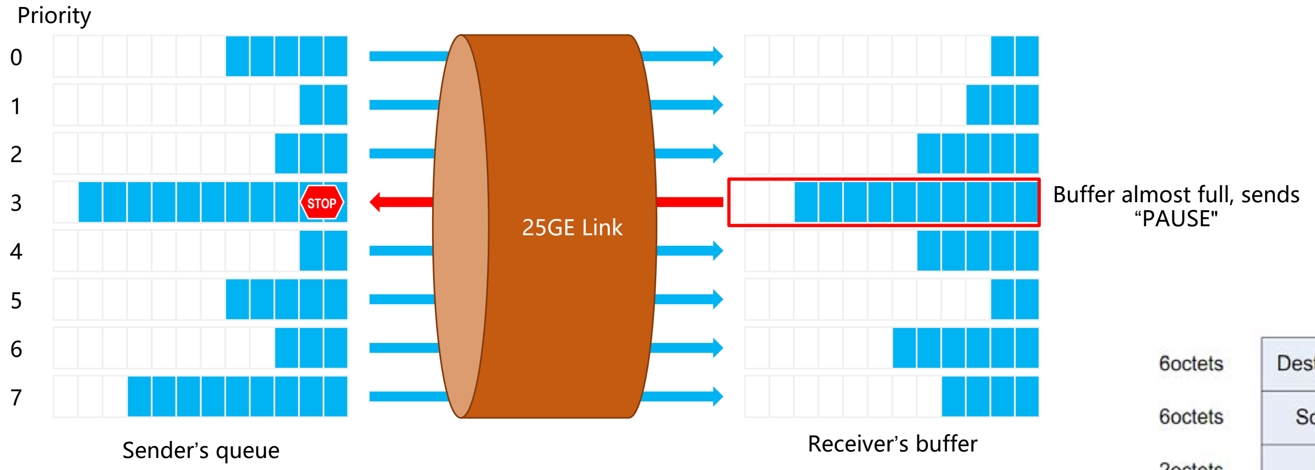
ECN – Explicit Congestion Notification

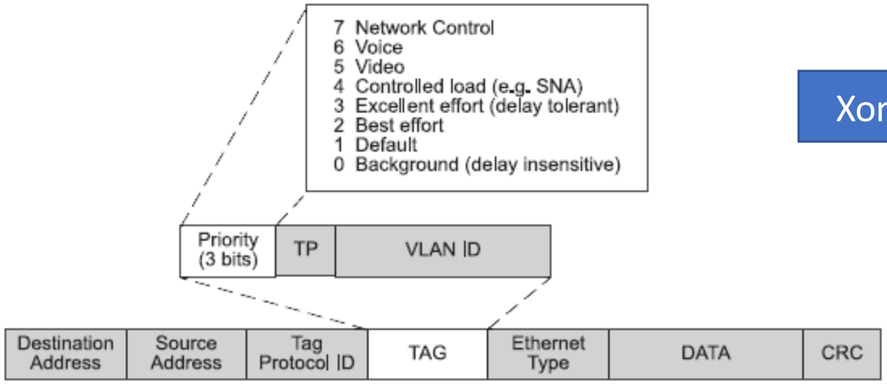




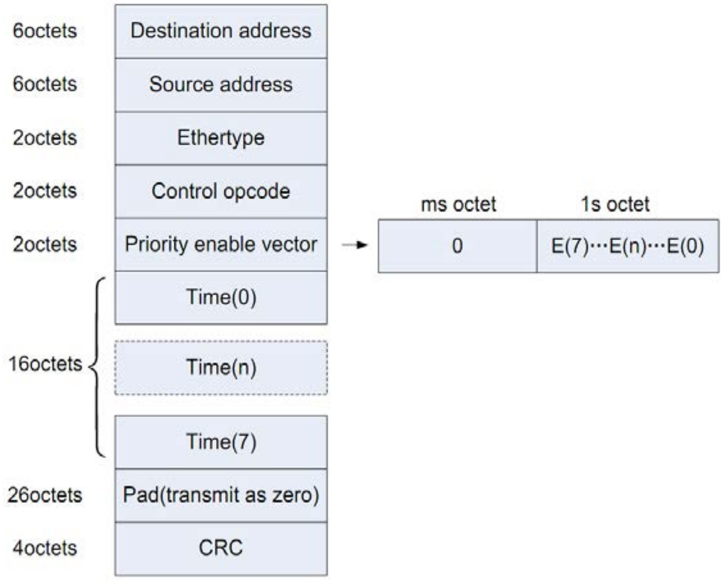
- **Forward Explicit Congestion Notification (FECN)**
- **Triggers must account for max BDP/RTT**

PFC - Priority base Flow Control





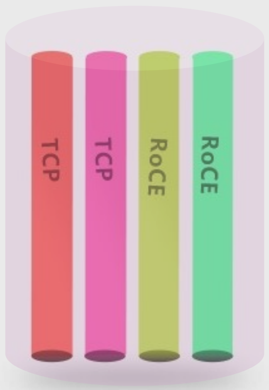
Xon/Xoff



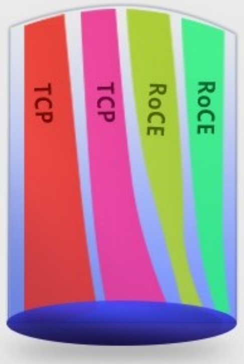
ETS – Enhanced Transmission Selection

Due to different congestion control differences, TCP and RoCE may preempt each other on egress

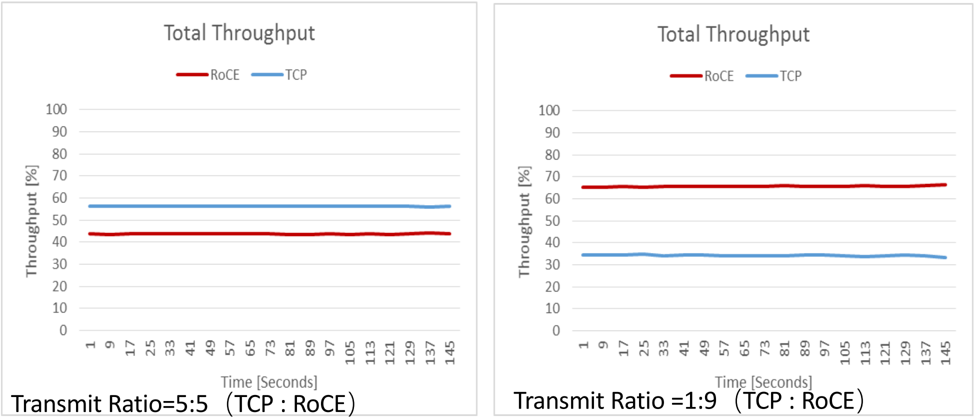
- Egress BW reservation per 802.1p Traffic Classes (TC)
  - Guaranteed minimum BW provided per TC
  - Typically via Distributed Weighted Round Robing scheduling



what we expect them transmit when they mix together



TCP/UDP may preempt RoCE



Transmit Ratio=5:5 （TCP : RoCE）

Transmit Ratio =1:9 （TCP : RoCE）

Experiments show that traffic preemption occurs in different traffic ratios.

Flow preemption problem will affect RDMA traffic transmission, which may result in degraded performance

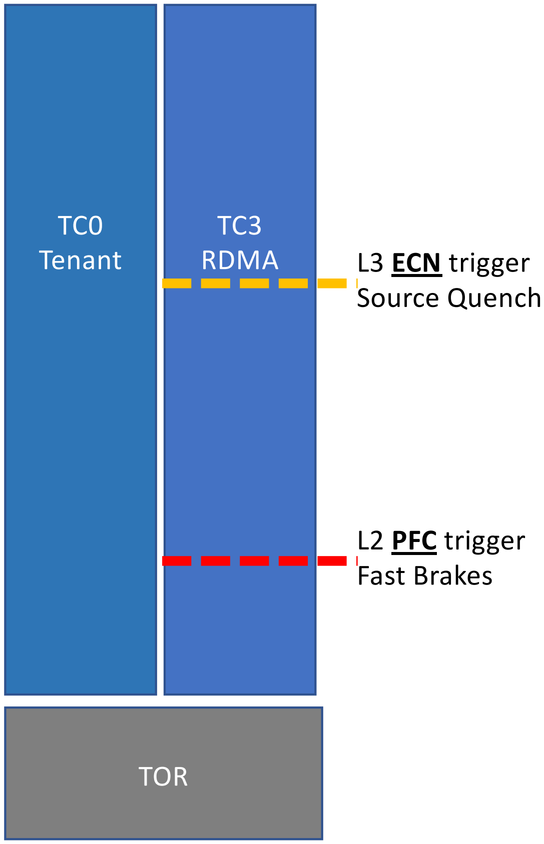
Configuration Challenges

{ETS, PFC, ECN} parameter settings

- A lossless environment requires thresholds
  - PFC and ECN buffer thresholds must be configured
  - Headroom, PFC-XOFF, OQ.Discard

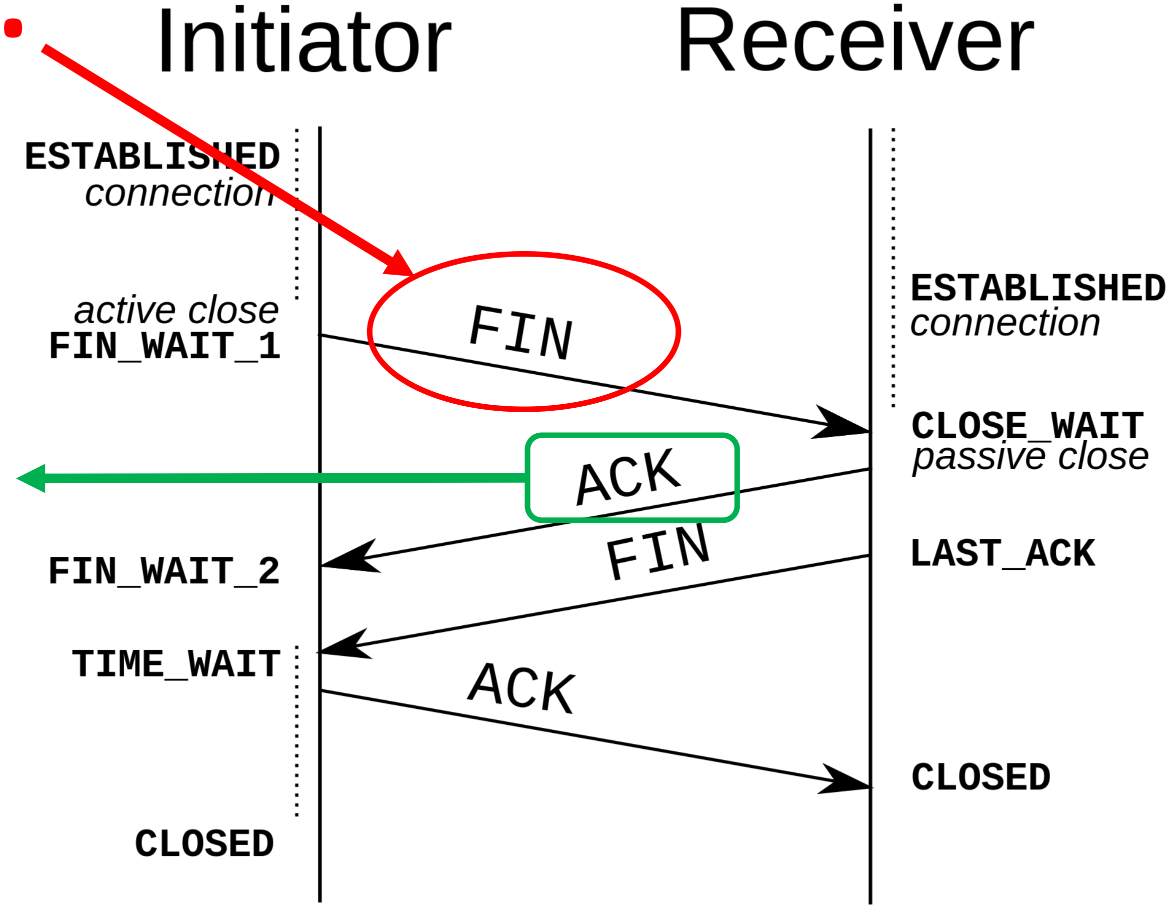
- ECN config must balance throughput and delay buffer thresholds
  - Kmin, Kmax, Pmax

- ECN must trigger before PFC
  - Emergency brakes!
  - Fabric density scaling will require reconfiguration



You are here…

Questions or Comments?



Backups

References

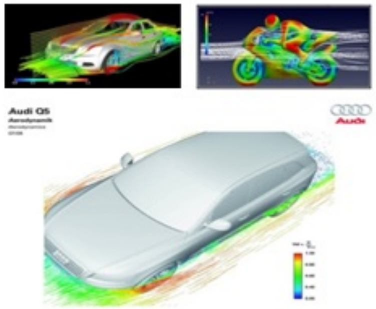
- iWARP and associated RFCs
  - http://www.rdmaconsortium.org/
- RoCE and associated Specs/RFCs
  - https://www.infinibandta.org/roce-initiative/
  - https://www.infinibandta.org/
  - ECN - https://tools.ietf.org/html/rfc3168
- Data Center Bridging and associated RFCs
  - Priority-based Flow Control (PFC): IEEE 802.1Qbb
  - Enhanced Transmission Selection (ETS): IEEE 802.1Qaz
  - IEEE 802.1p/Q provides 8 traffic classes for priority based forwarding.

Research on scaling

- Suggestions have been made to scale RDMA/HPC
  - RDMA over commodity Ethernet at scale, SIGCOMM 2016
  - iWARP Redefined: Scalable Connectionless Communication over High-Speed Ethernet, 2010 International Conference on High Performance Computing
  - Tuning ECN for Data Center Networks, CoNEXT '12
  - Revisiting Network Support for RDMA, SIGCOMM 2018

The benchmark for HPC services is Job Completion Time (JCT)

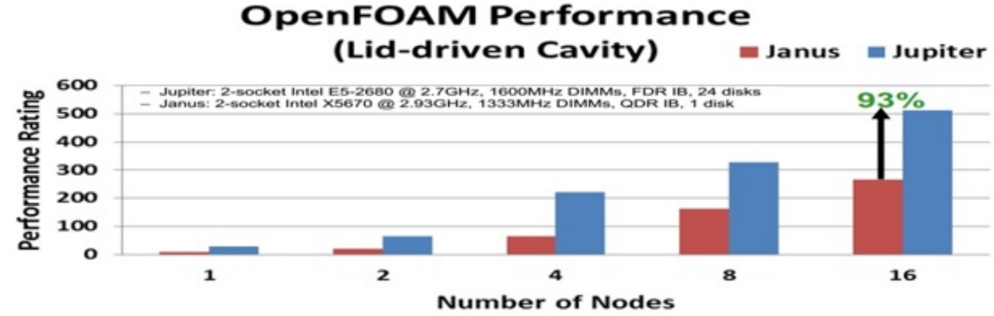
- **OpenFOAM® (Open Field Operation and Manipulation) CFD Toolbox in an open source CFD applications that can simulate**
  - Complex fluid flows involving
    - Chemical reactions
    - Turbulence
    - Heat transfer
  - Solid dynamics
  - Electromagnetics
  - The pricing of financial options

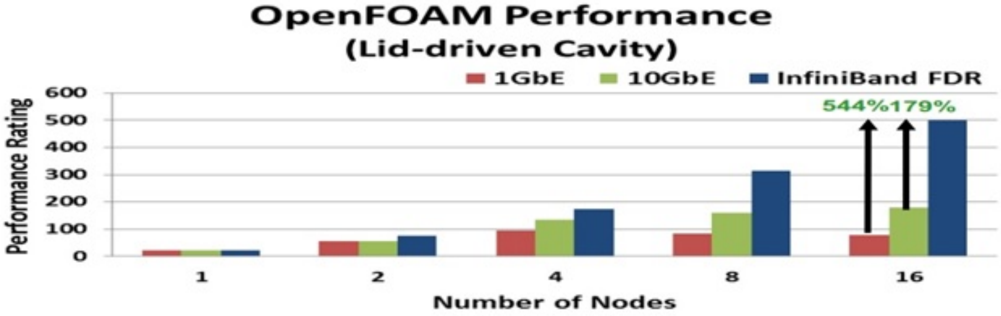


OpenFOAM Performance Benchmark and Profiling

HPC ADVISORY COUNCIL

Performance Rating = Jobs/Day , Higher is better


OpenFOAM Performance (Lid-driven Cavity)

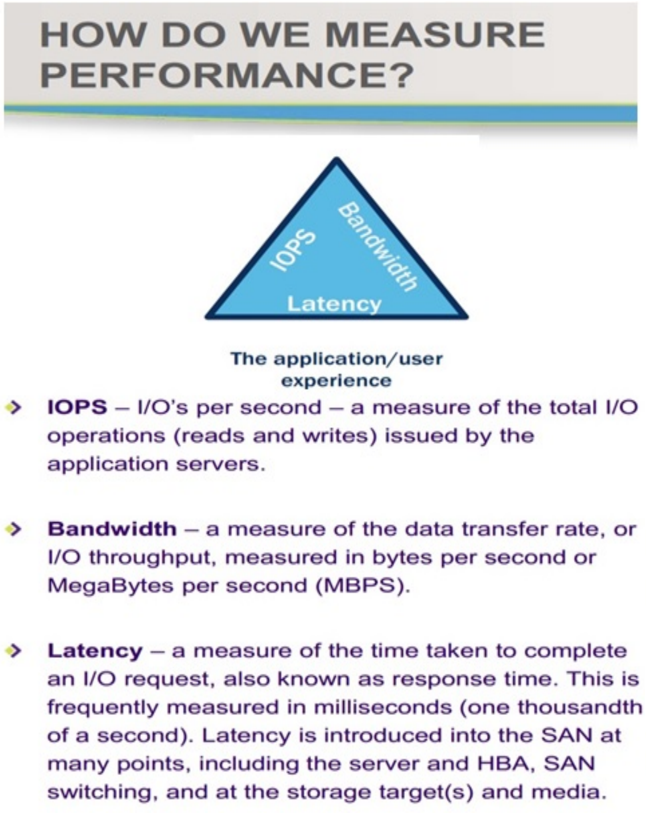

OpenFOAM Performance (Lid-driven Cavity)

1, OpenFOAM is a representative flow model of HPC, commonly used in fluid computing; it is widely used because of its open sourse and it's also used in financial sector nowdays.

2, The HPC Advisory Council specifies OpenFOAM performance benchmark. The only one is Jobs/Day, which is essentially the Job Completion Time (JCT). No matter measuring the difference in computing capability or network capability, this is the only benchmark.

Source: 1),OpenFOAM Performance Benchmark and Profiling, HPC Advisory Council, 2014.07; 2),Tackling Computational Fluid Dynamics in the Cloud, thePlatform, 2017.06; 3),The Need for Speed: Benchmarking DL Workloads, Baidu, 2016.09; 4),Baidu Targets Deep Learning Scalability Challenges, thePlatform, 2017.02

The benchmark for NOF services is IOPS and tail latency



HOW DO WE MEASURE PERFORMANCE?

IOPS
Bandwidth
Latency

The application/user experience

- **IOPS** – I/O's per second – a measure of the total I/O operations (reads and writes) issued by the application servers.
- **Bandwidth** – a measure of the data transfer rate, or I/O throughput, measured in bytes per second or MegaBytes per second (MBPS).
- **Latency** – a measure of the time taken to complete an I/O request, also known as response time. This is frequently measured in milliseconds (one thousandth of a second). Latency is introduced into the SAN at many points, including the server and HBA, SAN switching, and at the storage target(s) and media.

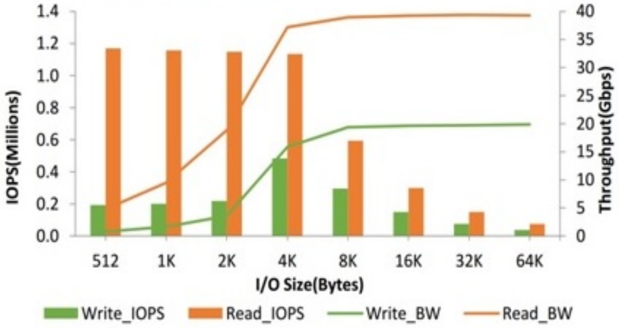


Figure 2 - IOPS & Throughput vs. I/O size

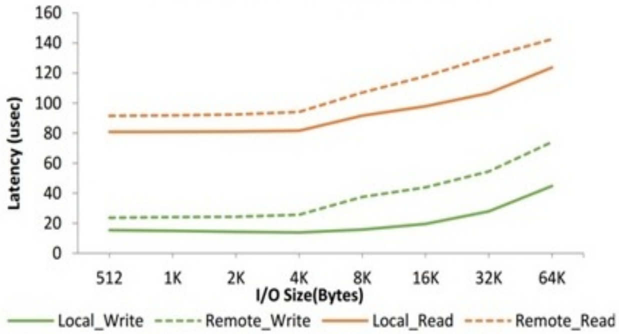
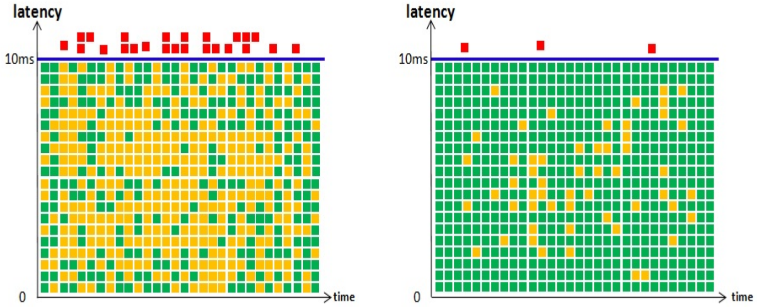


Figure 3 – Latency vs. I/O size (Local vs. Remote)



3.2, The *tail latency* (the completion time of the last IO operation is the completion time of the entire task) is used to measure the effectiveness of the throughput. This is also an important indicator for academic and industrial evaluation. A single task (Job) contains 20 IO accesses. If the single IO delay is greater than 10ms, the task fails. the left proportion of IO delay less than 10ms is 96%, and the right one is 99.6%. The overall effective IO is 44% vs. 92%.

1, From the perspective of the application (user) experience, NOF has clear performance benchmark: IOPS, bandwidth (IO throughput), latency (IO response time);

Source: 1),Next Generation Low Latency SAN, Qlogic@SNIA, 2015.04; 2),High performance NVMe over 40GbE iWARP, Intel & Chelsio, 2016.08; 3),Experiences with NVMe over Fabrics, Mellanox, 2017.03; 4),NVM Express Over Fabrics, Intel, 2015.03; 5),How to calculate your Disk I/O requirements, Microsoft, 2006.05