Mellanox High Performance Networks for Ceph

Building World Class Data Centers

Ceph Day, June 10th, 2014

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Leading Supplier of End-to-End Interconnect Solutions

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs

Adapter Cards

Switches/Gateways

Host/Fabric Software

Metro / WAN

Cables/Modules

The Future Depends on Fastest Interconnects

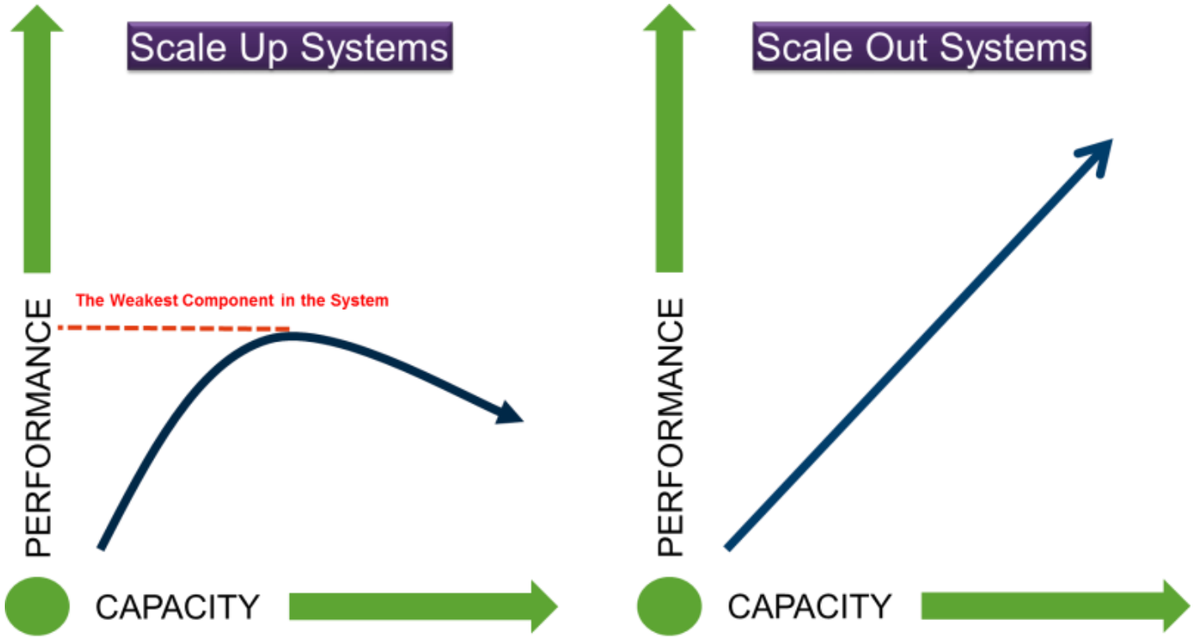
1Gb/s

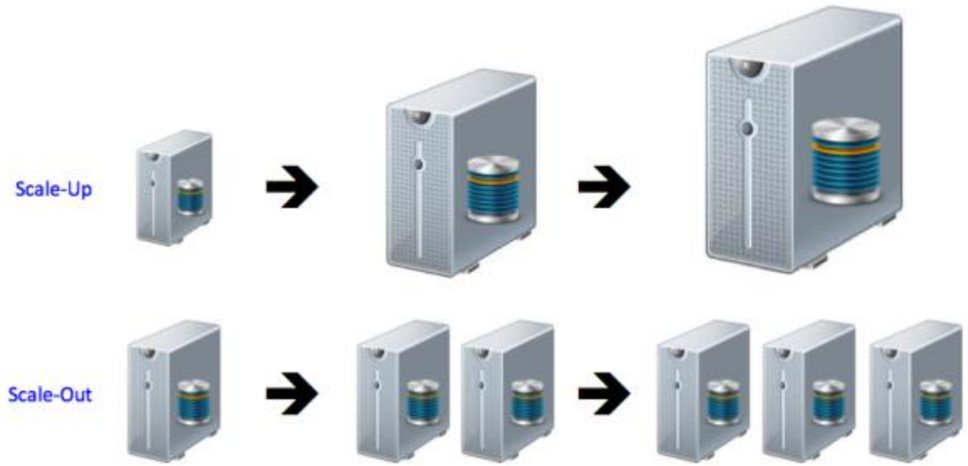
10Gb/s

40/56Gb/s

From Scale-Up to Scale-Out Architecture

- Only way to support storage capacity growth in a cost-effective manner
- We have seen this transition on the compute side in HPC in the early 2000s
- Scaling performance linearly requires “seamless connectivity” (ie lossless, high bw, low latency, cpu offloads)



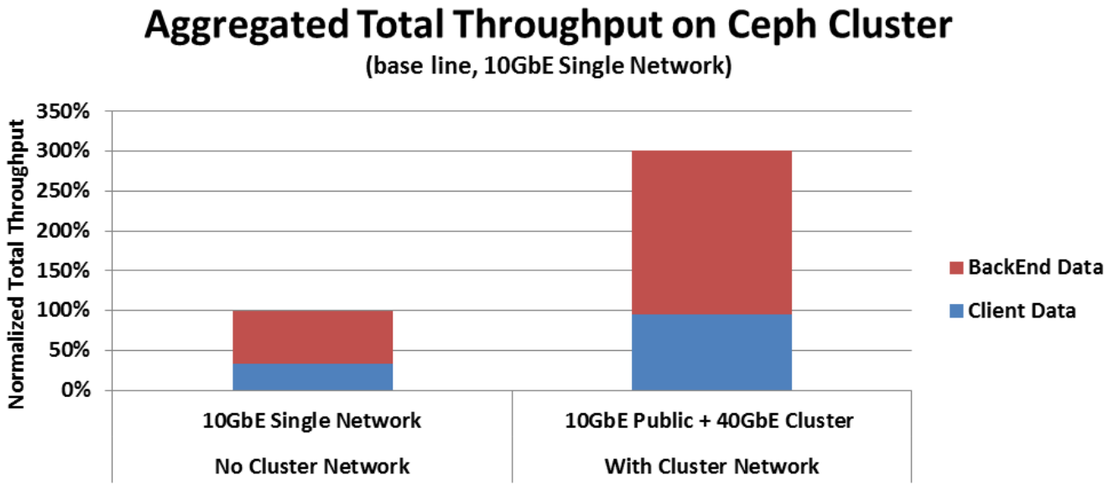


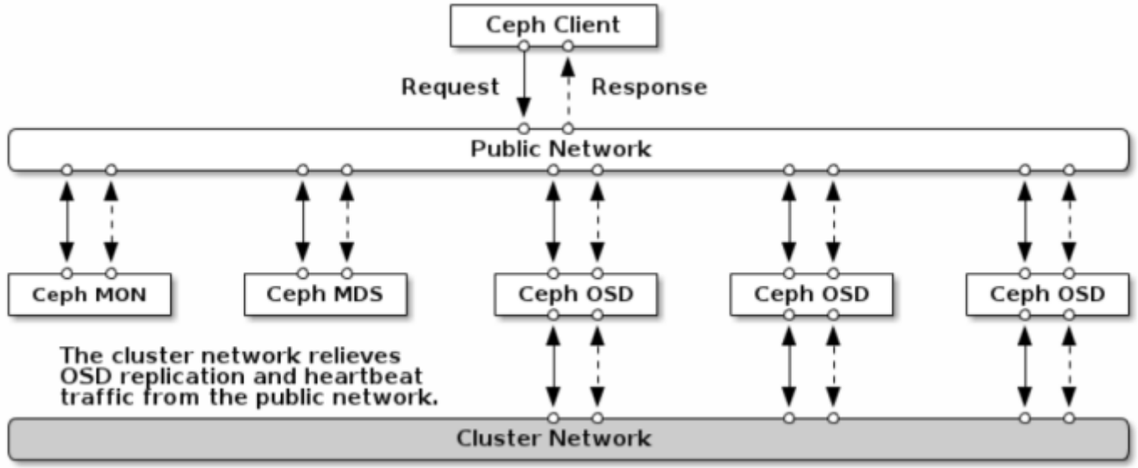
Interconnect Capabilities Determine Scale Out Performance

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

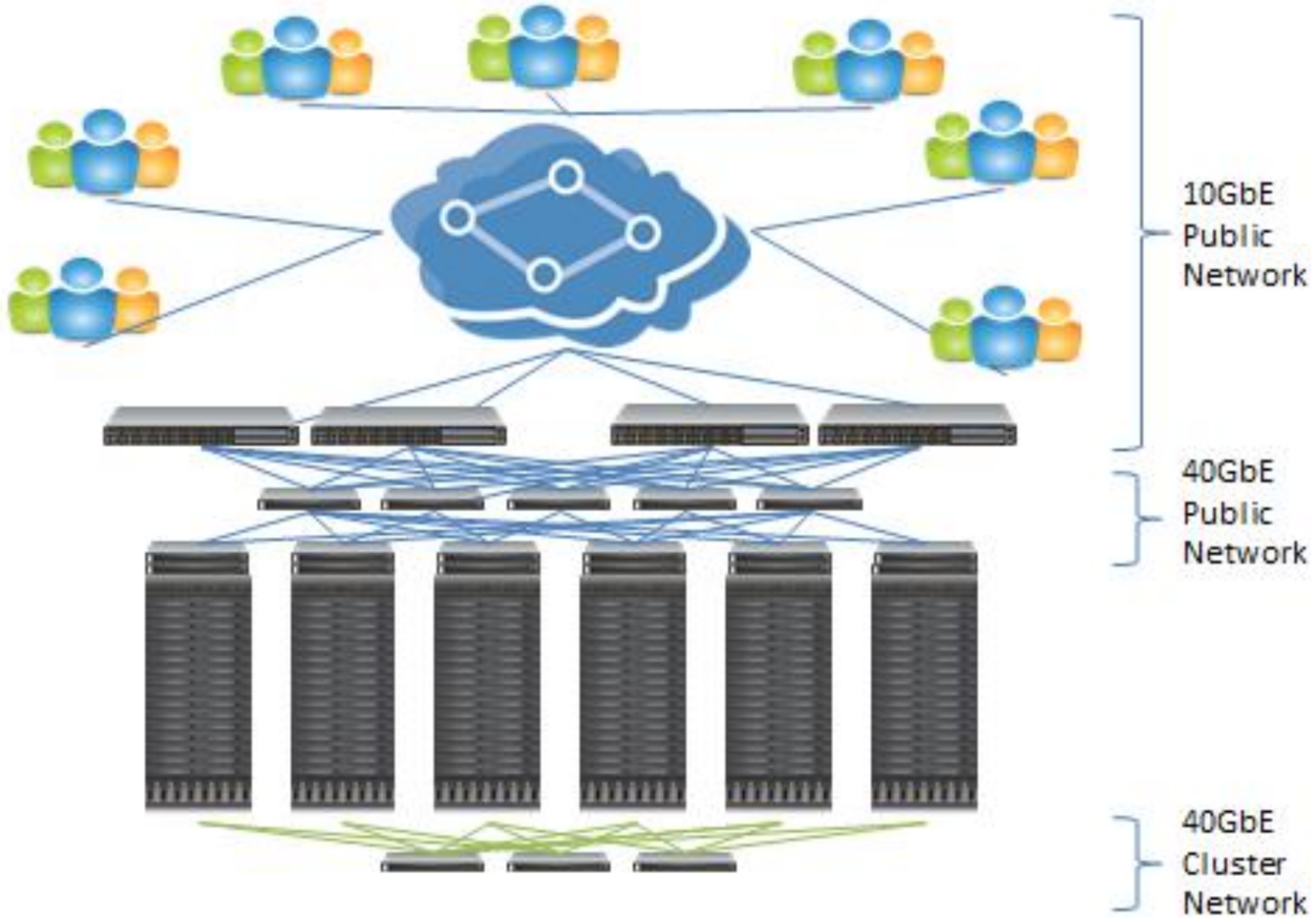
- Cluster (“backend”) network performance dictates cluster’s performance and scalability
  - “**Network load between Ceph OSD Daemons easily dwarfs the network load between Ceph Clients and the Ceph Storage Cluster**” (Ceph Documentation)





How Customers Deploy CEPH with Mellanox Interconnect

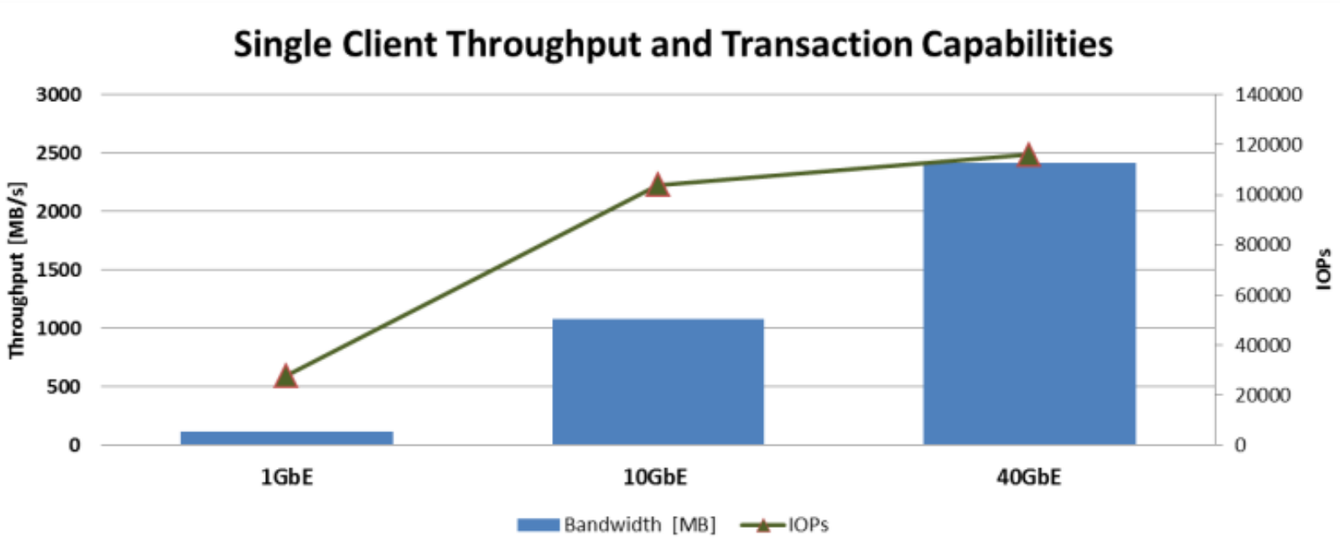
- Building Scalable, Performing Storage Solutions
  - Cluster network @ 40Gb Ethernet
  - Clients @ 10G/40Gb Ethernet
- Directly connect over 500 Client Nodes
  - Target Retail Cost: US$350/1TB
- Scale Out Customers Use SSDs
  - For OSDs and Journals

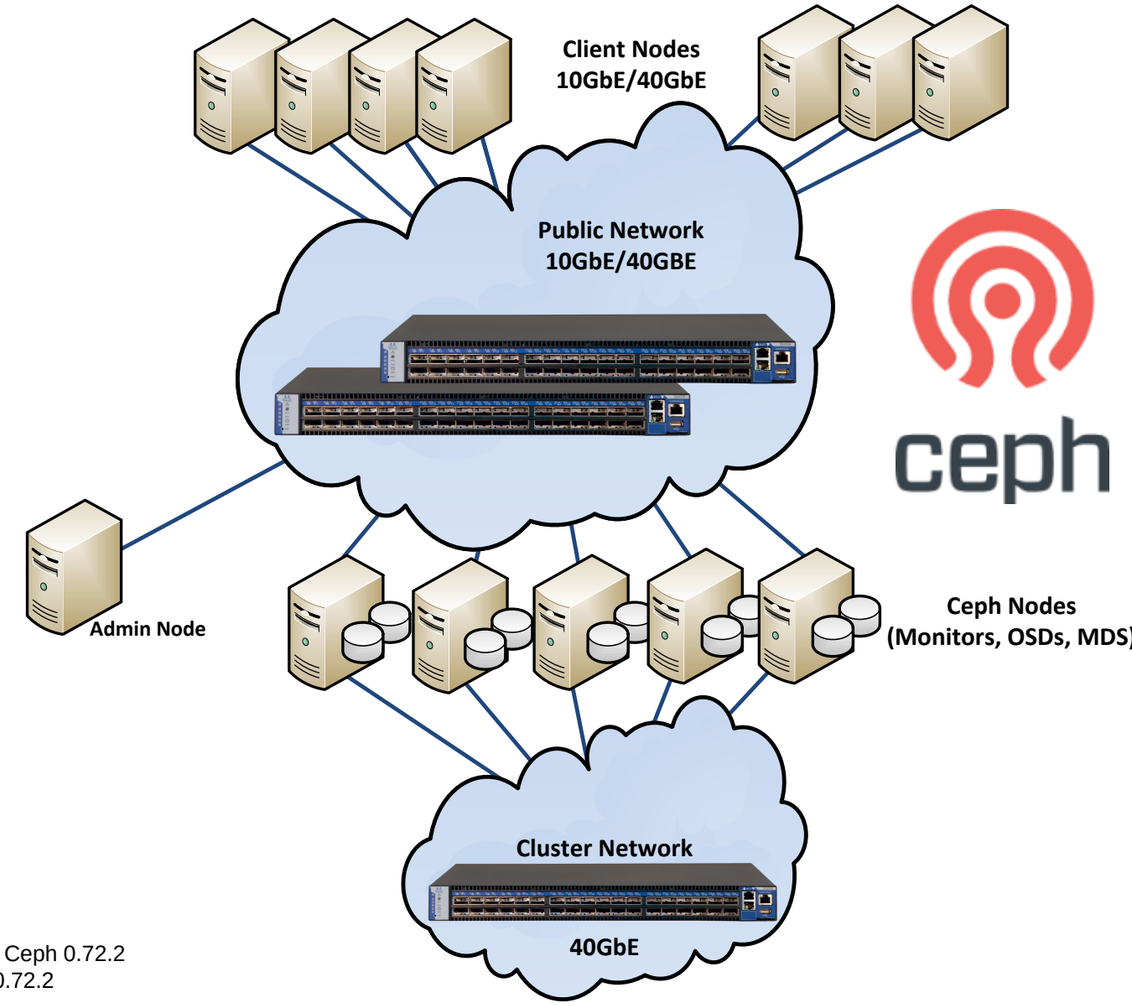


8.5PB System Currently Being Deployed

CEPH Deployment Using 10GbE and 40GbE

- 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



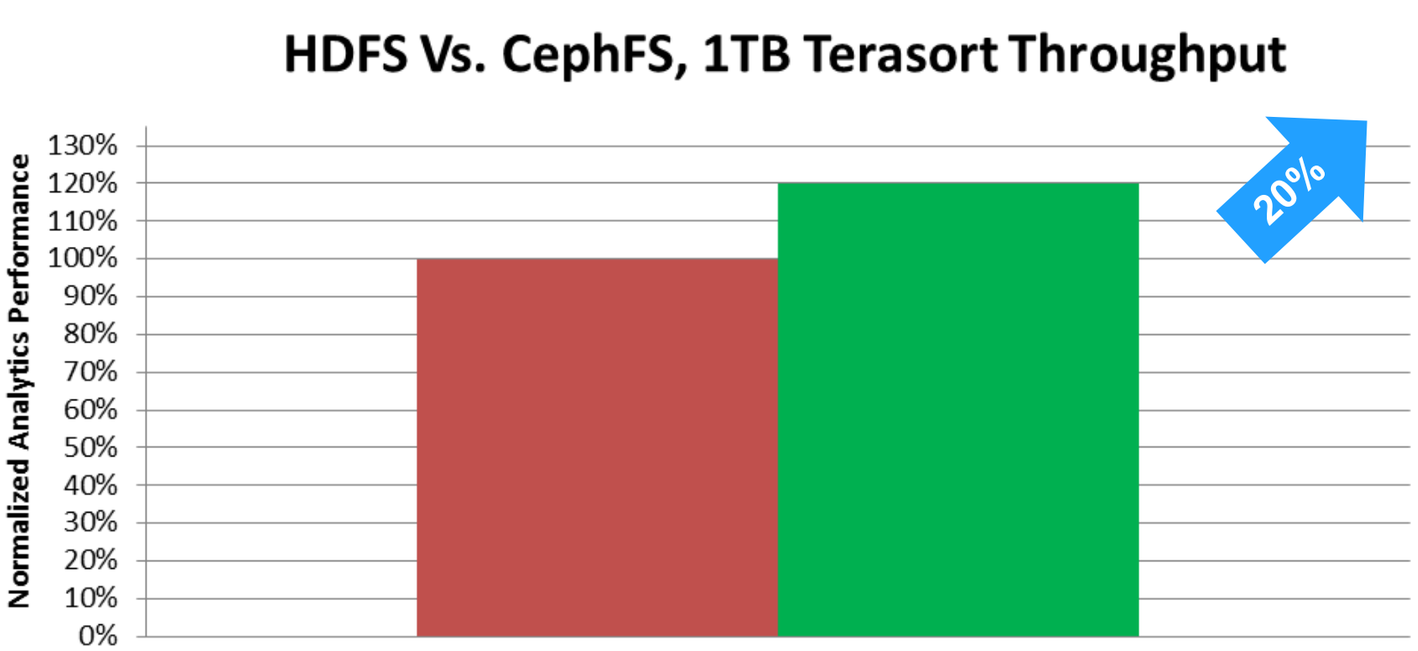


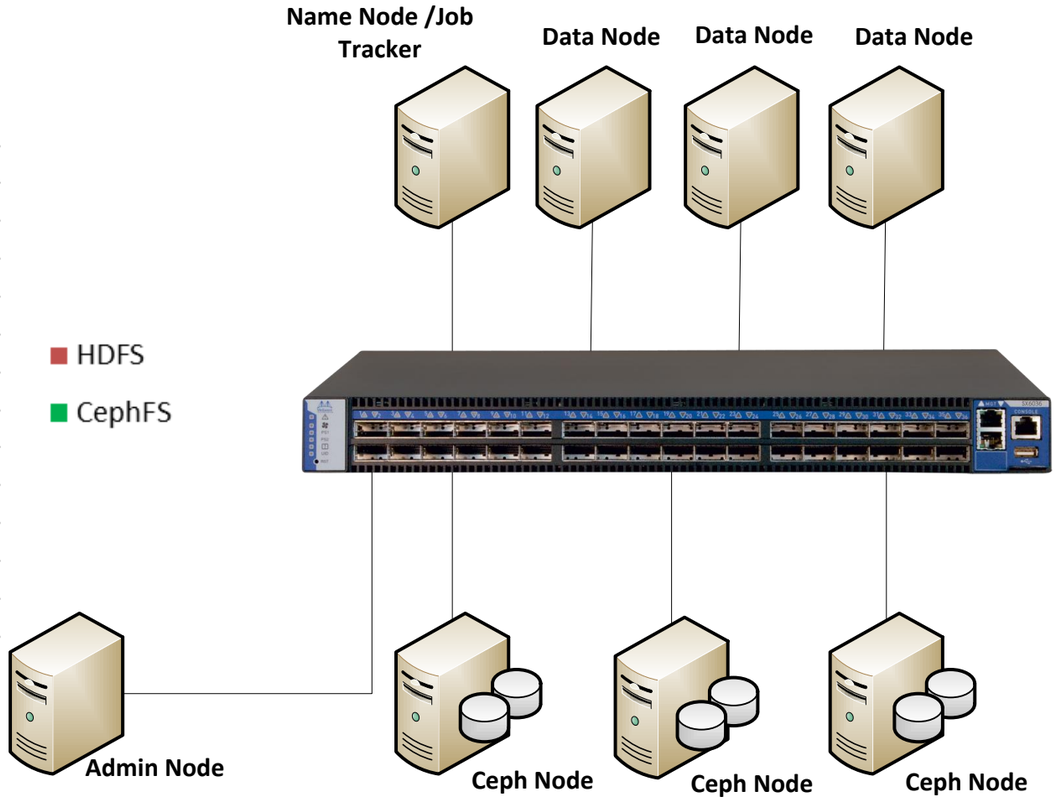
Throughput Testing results based on fio benchmark, 8m block, 20GB file,128 parallel jobs, RBD Kernel Driver with Linux Kernel 3.13.3 RHEL 6.3, Ceph 0.72.2
IOPs Testing results based on fio benchmark, 4k block, 20GB file,128 parallel jobs, RBD Kernel Driver with Linux Kernel 3.13.3 RHEL 6.3, Ceph 0.72.2

20x Higher Throughput , 4x Higher IOPs with 40Gb Ethernet Clients!
(http://www.mellanox.com/related-docs/whitepapers/WP_Deploying_Ceph_over_High_Performance_Networks.pdf)

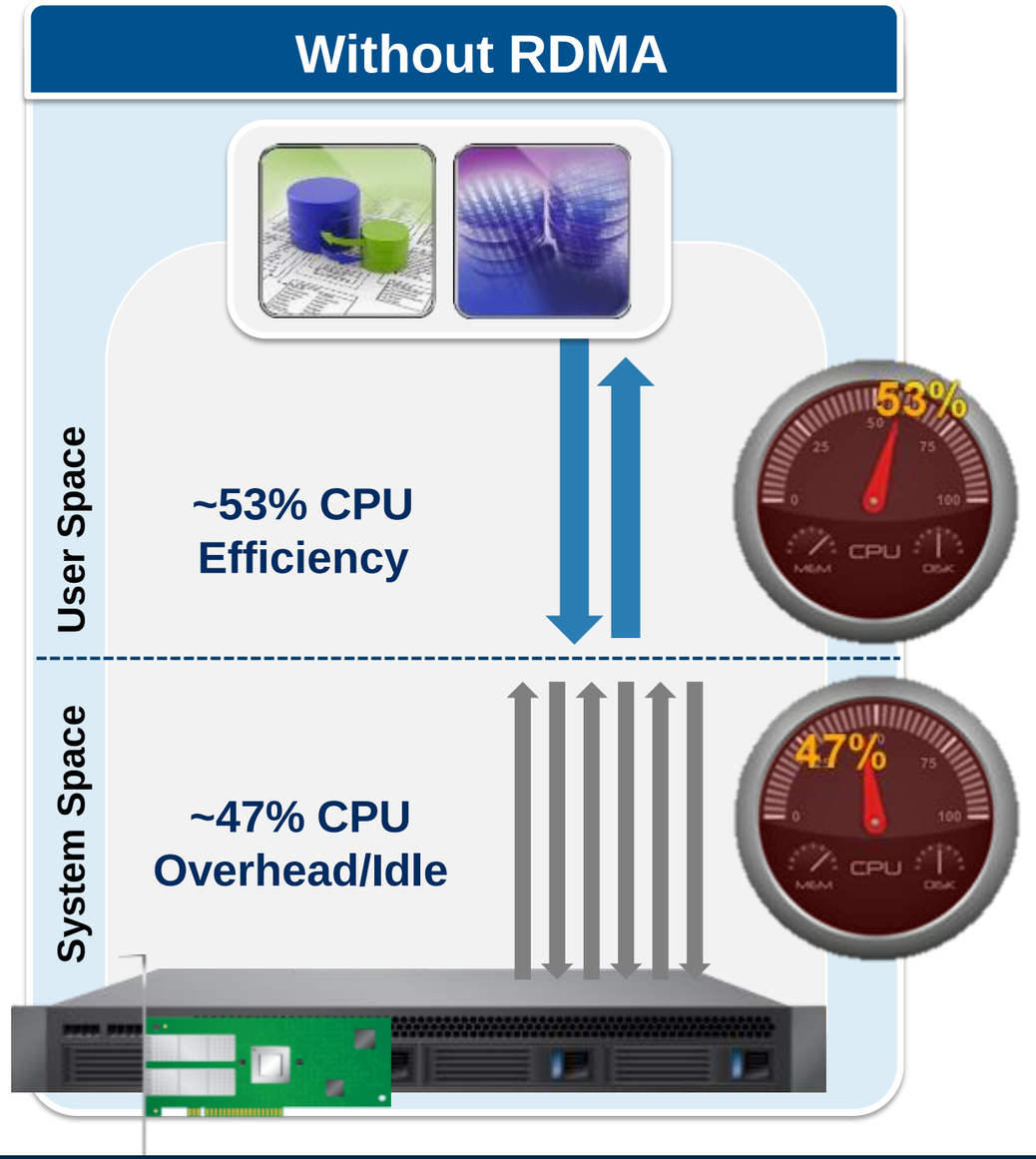
CEPH and Hadoop Co-Exist

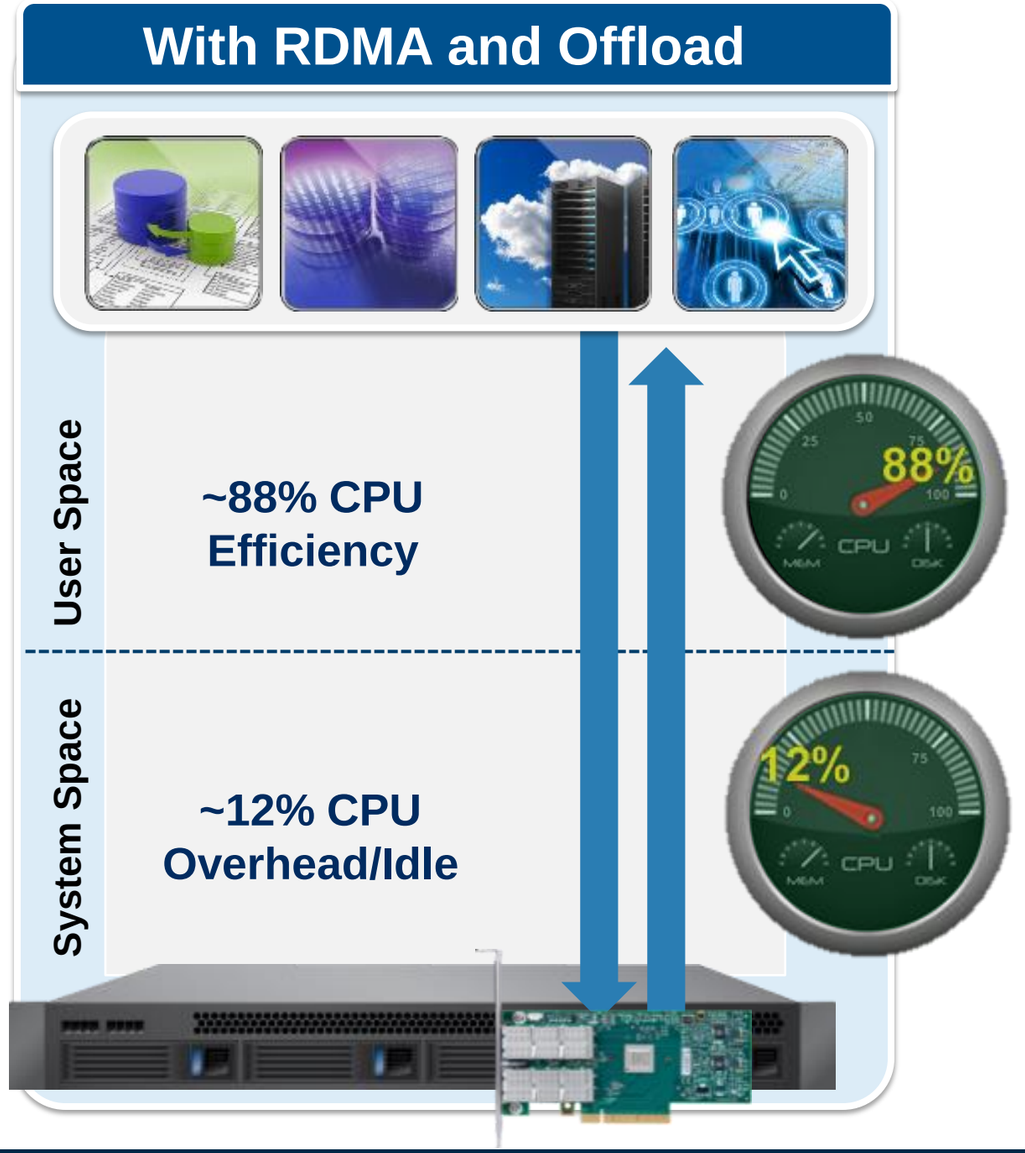
- Increase Hadoop Cluster Performance
- Scale Compute and Storage solutions in Efficient Ways
- Mitigate Single Point of Failure Events in Hadoop Architecture





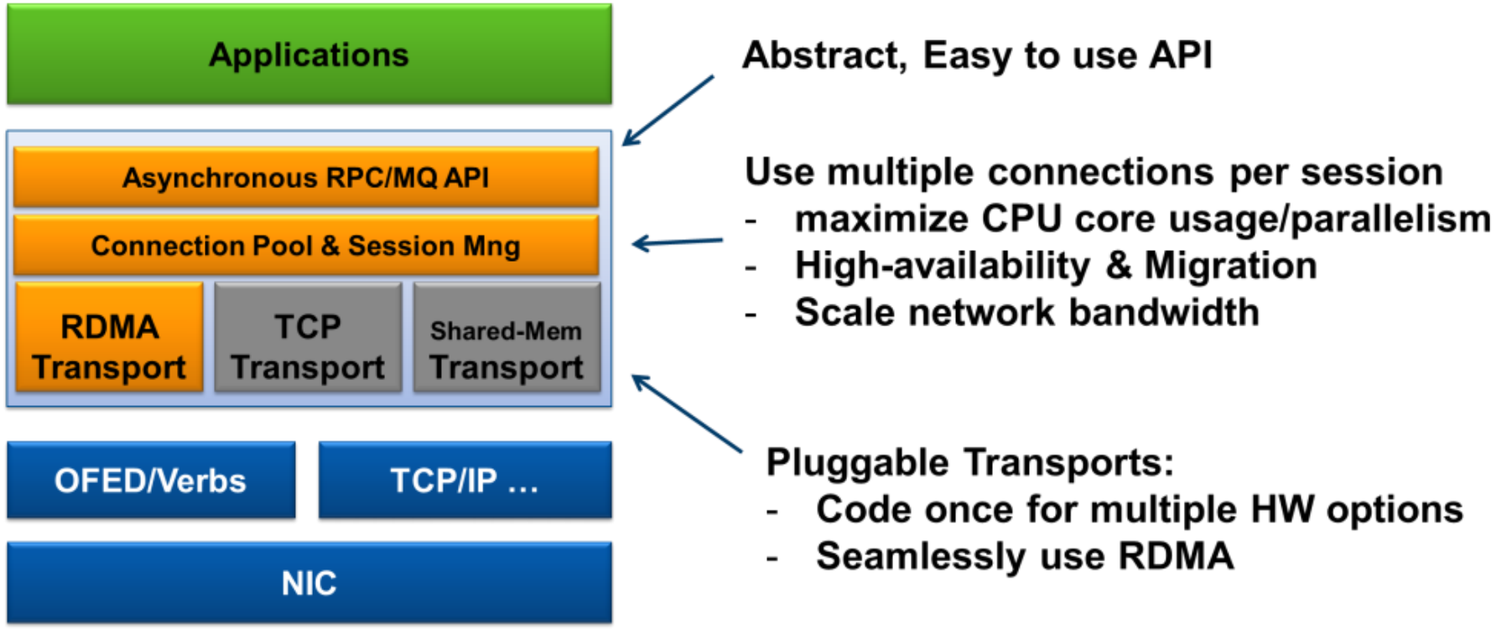
I/O Offload Frees Up CPU for Application Processing





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
- In Next Generation Blueprint (Giant)
  - http://wiki.ceph.com/Planning/Blueprints/Giant/Accelio_RDMA_Messenger



Summary

- CEPH cluster scalability and availability rely on high performance networks
- End to end 40/56 Gb/s transport with full CPU offloads available and being deployed
  - 100Gb/s around the corner
- Stay tuned for the afternoon session by CohortFS on RDMA for CEPH

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™