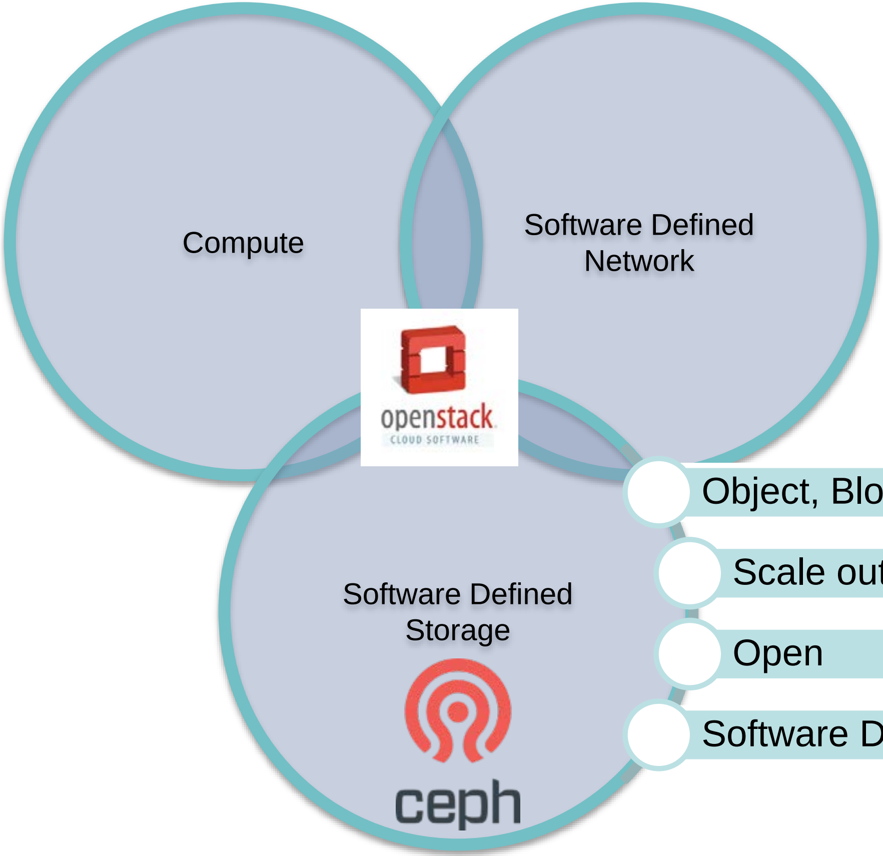
Accelerating Ceph with Flash and High Speed Networks

Dror Goldenberg

VP Software Architecture

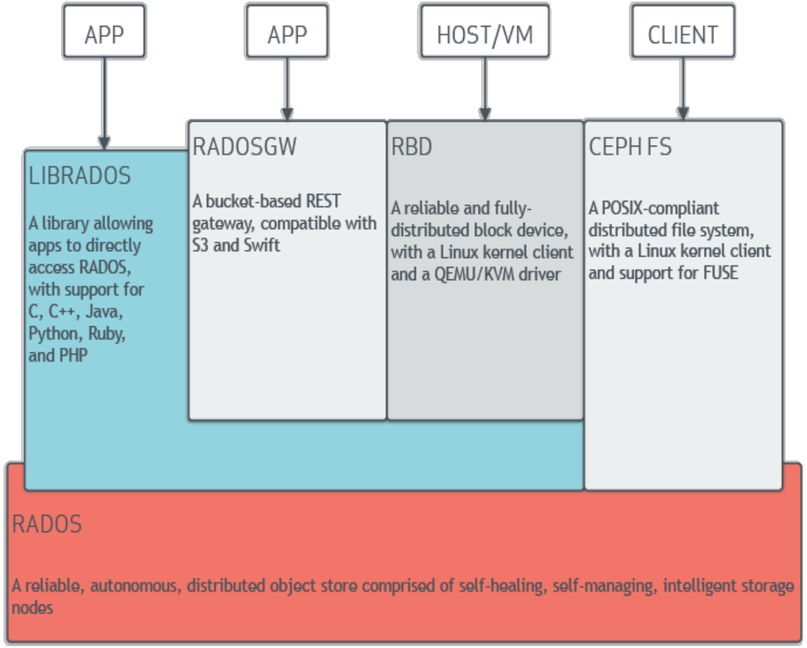
The New Open Cloud Era



- Object, Block
- Scale out
- Open
- Software Defined

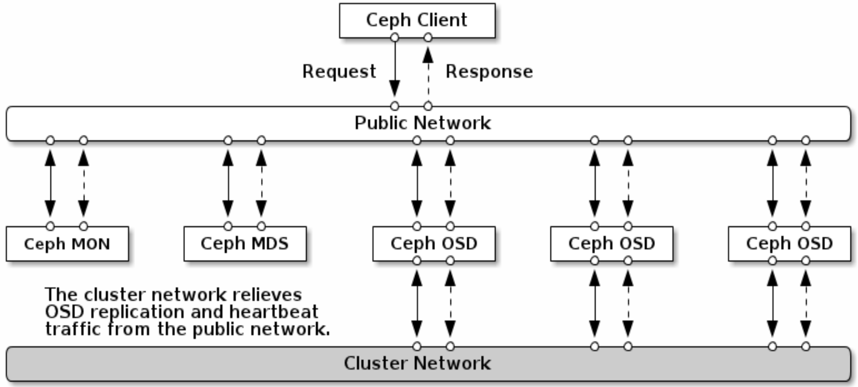
Ceph Architecture

Architecture enables object, block & file access


APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

Source: http://ceph.com/docs/master/architecture/

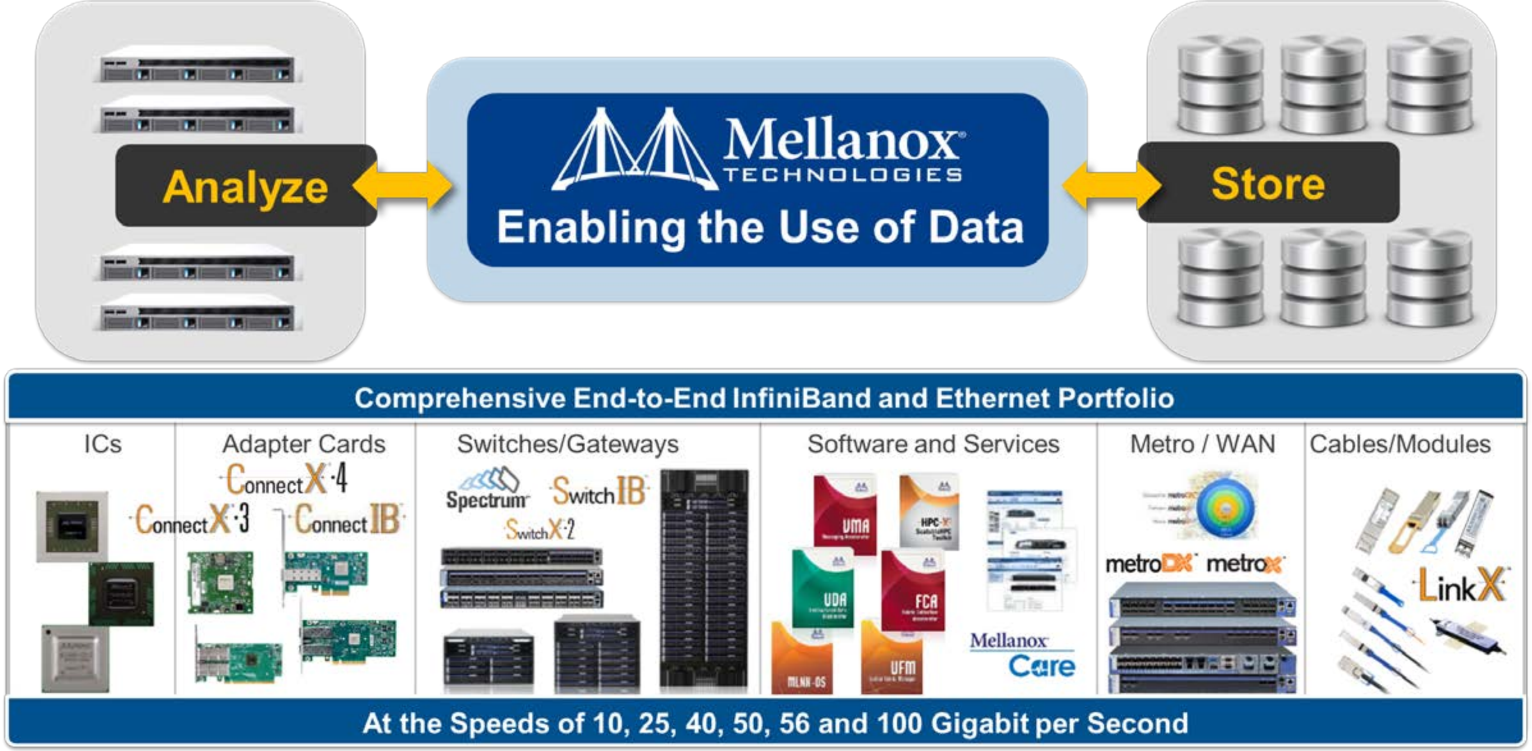
Fully distributed scale out



Source: http://ceph.com/docs/master/rados/configuration/network-config-ref/

Interconnect Capabilities Determine Scale Out Performance

Mellanox - Leading Supplier of End-to-End Interconnect Solutions



Analyze

Mellanox TECHNOLOGIES

Enabling the Use of Data

Store

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs | Adapter Cards | Switches/Gateways | Software and Services | Metro / WAN | Cables/Modules

At the Speeds of 10, 25, 40, 50, 56 and 100 Gigabit per Second

Flash Memory SUMMIT

Entering the Era of 100Gb/s Networks

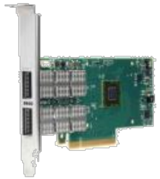
Adapters

ConnectX-4

100Gb/s Adapter, 0.7us latency, RDMA

150 million messages per second

(10 / 25 / 40 / 50 / 56 / 100Gb/s)



Switch

SwitchIB™

36 EDR (100Gb/s) Ports, <90ns Latency

Throughput of 7.2Tb/s

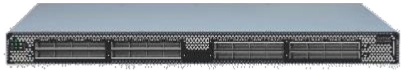
Switch

Spectrum™

32 100GbE Ports, 64 25/50GbE Ports

(10 / 25 / 40 / 50 / 100GbE)

Throughput of 6.4Tb/s



Interconnect

LinkX™

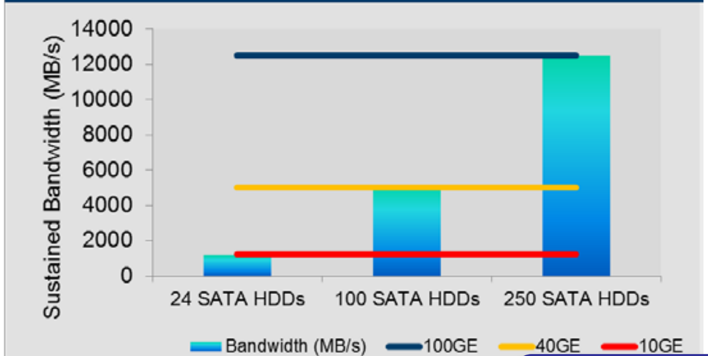
Copper (Passive, Active)

Optical Cables (VCSEL)

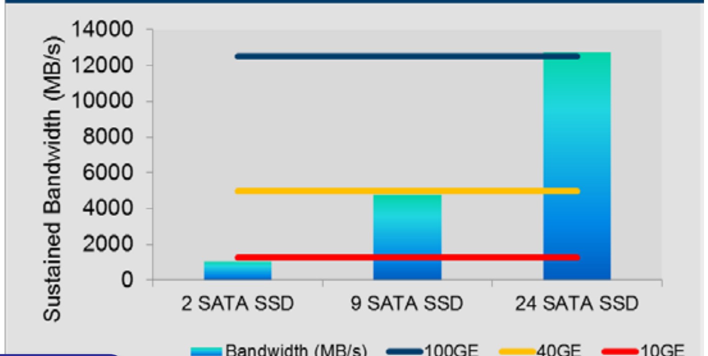
Silicon Photonics

Building a Balanced Element for the Scale Out Storage System

SATA HDDs

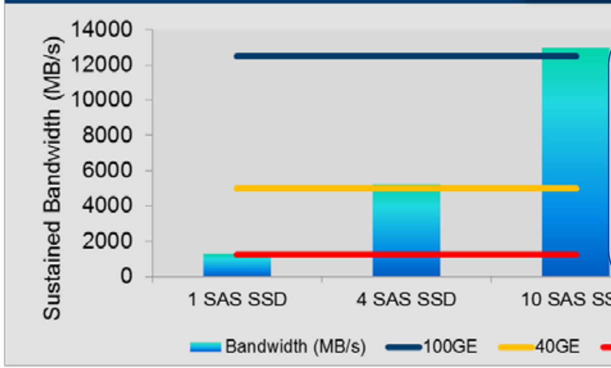


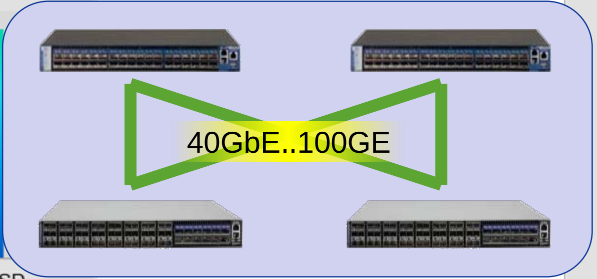
SATA SSDs



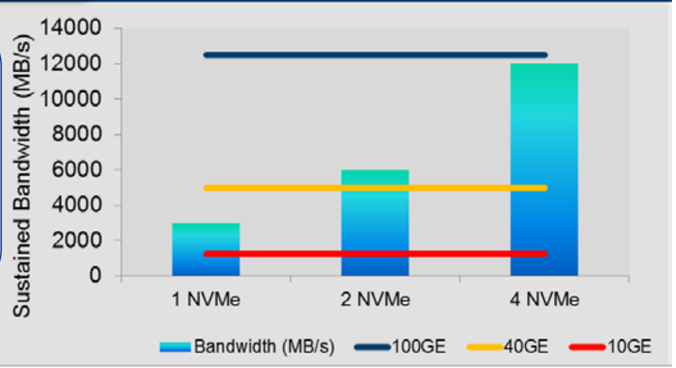
Network I/O should match Flash Drive I/O

SAS SSDs



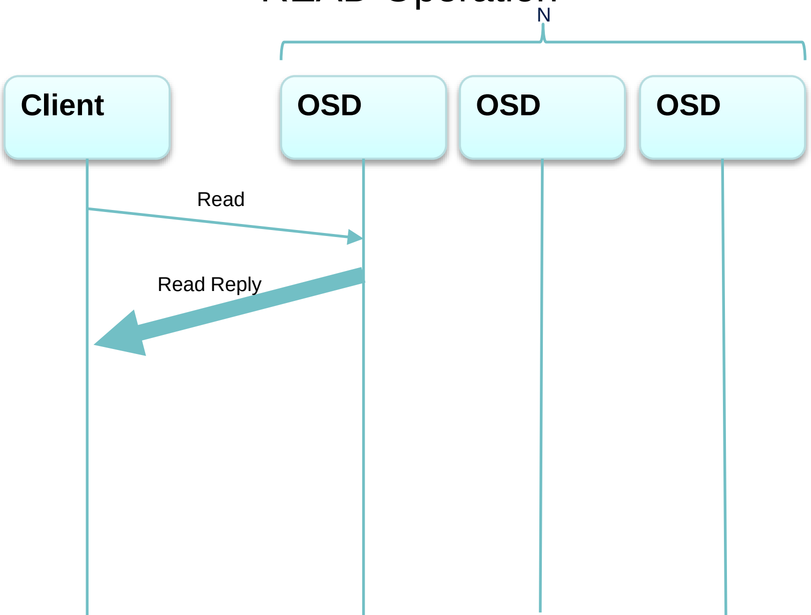


NVMe



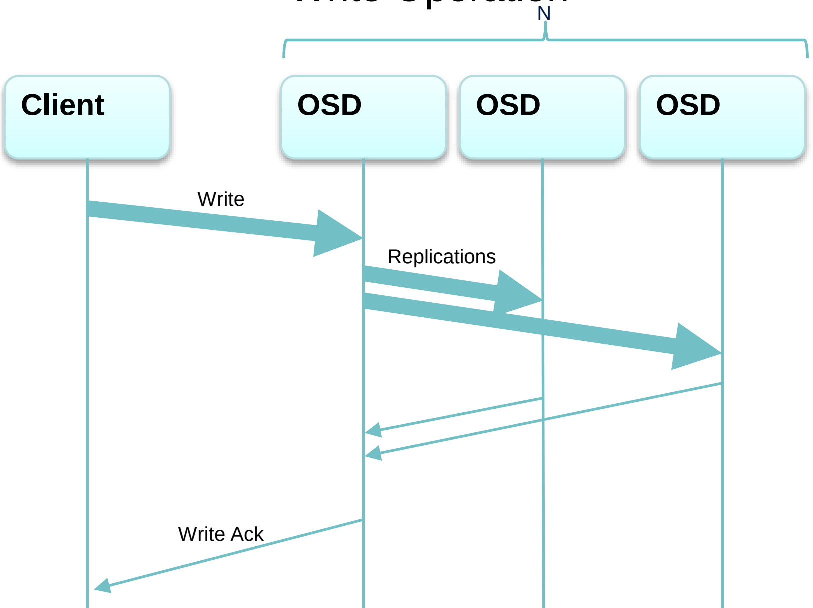
Cluster Network Traffic – Sunny Day Scenarios (Replication)

READ Operation



No extra cluster network traffic

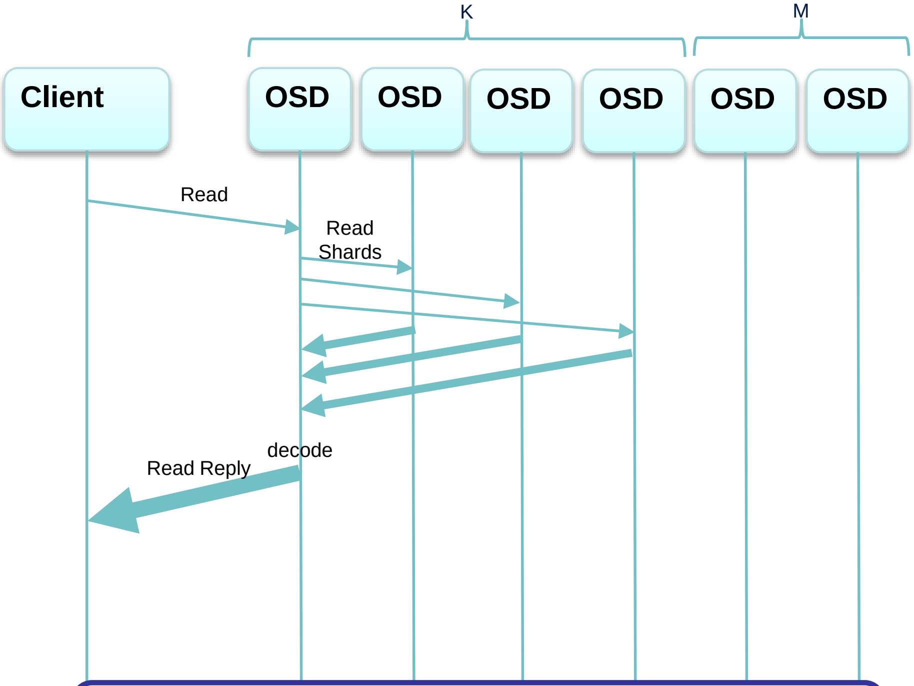
Write Operation



Typically 2x cluster network traffic
(N-1)·x

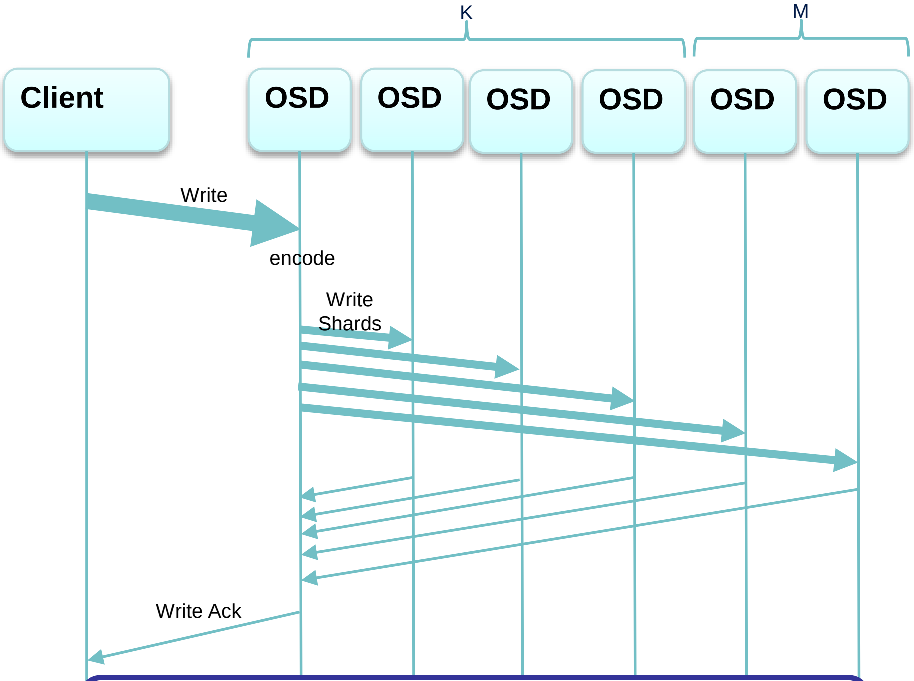
Cluster Network Traffic – Sunny Day Scenarios (Erasure Coding)

READ Operation



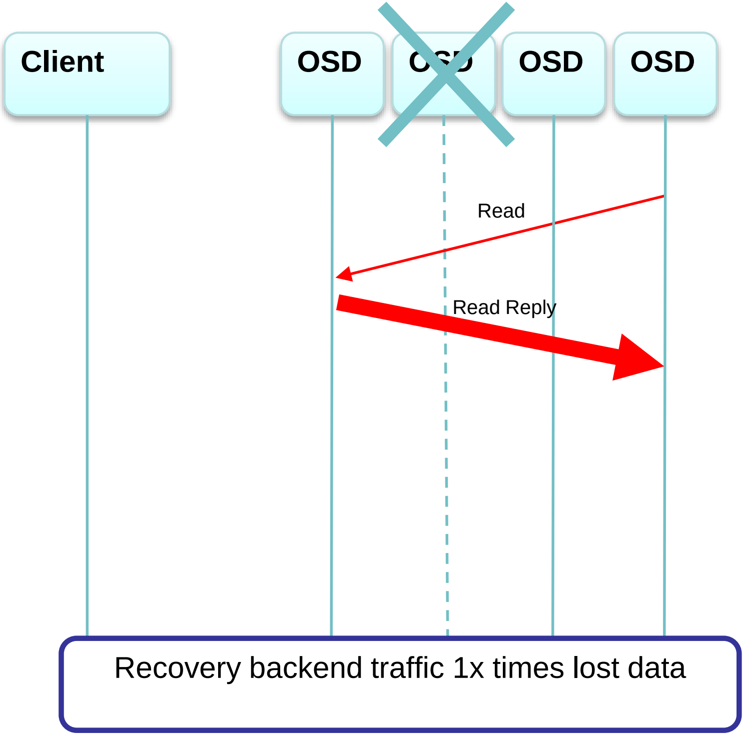
~1x cluster network traffic
((k-1)/k)·x

Write Operation



Typically ~1.4x cluster network traffic
((k+m-1)/k) ·x

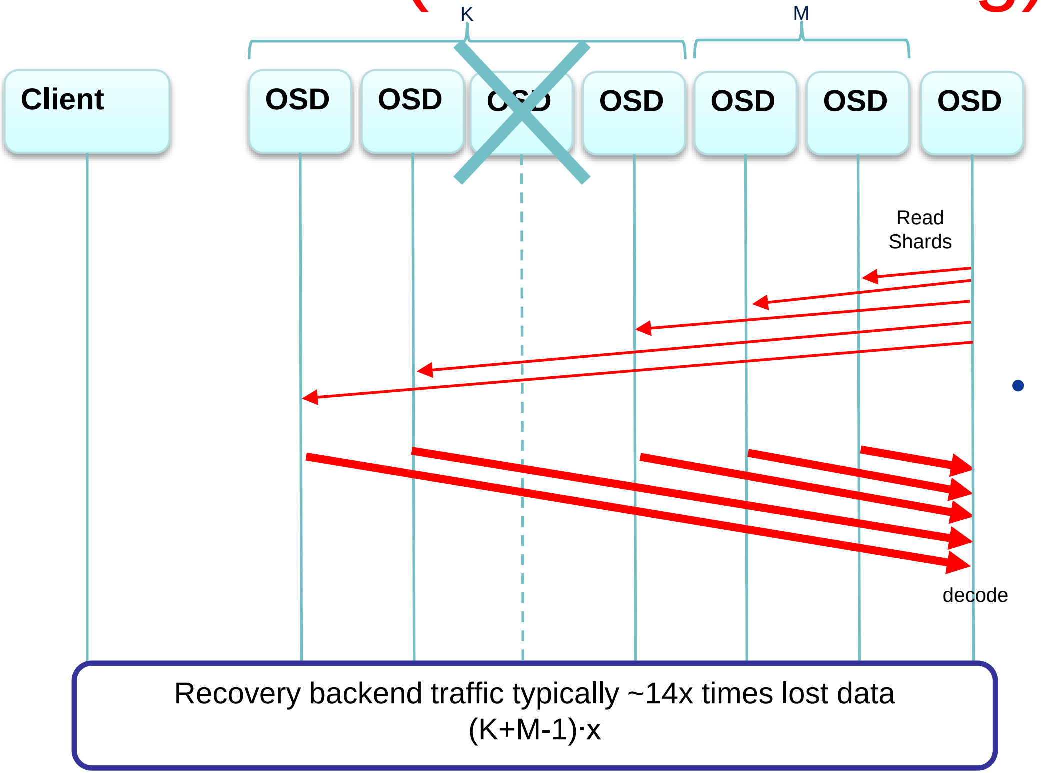
Cluster Network Traffic – Recovery (Replication)



Recovery backend traffic 1x times lost data

- Example - Time to recover
  - Net networking time to move data
  - 20TB system @40GE 1.1hrs
  - 200TB system @40GE 11.1hrs
- Similar flows for scrubbing
  - But more demanding in I/O

Cluster Network Traffic – Recovery (Erasure Coding)



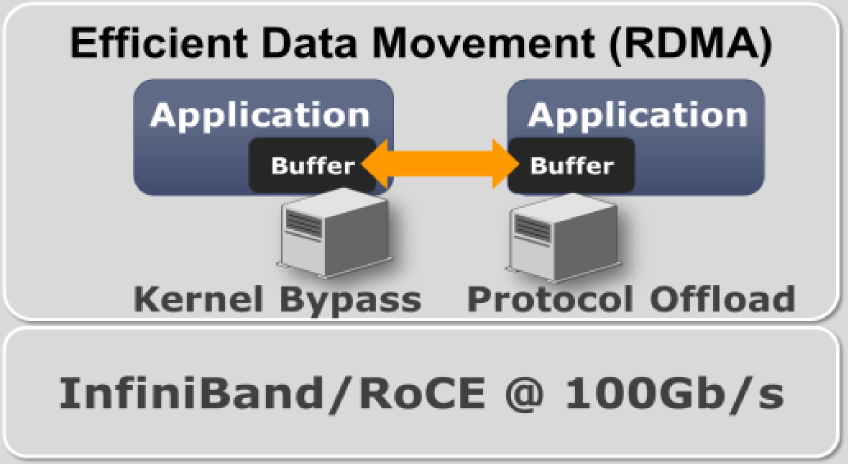
Recovery backend traffic typically ~14x times lost data
(K+M-1)·x

Example - Time to recover (10+4)

- Net networking time to move data
- 20TB system @40GE 14.4hrs
- 200TB system @40GE 144.4hrs

- Similar flows for scrubbing

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

Accelio, High-Performance Reliable Messaging and RPC Library

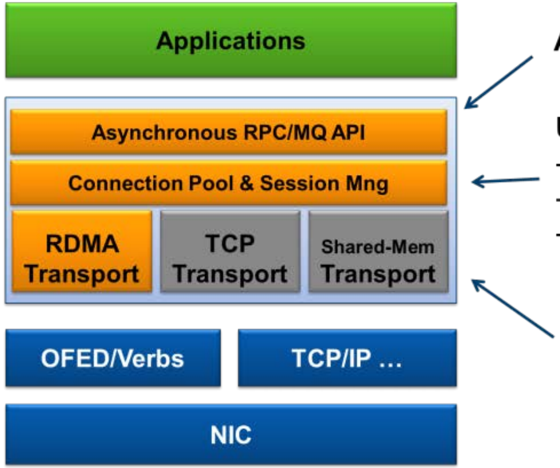
- Open source!
  - https://github.com/accelio/accelio/ && www.accelio.org

- Faster RDMA integration to application

- Asynchronous

- Maximize msg and CPU parallelism
  - Enable >10GB/s from single node
  - Enable <10usec latency under load

- Integrated with Ceph
  - Beta available in Hammer
  - Mellanox, Red Hat, CohortFS, and Community collaboration
  - XioMessenger built on top of Accelio (RDMA abstraction layer)

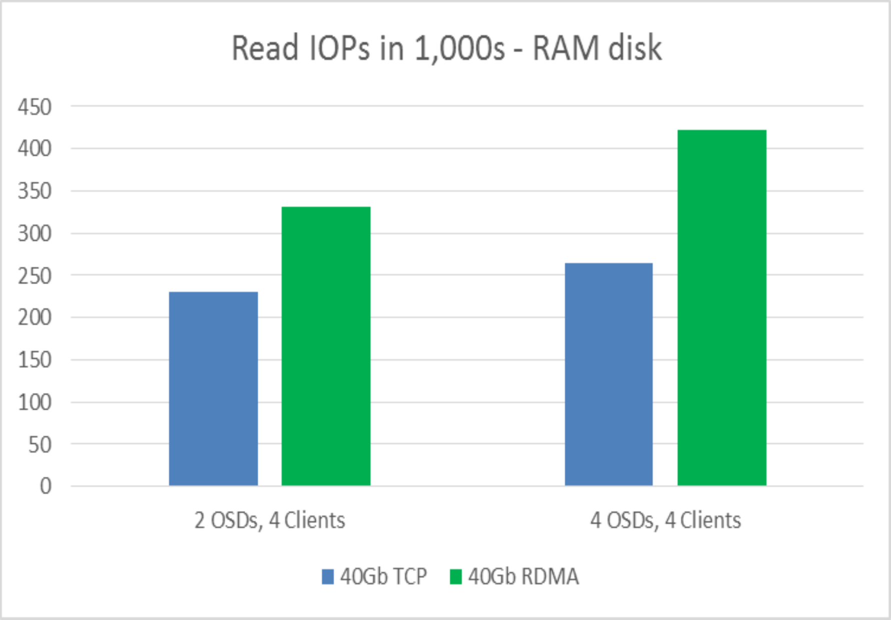


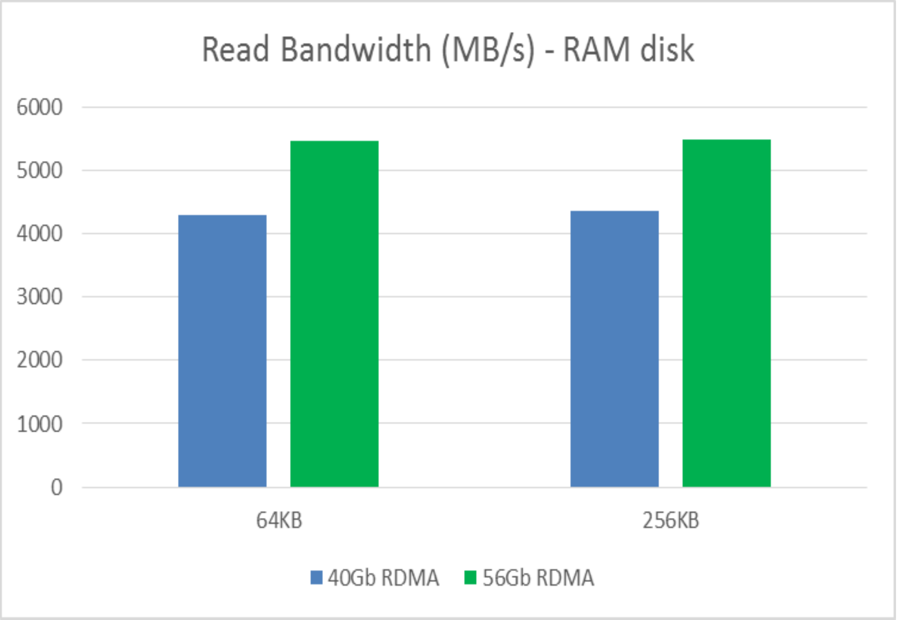
Abstract, Easy to use API

Use multiple connections per session
- **maximize CPU core usage/parallelism**
- **High-availability & Migration**
- **Scale network bandwidth**

Pluggable Transports:
- **Code once for multiple HW options**
- **Seamlessly use RDMA**

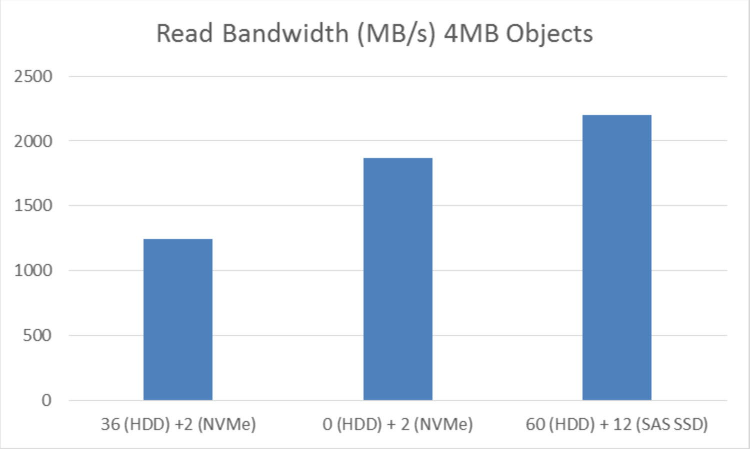
RDMA and 56GE Contribution to Performance

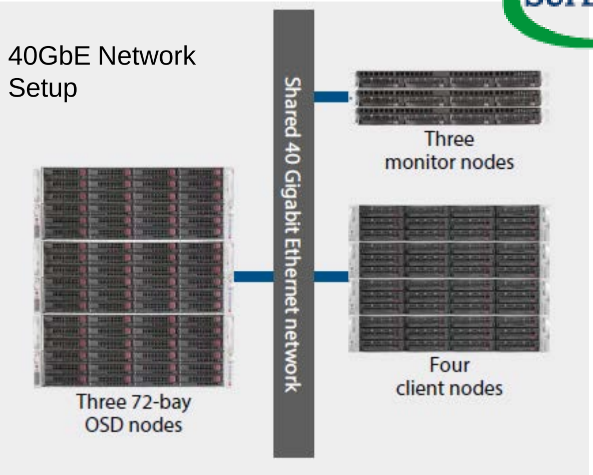




Optimizing Ceph For Throughput and Price/Throughput

Red Hat, Supermicro, Seagate, Mellanox, Intel

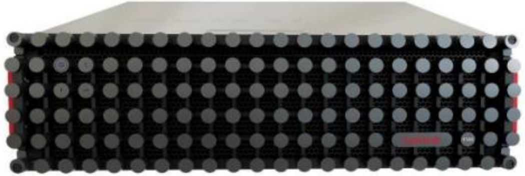




- 40GbE Advantages
  - Up to 2x read throughput per server
  - Up to 50% decrease in latency

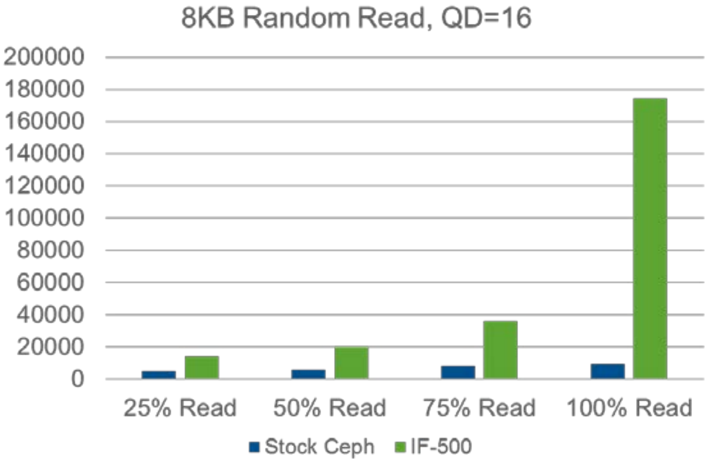
SanDisk InfiniFlash, Maximizing Ceph Random Read IOPS

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

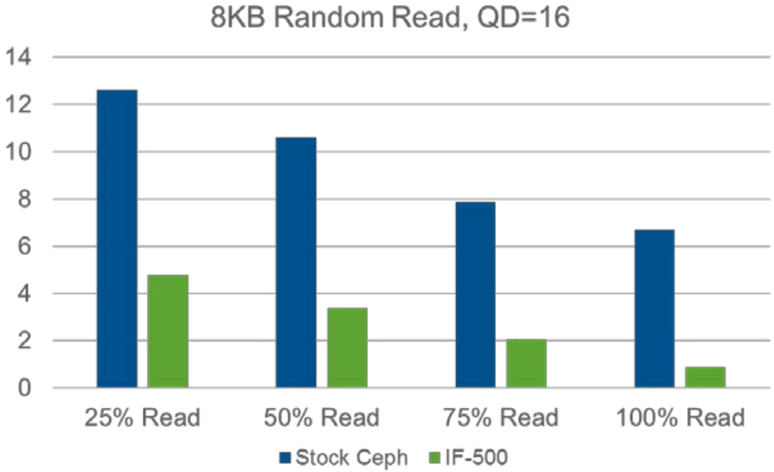


SanDisk InfiniFlash

Random Read IOPs



Random Read Latency (ms)



Ceph Optimizations for Flash

Setup	SanDisk InfiniFlash	Scalable Informatics	Supermicro	Mellanox
OSD Servers	Dell R720	SI Unison	Supermicro	Supermicro
OSD Nodes	2	2	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

High Speed Efficient RDMA Networks – Ceph Benefits

- Balanced systems for true scale out
  - Storage and network bandwidth match per system element
- Optimal networking performance for key scenarios
  - Replicaton, erasure coding, rebuild, scrubbing and cache tiering
  - Scale-out non blocking network
- Avoid traffic jams
  - I/O at lowest latency
  - Efficient fabric drain on incast scenarios
- Efficient data movement with RDMA – CPU offload

Future

- QoS improves degraded state behavior, converged networks
- Hyper-converged systems
- Advanced features offload – erasure coding
- Optimizations, optimizations, optimizations

Thank You !

gdror at mellanox.com