ceph

Ceph中国社区

IT大咖说 知识共享平台

CEPHALOCON APAC 2018

THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

Accelerating Ceph Performance with High Speed Networks and Protocols

Qingchun Song

Sr. Director of Market Development-APJ & China

Mellanox TECHNOLOGIES

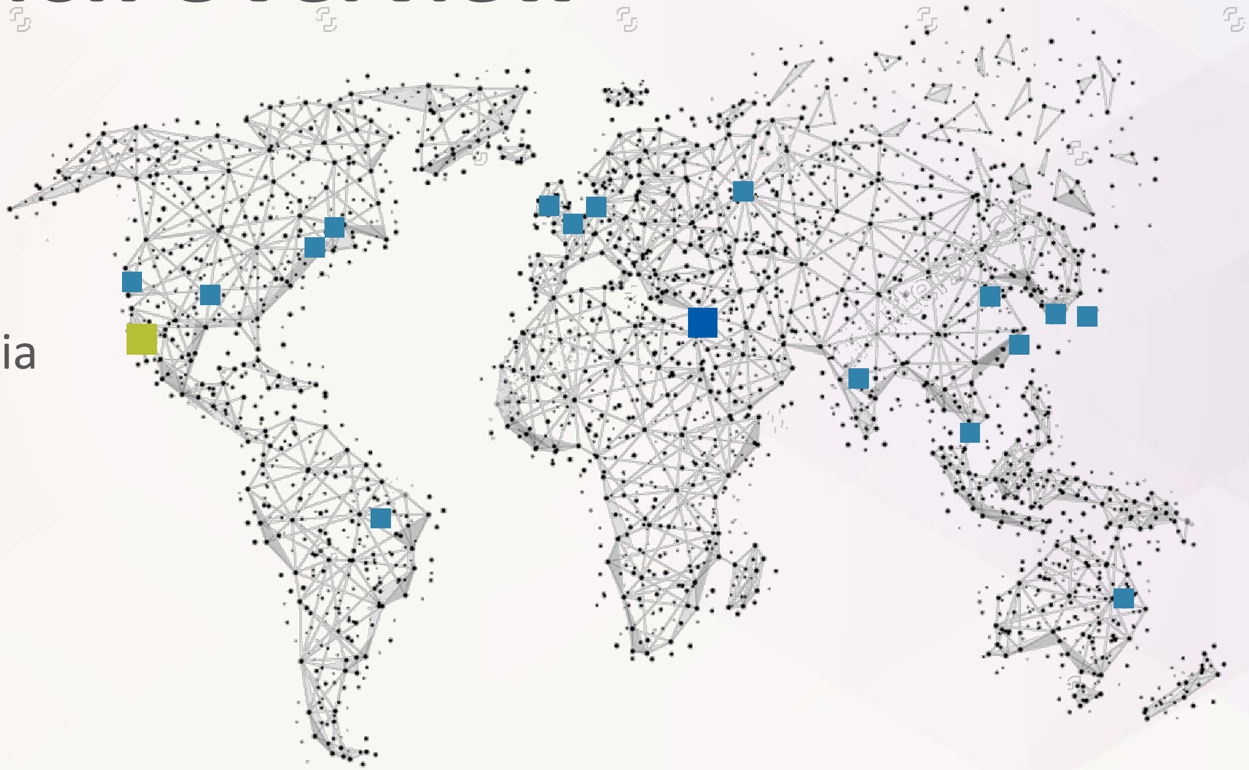
Mellanox Overview

Company Headquarters

🟦 Yokneam, Israel

🟩 Sunnyvale, California

🟦 Worldwide Offices



~2,900

Employees worldwide

NASDAQ®

Ticker: MLNX

ceph

Leadership in Storage Platforms

Delivering the Highest Data Center Return on Investment

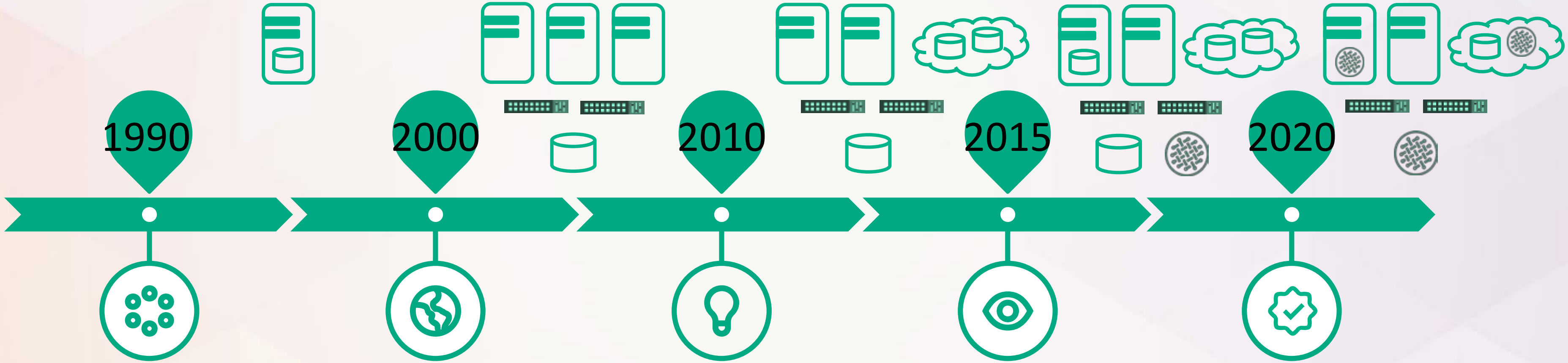
Microsoft SMB Direct | Hewlett Packard Enterprise | SEAGATE xyratex | IBM xiv tms

ORACLE | TOSHIBA | DataDirect NETWORKS | FUJITSU

DELL EMC | Micron | PURESTORAGE

INFINIDAT | X-IO technologies | TERADATA | WD Western Digital VIRIDENT | NetApp

Storage & Connectivity Evo



Internal Storage

Distributed Storage

Shared Network

File & Block Data

External Storage

Storage Consolidation

Dedicated FC SAN

Block/Structured Data

Virtualization/ Cloud

Data Specific Storage

Lossless IB/Eth

Unstructured Data

Flash/ Convergence

Media Migration

Dedicated Eth/IB/FC

All Data Types

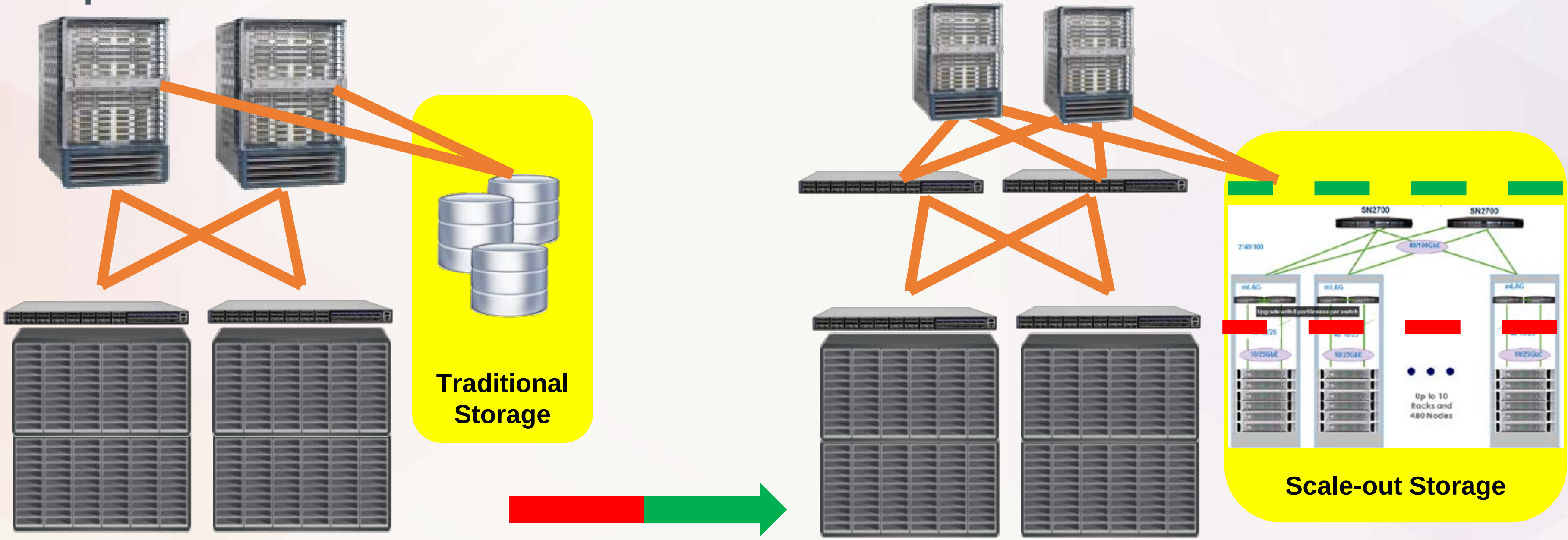
NVMe/ Big Data

Many Platforms

Dedicated Eth/IB

Right Data for Platform

Where to Draw the Line?



Traditional Storage

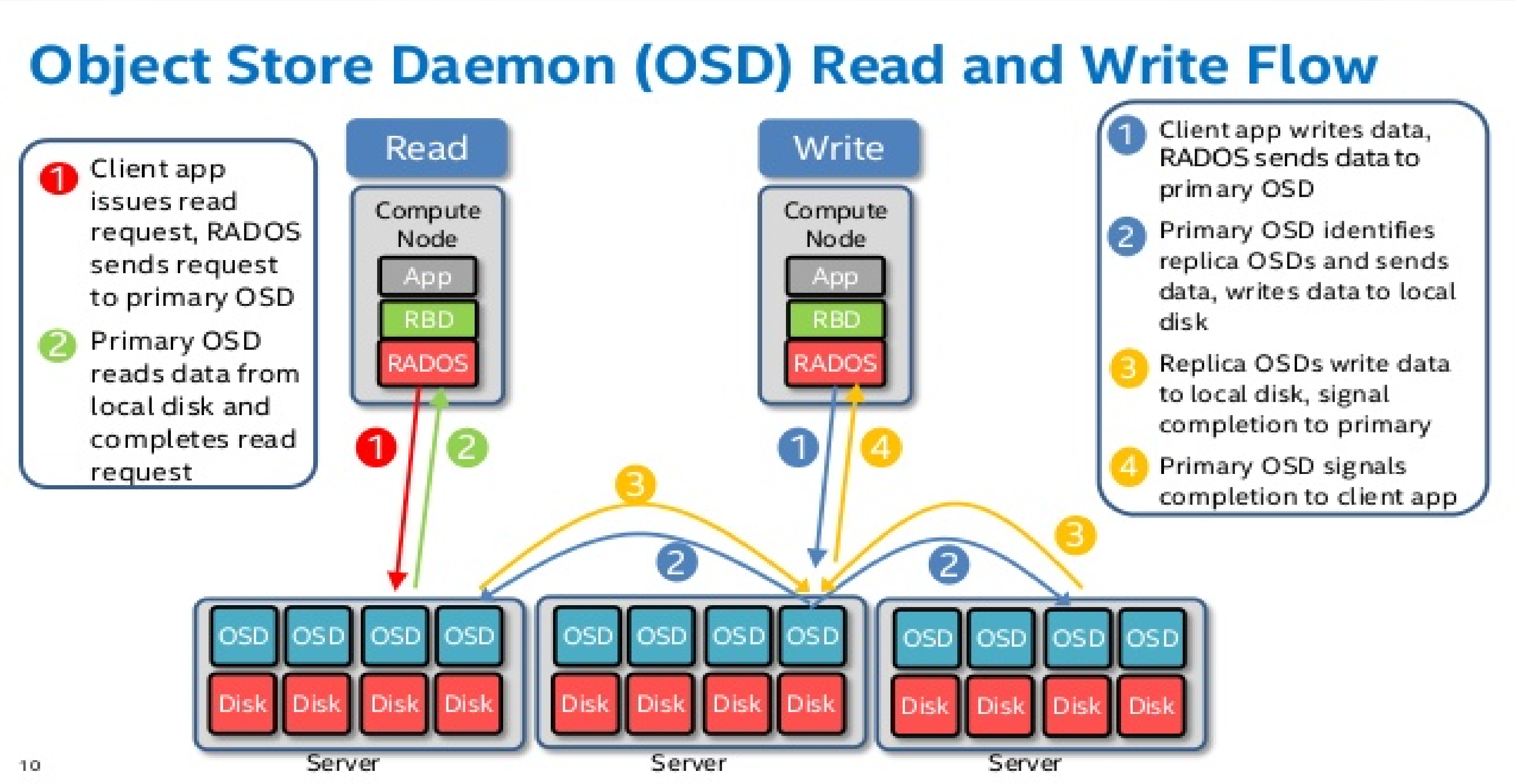
Scale-out Storage

SN2700
SN2700
2*40/100
40/100GbE
mLAG
mLAG
mLAG
Upgrade with 8 port license per switch
10/25GbE
10/25GbE
10/25GbE
Up to 10 Racks and 480 Nodes

Legacy DC – FC SAN

Modern DC – Ethernet Storage Fabric

CePH Work Flow



Object Store Daemon (OSD) Read and Write Flow

1 Client app issues read request, RADOS sends request to primary OSD

2 Primary OSD reads data from local disk and completes read request

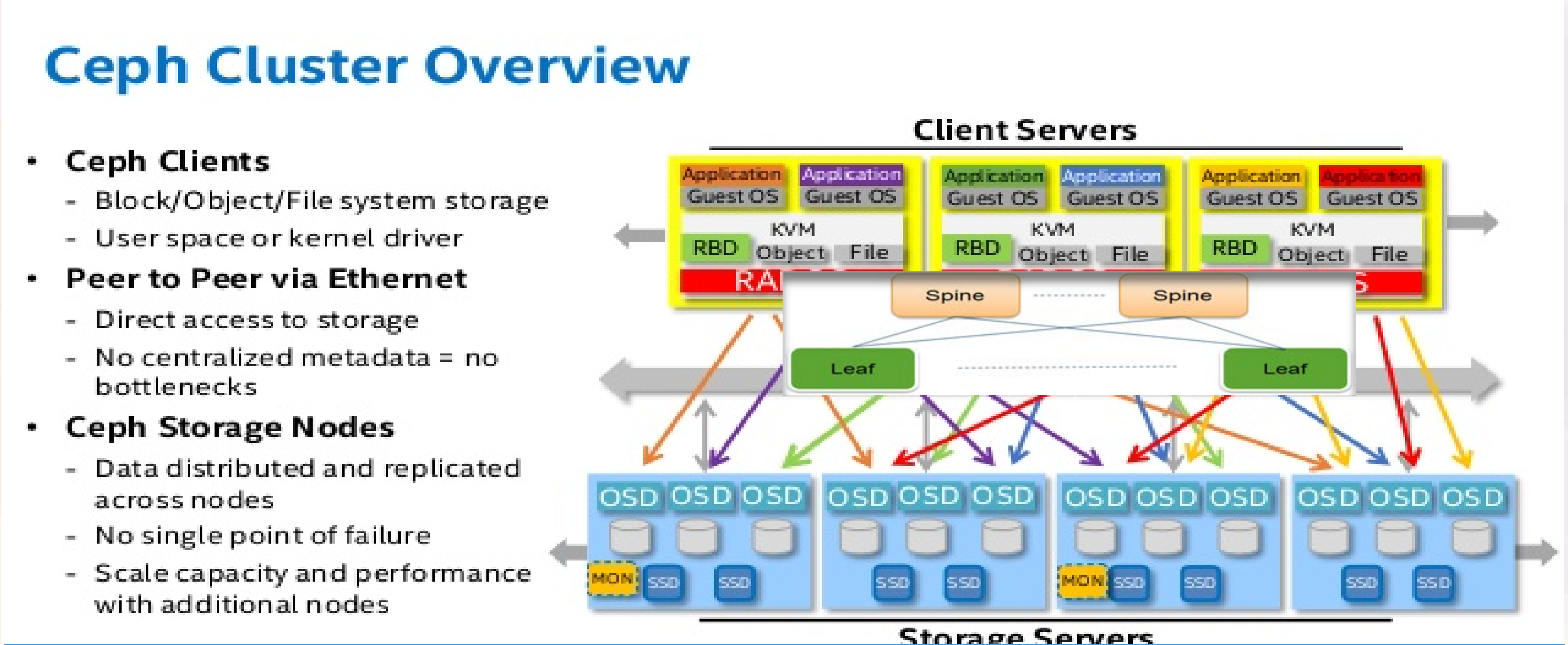
1 Client app writes data, RADOS sends data to primary OSD

2 Primary OSD identifies replica OSDs and sends data, writes data to local disk

3 Replica OSDs write data to local disk, signal completion to primary

4 Primary OSD signals completion to client app

Storage or Data Bottlene Bandwidth



Ceph Cluster Overview

- Ceph Clients
  - Block/Object/File system storage
  - User space or kernel driver
- Peer to Peer via Ethernet
  - Direct access to storage
  - No centralized metadata = no bottlenecks
- Ceph Storage Nodes
  - Data distributed and replicated across nodes
  - No single point of failure
  - Scale capacity and performance with additional nodes

OSD read:

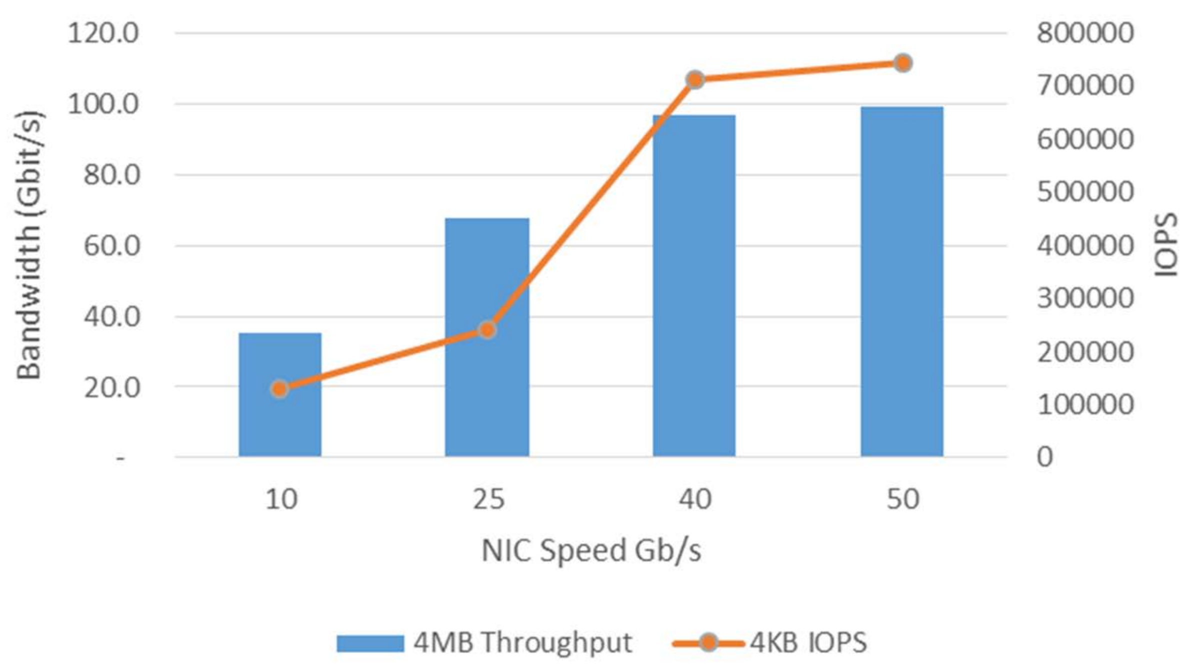
- Client(App <-> RBD <-> RADOS) <-> NIC <-> Leaf <-> Spine <-> Leaf <-> NIC <->OSD <-> NVMe

OSD write:

- Client(App <-> RBD <-> RADOS) <-> NIC <-> Leaf <-> Spine <-> Leaf <-> NIC <->OSD <-> NVMe <-> OSD <-> NIC <-> Leaf <-> Spine <-> Leaf <-> NIC <->OSD <-> NVMe

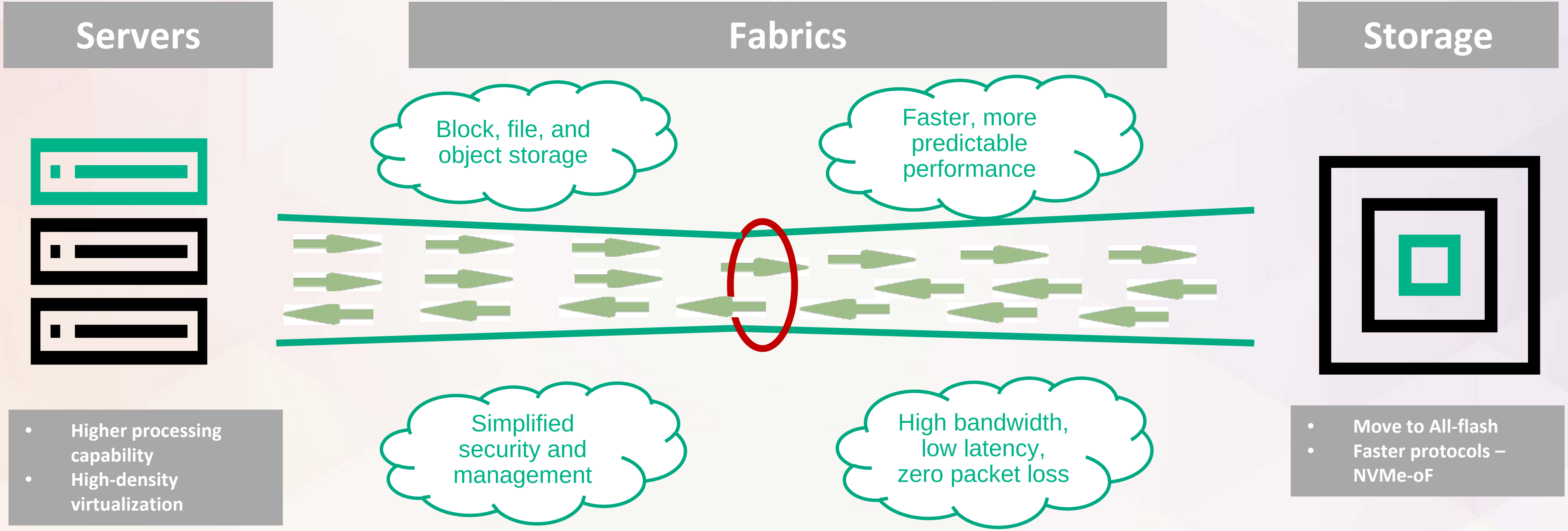
Ceph Bandwidth Perform Improvement

- Aggregate performance of 4 Ceph servers
  - 25GbE has 92% more bandwidth than 10GbE
  - 25GbE has 86% more IOPS than 10GbE
- Internet search results seem to recommend one 10GbE NIC for each ~15 HDDs in an OSD
  - Mirantis, Red Hat, Supermicro, etc.



Storage or Data Bottleneck: L

Predictable Performance, Deterministic & Secure Fabrics



Servers

Fabrics

Storage

Block, file, and object storage

Faster, more predictable performance

Simplified security and management

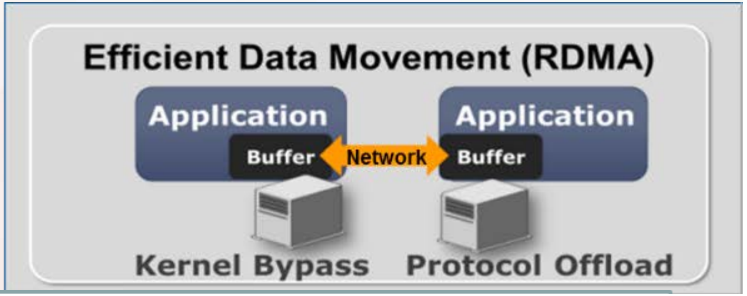
High bandwidth, low latency, zero packet loss

- Higher processing capability
- High-density virtualization

- Move to All-flash
- Faster protocols – NVMe-oF

Data Center modernization requires Future Proof, faster, lossless Ethernet Storage Fabrics

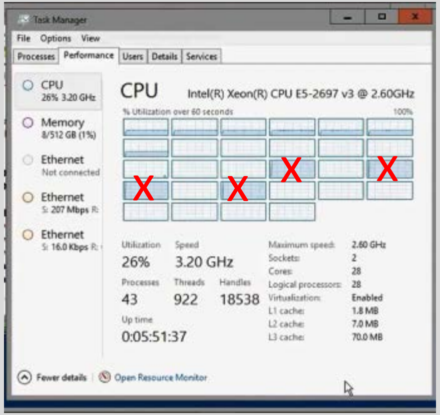
RDMA Is The Key For Storage Latency



RDMA Enables Efficient Data Movement

100GbE With CPU Onload

100 GbE With Network Offload



CPU 26% 3.20 GHz

Memory 8/512 GB (1%)

Ethernet Not connected

Ethernet S: 207 Mbps R:

Ethernet S: 16.0 Kbps R:

CPU Intel(R) Xeon(R) CPU E5-2697 v3 @ 2.60GHz

% Utilization over 60 seconds 100%

Utilization 26%

Speed 3.20 GHz

Processes 43

Threads 922

Handles 18538

Up time 0:05:51:37

Maximum speed: 2.60 GHz
Sockets: 2
Cores: 28
Logical processors: 28
Virtualization: Enabled
L1 cache: 1.8 MB
L2 cache: 7.0 MB
L3 cache: 70.0 MB

CPU Onload Penalties

- *Half the Throughput*
- *Twice the Latency*
- *Higher CPU Consumption*

2X Better Bandwidth

Half the Latency

33% Lower CPU

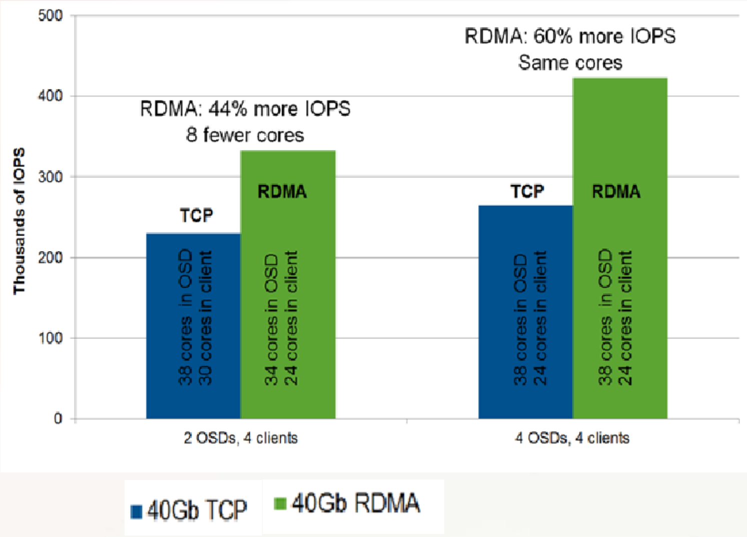
See the demo: https://www.youtube.com/watch?v=u8ZYhUjSUoI

- Without RDMA
  - 5.7 GB/s throughput
  - 20-26% CPU utilization
  - 4 cores 100% consumed by moving data
- With Hardware RDMA
  - 11.1 GB/s throughput at half the latency
  - 13-14% CPU utilization
  - More CPU power for applications, better ROI

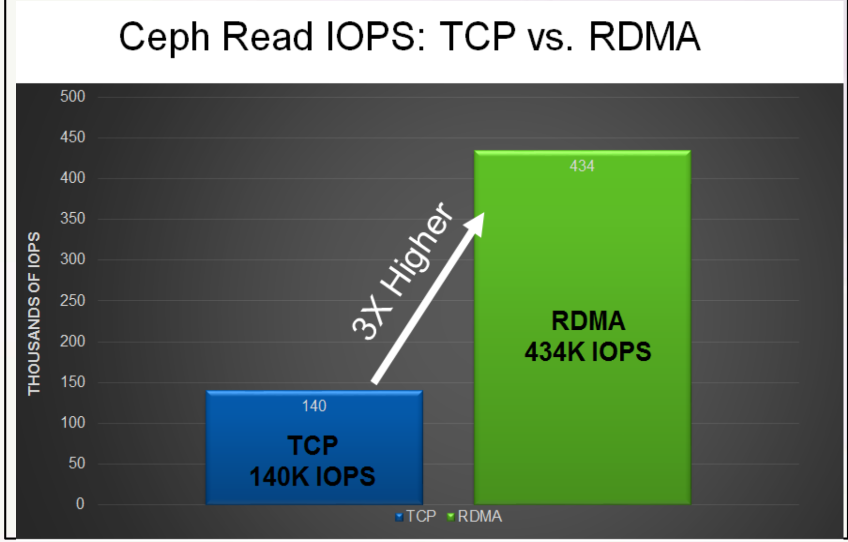
Ceph RDMA Performance Improv

- Conservative Results: 44%~60% more IOPS

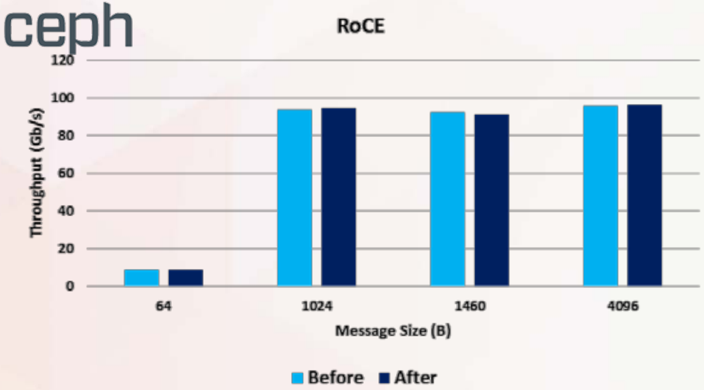
- RDMA offers significant benefits to Ceph performance for small block size (4KB) IOPS.
  - 2 OSDs with 4 clients, RDMA allowed 44% more IOPS.
  - 4 OSDs and 4 clients, RDMA allowed 60% more IOPS.



- Best Results: 3x Higher IOPS
- RDMA's biggest benefit for Ceph block storage
  - High IOPS workloads
  - Small block sizes (<32KB)
- Enable > 10GB/s from single node
- Enable < 10usec latency under load

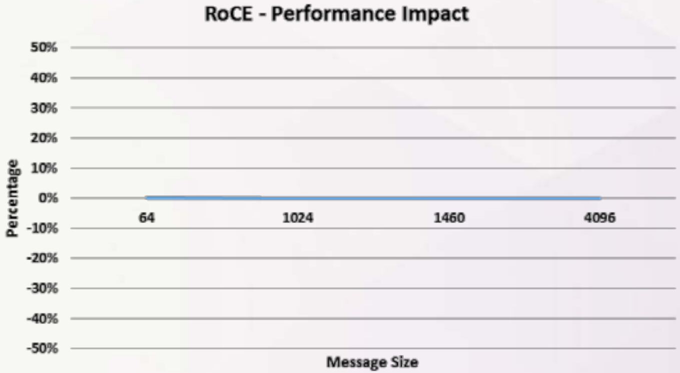

Ceph Read IOPS: TCP vs. RDMA

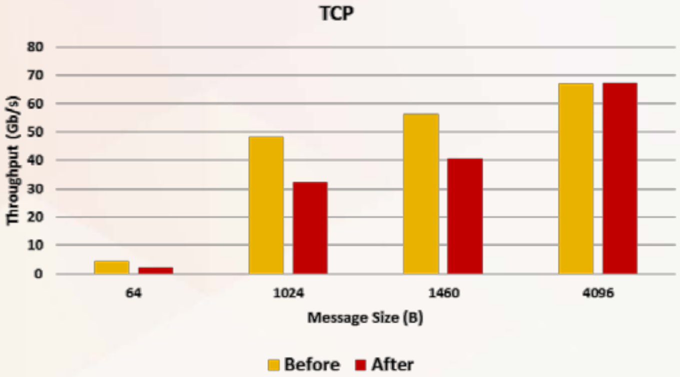
RDMA: Mitigates Meltdown Mess Stops Spectre Security Slowdown



Performance Impact: 0%

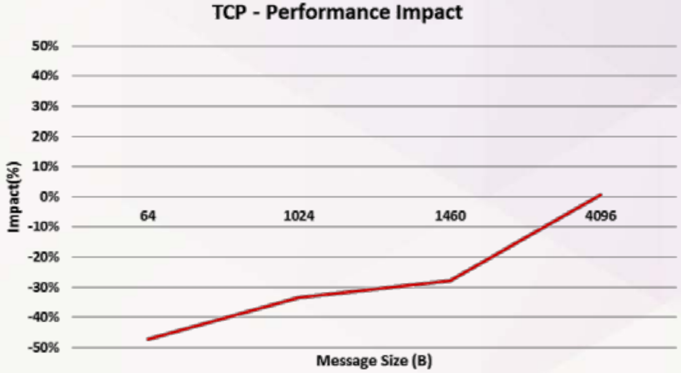
0%





Performance Impact: -47%

-47%



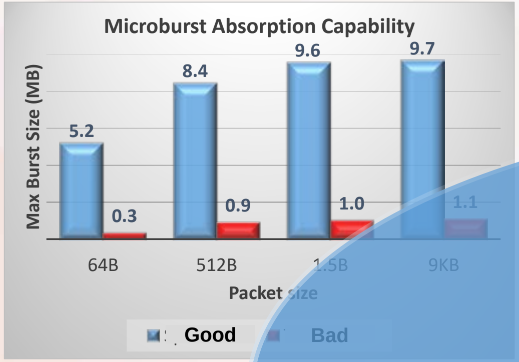
Before – Before applying software patches of Meltdown & Spectre **After –** After applying software patches of Meltdown & Spectre

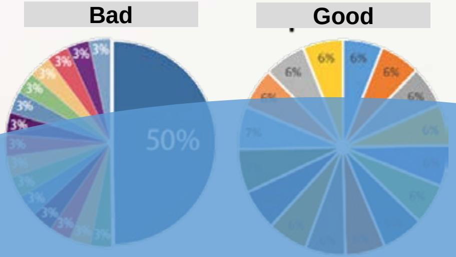
CePH RDMA Status

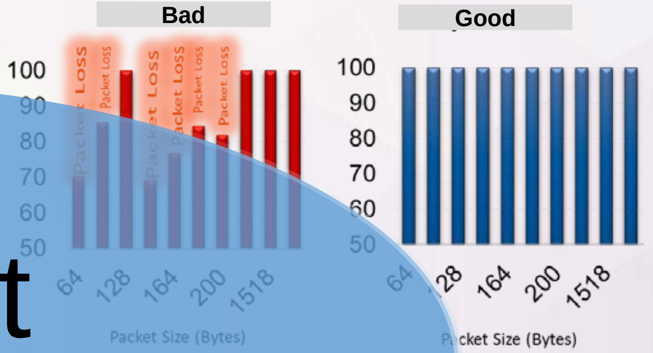
- CePH RDMA working group
    - Mellanox
    - Xsky
    - Samsung
    - SanDisk
    - RedHat
- The latest stable CePH RDMA version
    - https://github.com/Mellanox/ceph/tree/luminous-12.1.0-rdma
- Bring Up Ceph RDMA - Developer's Guide
    - https://community.mellanox.com/docs/DOC-2721
- RDMA/RoCE Configuration Guide
    - https://community.mellanox.com/docs/DOC-2283

Storage or Data Bottleneck: Storage Fabric

Ceph中国社区 | IT大咖说 知识共享平台







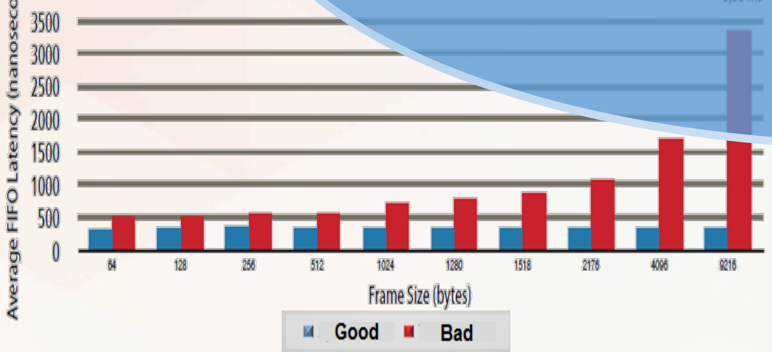
Congestion Management

Fairness

Avoidable Packet Loss

Ethernet
Storage Fabric

Latency Test Results



Spectrum™

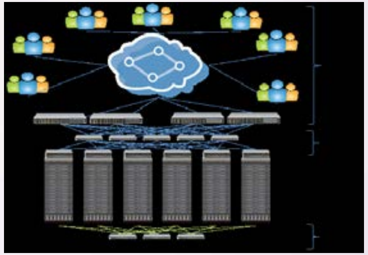
TEST REPORT Tolly.

www.zeropacketloss.com

www.Mellanox.com/tolly

Summary

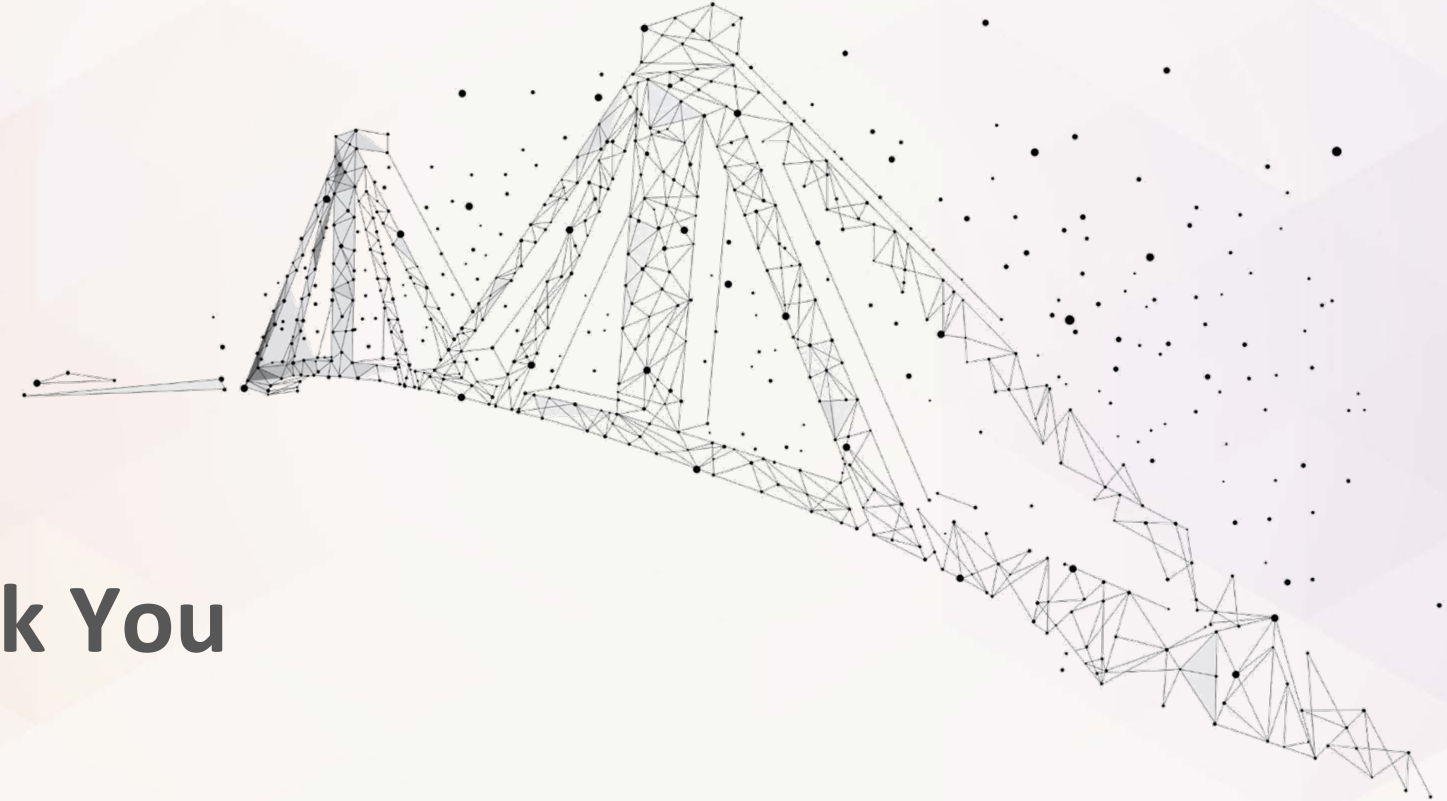
- Ceph Benefits from Faster Network
  - 10GbE is not enough!
- RDMA further optimizes Ceph performance
- Reduce the impact from Meltdown/Spectre fixes
- ESF(Ethernet Storage Fabric) is trend



ceph
Mellanox
TECHNOLOGIES

ceph

Ceph中国社区 | IT大咖说 知识共享平台



Thank You