SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2014

Deploying Ceph With High Performance Networks

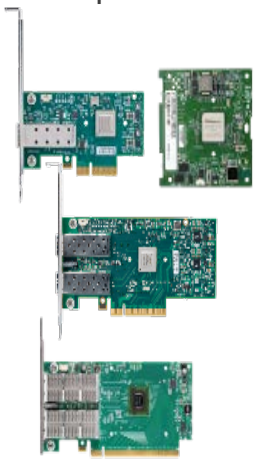
John F. Kim
Mellanox Technologies

Leading Supplier of End-to-End Interconnect Solutions

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs

Adapter Cards

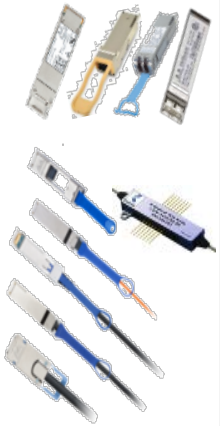


Switches/Gateways

Host/Fabric Software

Metro / WAN

Cables/Modules



The Future Needs Faster Interconnects

1Gb/s

10Gb/s

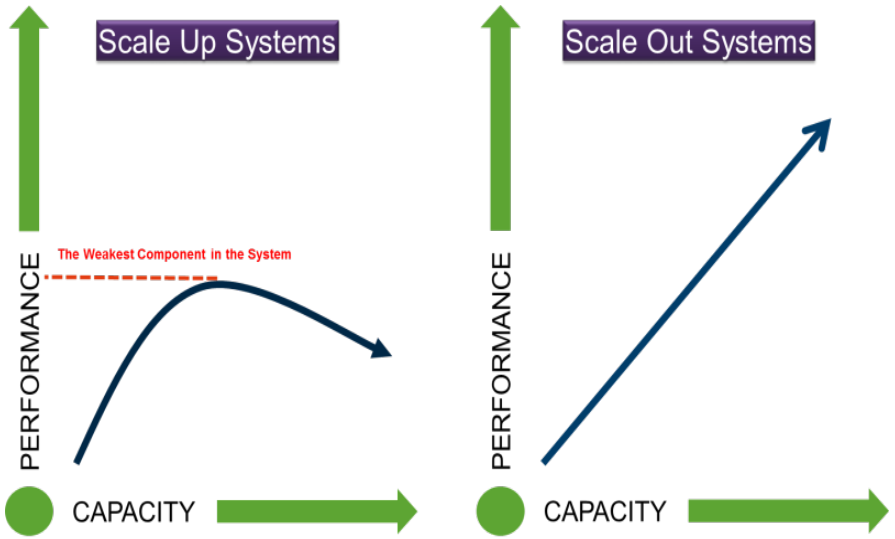
40/56Gb/s

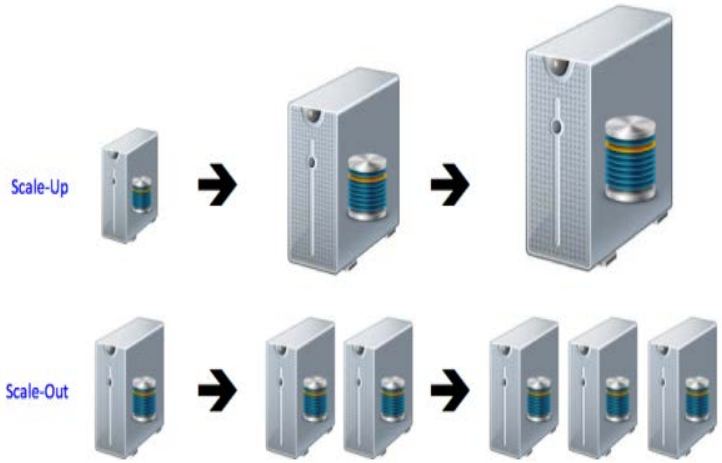
Cloud Storage Requires Massive Scale

- Preference for open-source/software-defined
- Tend to choose scale-out architecture
- Want converged network: storage + server
- Scale-out networks are Ethernet or InfiniBand
  - No Fibre Channel in the Cloud

From Scale-Up to Scale-Out Architecture

- Only cost-effective way to support storage growth
- This happened on the HPC compute side in early 2000s
- Scaling performance linearly requires “seamless connectivity” (i.e. lossless, high bw, low latency, network)

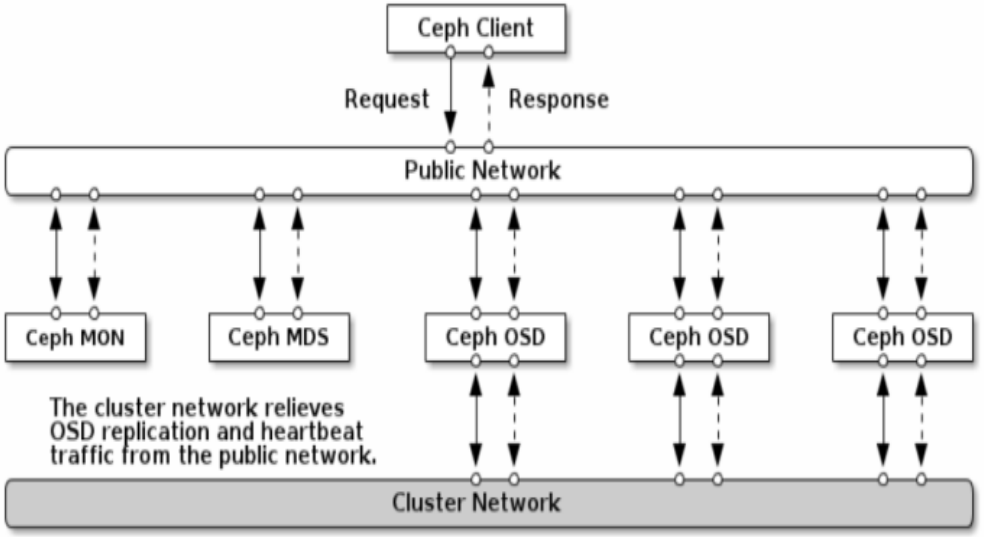




Interconnect Capabilities Determine Scale Out Performance

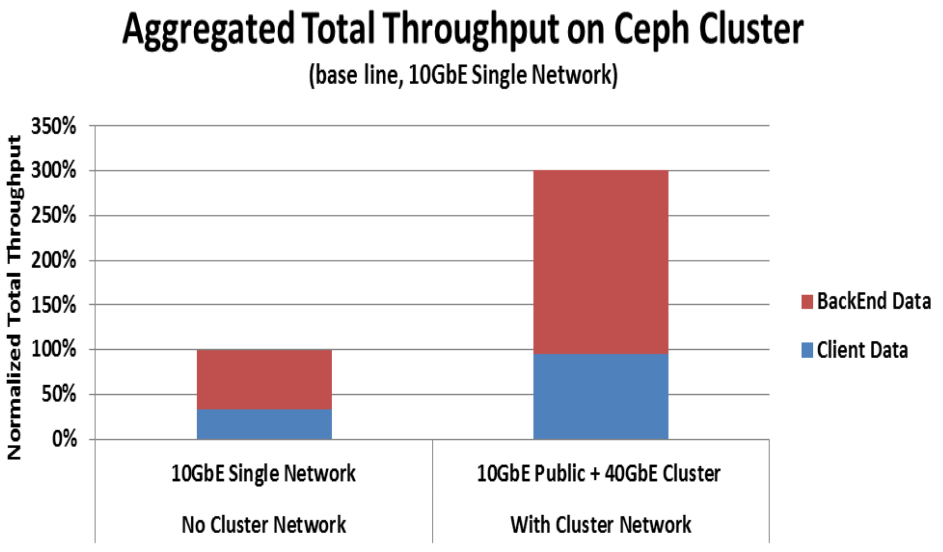
CEPH and Networks

- Two logical networks
  - Client traffic
  - OSD, Monitors and Metadata communication
  - Heartbeat, replication, recovery and re-balancing



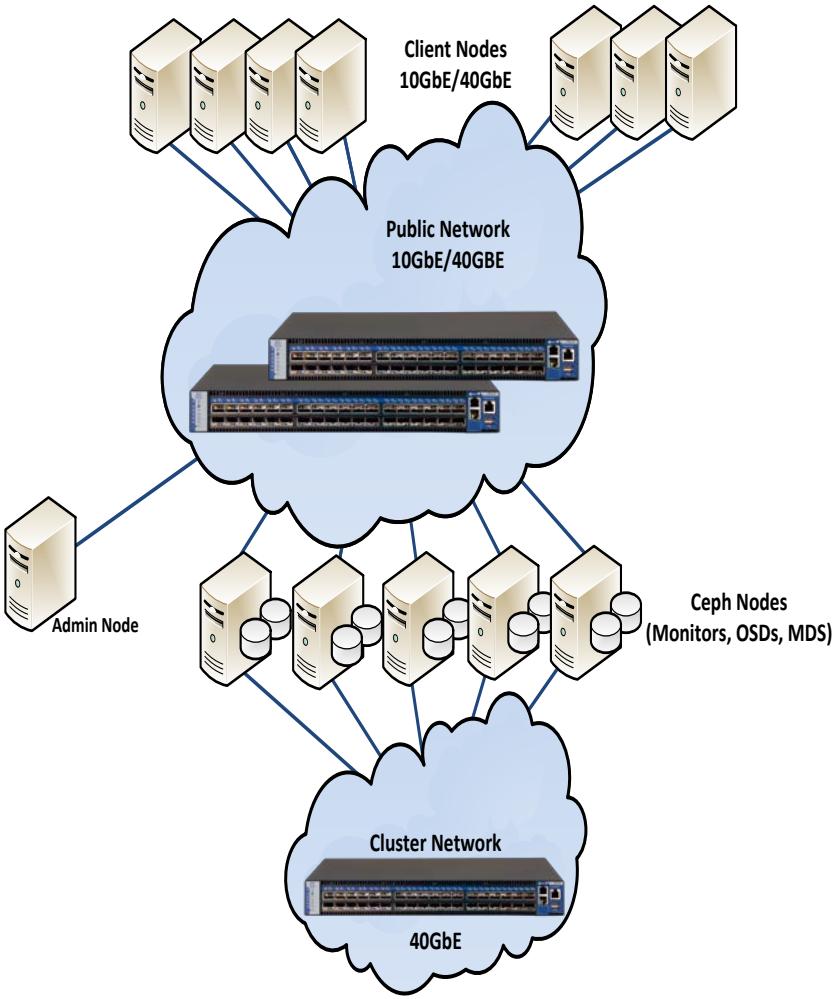
Using Two Networks for Ceph

- Cluster ("backend") network performance can dictates cluster's performance and scalability
- Network load between OSD Daemons easily dwarfs the Ceph Client-to-Storage load
- Separate network maximizes performance



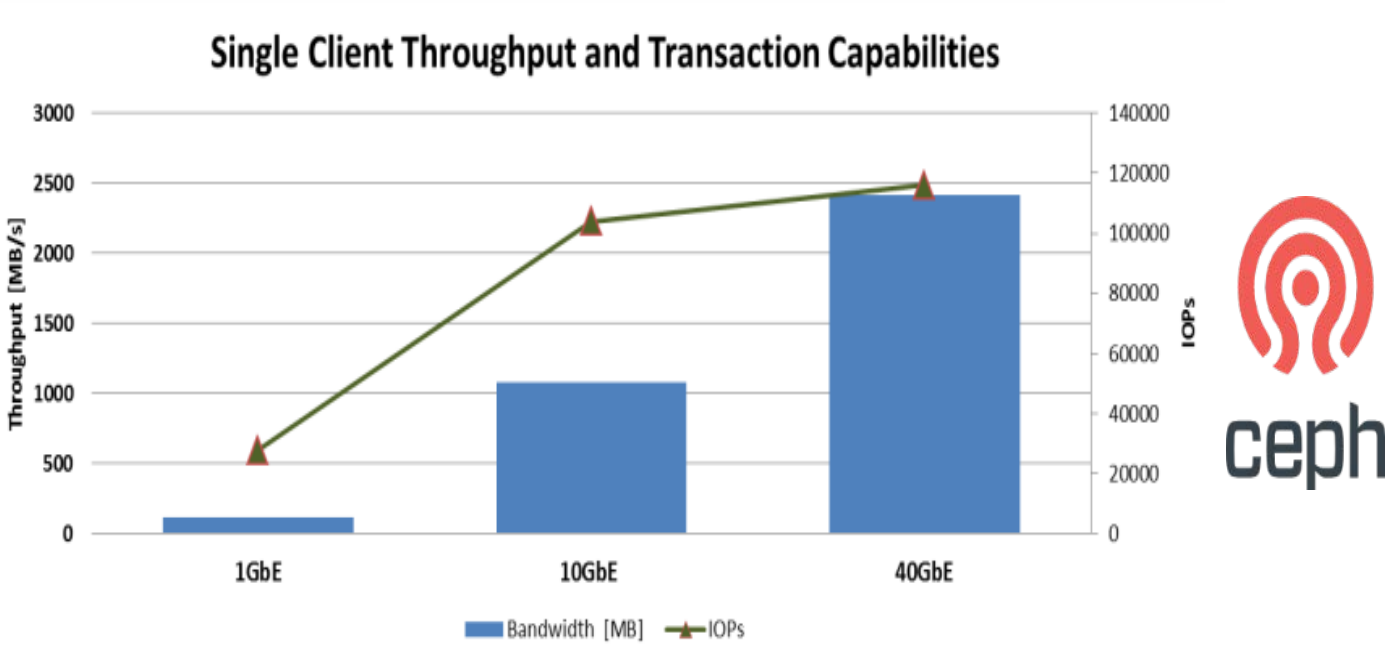
CEPH Deployment Using 10GbE & 40GbE

- 10 or 40GbE public network
- 40GbE Cluster (Private) Network
  - Smooth HA, unblocked heartbeats
  - Efficient data balancing
  - Supports erasure coding



20x Throughput, 4x IOPS using 40GbE

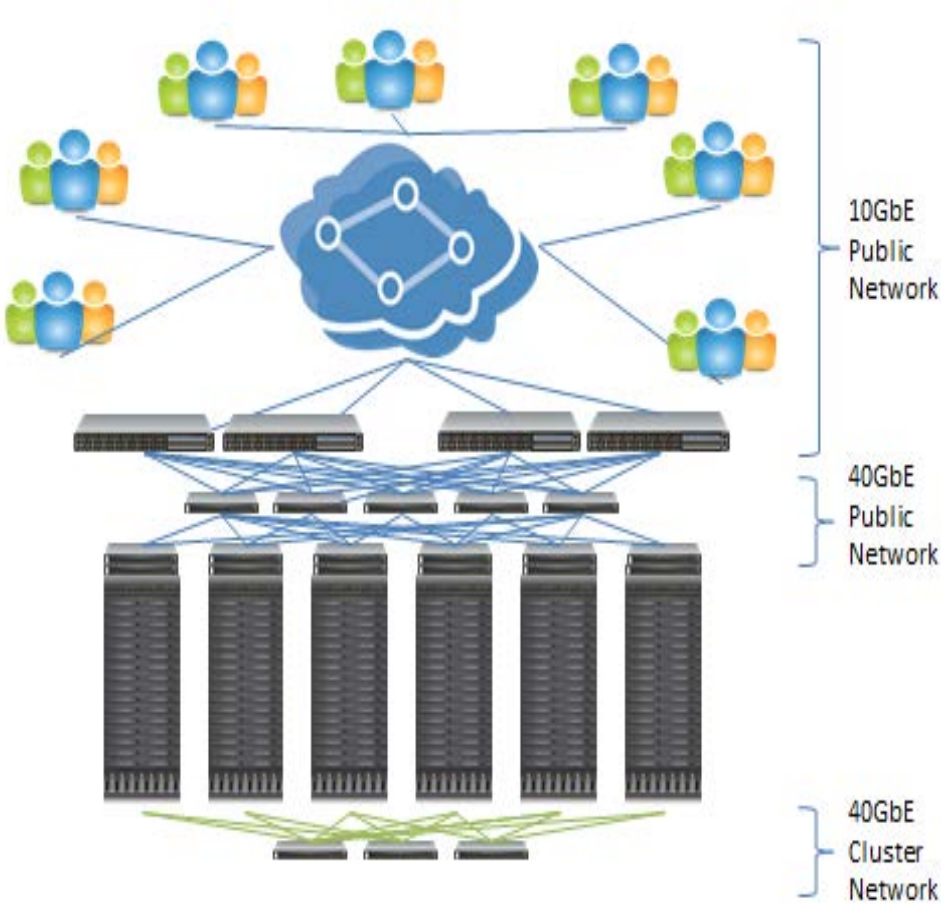
- Line rate for high ingress/egress clients
- 100K+ IOPs/Client @4K blocks



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

Deploying CEPH At Scale with Mellanox

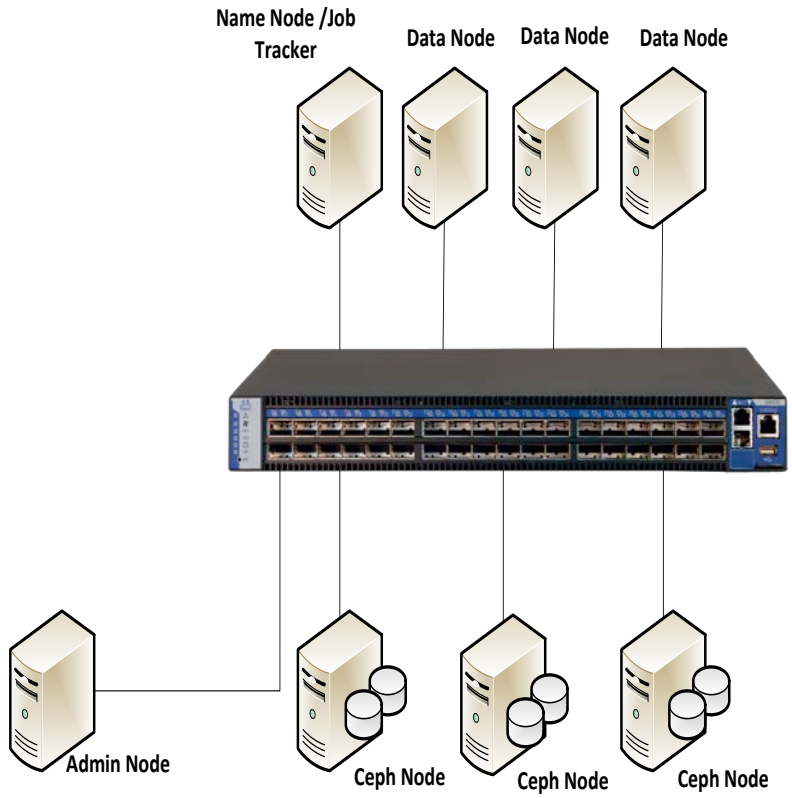
- Cluster network @ 40Gb Ethernet
- Clients @ 10G/40Gb Ethernet
- >500 Client Nodes
- SSDs For OSDs and Journals
- Target Retail Cost: US $350/1TB

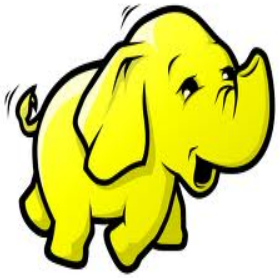


8.5PB System Currently Being Deployed

CEPH and Hadoop Can Co-Exist

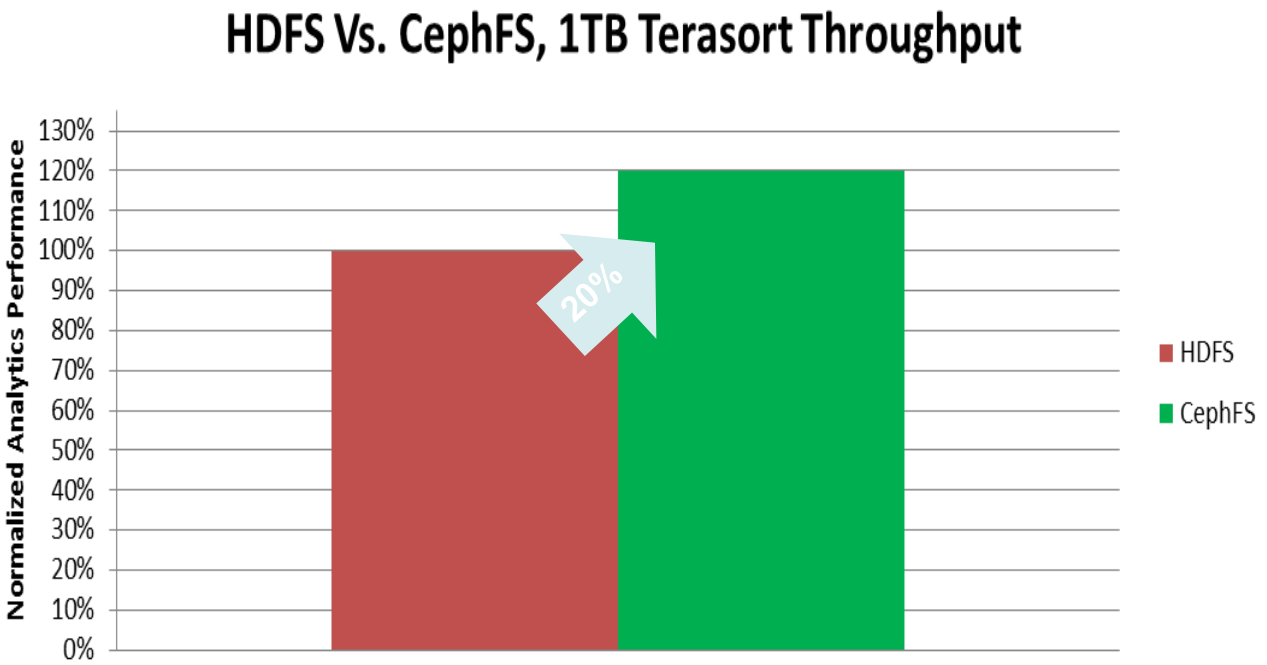
- Increase Hadoop Cluster Performance
- Scale Compute and Storage Efficiently
- Mitigate Hadoop Single Point of Failure



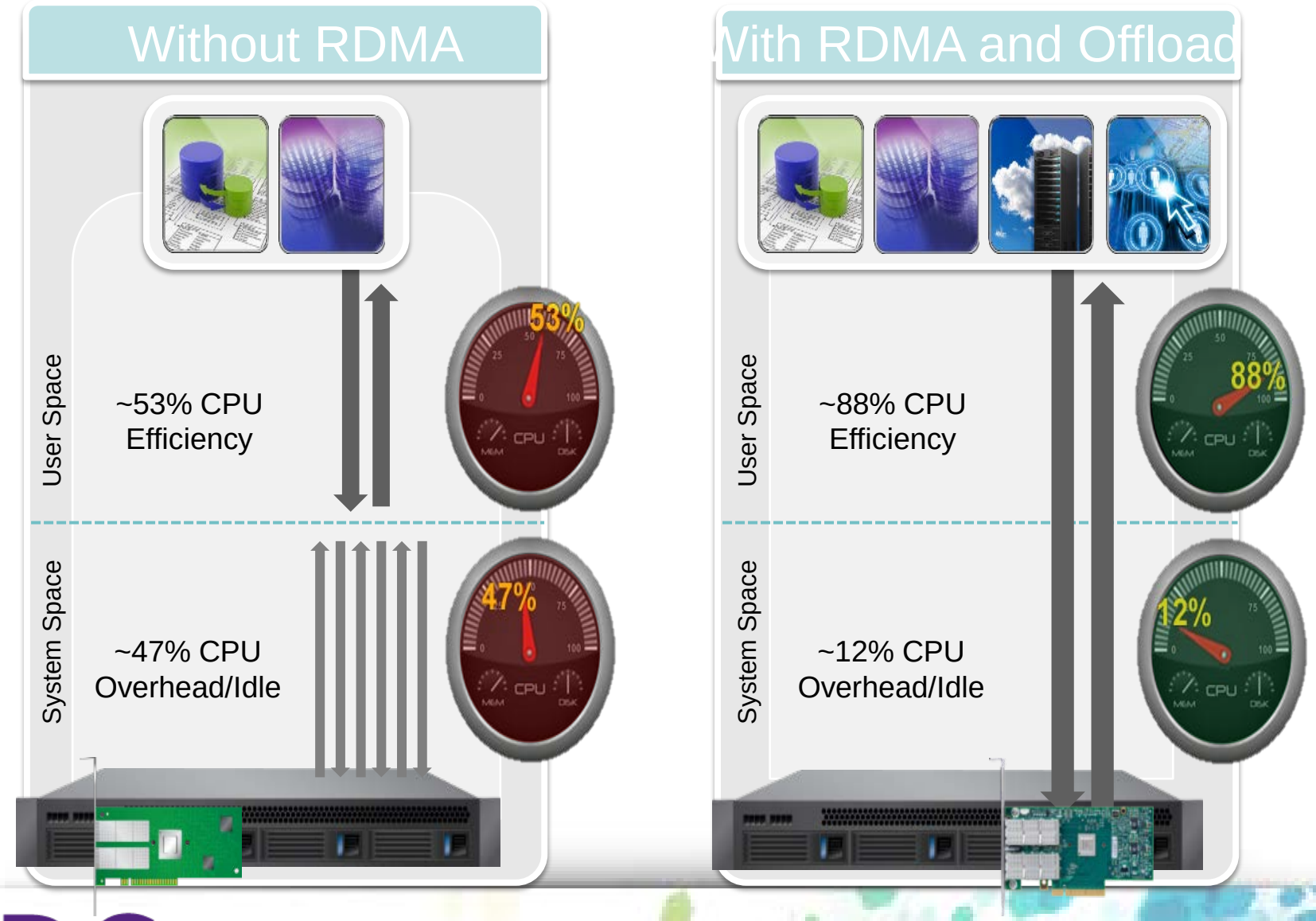


Improves Hadoop Performance 20%

- Ceph plug-in for Hadoop
- Ceph on InfiniBand network
- White paper:
  http://www.mellanox.com/related-docs/whitepapers/wp_hadoop_on_cephfs.pdf



Next for Ceph: RDMA

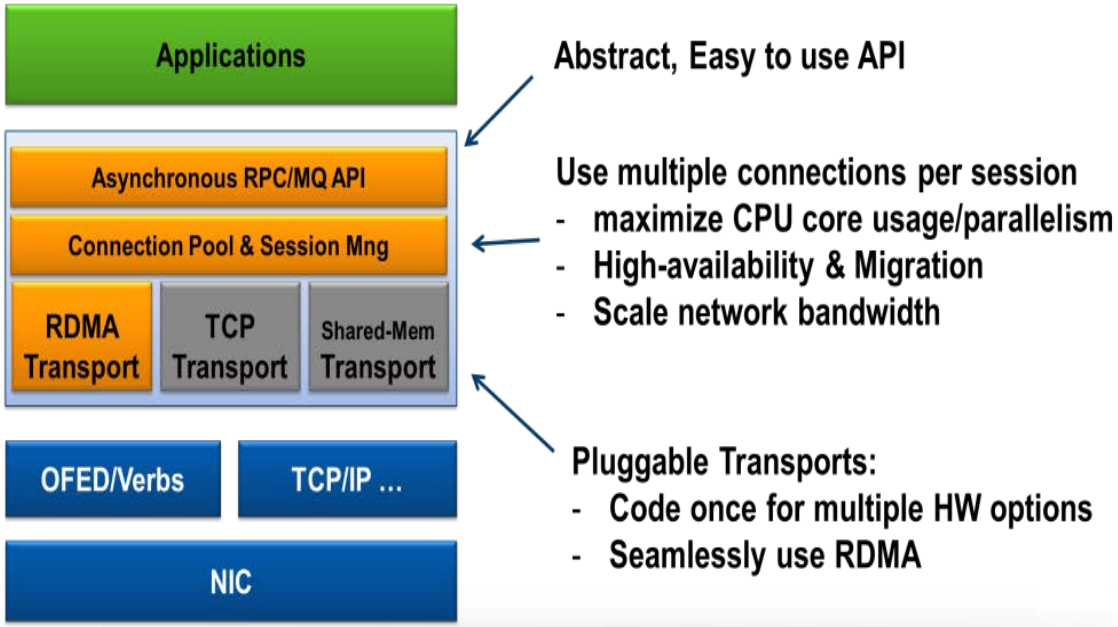


Accelio: Gets You to Faster, More Quickly

- Accelio provides
  - High-Performance Reliable Messaging
  - RPC Library
- Faster RDMA integration to application
- Asynchronous; Maximize message and CPU parallelism
- Enable > 10GB/s from single node
- Enable < 10usec latency under load
- Open source!
  - https://github.com/accelio/accelio/
  - www.accelio.org

Using Accelio to Enable RDMA for Ceph

- In next generation blueprint (Giant)
  - http://wiki.ceph.com/Planning/Blueprints/Giant/Accelio_RDMA_Messenger\
- Encourages additional optimizations to Ceph



Summary

- CEPH cluster scalability and availability rely on high performance networks
- End to end 40/56 Gb/s transport available now
  - RDMA being added to Ceph
  - 100Gb/s around the corner

Thank You