Improving Ceph Performance With Networking

Ceph Networking—When 10GbE Is Not Enough

Tong Liu

tong@mellanox.com

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

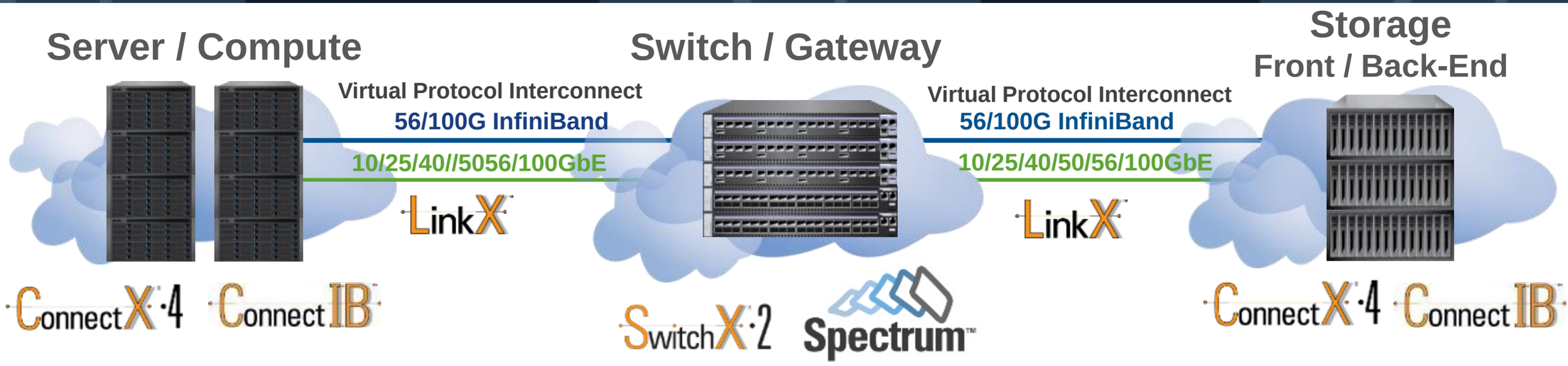
Mellanox Overview

Ticker: MLNX NASDAQ®

- Company headquarters:
  - Yokneam, Israel; Sunnyvale, California
  - ~1,900 employees* worldwide
- Solid financial position
  - FY14 revenue of $463.6M
  - 2Q15 revenue of $163.1M
  - 3Q15 guidance of $169-$171M
  - Cash + investments @June 30, 2015 = $467.2M

* As of June 2015

Leading Supplier of End-to-End Interconnect Solutions



Comprehensive End-to-End InfiniBand and Ethernet Portfolio (VPI)

ICs

Adapter Cards

Switches/Gateways

Host/Fabric Software

Metro / WAN

Cables/Modules

Ethernet that Delivers: 25, 50, 100GbE Switch & Adapters

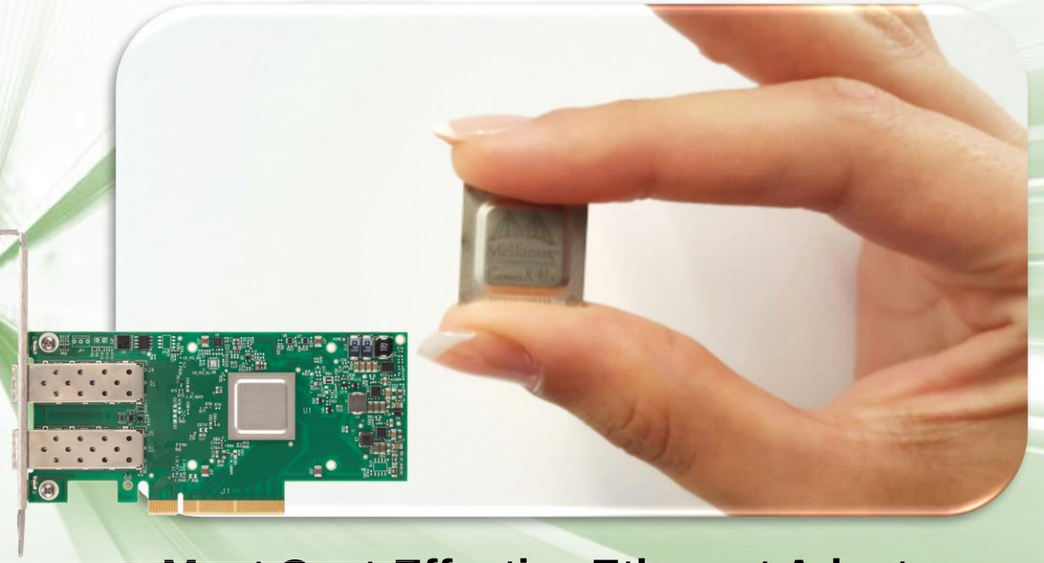
Spectrum™

One Switch. A World of Options.

Flexibility, Opportunities, Speed

Open Ethernet, Zero Packet Loss

ConnectX®-4 ConnectX®-4 Lx



Most Cost-Effective Ethernet Adapter

2.5X the Network Performance

Same Infrastructure, Same Connector

One Switch - A World of Opportunities

25, 50, & 100G at Your Fingertips

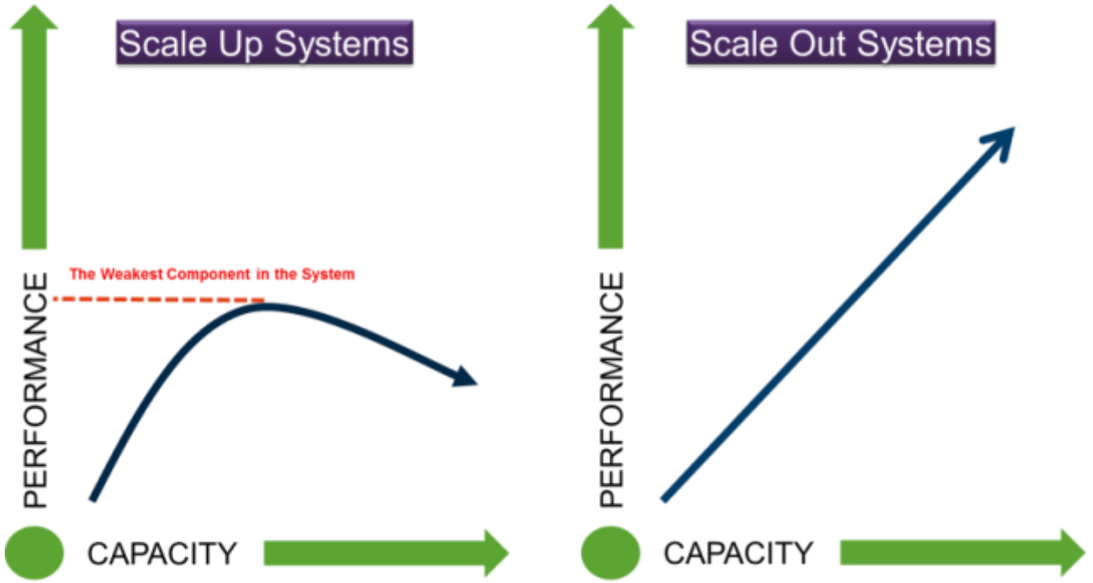
Why is Ceph Popular?

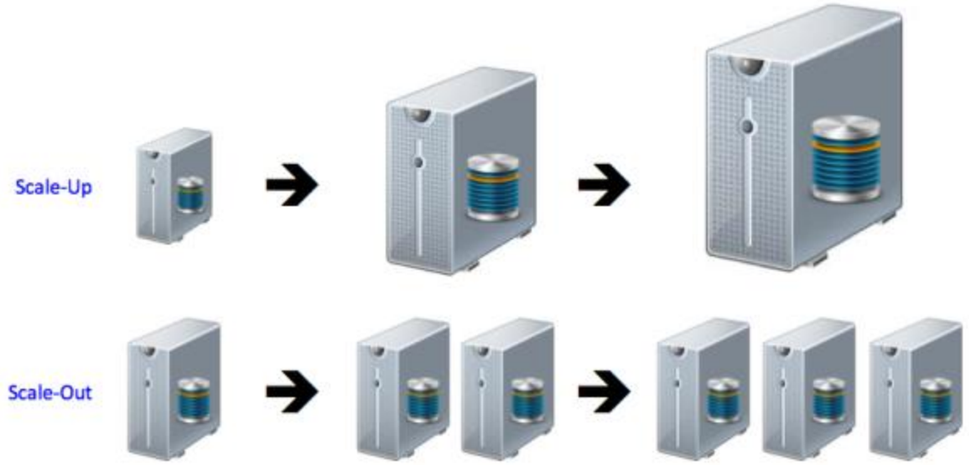
The only open-source, software-defined, scale-out enterprise storage

- Scale-out Block and Object
  - Up to hundreds of nodes, petabytes of storage
  - Distributed architecture for performance and availability
- Enterprise Features
  - High availability: redundancy, replication, failover, rebalancing
  - Capacity: Multi-site, tiering, erasure coding
  - Data management: Cloning, snapshots, thin provisioning
- Software-Defined and Open Source
  - Flexible hardware options
  - Build your own or buy as pre-integrated Ceph appliances
  - Integrated With OpenStack

Scale-Out Architecture Requires A Fast Network

- Scale-out grows capacity and performance in parallel
- Requires fast network for replication, sharing, and metadata (file)
  - Throughput requires bandwidth
  - IOPS requires low latency
- Proven in HPC, storage appliances, cloud, and now… Ceph

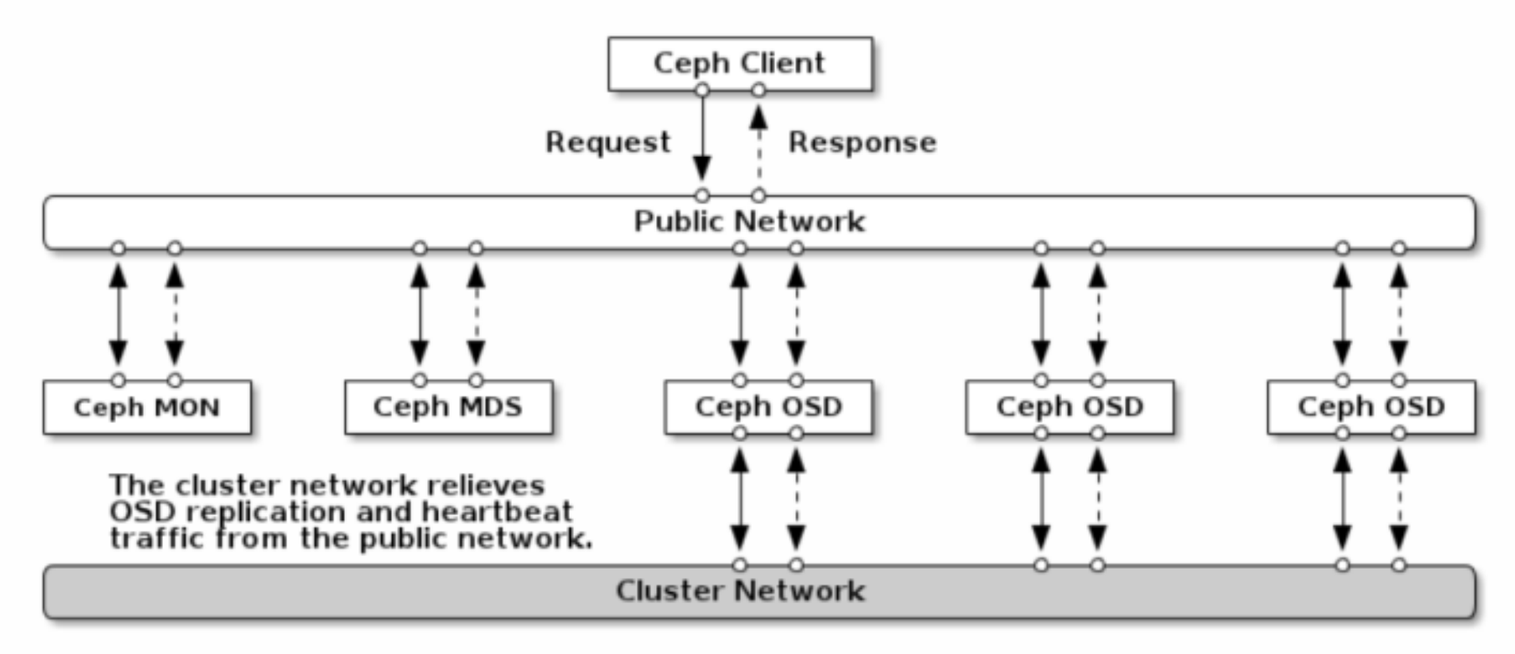




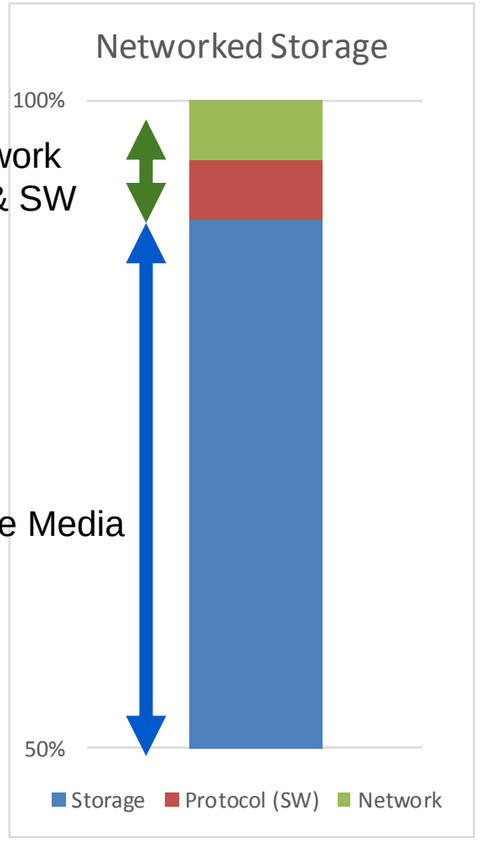
Interconnect Capabilities Determine Scale Out Performance

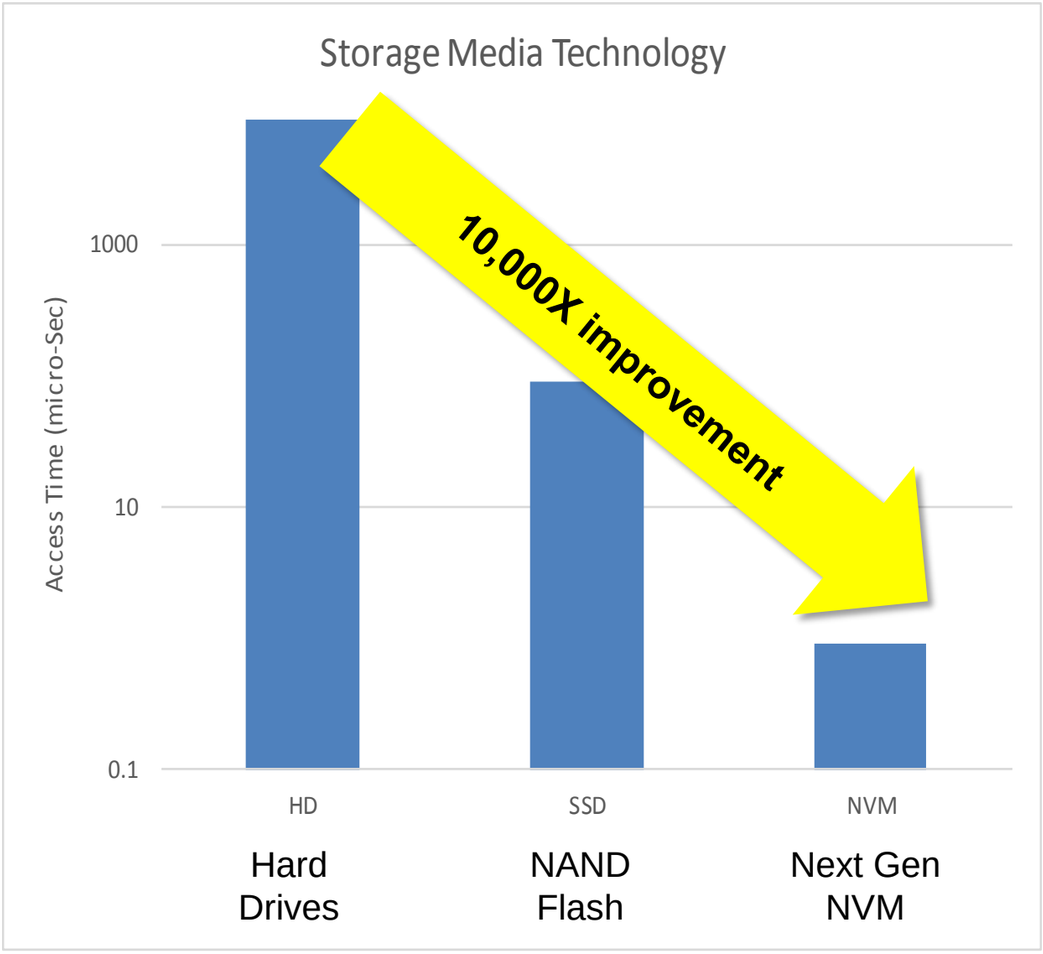
Ceph Has Two Logical Networks

- High performance networks enable maximum cluster availability
  - Clients, OSD, Monitors and Metadata servers communicate over multiple network layers
  - Real-time requirements for heartbeat, replication, recovery and re-balancing
  - Cluster write traffic is 3x (replication) or 1.5x (EC) more than public network
- Cluster ("backend") network performance dictates cluster's performance and scalability
  - "**Network load between Ceph OSD Daemons easily dwarfs the network load between Ceph Clients and the Ceph Storage Cluster**" (Ceph Documentation)



Solid State Storage – Faster Storage Needs Faster Networks

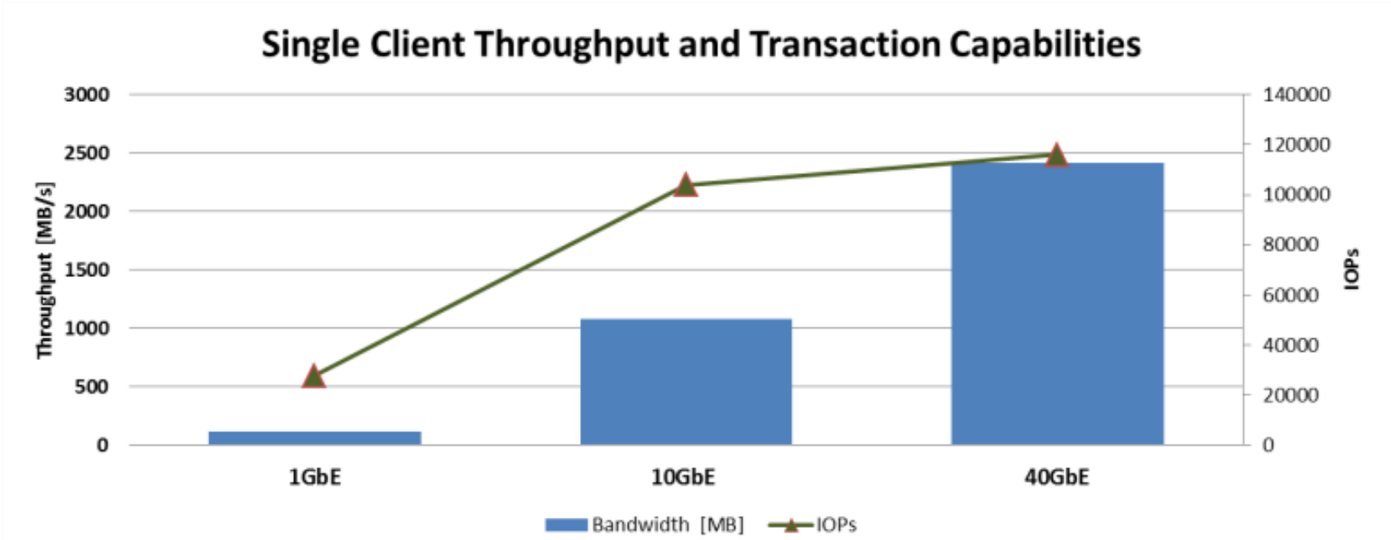


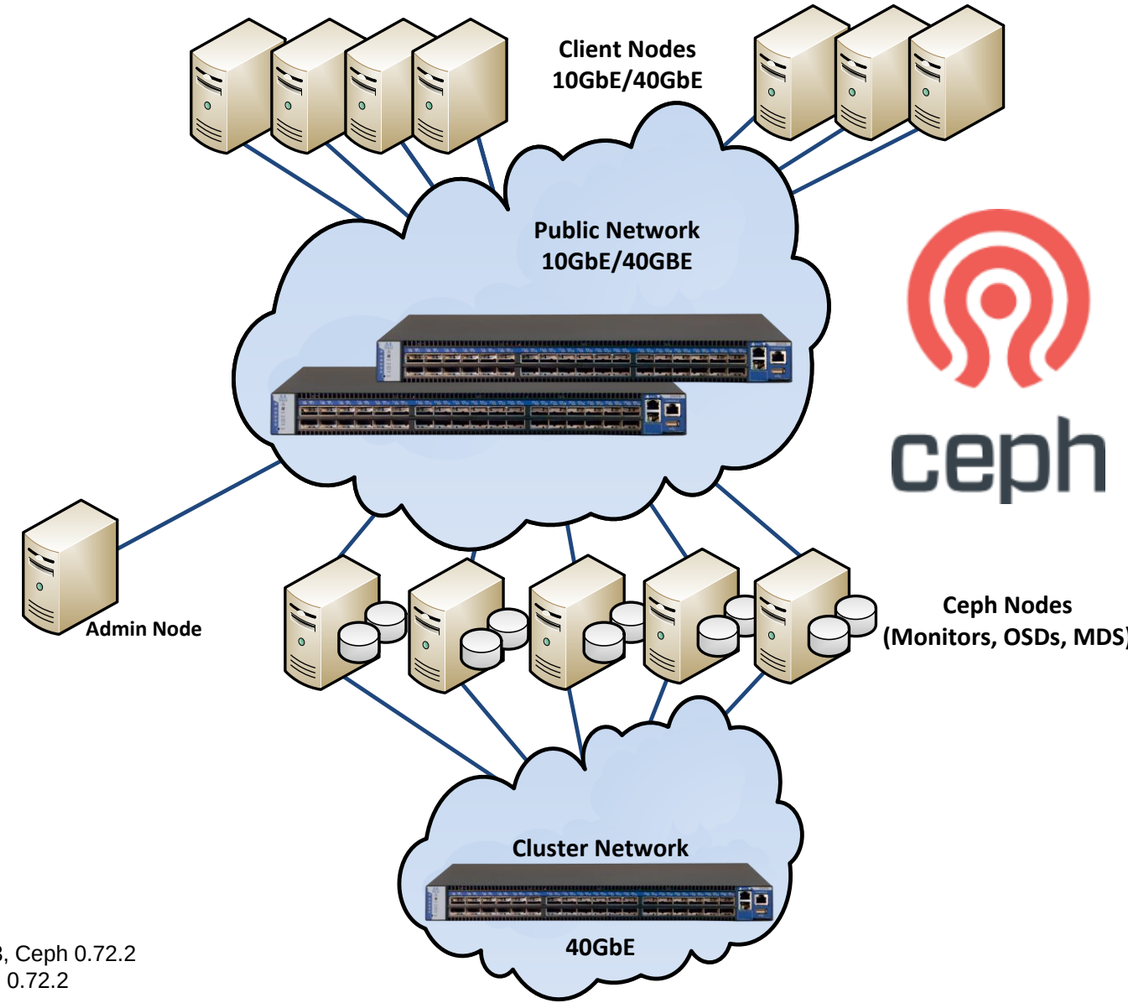


Advanced Networking and Protocol Offloads Required to Match Storage Media Performance

Ceph Deployment Using 10GbE and 40GbE

- Cluster (Private) Network @ 40/56GbE
  - Smooth HA, unblocked heartbeats, efficient data balancing
- Throughput Clients @ 40/56GbE
  - Guaranties line rate for high ingress/egress clients
- IOPs Clients @ 10GbE or 40/56GbE
  - 100K+ IOPs/Client @4K blocks



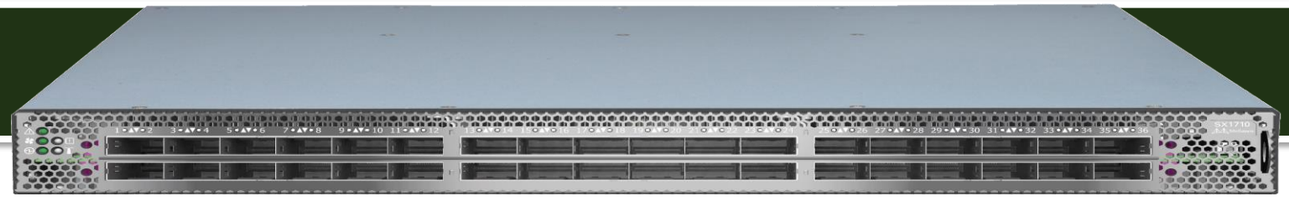


Throughput Testing results based on fio benchmark, 8MB block, 20GB file,128 parallel jobs, RBD Kernel Driver with Linux Kernel 3.13.3 RHEL 6.3, Ceph 0.72.2
IOPs Testing results based on fio benchmark, 4KB block, 20GB file,128 parallel jobs, RBD Kernel Driver with Linux Kernel 3.13.3 RHEL 6.3, Ceph 0.72.2

2.5x Higher Throughput , 15% Higher IOPs with 40Gb Ethernet vs. 10GbE!
(http://www.mellanox.com/related-docs/whitepapers/WP_Deploying_Ceph_over_High_Performance_Networks.pdf)

Mellanox Switches Maximize Ceph Performance & Efficiency

SX1036/1710(x86) – 36x 40/56GbE ports
Ideal 40GbE Aggregation Switch



SX1024/1400(x86) 48x 10GbE + 12x 40GbE
Non-blocking 10GbE➔ 40GbE ToR

SX1012 – 12x 10/40/56GbE ports
Ideal storage/Database 10/40GbE Switch

SX1016 – 64x 10GbE ports
Highest density 10GbE ToR

energy efficiency

Lowest Power Consumption

SX1710 – 91W
SX1036 – 83W
SX1016 – 62W
SX1024 – 75W
SX1012 – 50W

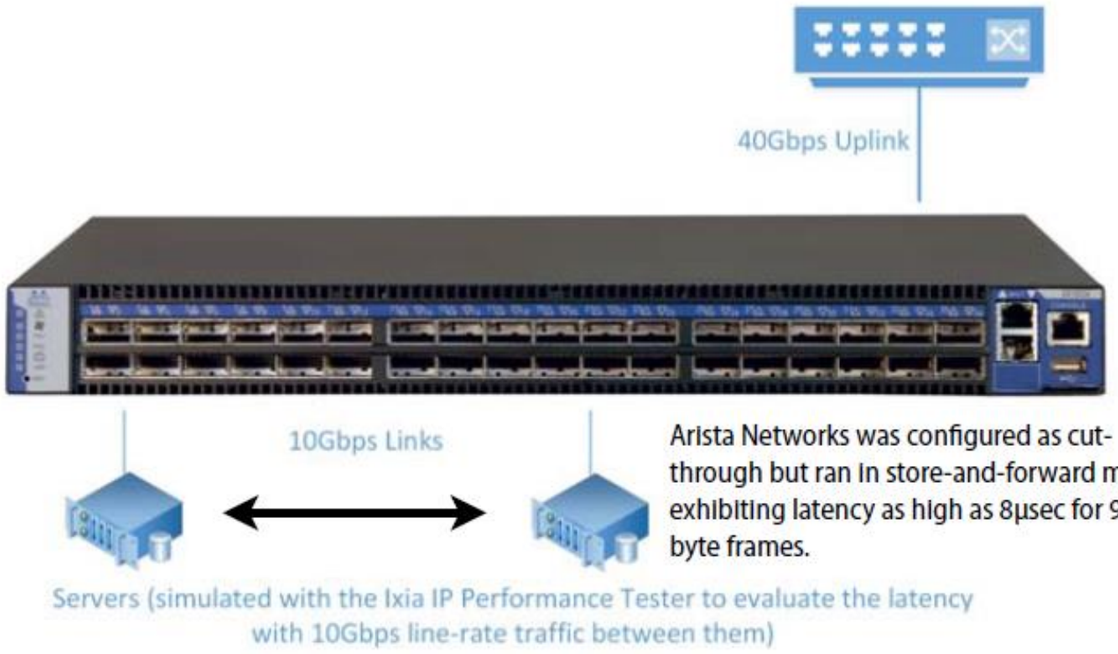
Lowest Latency

220ns

True Cut Through 10/40GbE Switches: 220ns vs. 8000ns

10GbE Port to 10GbE Port Latency Test Bed

Typical Data Center ToR Switch User Scenario



Arista Networks was configured as cut-through but ran in store-and-forward mode exhibiting latency as high as 8µsec for 9212-byte frames.

Servers (simulated with the Ixia IP Performance Tester to evaluate the latency with 10Gbps line-rate traffic between them)

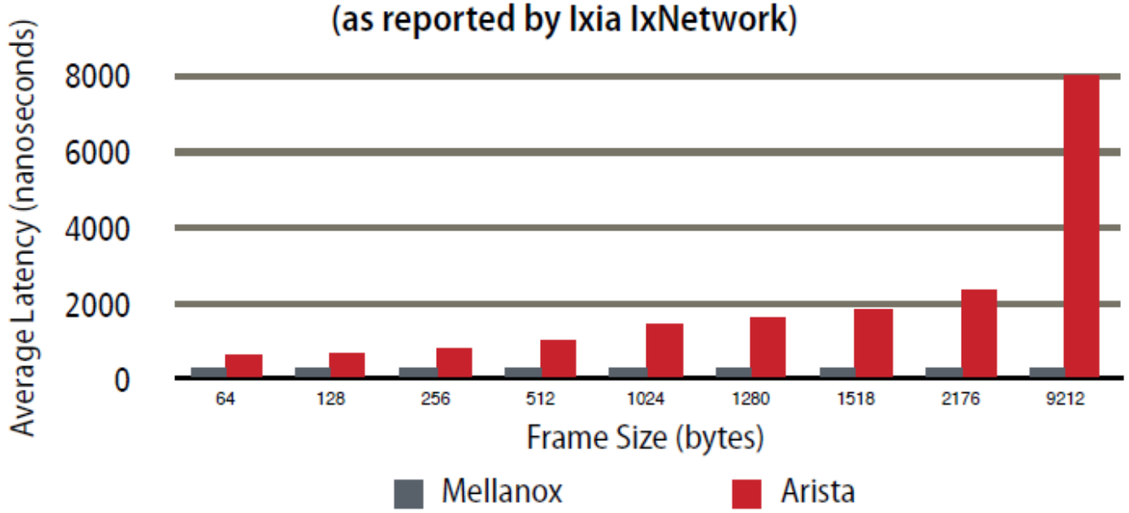
Note: 10GbE connectivity was achieved through break-out cables.

Source: Tolly, January 2015

Figure 3

Typical 10-40GbE ToR RFC 2544 Latency Results: Mellanox SX1036 vs. Arista DCS-7050QX

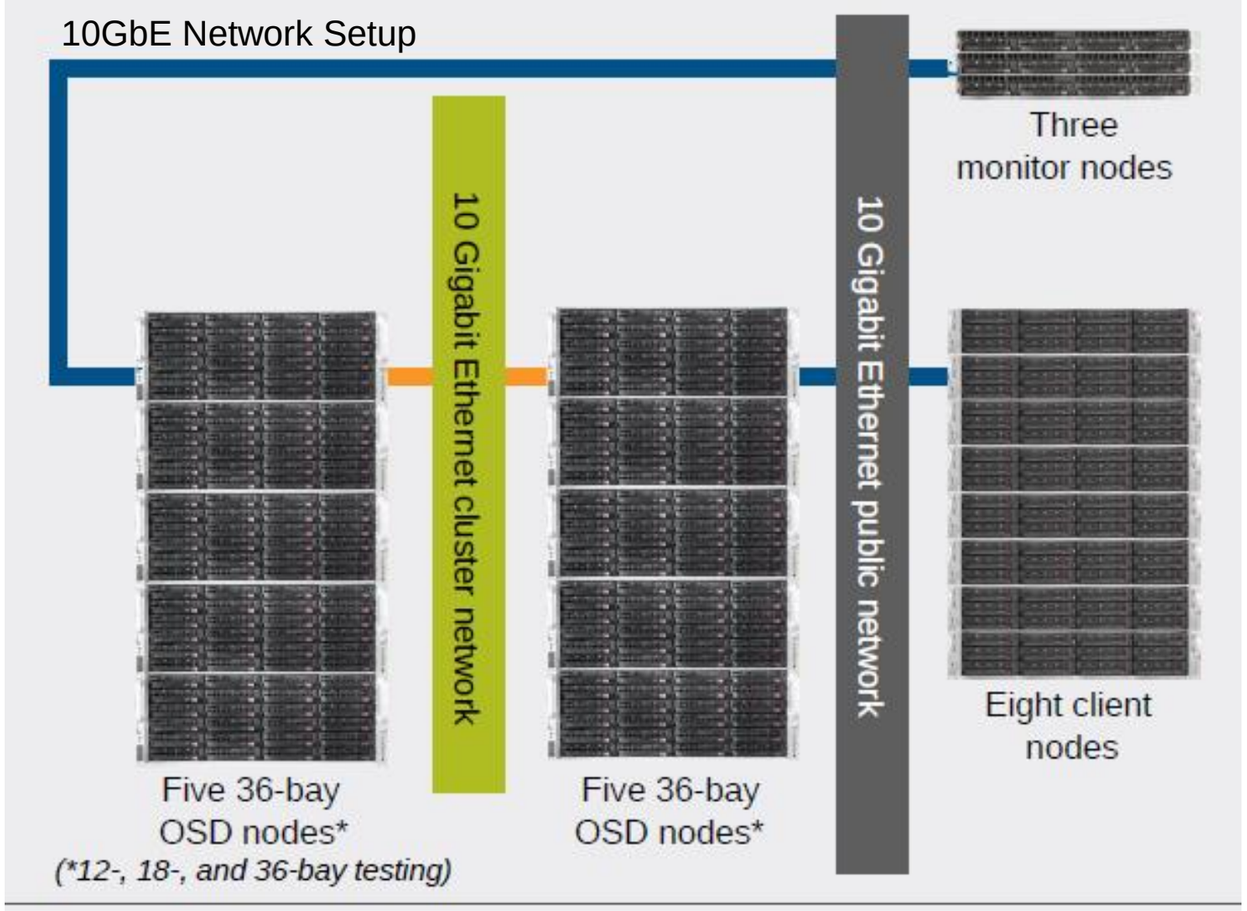
Layer 2 Two 10GbE Ports (as reported by Ixia IxNetwork)

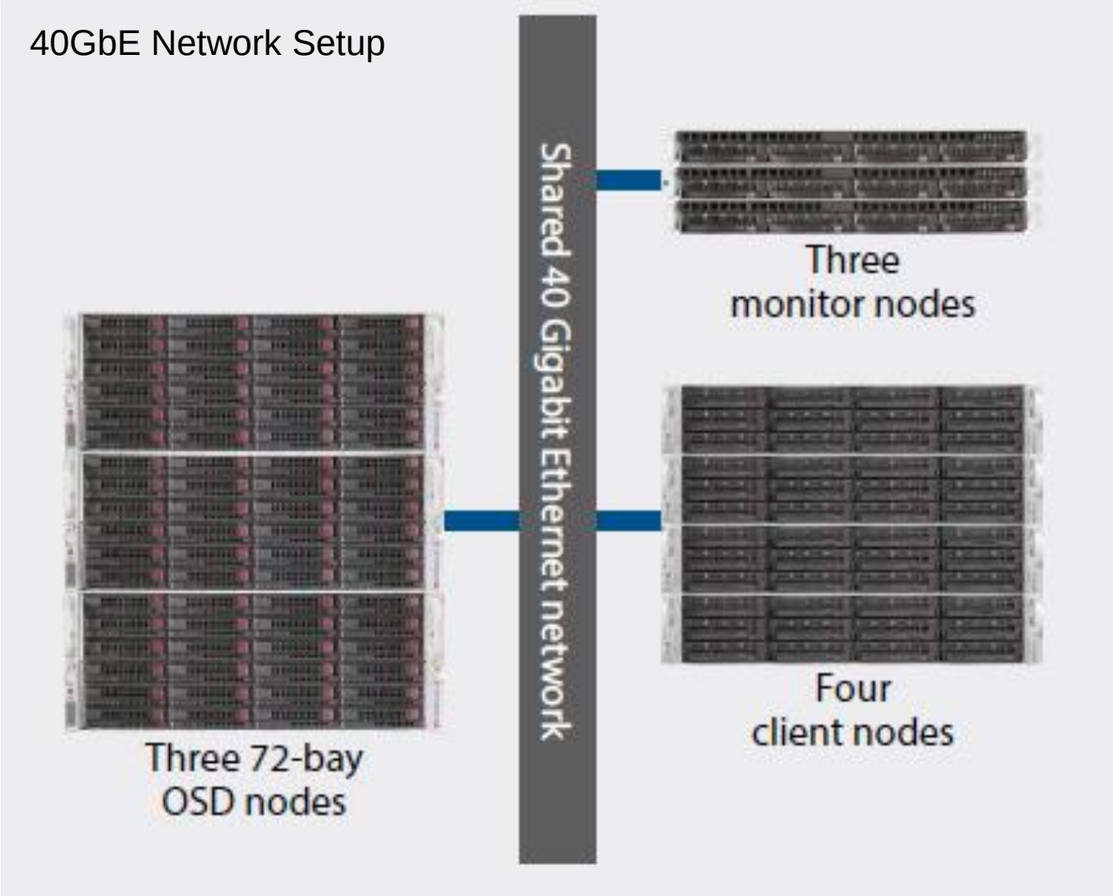


Ceph Performance Testing Using Hard Drives

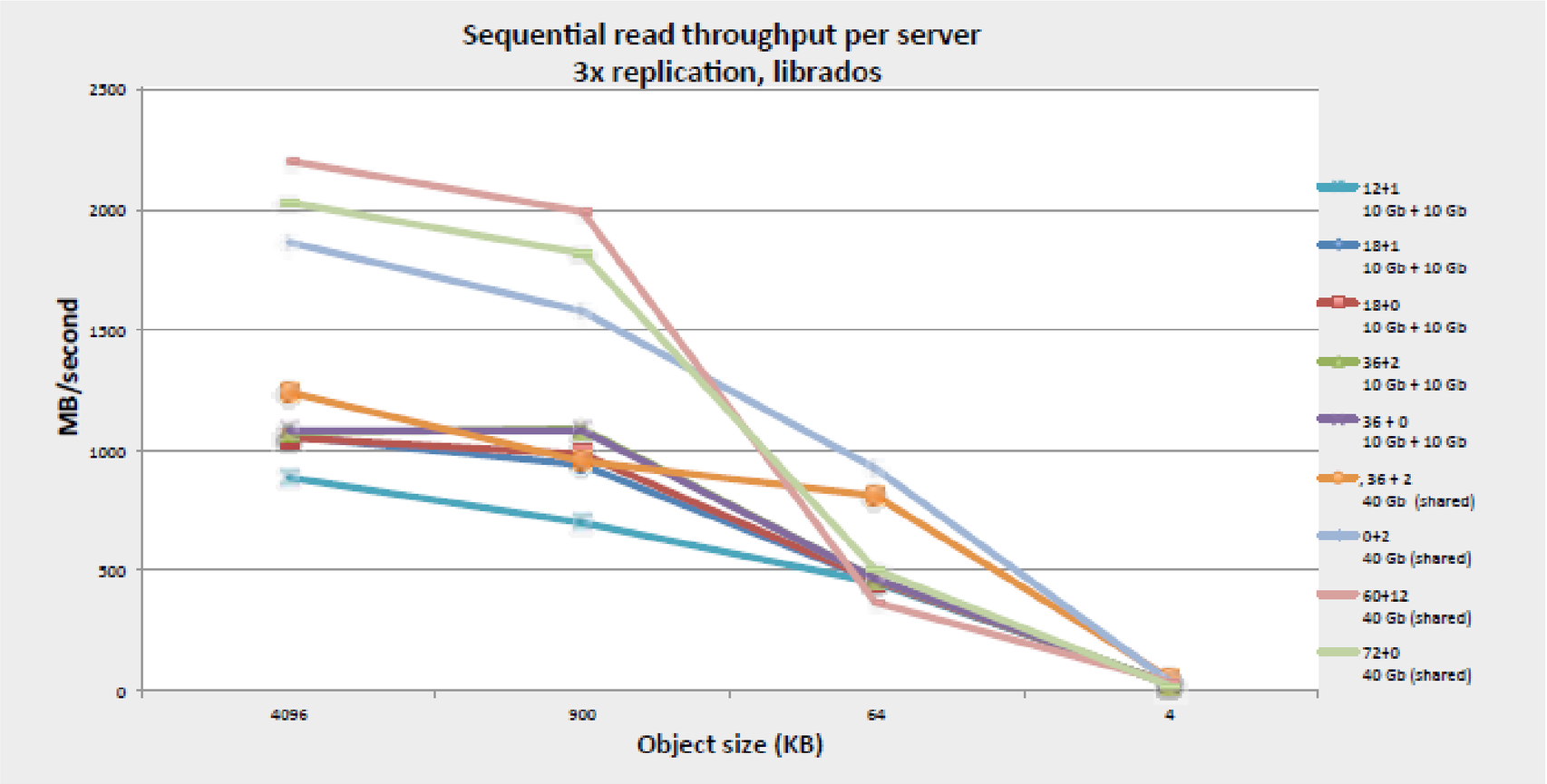
Supermicro: Optimizing For Throughput and Price/Throughput

- Red Hat, Supermicro, Seagate, Mellanox, Intel
- Extensive Performance Testing: Disk, Flash, Network, CPU, OS, Ceph
- Reference Architecture Published on the Red Hat site



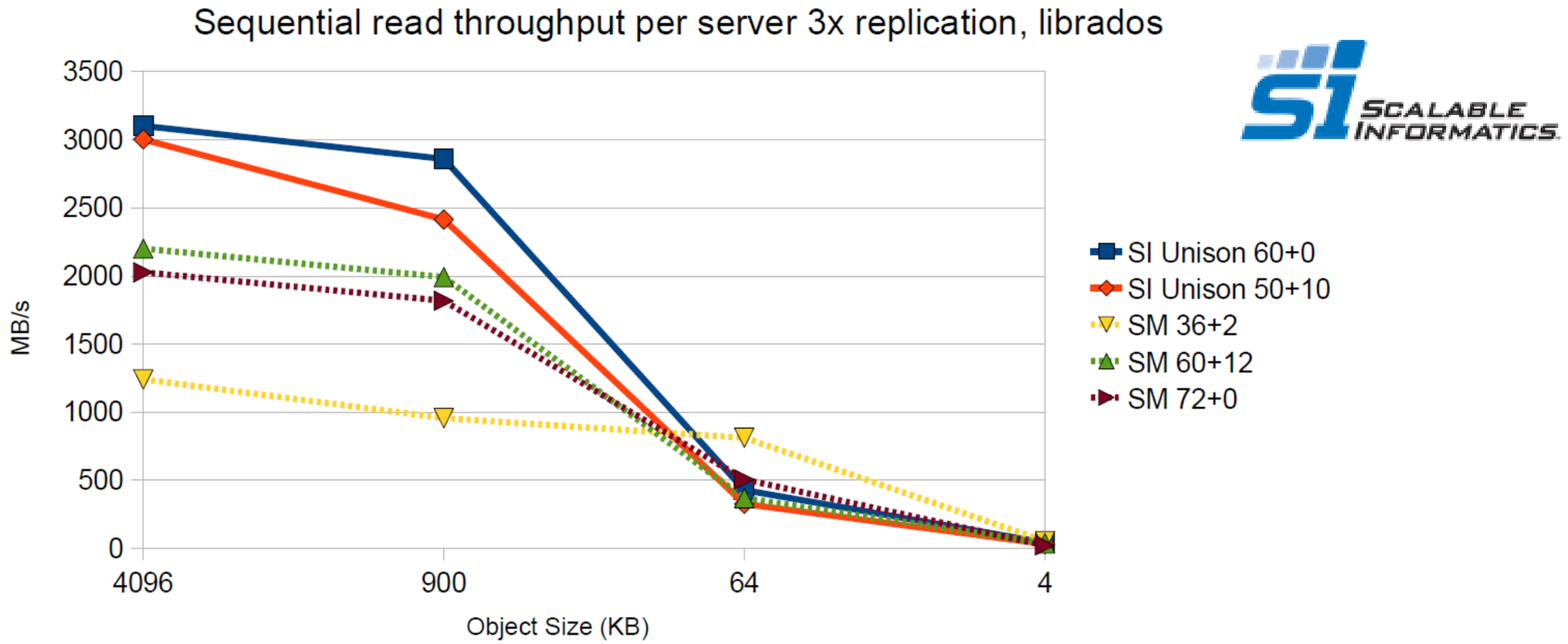


Supermicro Testing 12 -72 Disks Per Node, 2x10GbE vs. 40GbE



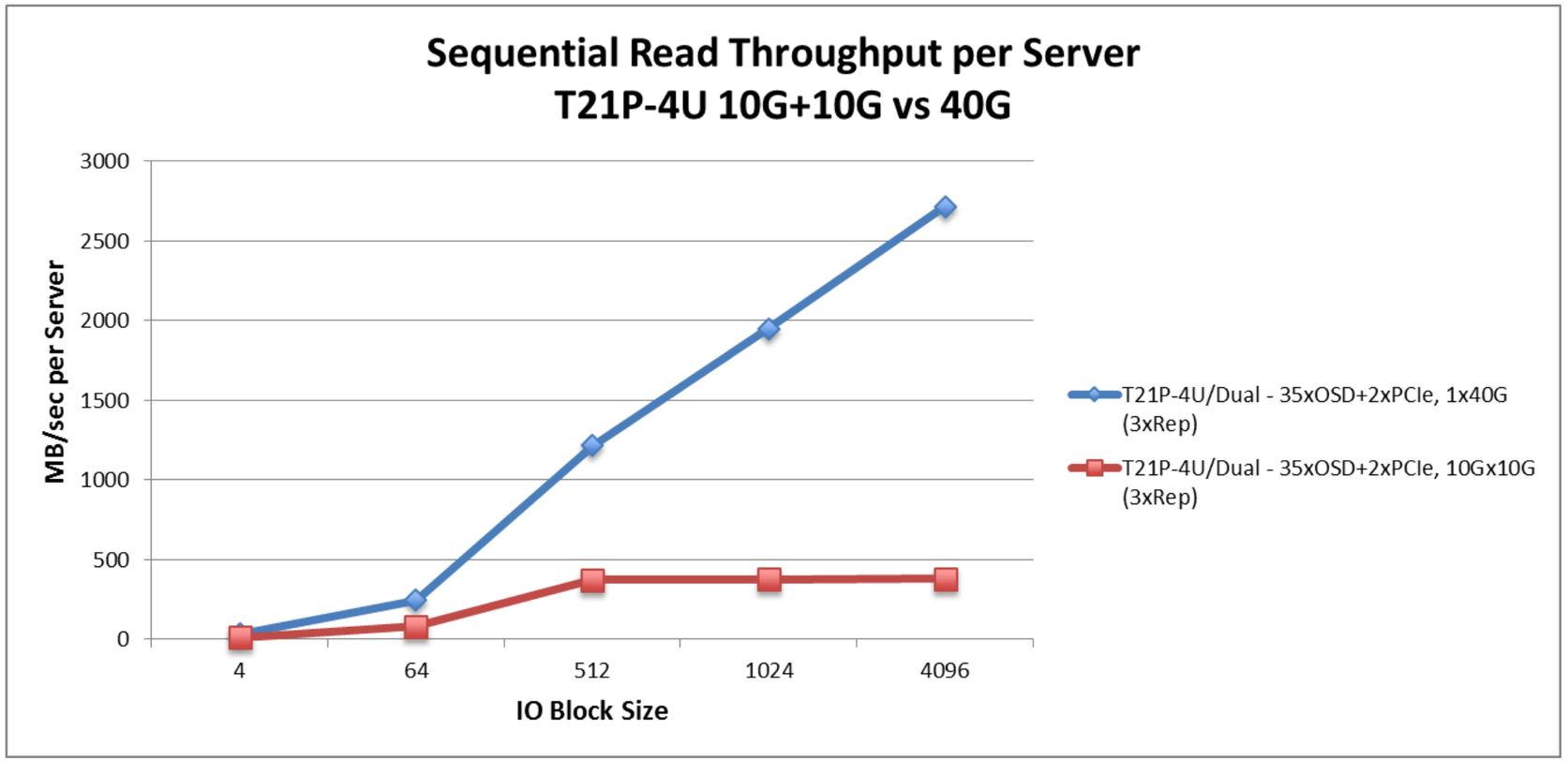
- Key Test Results
  - More disks = more MB/s per server, less/OSD
  - More flash is faster (usually)
  - All-flash 2-SSDs node faster than 36 HDDs
- 40GbE Advantages
  - Up to 2x read throughput per server
  - Up to 50% decrease in latency
  - Easier than bonding multiple 10GbE links

Scalable Informatics Ceph Testing at 40GbE



- Scalable Informatics Test Setup
  - 50 * 3TB SATA HDD + 10 * SATA SSD, or…
  - 60 * 3TB SATA HDD (no SSDs)
  - Mellanox 40GbE network
- Key Test Results
  - Read 3000 MB/s per server with 3x rep (24 Gb/s)
  - Read 2500 MB/s per server with 6+2 EC (20 Gb/s)
  - Definitely need 40GbE

QCT (Quanta Cloud Technology) & Red Hat Ceph Storage Testing



- QuantaPlex T21P-4U Dual-Node
  - 2 OSD nodes, 70 HDD & 4 SSD per server
  - 35x 8TB HDD + 2x PCIe SSD per node
  - 10GbE or Mellanox 40GbE NIC
- Key 40GbE Test Results
  - Up to 2700MB/s read per node
  - Up to 7x faster reads than 10GbE

Optimizing Ceph for Flash

Ceph Flash Optimization By SanDisk

Highlights Compared to Stock Ceph

- Read performance up to 8x better
- Write performance up to 2x better with tuning

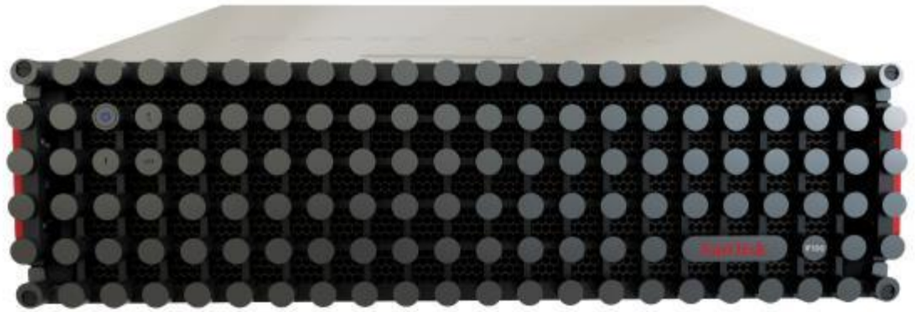
Optimizations

- All-flash storage for OSDs
- Enhanced parallelism and lock optimization
- Optimization for reads from flash
- Improvements to Ceph messenger

Test Configuration

- InfiniFlash Storage with IFOS 1.0 EAP3
- Up to 4 RBDs
- 2 Ceph OSD nodes, connected to InfiniFlash
- 40GbE NICs from Mellanox

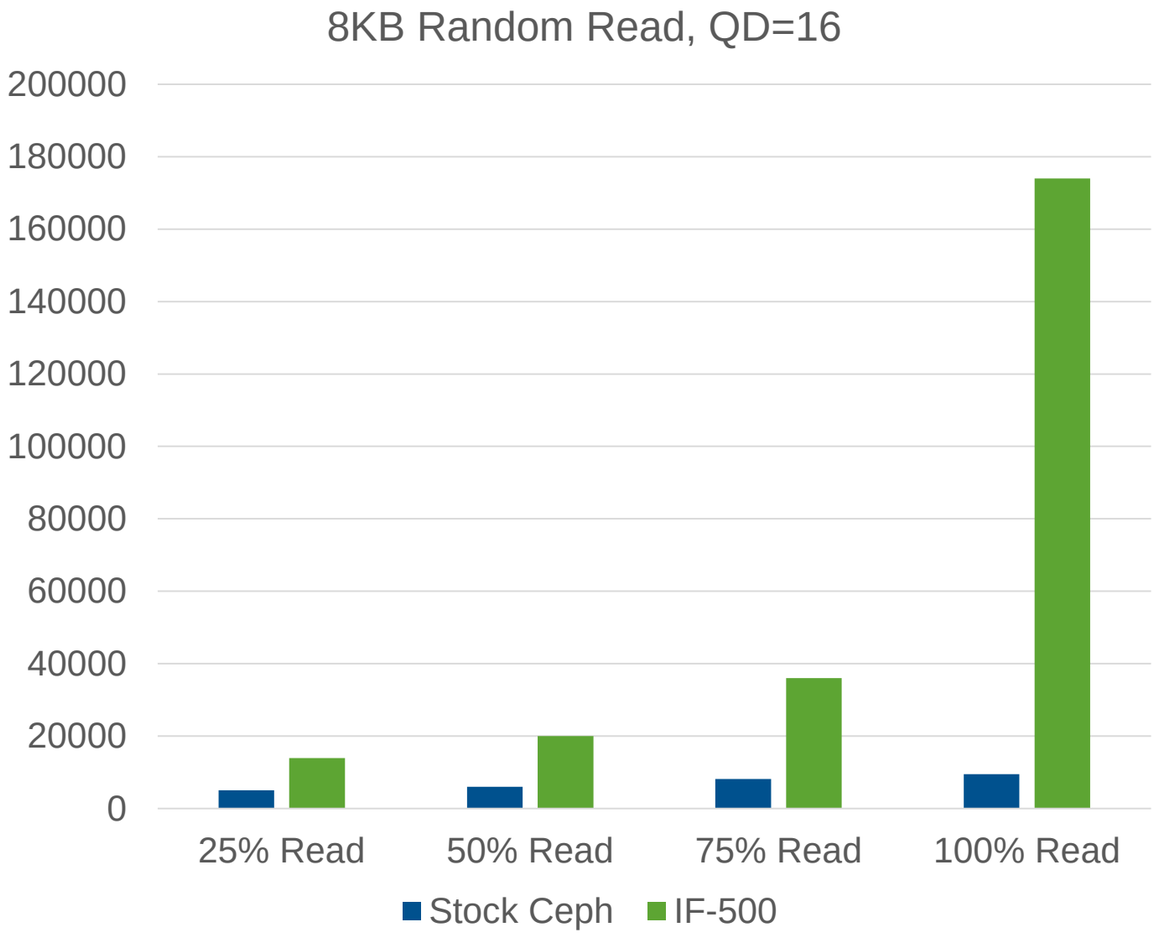
SanDisk



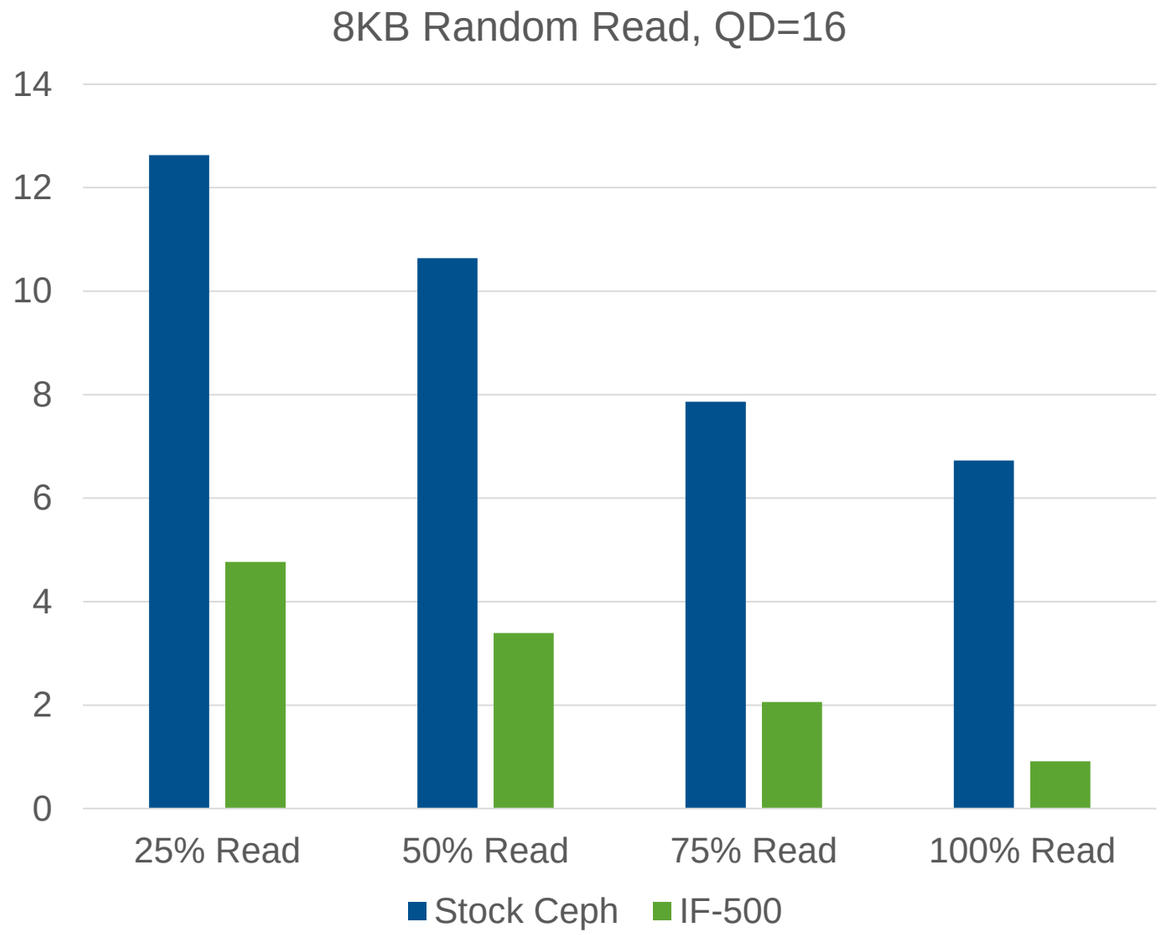
SanDisk InfiniFlash

SanDisk InfiniFlash, Maximizing Ceph Random Read IOPS

Random Read IOPs



Random Read Latency (ms)



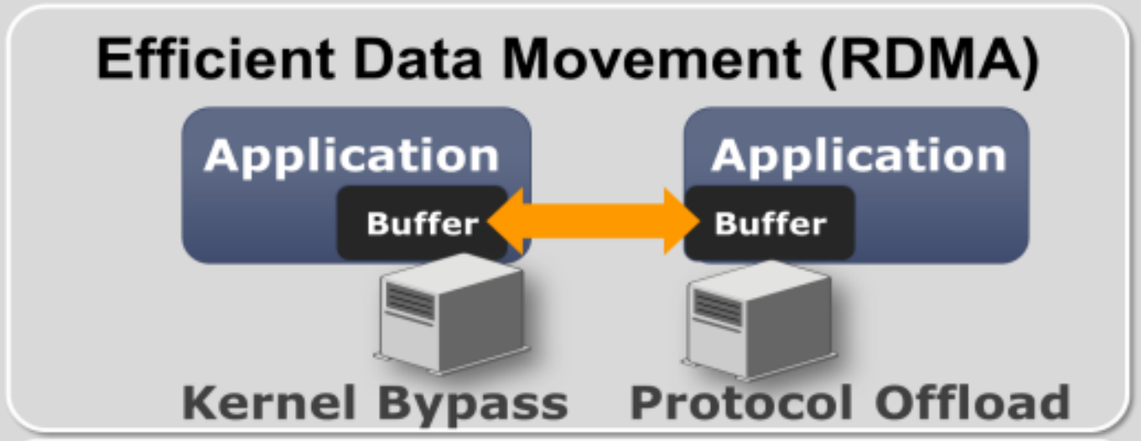
Ceph On Flash Needs 40GbE (or 100GbE)

Setup	**SanDisk InfiniFlash**	**Scalable Informatics**	**Supermicro** (more all-flash tests coming)	**Mellanox** (more all-flash tests coming)
OSD Servers	Dell R720	SI Unison	Supermicro	Supermicro
OSD Nodes	2	1	3	2
Flash	1 InfiniFlash 64x8TB = 512TB	24 SATA SSDs per node	2x PCIe SSDs per node	12x SAS SSDs per node
Cluster Network	40GbE	100GbE	40GbE	56GbE
Total Read Throughput	71.6 Gb/s	70 Gb/s	43 Gb/s	44 Gb/s
Per-Server Read Throughput	35 Gb/s	70 Gb/s	14 Gb/s	22 Gb/s

Adding RDMA To Ceph

XioMessenger

RDMA Enables Efficient Data Movement



Efficient Data Movement With RDMA

Higher Bandwidth

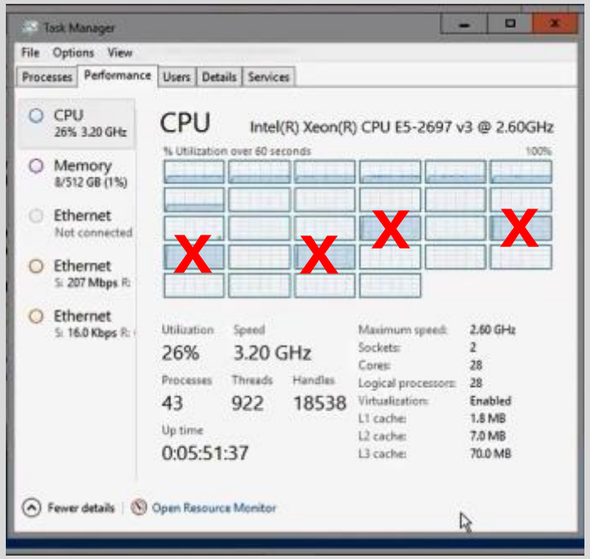
Lower Latency

More CPU Power For Applications

- Hardware Network Acceleration → Higher bandwidth, Lower latency
- Highest CPU efficiency → more CPU Power To Run Applications

RDMA Enables Efficient Data Movement At 100Gb/s

100GbE With CPU Onload



CPU Onload Penalties

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

100 GbE With Network Offload

2X Better Bandwidth

Half the Latency

33-50% Lower CPU

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

Microsoft

- 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

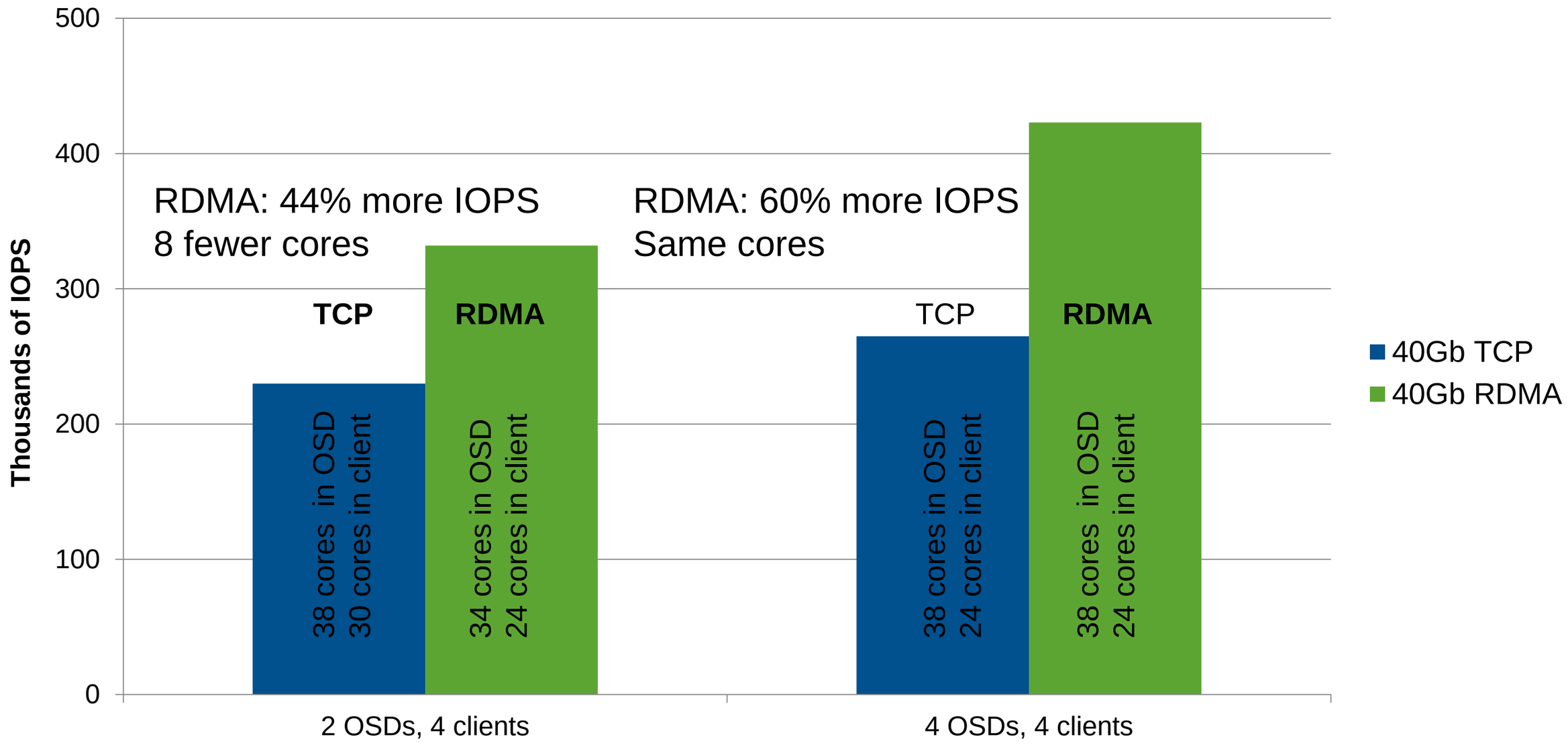
Adding RDMA to Ceph

- RDMA Beta Included in *Hammer*
  - Mellanox, Red Hat, and CohortFS (now part of Red Hat)
  - Full RDMA expected in *Infernalis* release

- Refactoring of Ceph Messaging Layer
  - New RDMA messenger layer called XioMessenger
  - New class hierarchy allowing multiple transports (simple one is TCP)
  - Async design that leverages Accelio
  - Reduced locks; Reduced number of threads

- XioMessenger built on top of Accelio (RDMA abstraction layer)
  - Integrated into all CEPH user space components: daemons and clients
  - Both “public network” and “cloud network”

Ceph 4KB Read IOPS: 40Gb TCP vs. 40Gb RDMA



Deployment Examples

Appliances, Integrators, and Customers

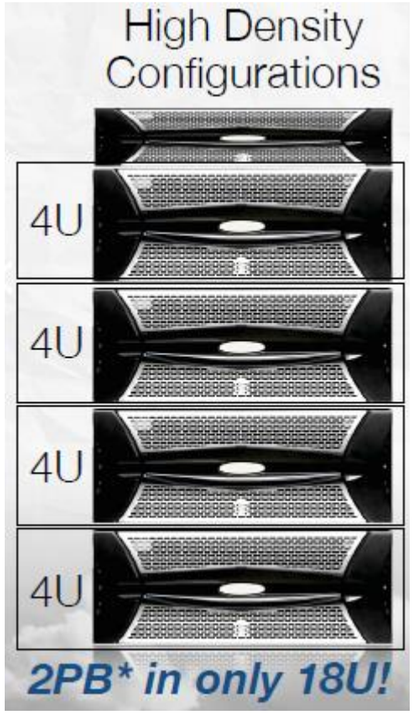
Ceph For Large Scale Storage– Fujitsu Eternus CD10000

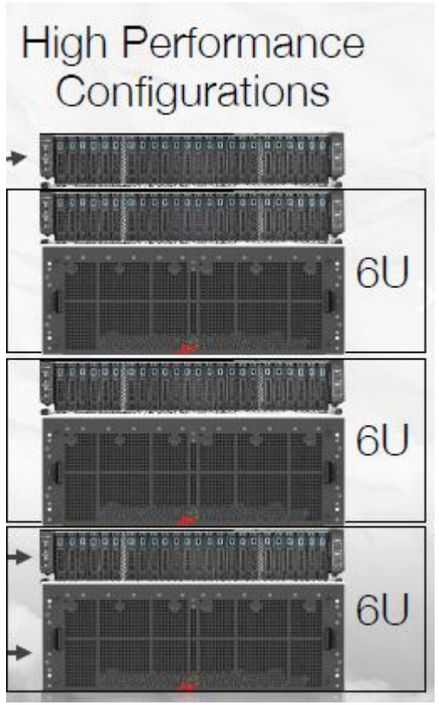
- Hyperscale Storage
  - 4 to 224 nodes
  - Up to 56 PB raw capacity
- Runs Ceph with Enhancements
  - 3 different storage nodes
  - Object, block, and file storage
- Mellanox InfiniBand Cluster Network
  - 40Gb InfiniBand cluster network
  - 10Gb Ethernet front end network

Media & Entertainment Storage – StorageFoundry Nautilus

- Turnkey Object Storage
  - Built on Ceph
  - Pre-configured for rapid deployment
  - Mellanox 10/40GbE networking
- High-Capacity Configuration
  - 6-8TB Helium-filled drives
  - Up to 2PB in 18U
- High-Performance Configuration
  - Single client read 2.2 GB/s
  - SSD caching + Hard Drives
  - Supports Ethernet, IB, FC, FCoE front-end ports
- More information: www.storagefoundry.net

powered by ceph
NAUTILUS



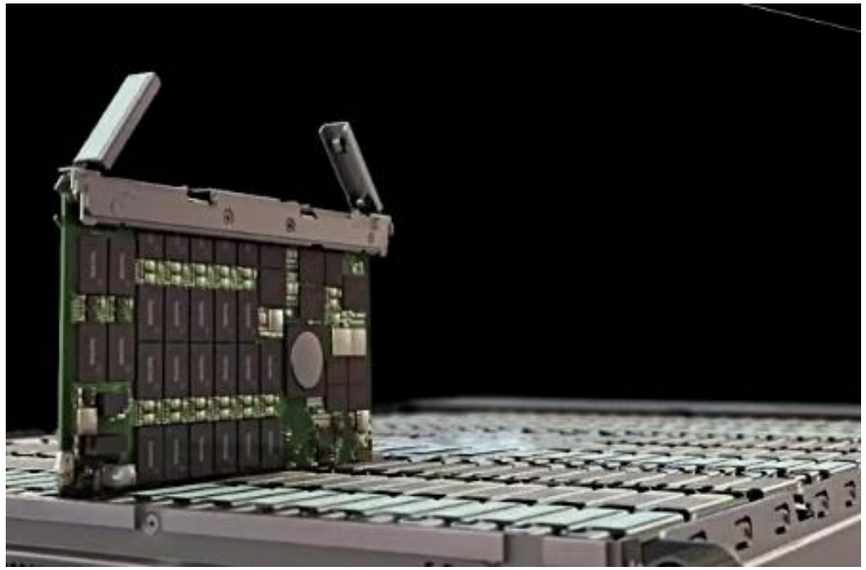


Fast Ceph Storage – Scalable Informatics Unison

- For Multi-tenancy and Cloud Deployments
  - Powered by Ceph
  - Object, block, file and Hadoop storage
  - Mellanox 10/40GbE networking
- High-Availability Cluster
  - 60 HDD in 4U or 50HDD + 10 SSD
  - Available in all-flash configuration
  - 40, 56, or 100Gb/s cluster network
- More information:
  - www.scalableinformatics.com/unison.html
  - See the white paper

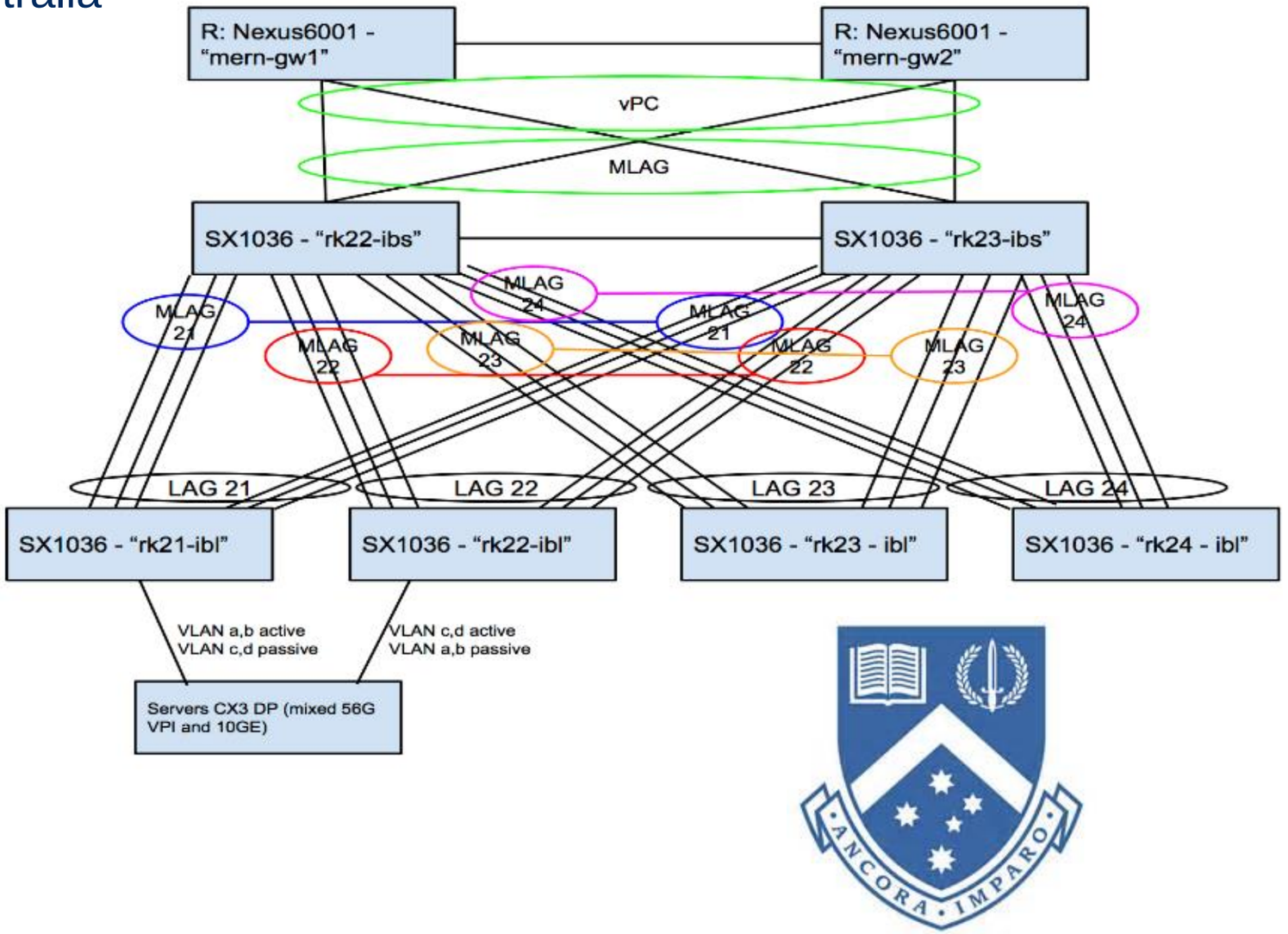
SanDisk InfiniFlash

- Flash Storage System
  - Announced March 2015
  - 512 TB (raw) in one 3U enclosure
  - Tested with 40GbE networking
- High Throughput
  - 8 SAS ports, up to 7GB/s
  - Connect to 2 or 4 OSD nodes
  - Up to 1M IOPS with two nodes
- More information:
  - http://bigdataflash.sandisk.com/infiniflash



Ceph Customer: Monash University

- Research University in Melbourne, Australia
  - 67,000 students and 15,000 staff
  - 9 locations in 5 countries
- 3 Ceph Clusters
  - 8, 17, and 37 nodes
  - >6PB of storage
  - Hybrid OSD nodes (mostly HDD)
- OpenStack Storage
  - Cinder for block, S3/Swift for object
- Mellanox Networking
  - SX1036 switches
  - 10GbE to each Ceph node
  - 56GbE links between switches



Summary

- Ceph Benefits from Faster Networks – 10GbE is not enough!
  - If using >20 HDD per server or all-flash
- End-to-end 40/56 Gb/s transport accelerates Ceph today
  - 100GbE testing has begun!
  - Available in various Ceph solutions and appliances
- What's Coming Next for Ceph
  - RDMA to optimize flash performance—beta in *Hammer*
  - Erasure Coding hardware offload
  - 25GbE Testing

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™