SNIA DSI '16 DATA STORAGE INNOVATION CONFERENCE
Innovation in Storage Products, Services, and Solutions
June 13-15, 2016 | Marriott San Mateo | San Mateo, CA

Benefits of 25, 40, and 50GbE Networks for Ceph and Hyper-Converged Infrastructure

John F. Kim

Mellanox Technologies

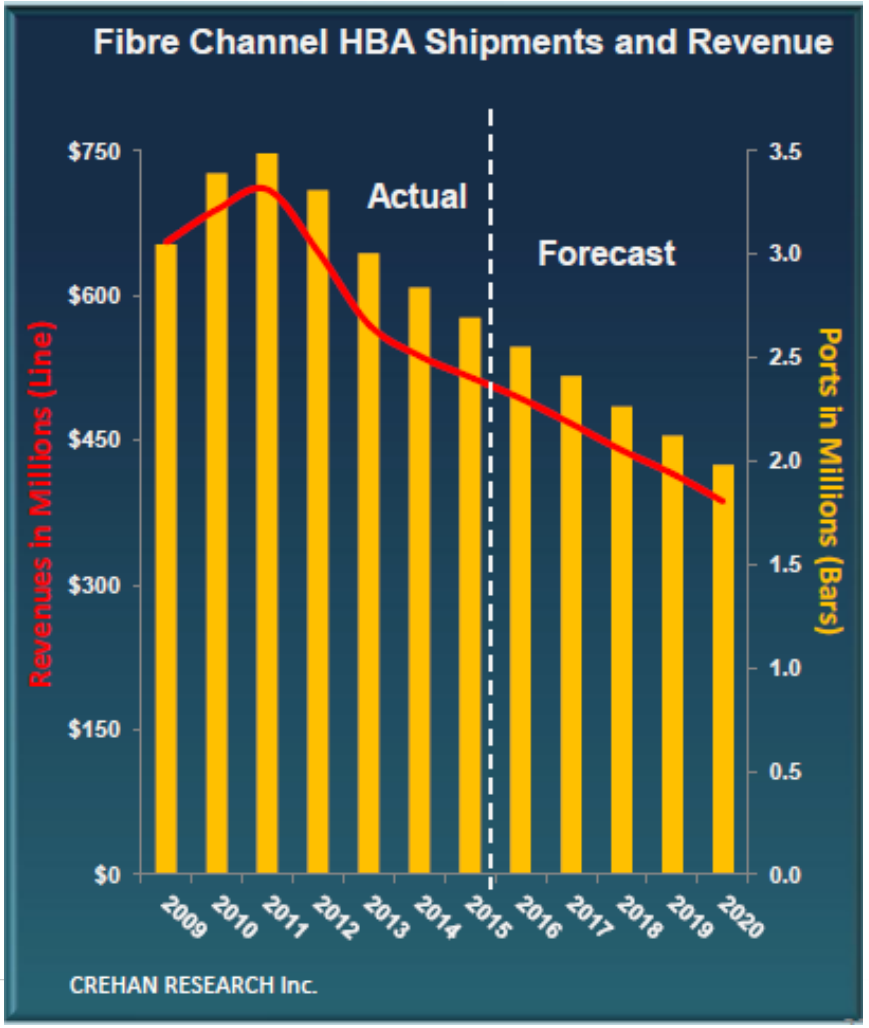
Storage Transitions Change Network Needs

HYPERCONVERGED INFRASTRUCTURE

- Major storage technology transitions
  - Software Defined Storage
  - SSDs replacing hard disk drives
  - Hyper-converged and Cloud
- All require faster networks, and efficient transport
  - More east-west traffic
  - Faster storage needs faster networks
  - Higher bandwidth needs better offloads

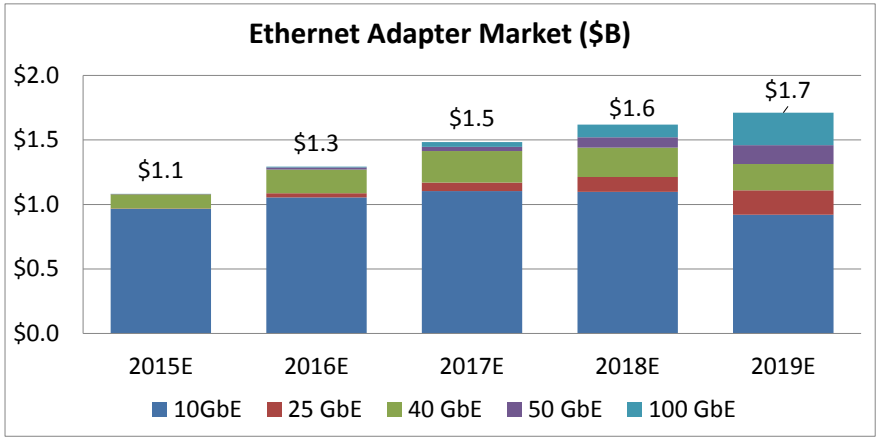
Fibre Channel in a Slow Decline

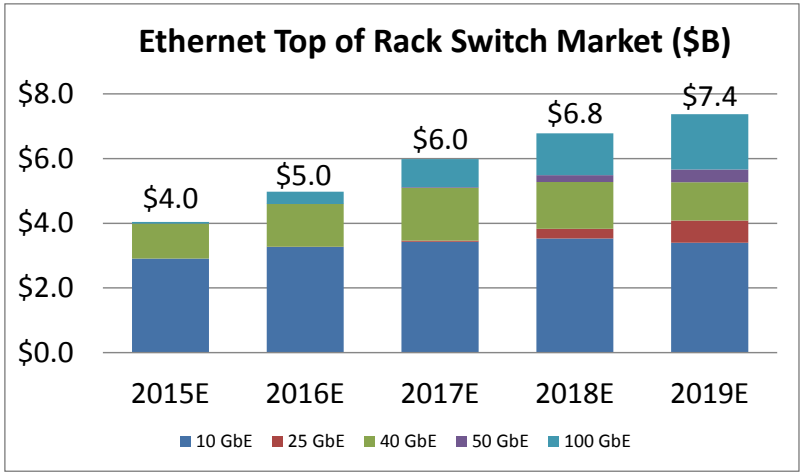
- FC being outgrown
  - Ethernet Block & NAS
  - Distributed Storage
  - Big Data
  - Hyper-converged
  - 4K video editing
- FC-SANs do not go away
  - But they don’t grow
  - New storage is Ethernet



25/50/100 GbE Rapid Growth Projected

- 25GbE to grow rapidly
  - Fastest ramp in history
  - Passes 40GbE in 2019
- 50 & 100GbE Also Grow
  - Cloud/Web 2.0 customers
  - Large enterprises



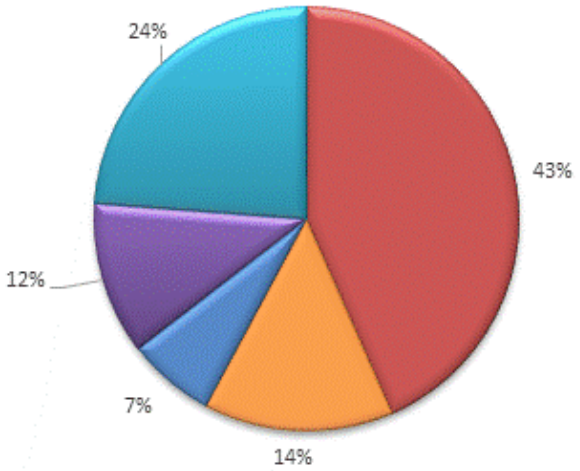


Crehan Research: Q2'15 Server Class Adapter Market Share (August 2015), Long Range Forecast – Server-class Adapters & LOM (July 2015), Adapter forecast includes LOM;

Ethernet Growth in 25Gb/s & Faster

- By 2020, Over 50% of market
  - Fastest ramp in history
  - Passes 40GbE in 2019
- 50 & 100GbE Also Grow
  - Cloud/Web 2.0 customers
  - Large enterprises

High Speed Adapter Market Forecast
2020 ($1.8B)



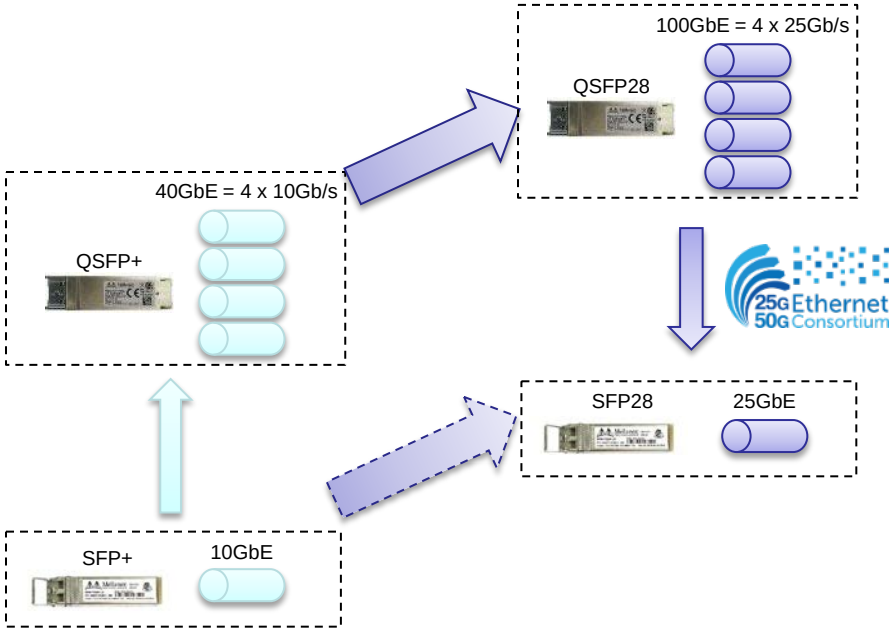
25/40/50/100 GbE speeds will make up 57% of high-speed adapter sales by 2020

Crehan Research: Q3'15 Server Class Adapter Market Share (Nov 2015), Long Range Forecast – Server-class Adapters & LOM (Jan 2016), Adapter forecast includes LOM;

25 Is the New 10; 50 is the new 40

- 25GbE = 2.5x bandwidth
  - Only 1.5x the price
  - 40% lower cost/BW
  - Compatible with 10GbE ports
- 50GbE = 1.25x BW at same cost
  - Upgrade from 40GbE
  - 20% lower cost/BW
  - Cheaper optical cables
- 100GbE for switch links
  - 60% fewer links
- Re-use existing fiber cables!

New Ethernet Drives 25Gb/s Per Lane
More Data on the Same Links



High Speed Network Cables: Quick Definition Review

Direct Attach Copper (DAC)

- Uses pulsing electrical signals sent into shielded copper wires.
- At high data rates, wire acts like a radio antenna.
- Data signal leaves the wire and lost, so length becomes shorter.
- Longer the cable, thicker it becomes.

- Maximum length for 10/40GbE – 5m
- Maximum length for 25/100GbE –3m

Key Benefits: Lowest Cost, zero power
Typical usage: Inside rack, 0.5-3m reach

Optical Transceiver

"Transceiver"
n*channels Transmit
n*channels Receiver

- Converts electrical signals to optical.
- Transmits blinking laser light over tiny glass optical fibers.

- 2 fiber types:
  - Multi-mode – up to 100m
  - Single-mode – up to 2Km

- Applies to both 10G/40G and 25G/100G
- More expensive than DAC copper

Key Benefits: Long reaches
Uses: Linking switches up to 2Km.

Active Optical Cable

2 Transceivers with optical fiber glued in.
\+ *Lowest Cost Optical for <200m*

Ceph

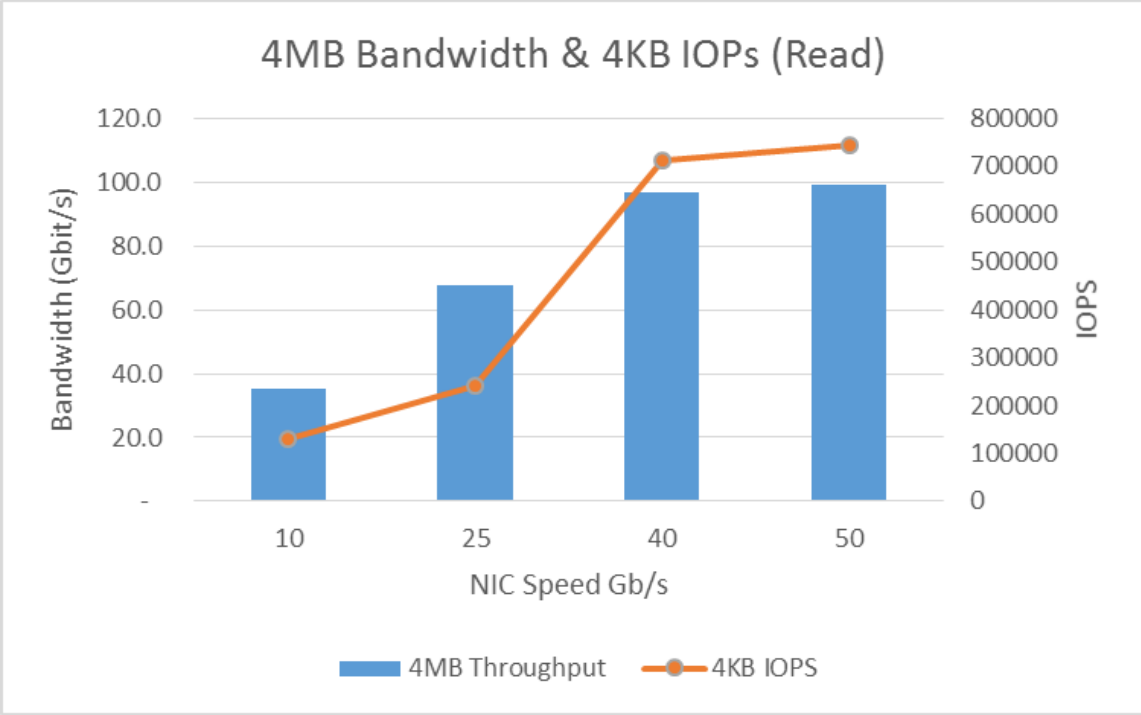
Ceph Performance Testing

- Ceph is a scale-out system
  - Two logical networks
  - 3-way replication or erasure-coding
  - Metadata, monitoring, and client traffic
  - Rebuild or rebalance traffic
- Faster network = faster performance
  - Higher bandwidth and lower latency
  - Increases small block IOPS
  - Increases large block throughput

Mellanox Testing 10/25/40/50 GbE

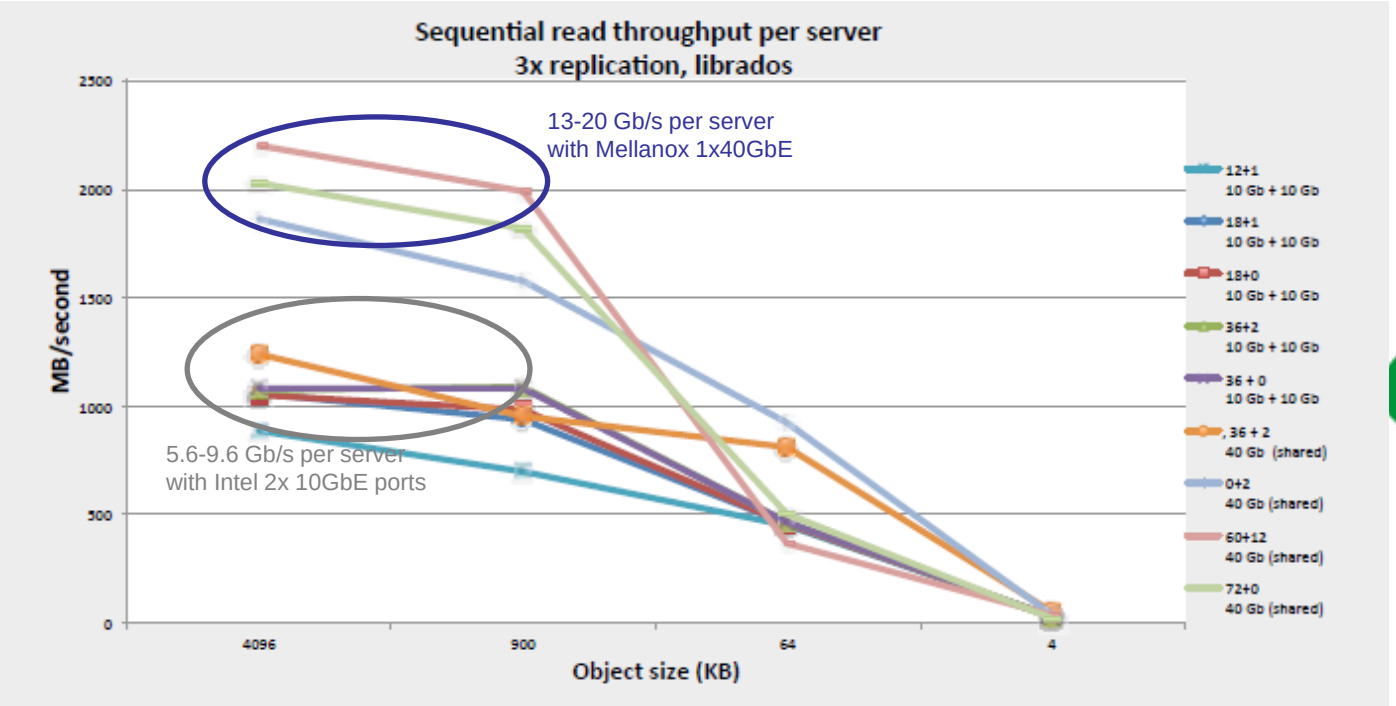
Ceph Performance Testing vs. 10GbE

- 4 Ceph OSD servers
- 3 NVMe SSDs each
- ConnectX-4 Lx
- Set network to 10, 25, 40, and 50GbE speeds



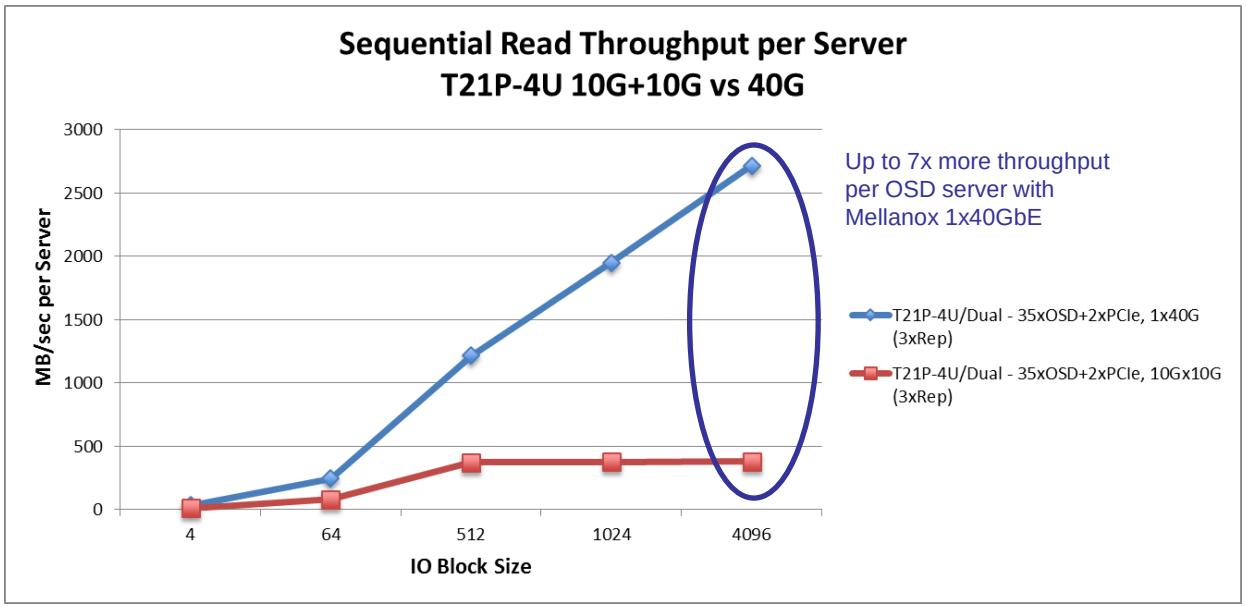
- Aggregate performance of 4 Ceph servers
  - 25GbE: 67Gb/s & 242K IOPS (vs. 35Gb/s & 130K IOPS at 10GbE)
  - 25GbE has 92% more throughput than 10GbE
  - 25GbE has 86% more IOPS than 10GbE

Supermicro Testing: 2x10GbE vs. 40GbE



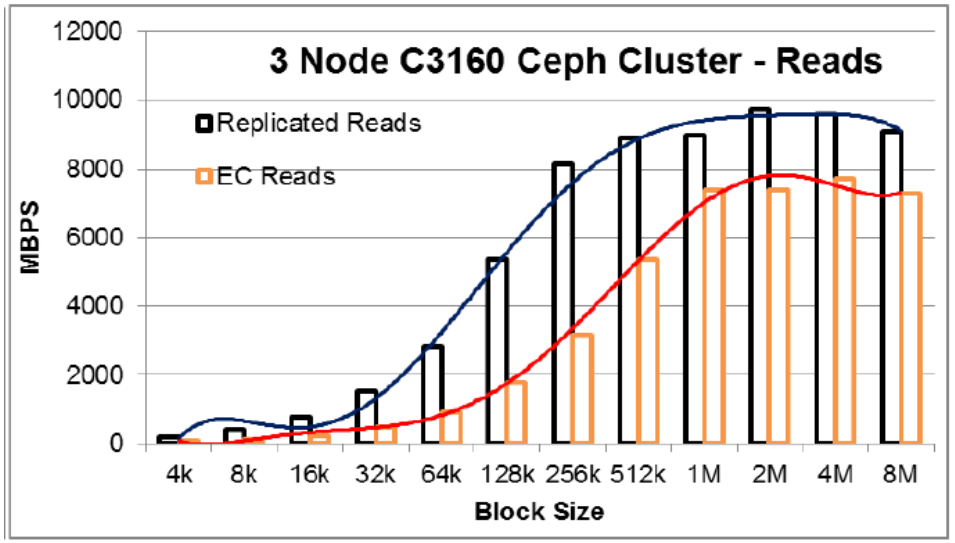
- Key Test Results
  - More disks = more MB/s per server, less/OSD
  - More flash is faster (usually)
  - All-flash 2-SSDs node faster than 36 HDDs
- 40GbE Advantages
  - Up to 2x read throughput per server
  - Up to 50% decrease in latency
  - Easier than bonding multiple 10GbE links

QCT & 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 (21.6Gb/s)
  - Up to 7x faster reads than 10GbE
  - Also faster write throughput, even when <10Gb/s

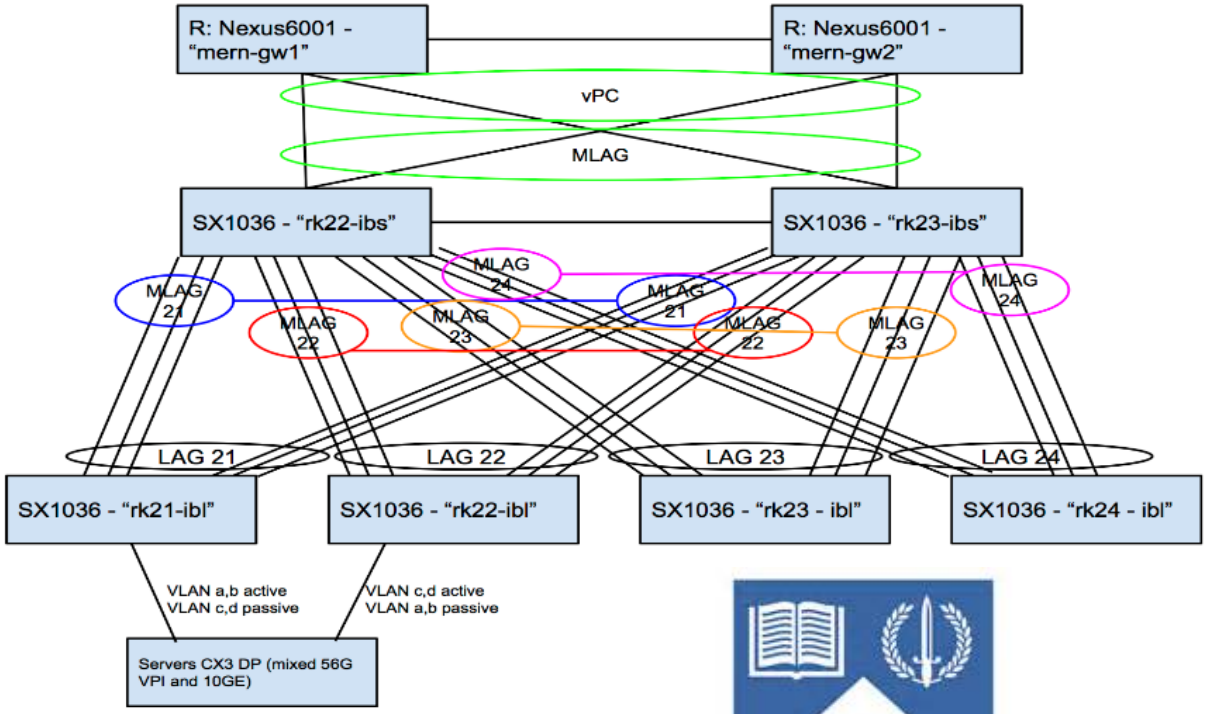
Cisco Ceph Testing: 10GbE Not Enough



- Cisco Test Setup
  - UCS C3160 servers, Nexus 9396PX switch
  - 28 or 56 6TB SAS disks; Replication or EC
  - 4x10GbE per server, bonded
  - (Easier if had used Mellanox 40GbE NICs)
- Results with 3x Replication
  - One node read: 3700 MB/s (29.6 Gb/s)
  - One node write: 860 MB/s rep (6.8 Gb/s)
  - 3 nodes read: 9,700 MB/s (77.6 Gb/s)
  - 8 nodes read: 20,000 MB/s (160 Gb/s)

Ceph Customer Network: Monash University

- Research University
  - 67,000 students
  - 9 locations
  - 3 Ceph clusters
  - >60 nodes
  - >6PB storage
- First network:
  - 10GbE to nodes
  - 56GbE inter-switch
- Second Ceph network
  - 25GbE to nodes
  - 100GbE inter-switch



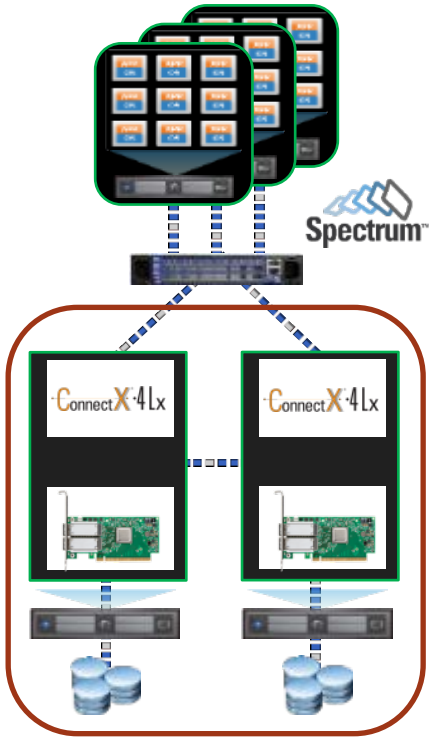
Hyper-Converged

Faster Interconnect Enables Higher ROI

*Faster Networking Enables Savings in VDI Deployments over vSphere**

Interconnect	# Virtual Desktop per Server	# Servers	# Switches	CapEx	CapEx per Virtual Desktop
10GbE	60	84	2	$ 540,545	$108
25GbE**	140	36	1	$ 279,214	$ 55
40GbE	167	30	1	$ 207,890	$ 41

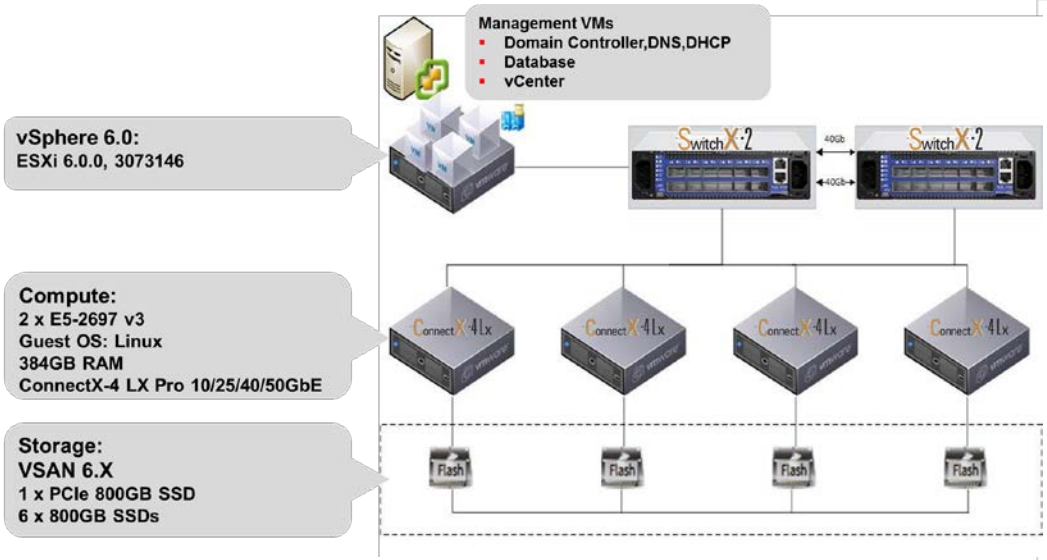
- 50GbE,100GbE and RoCE will enable higher efficiency
- Significant savings when running 5000 virtual desktops

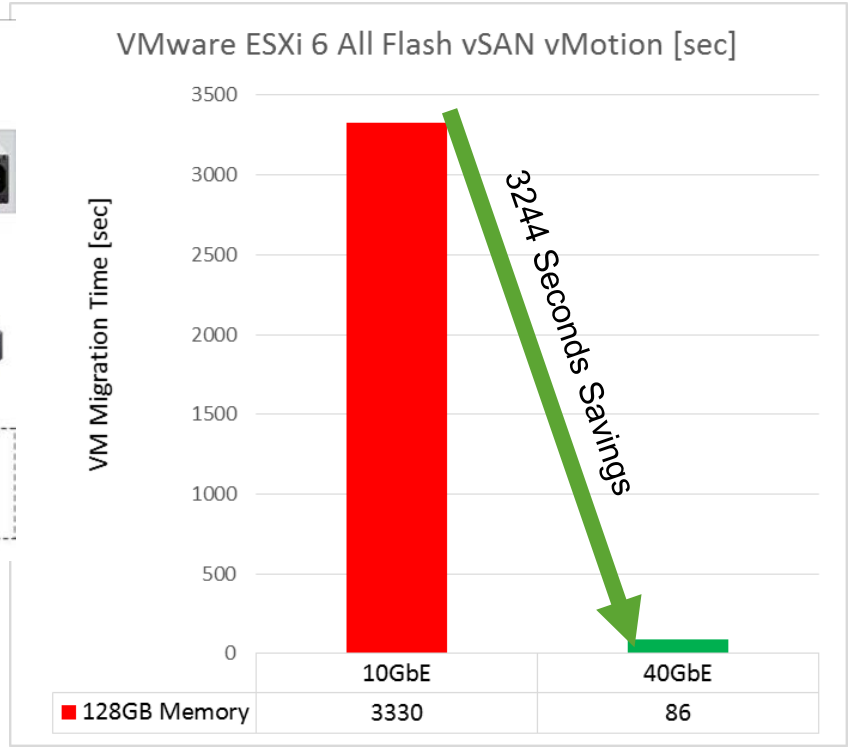


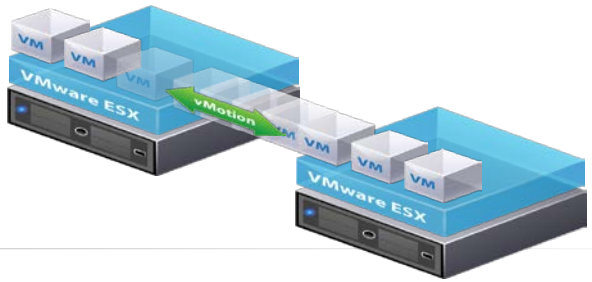
* Hardware savings only (not including VMware, guest OS license) and OpEx
** Interpolation based on the 10GbE and 40GbE results

Higher Performance Interconnect Enables >2X Higher Efficiency

38x Faster Live Migration on 40GbE

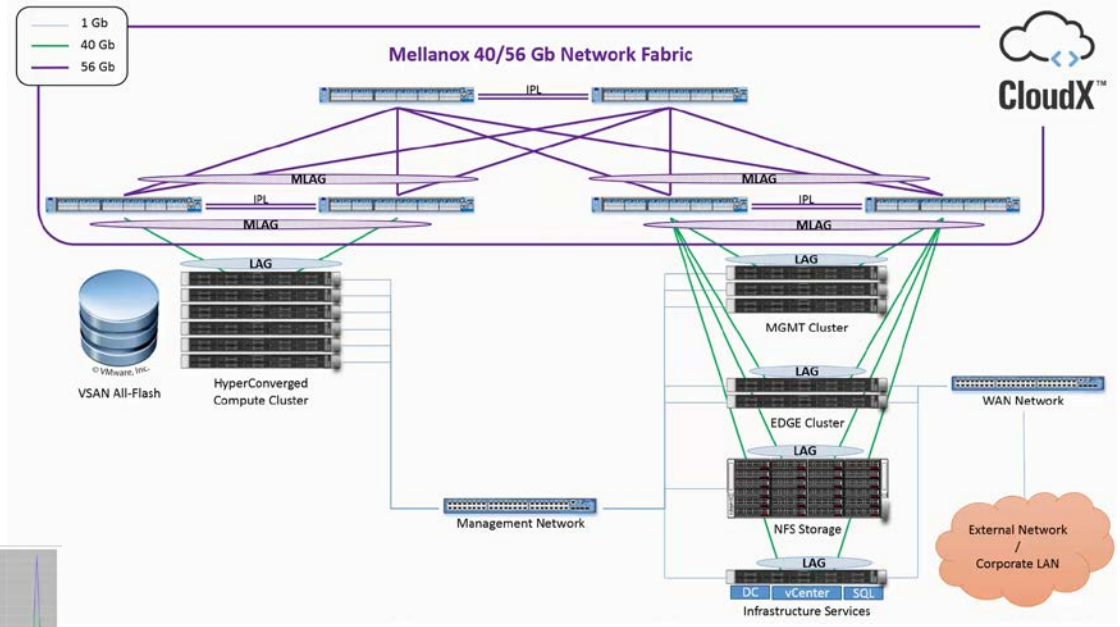


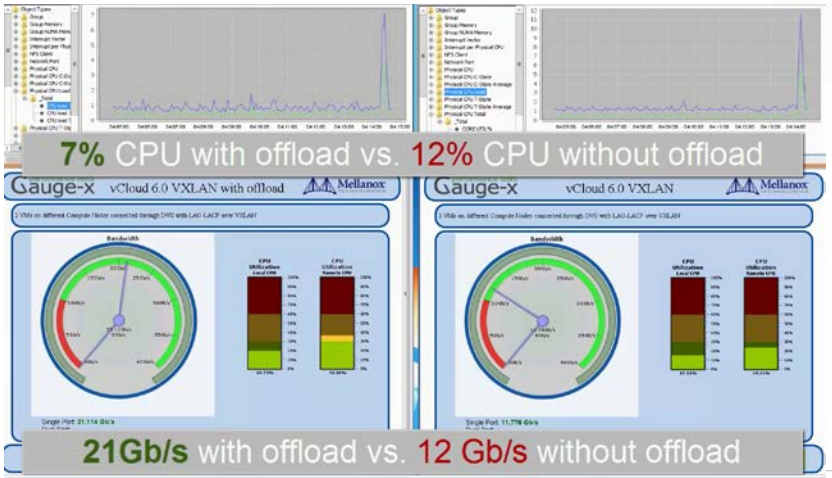




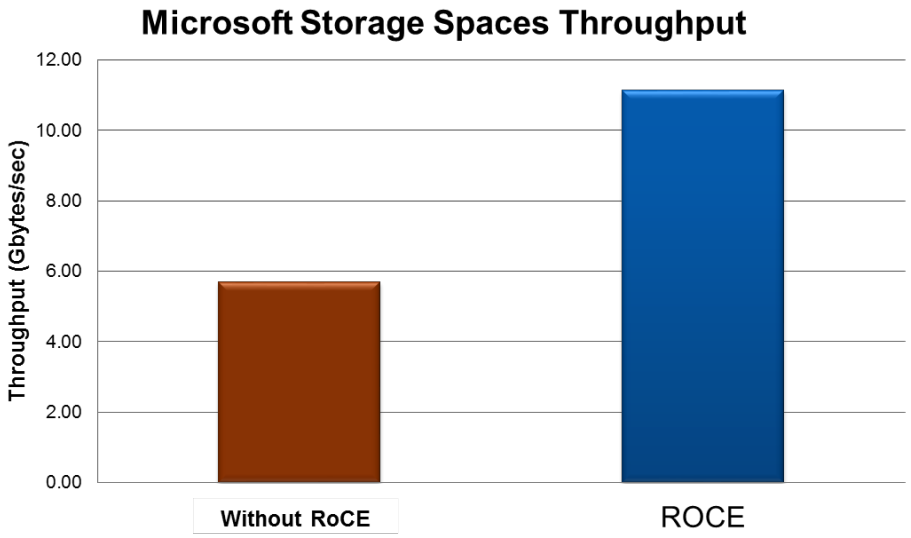
Higher Efficiency with VSAN on 40GbE

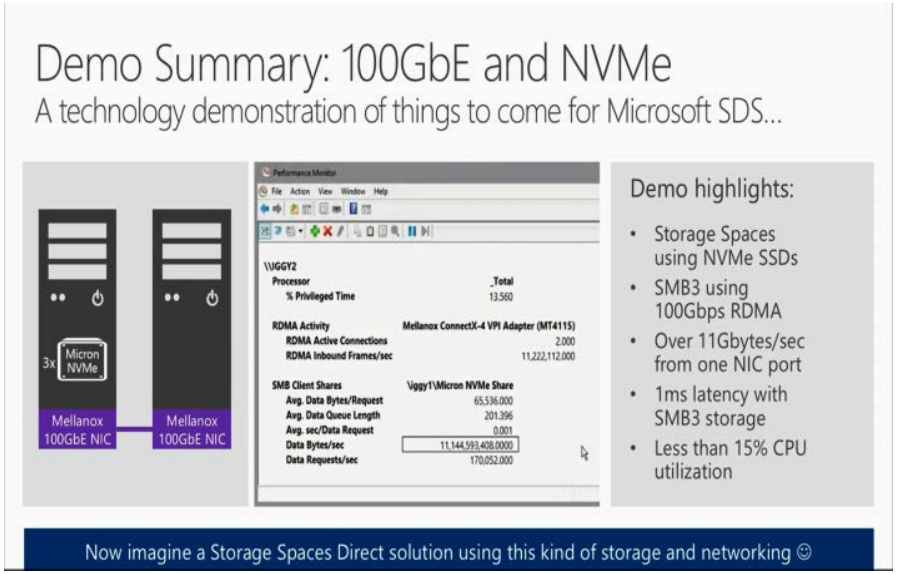
- VMware VSAN
  - All-flash storage
  - vCloud Suite 6.0
- Network:
  - 40/56 GbE
  - NSX





Windows Storage on 100GbE With RDMA





- 2X better performance with RoCE
  - 2X higher bandwidth & 2X better CPU efficiency
- RoCE achieves full Flash storage bandwidth
  - Remote storage without compromises

Microsoft Storage Spaces Direct @ 100Gb/s

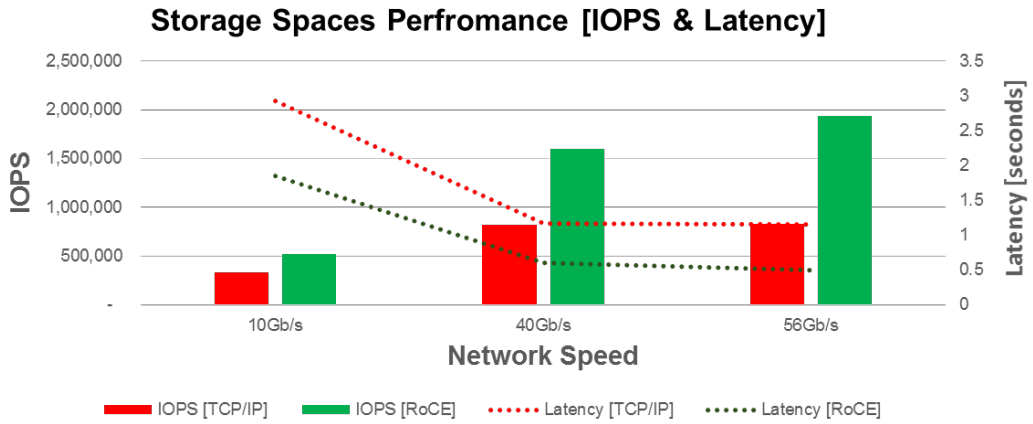
- Microsoft recently benchmarked 100Gb/s Storage Spaces Direct
  - 4 Dell R730XD Servers
  - Samsung NVMe Flash
  - ConnectX-4 100GbE Adapters
  - RoCE for RDMA
- 60 GByte/sec aggregate throughput
  - Could transmit entire content of Wikipedia in 5 seconds

4 R730XD Servers

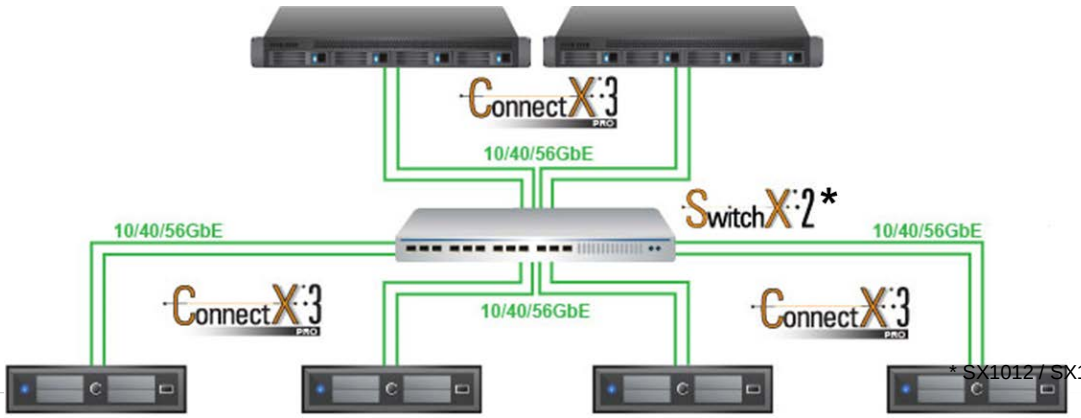
https://blogs.technet.microsoft.com/filecab/2016/05/11/s2dthroughputtp5/

Windows Storage Spaces –Faster Networks

- RoCE vs. TCP IOPS
  - 10Gb/s: +58%
  - 40Gb/s: +94%
  - 56Gb/s: +131%
- RoCE vs. TCP Latency
  - 10Gb/s: -63%
  - 40Gb/s: -51%
  - 56Gb/s: -43%
- 40GbE vs. 10GbE IOPS
  - TCP/IP: +151%
  - RoCE: +208%
- 56Gb/s vs. 40Gb/s
  - TCP/IP: None
  - RoCE: +20%



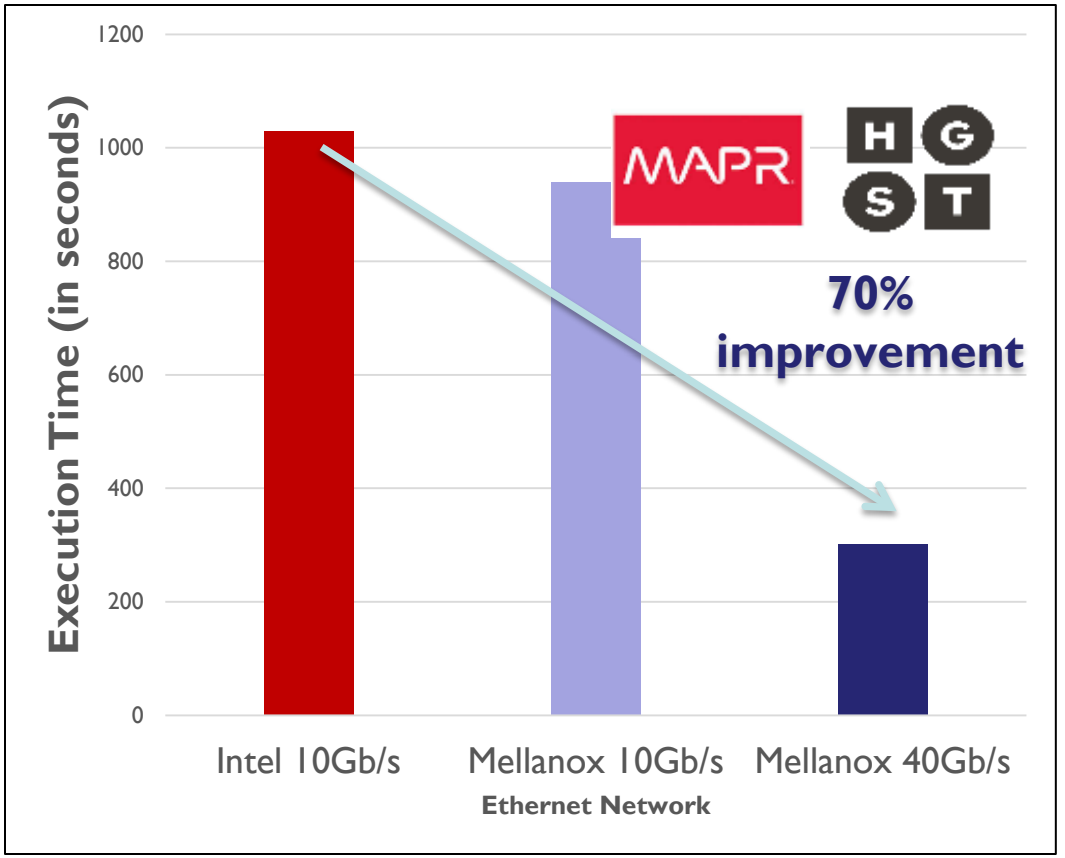
Windows Server 2016 Storage Spaces Direct Setup



* SX1012 / SX1036

Other Performance Testing

- MapR Comparison
  - 40GbE up to 70% Better
  - HGST SSDs
- Nexenta Edge
  - 128KB random writes
  - 25GbE: 14.4Gb/s
  - 50GbE: 23 Gb/s



Thank You