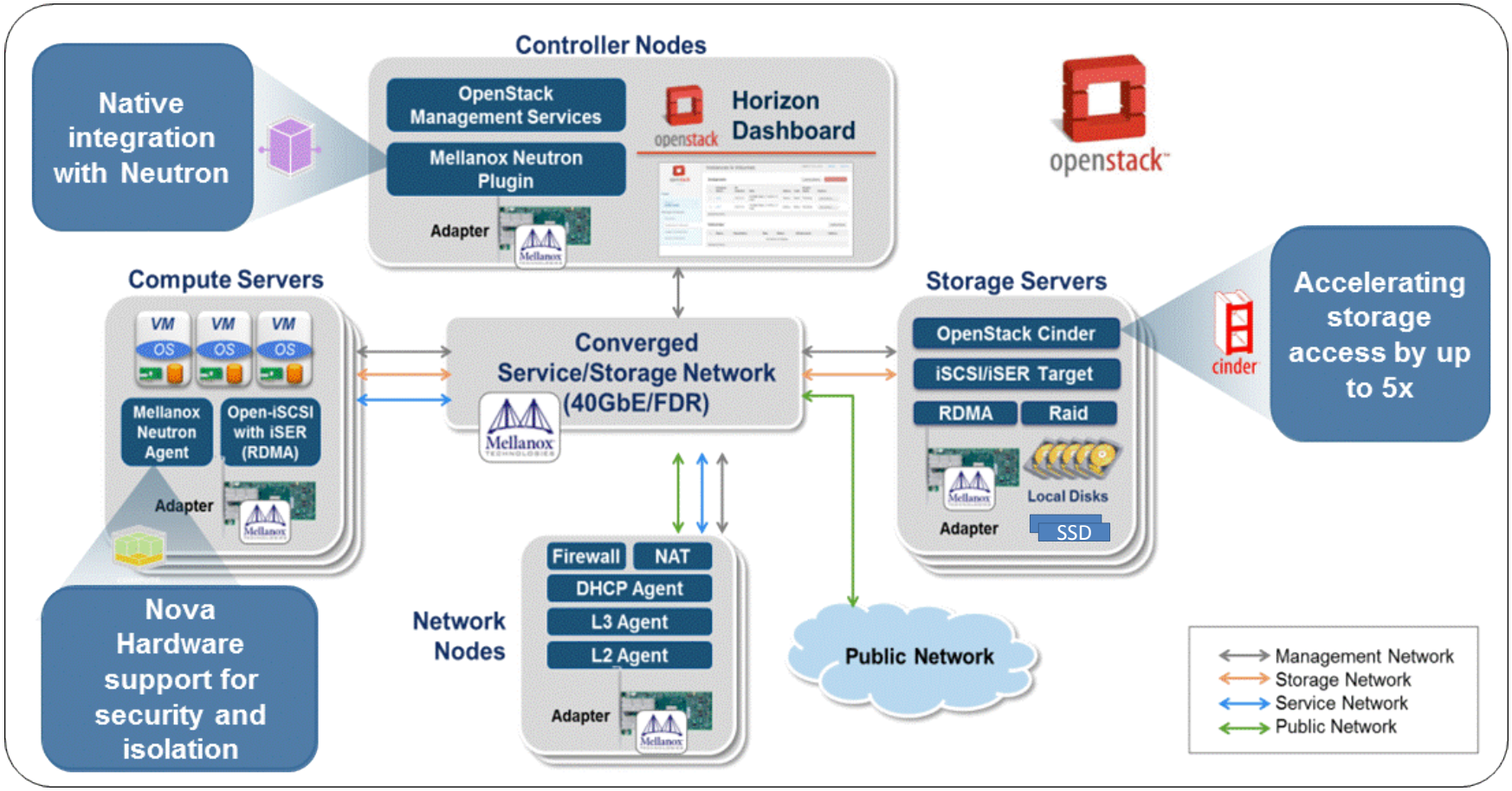
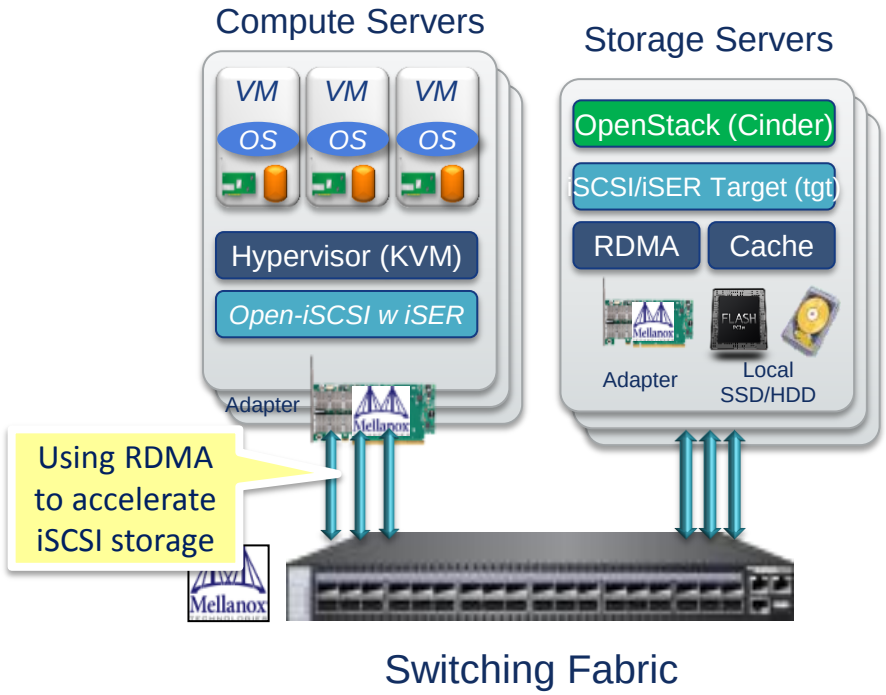
Flash Storage & 40GbE for OpenStack

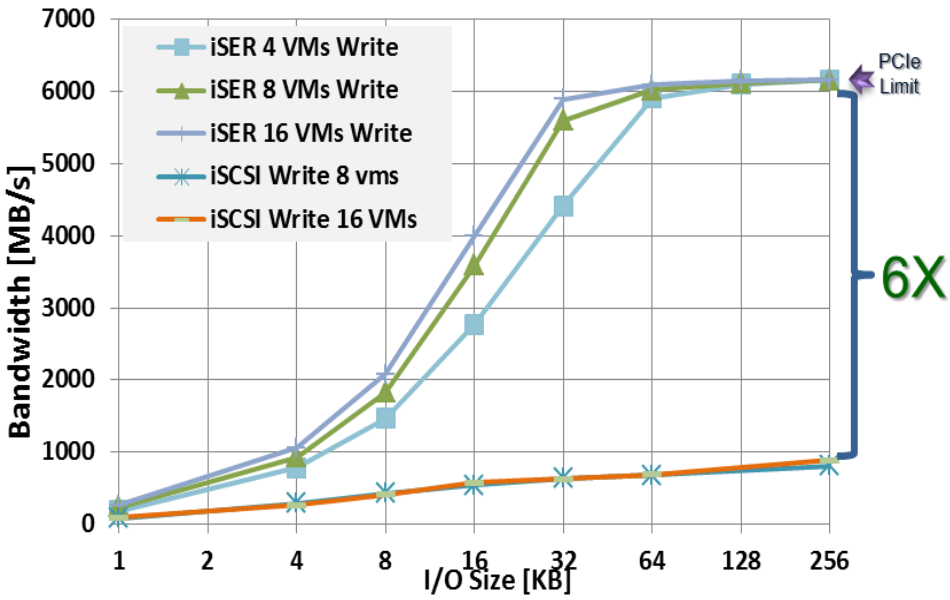
Kevin Deierling
Vice President Mellanox Technologies
kevind AT mellanox.com

OpenStack Integration



Fastest OpenStack Storage Access





- Using OpenStack Built-in components and management
  - No additional software is required
  - RDMA is already inbox and used by OpenStack customers
- RDMA enables faster Flash performance & lower CPU consumption
- More efficient network saves CapEx and OpEx to perform a given workload

Flash Memory SUMMIT

High-Performance Flash Storage Systems

NetApp™

EF540

- Up to 38 Terabyte (24 +24 SSDs)
- 24 Drives in 2U
- 40Gb/s iSCSI

Violin MEMORY

Mellanox ConnectX

- Up to 1 Million IOPS
- Latency as low as 100us
- Up to 70TB of raw flash
- Only 3 Rack Units (3U)

NIMBUS DATA

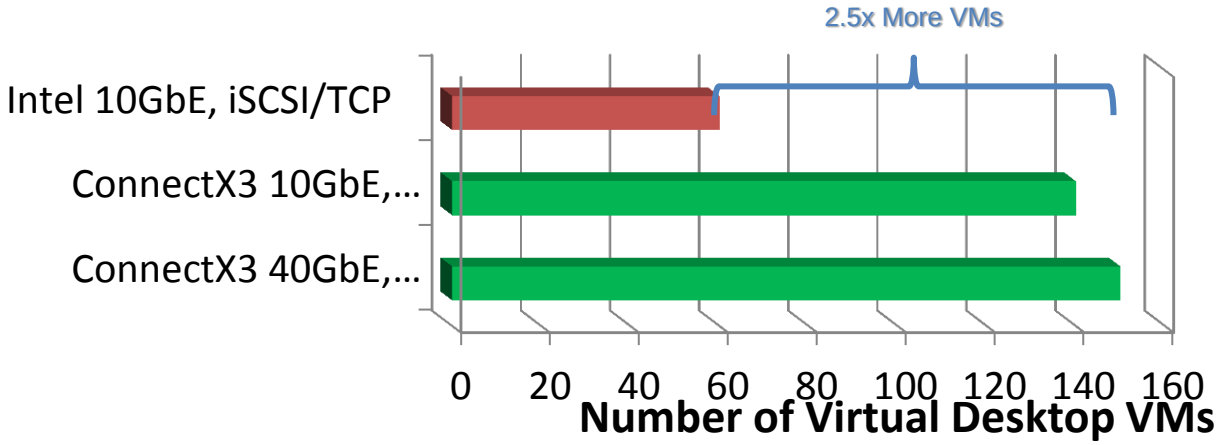
Gemini F600

40GbE or InfiniBand FDR Ports

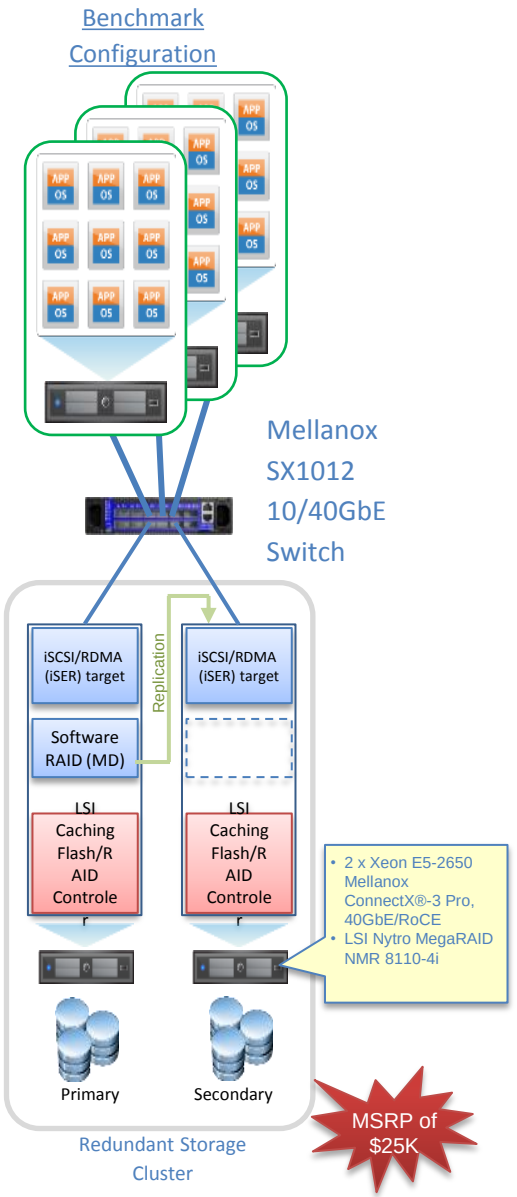
- 8 x 56 Gb FDR IB / 40 GbE
- From 3 TB to 48 TB (in 2U)
- Up to 385 TB usable
- 1M Write IOPs, 2M Read IOPs
- 50us IO Latency

Flash Memory SUMMIT

Flash & 40GbE Enable 2.5x More VMs

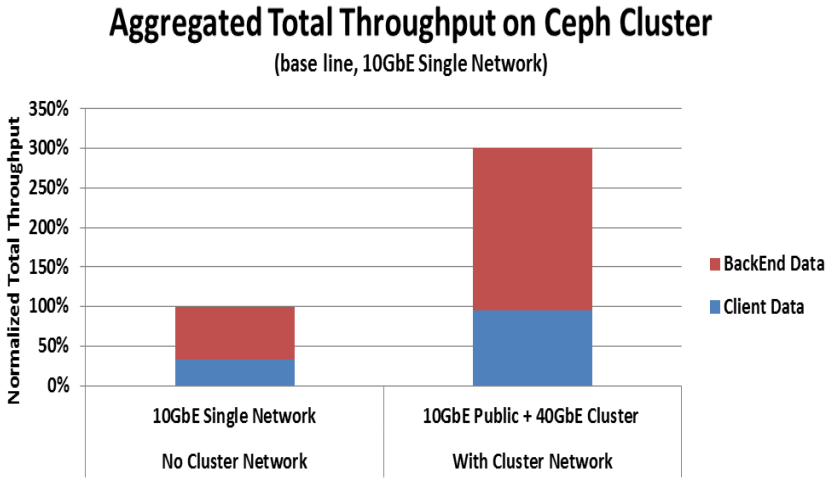


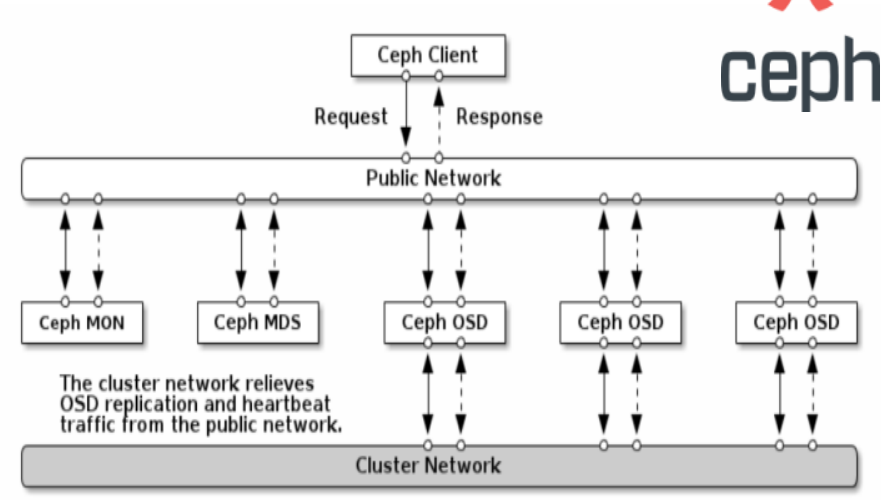
- Flash & iSER overcome IOPs bottlenecks in Virtual Desktop Infrastructure (VDI)
  - LSI Nytro MegaRAID accelerate disk access through SSD based caching
  - Mellanox ConnectX®-3 10/40GbE Adapter with RDMA
  - Accelerate Hypervisor access to fast shared Flash over 40G Ethernet
  - Zero-overhead replication
- Unmatched VDI density of 150 VMs per server
  - Using iSCSI/RDMA (iSER) enabled 2.5x more VMs compared to using iSCSI with TCP/IP over the exact same setup



Flash Memory SUMMIT

CEPH and Networks

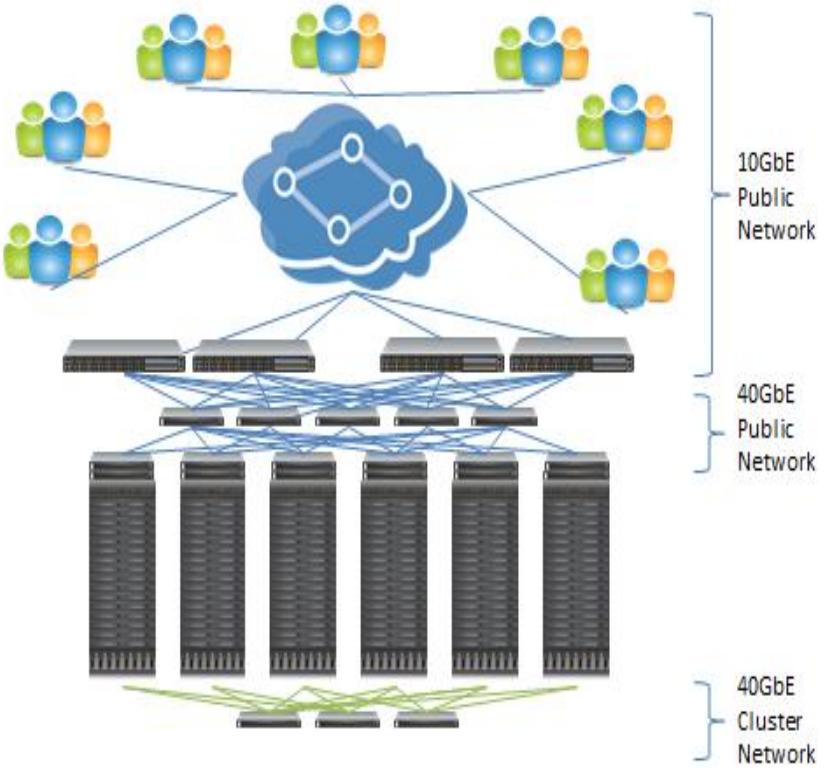




- High performance networks enable maximum cluster availability
  - Clients, OSD, Monitors and Metadata servers communicate over multiple network layers
  - Real-time requirements for heartbeat, replication, recovery and re-balancing
- Cluster ("backend") network performance dictates cluster's performance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

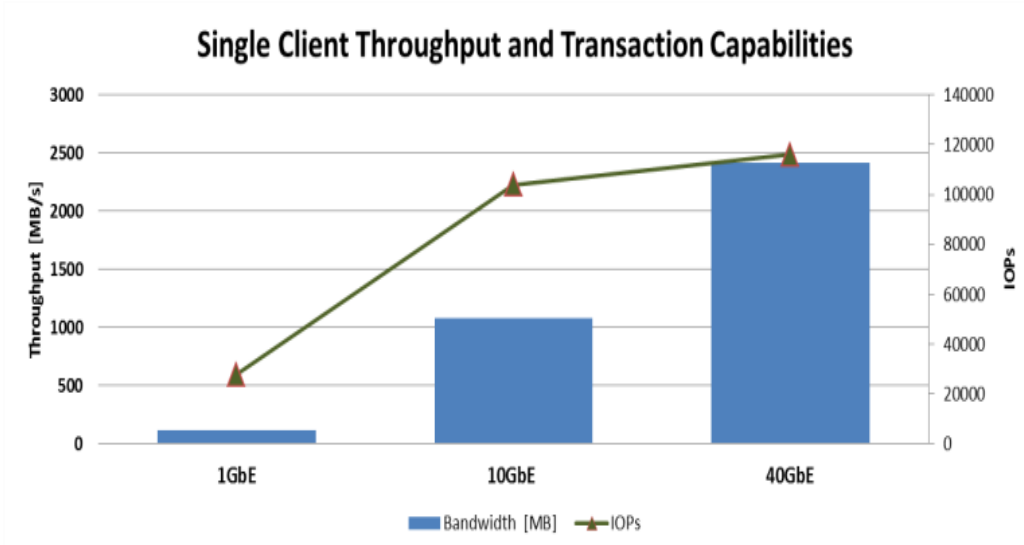
Deploy CEPH with 40GbE Interconnect

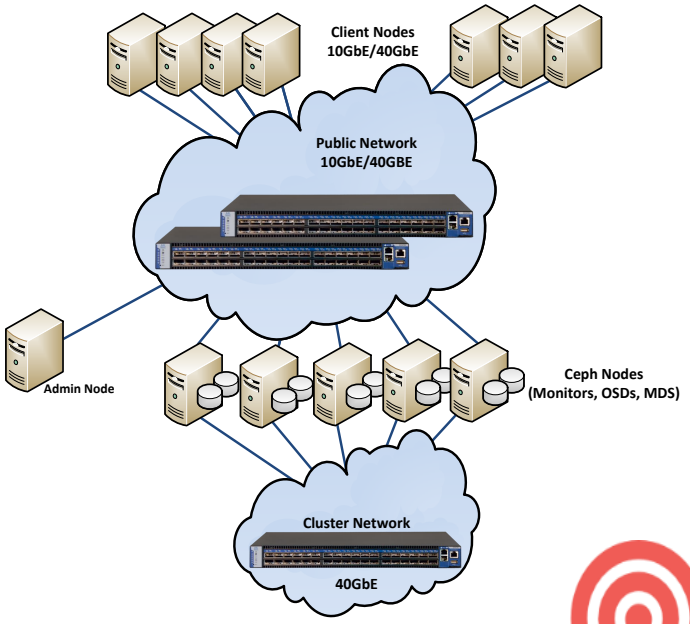
- Building Scalable, Performing Storage Solutions
  - Cluster @ 40Gb Etehrent
  - Clients @ 10G/40Gb Ethernet
- Directly connect over 500 Client Nodes
  - Target Retail Cost: US$350/1TB
- Scale Out Customers Use SSDs
  - For OSDs and Journals



CEPH Performance with 40GbE

20X Higher Throughput
4X Higher IOPS with 40GbE clients





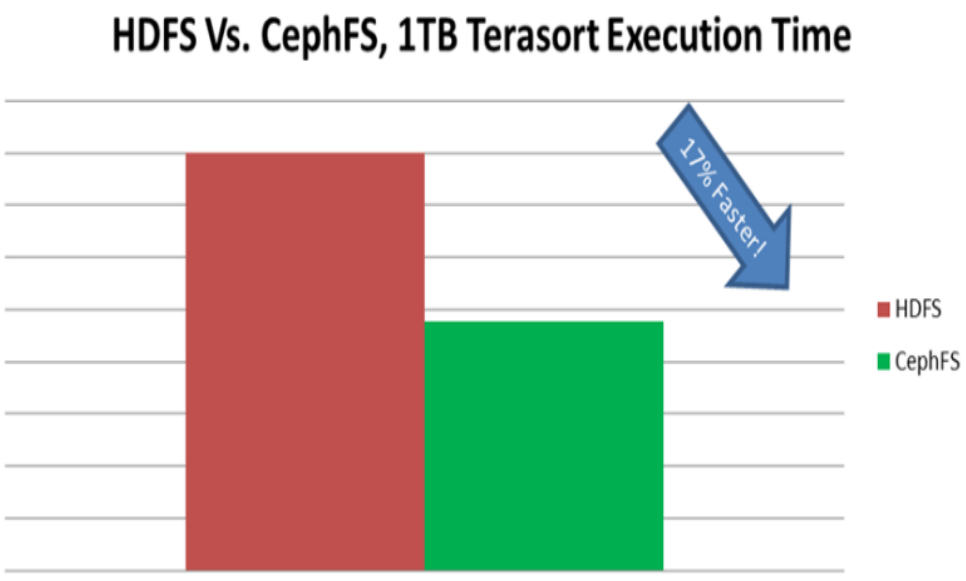
- Cluster (Private) Network @ 40GbE
  - Smooth HA, unblocked heartbeats, efficient data balancing
- Throughput Clients @ 40GbE
  - Guaranties line rate for high ingress/egress clients
- IOPs Clients @ 10GbE / 40GbE
  - 100K+ IOPs/Client @4K blocks

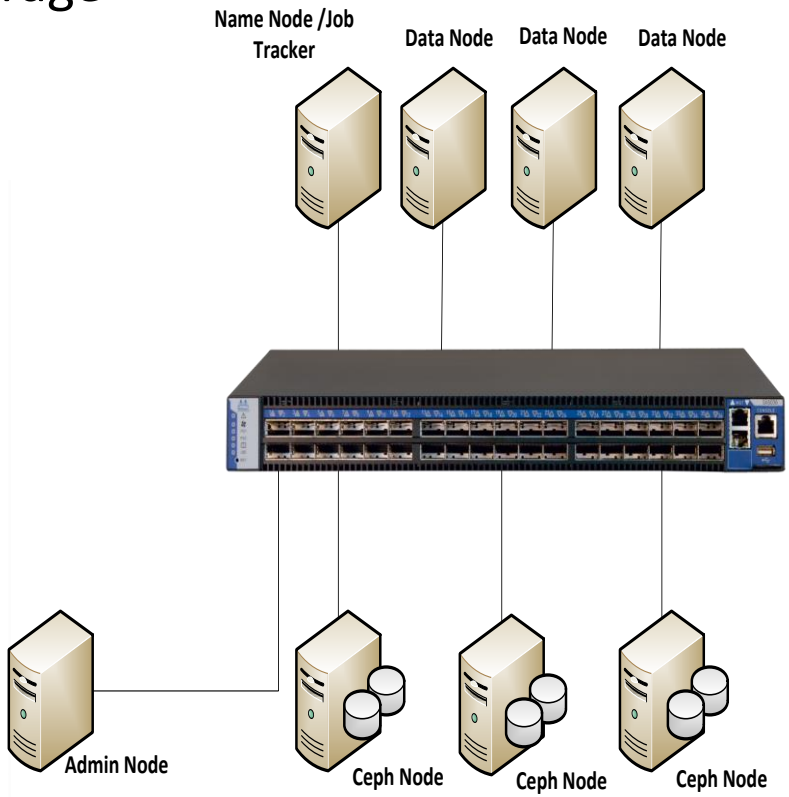
http://www.mellanox.com/related-docs/whitepapers/WP_Deploying_Ceph_over_High_Performance_Networks.pdf

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

Hadoop with Ceph, Flash, & 40GbE

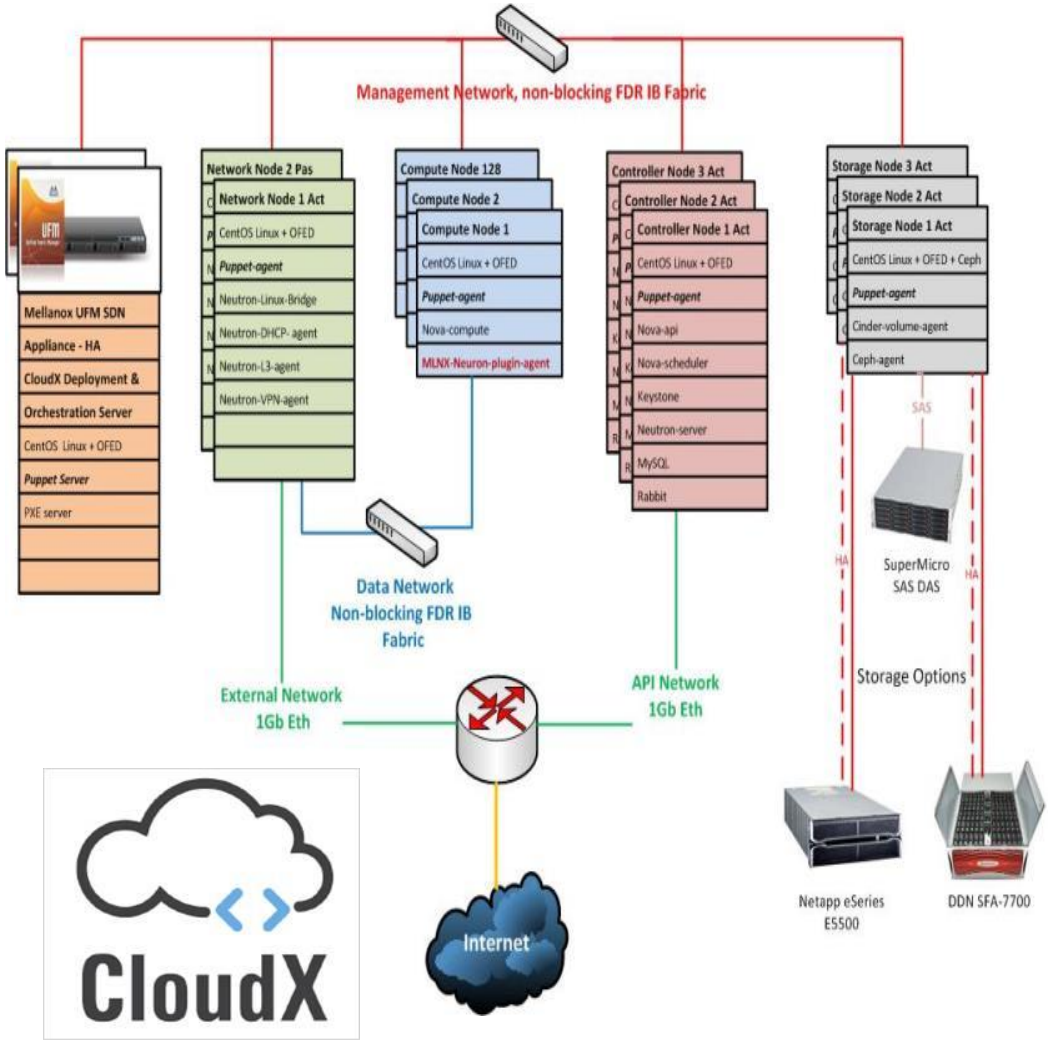
- Increase Hadoop Cluster Performance
- Accelerate Hadoop over Ceph
  - Flash based OSD
  - High Performance Interconnect
- Dynamically Scale Compute and Storage
- Eliminate Single Point of Failure





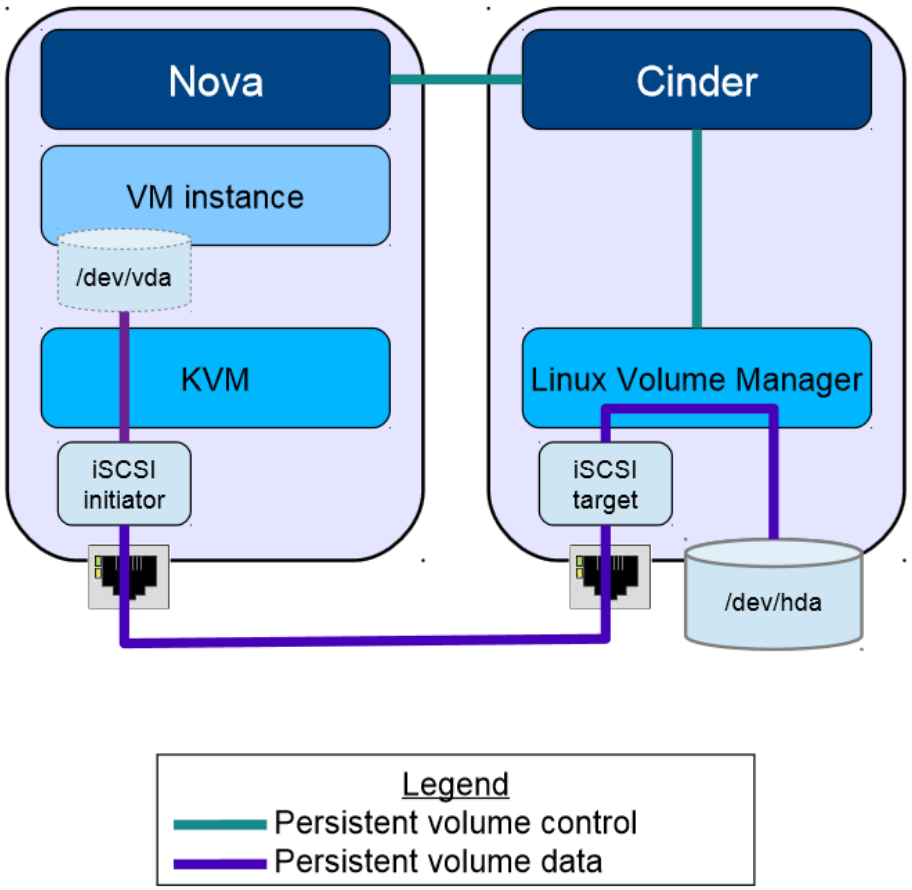
CloudX Architecture Overview

- Hardware
  - Industry standard components
    - Servers, storage, interconnect, software
  - Mellanox 40GbE interconnect
  - Solid State Drives (SSD)
- OpenStack components
  - Nova compute
  - Network controller node
  - Cinder/iSER & Ceph Storage
  - Mellanox OpenStack plug-ins
- Toolkit for automated switch and server deployment



Thanks!
Questions

Cinder Block Storage Integration



- Cinder (Target)
  - Volume Management
  - Persistence independent of VM
- Nova (Initiator)
  - Boot VM instance
  - Attach volume to VM