WHITE PAPER

March 2014

Installing Hadoop over Ceph, Using High Performance Networking

Contents

Background

Hadoop[1] is a Java-based programming framework that supports the processing and storage of large data sets in a distributed computing environment. The framework enables the system to continue operating uninterrupted in case of a node failure. This approach lowers the risk of catastrophic system failure. The Hadoop core includes the analytical Map-Reduce engine, and the distributed file system.

Hadoop has become a leading programming framework in the big data space. Organizations are replacing several traditional architectures with Hadoop and use it as a storage, data base, business intelligence and data warehouse solution. Enabling a single file system for Hadoop and other programming frameworks, benefits users who need dynamic scalability of compute and or storage capabilities. However, native Hadoop Distributed File System (HDFS) has several weaknesses and inefficiencies:

Single point of failure (SPOF)
Centralized name server resulting in scalability and cold restart challenges
Not POSIX compliant
Stores at least 3 copies of the data

System manageability and security challenges should also be taken into consideration. The Apache™ Hadoop® project and storage software vendors are independently developing alternative distributed file system solutions that address these limitations.

Different solutions including IBM's GPFS [4], Lustre[5], Parallel Virtual File System (PVFS)[6] and QFS[7] have addressed the inefficiencies of HDFS. Ceph[2] , an emerging software storage solution, mainly for cloud based installations has a file system plugin for Hadoop. Ceph, in conjunction with high performance InfiniBand network, provides an innovative way to store and process Peta and Exa bytes of information for Big Data applications.

The Hadoop architecture distributes not only the data but also the processing of this data across many different compute-storage servers that all must communicate rapidly with each other. This distributed data repository and compute architecture drives the requirement for interconnect performance and scalability in ways previously seen only in the largest super computers in the world. InfiniBand, with FDR 56Gb/s throughput, fits the bill as the interconnect of choice for distributed compu-storage solutions like Hadoop.

Project Scope

The objective of this test is to examine the performance and scalability capabilities of Hadoop over CephFS installation. The test use Hortonworks Data Platform (HDP) 1.3 [8] given this Hadoop distribution is the closest to Apache Hadoop. HDP 1.3 was mounted on Ceph Emperor Version 0.72.2.

The test results show CephFS performed similar or better than the native HDFS. Data centers can deploy Hadoop clusters in conjunction with other applications on a single file system, without degrading cluster or application performance. Moreover, dynamic scalability of compute and or storage can be applied to Hadoop specific workloads. Adding more power to the network infrastructure enables scalability, ease of use and performance.

Ceph

Ceph is an emerging storage solution with object and block storage capabilities. Mainly deployed in cloud based installations and provides a scalable and reliable alternative to traditional storage applications. When used in conjunction with high-performance networks, Ceph can provide the needed throughput and input/output operations per second (IOPs) to support a multi-user Hadoop or any other data intensive application. Ceph is using RADOS [9] , a reliable autonomic distributed object store to enable client reach to the stored data. With object and block storage capabilities, Ceph is the storage solution of choice for cloud based deployments. The Cinder [10] or S3[11] based application protocol interfaces enable users to use Ceph in OpenStack and public cloud based solutions.

Ceph File System (CephFS)

The installation and configuration details of a Ceph cluster is available on Ceph's website at www.ceph.com. The Ceph installation and architecture should be reviewed prior to referring to this document for a deployment.

CephFS provides users access to file system storage based on Ceph's object storage solutions. Benefits to using CephFS are listed below.

Benefits:

- Stronger data safety for mission-critical applications
- Virtually unlimited storage to file systems
- Self-rebalancing for performance maximization
- Support for POSIX semantics.

CephFS is still evolving, but this test shows it is already useful for improving the scalability and ROI of Hadoop cluster builds.

Hadoop clients can access CephFS through a Java-based plugin named hadoop-cephfs.jar. Visit the Ceph website at http://ceph.com/download/hadoop-cephfs.jar to download the plugin. The two Java classes below are required to support Hadoop connectivity to CephFS.

Java Classes for Hadoop connectivity to CephFS:

1. libcephfs.jar that should be placed in /usr/share/java/ and the path should be added to: HADOOP_CLASSPATH in Hadoop_env.sh file

   Download website: ftp://ftp.pbone.net/mirror/ftp5.gwdg.de/pub/opensuse/repositories/home:/H4T:/filesystem:/Testing/RedHat_RHEL-6/x86_64/cephfs-java-0.67.2-2.1.x86_64.rpm

2. libcephfs_jni.so that should be added to the LD_LIBRARY_PATH environment parameter and placed in /usr/lib/hadoop/lib we also soft linked the package to: /usr/lib/hadoop/lib/native/Linux-amd64-64/ libcephfs_jni.so

   Download website: ftp://ftp.pbone.net/mirror/ftp5.gwdg.de/pub/opensuse/repositories/home:/H4T:/filesystem:/Testing/RedHat_RHEL-6/x86_64/libcephfs_jni1-0.67.2-2.1.x86_64.rpm

InfiniBand

The ever increasing demand for higher performance, efficiency and scalability in data centers drives the need for faster server and storage connectivity. InfiniBand (IB) [12] is the leading standardized interconnect that provides the highest bandwidth, lowest latency, lowest power consumption and lowest CPU overhead for maximizing compute systems productivity and overall return on investment.

The high-speed InfiniBand server and storage connectivity has become the de facto scalable solution for systems of any size – ranging from small, departmental-based compute infrastructures to the world's largest PetaScale systems. The rich feature set and design flexibility enable users to deploy the InfiniBand connectivity between servers and storage in various architectures and topologies to meet performance and or productivity goals.

InfiniBand currently supports bandwidth data rate of up to 56Gb/s, with a roadmap to 100Gb/s and 200Gb/s, in the near future. Latency for Remote Direct Memory Access (RDMA) based applications is less than 5us. Some applications will be able to demonstrate latency lower than 1us for a data transfer from one server to another.

The Test Environment

The test environment has a total of 8 servers. Ceph server's configuration can be found in appendix 3. Hadoop server's configuration can be found in appendix 4.

Server Count:

- (4) Nodes Ceph Cluster
    - o (1) Admin node
    - o (3) Ceph storage nodes

- (4) Nodes Hadoop cluster
    - o (1) Name node and job tracker
    - o (3) data/task node

All servers are installed with Red Hat Enterprise Linux version 6.3 with kernel version 3.10.26. A newer version of kernel is recommended. The Ceph cluster was approached by the clients using the kernel RBD driver and achieved over 600MB/s of throughput from every node.

Hadoop over HDFS Performance Testing

During the testing, the Hadoop over CephFS installation was assembled according to the instructions in the HDP installation manual. Each data node had 4 hard drives, 7.2K rpm, 1 Terabyte each. Mellanox FDR 56Gb/s InfiniBand ConnectX-3 host adapter card with Mellanox OFED version. 2.0.3 were used inside the cluster. Based on iperf [13] benchmarking standards, the test yielded a constant 47Gb/s of throughput between the servers. For metadata records, a single name node with a job tracker on the name node server was deployed, resulting in 3 servers reserved as data nodes.

The name node is configured with a 256GB SAS SSD boot drive and a 800GB PCIe SSD card, used for the name node metadata storage. The name node is using a Mellanox FDR 56Gb/s InfiniBand ConnectX-3 host adapter card installed with Mellanox OFED version 2.0.3.

The switching solution used is a single 1RU, Mellanox 36-port FDR 56gb/s InfiniBand MSX6036 switch.

The intrinsic Terasort[14] benchmark was used and configured to analyze 1 Terabyte of synthetic data and the replication factor was set to 3 (default). The testing conducted under the default settings from HDP installation manual.

The result achieved for the HDFS testing was: 5338 Seconds for 1TB Terasort benchmarking.

The Teragen and Terasort execution command lines are below:

Teragen:

```
bin/hadoop jar /usr/lib/hadoop/hadoop-examples.jar teragen 10000000000 /user/root/teragen
```

Terasort:

```
bin/hadoop jar /usr/lib/hadoop/hadoop-examples.jar terasort /user/root/teragen /user/root/terasort
```

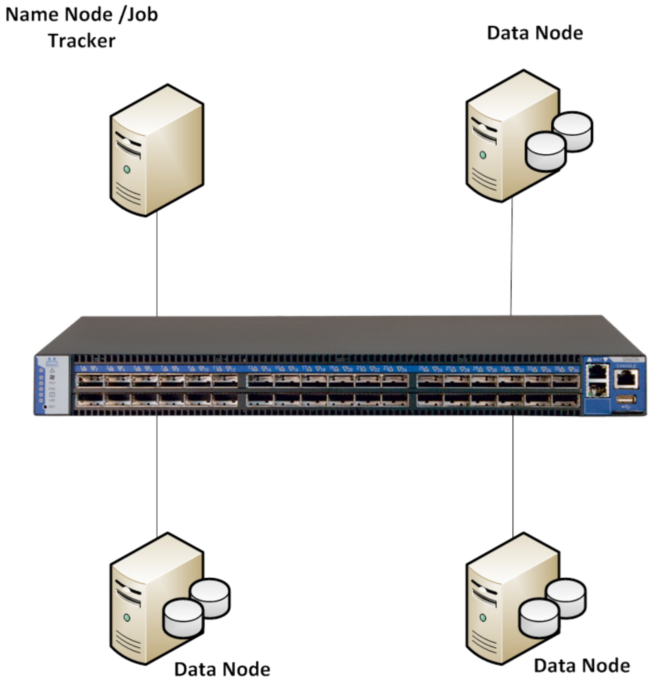


Figure 1: *Hadoop over HDFS Testing Enviornment*

Hadoop over CephFS Performance Testing

The Ceph cluster used during this test was installed according to the instructions in Ceph installation manual located at http://ceph.com/docs/master/install. The following hardware was used for the Ceph cluster deployment.

Hardware:

- (1) Admin node (ceph_deploy)
- (3) Object Storage Devices (OSD) nodes
- (1) Metadata node
- (3) Monitor nodes
- (12) OSDs
  - 400 placement groups per OSD
- (1) 800GB PCIe SSD, operating on a PCIe Gen2 x4 (for journaling)
  - One PCIe SSD card in each OSDs node, partitioned to 4 sections, one for every OSD

Ceph-S Output:

```
cluster 1cf76947-d348-4213-ac01-0a890a9097e2
health HEALTH_OK
monmap e3: 3 mons at {apollo002-ib=192.168.30.190:6789/0,apollo003 ib=192.168.30.191:6789/0,apol
lo008-ib=192.168.30.196:6789/0}, election epoch 18, quorum 0,1,2 apollo002-ib,apollo003-ib,apollo008-ib
mdsmap e18: 1/1/1 up {0=apollo002-ib=up:active}
osdmap e71413: 12 osds: 12 up, 12 in
pgmap v187340: 1200 pgs, 3 pools, 2851 GB data, 267 kobjects
5705 GB used, 5454 GB / 11159 GB avail
1200 active+clean
```

The configuration in Figure 2 below shows the Ceph and Hadoop cluster deployment. Changing the reference mounting point from the local hard drives on the Hadoop name and data node to the CephFS primary monitor node provides the connectivity to the Ceph cluster. The required changes are described in appendixes 1 and 2.

The intrinsic Terasort benchmark was used and configured to analyze 1 Terabyte of synthetic data and the Ceph replication factor is set to 2 (default).

The result achieved for the CephFS testing is: 4474 seconds for 1TB Terasort benchmarking.

The Teragen and Terasort execution command lines are below.

Teragen:

```
bin/hadoop jar /usr/lib/hadoop/hadoop-examples.jar teragen 10000000000 /user/root/teragen
```

Terasort:

```
bin/hadoop jar /usr/lib/hadoop/hadoop-examples.jar terasort /user/root/teragen /user/root/terasort
```

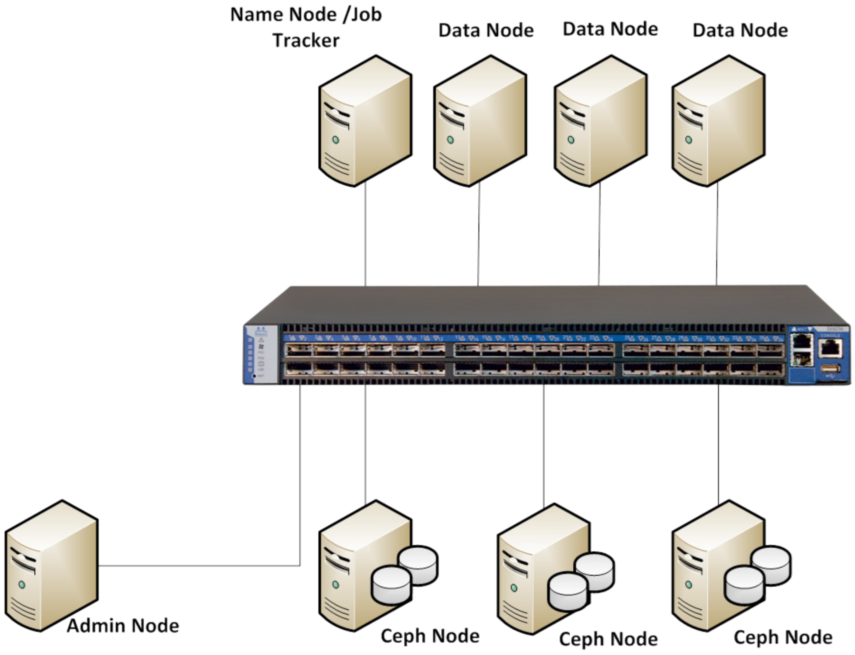


Figure 2: *Hadoop over Ceph Testing Environment*

The performance increase of 20% shows that HDFS can be replaced by an external file system without degrading the analytics engine performance.

Conclusion

1. Setting a Hadoop cluster on top of a dedicated file system enables flexibility, better redundancy and security for user's data.

    a. Each portion of the cluster can be scaled dynamically. The storage capacity of CephFS OSD nodes can be increased without investing in additional server or vice versa.

    b. With Hadoop over CephFS, the HDFS inefficiencies are completely addressed and performance levels are similar or better than the traditional Hadoop model as shown in

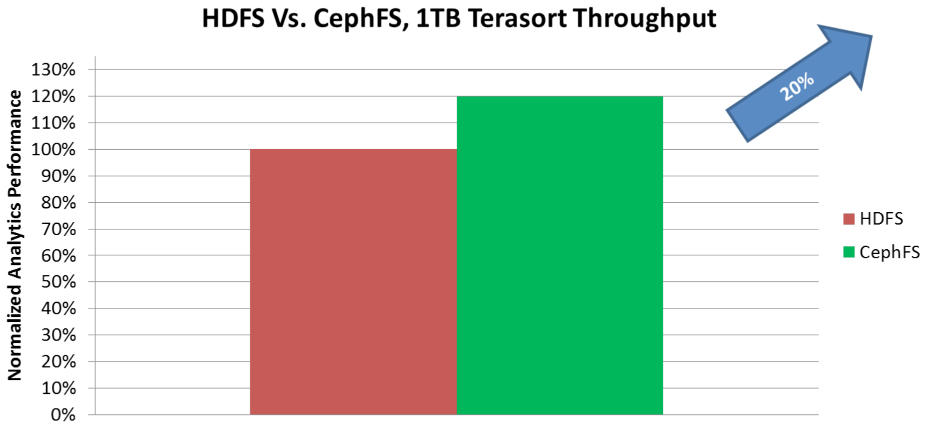


Figure 3: *Normalized Performance Results*

c. The release of Ceph with erasure coding, will significantly reduce the number of storage devices required to store data in a reliable manner. Thus, making the usage of Ceph as a storage solution for big data applications a cost reduction opportunity.

2. The network setup has a crucial impact on the ability to scale and provide the needed performance.

a. Architecting a private network for data balancing and heart beats provides a lower "noise" level in the storage system and less false alert on OSDs commission and decommission.
b. With erasure coding, a customer will require a more efficient networking solution to sustain the coding overhead without impacting cluster performance. RDMA provides an efficient, low CPU overhead solution for the data movement in clusters.
c. Bandwidth is imperative to storage capabilities. Currently, magnetic SATA drives can generate a steady ~120MB/s of data throughput. With servers allocating 20+ drives in a single chassis, the steady data bandwidth generated by a single server exceeds 2.4 GB/s or nearly 20Gbits/s and bursts of 25Gbit/s are real life occurrences. 40GbE and FDR 56Gb/s InfiniBand are the only available solutions in the market to support such bandwidths at a reasonable cost.

3. Other applications in the big data space can benefit from similar deployment solutions.

a. Many of the NoSQL, real-time analytics solutions are using a dedicated storage scheme, in many cases a proprietary one. Storing data and metadata for the different application on a single, cost effective, dedicated to all big data applications can solve a continuous operations and management headache.
b. Kernel based drivers as the ones available for Ceph, make the installation and integration an easy path for users' adoption.

4. The table below summarizes the benefits of using CephFS vs. HDFS

Feature	Hadoop over CephFS	Hadoop over HDFS
Enables heterogeneous storage and compute	Yes, compute servers are not coupled to storage	No. Typically, requires homogenous server architecture
Support for POSIX Protocol	Yes	No
Fault Tolerant: Data storage has no single point of failure	Yes, data redundancy and fault tolerance is maintained under the Ceph domain	No, Name Node crash will limit or prevent access to data
Native support for Erasure Coding	Yes, Building a more efficient storage solution	No, replication is required; data storage efficiency is less than 33%.
Optional dedicated balancing and heart beat infrastructure	Yes, eliminates bottlenecks on the user access network, minimizing false decommissions due to network latency.	No, user and HDFS servers share the same network infrastructure
Hadoop Performance Capabilities	Same or better than HDFS	Baseline

Appendix 1: Changes to incorporate into core-site.xml file on Hadoop nodes to support CephFS

```
<!-- file system properties -->
<property>
   <name>fs.ceph.impl</name>
   <value>org.apache.hadoop.fs.ceph.CephFileSystem</value>
</property>
<property>
   <name>fs.default.name</name>
   <value>ceph:///</value>
</property>
<property>
   <name>ceph.conf.file</name>
   <value>/etc/ceph/ceph.conf</value>
</property>
<property>
   <name>ceph.root.dir</name>
   <value>/</value>
</property>
<property>
   <name>ceph.mon.address</name>
   <value>192.168.30.190:6789</value>
<description>This is the primary monitor node IP address in our installation.</description>
</property>
<property>
   <name>ceph.auth.id</name>
   <value>admin</value>
</property>
<property>
   <name>ceph.auth.keyring</name>
   <value>/etc/ceph/ceph.client.admin.keyring</value>
</property>
```

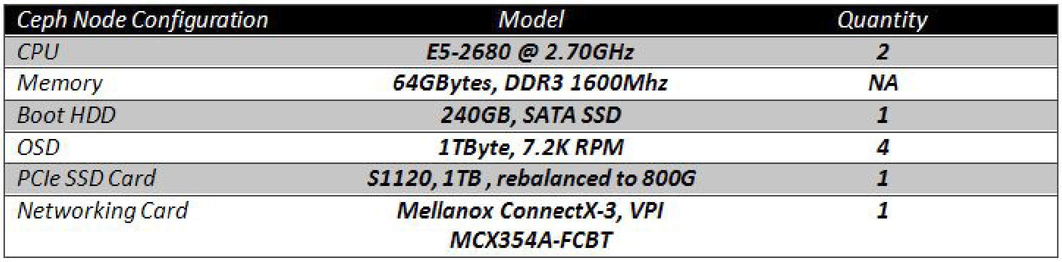
Appendix 2: Changes to Incorporate into Hadoop_env.sh file on Hadoop nodes to support CephFS

```
export LD_LIBRARY_PATH=/usr/lib/hadoop/lib

export HADOOP_CLASSPATH=/usr/lib/hadoop/lib/Hadoop-cephfs.jar:/usr/share/java/libcephfs.
jar:$HADOOP_CLASSPATH
```

Appendix 3: Ceph Cluster Configuration

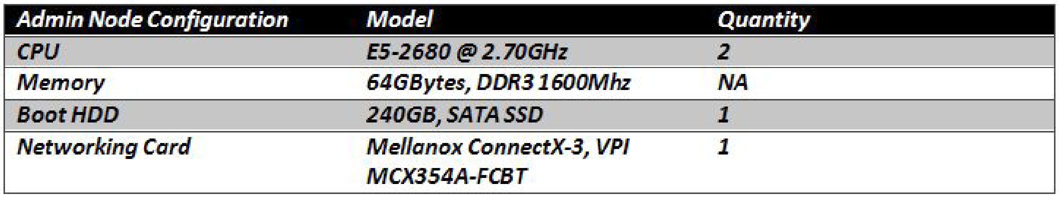
Ceph Nodes:



Ceph Node Configuration	Model	Quantity
CPU	E5-2680 @ 2.70GHz	2
Memory	64GBytes, DDR3 1600Mhz	NA
Boot HDD	240GB, SATA SSD	1
OSD	1TByte, 7.2K RPM	4
PCIe SSD Card	S1120, 1TB , rebalanced to 800G	1
Networking Card	Mellanox ConnectX-3, VPI MCX354A-FCBT	1

We used 3 nodes of the above configuration as OSD nodes, total of 12 OSDs.

Admin Nodes:



Admin Node Configuration	Model	Quantity
CPU	E5-2680 @ 2.70GHz	2
Memory	64GBytes, DDR3 1600Mhz	NA
Boot HDD	240GB, SATA SSD	1
Networking Card	Mellanox ConnectX-3, VPI MCX354A-FCBT	1

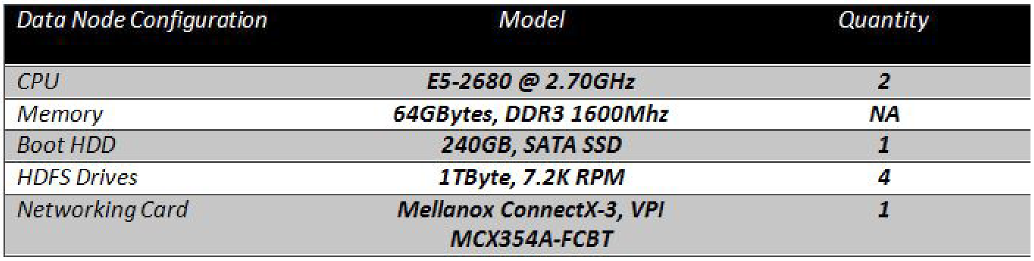
We used one admin node to install and deploy Ceph clients.

Cluster is connected using FDR 56Gbps Infiniband connectivity thorough Mellanox MSX6036, 36 ports, QSFP switch. Servers are connected to the switch using Mellanox FDR 56Gbps, copper cables. All servers are connected to the same switch and using the same subnet.

Only one public network is configured for the testing. Hadoop nodes, described in Appendix 4 are connected to the same public network for testing.

Appendix 4: Hadoop Cluster Configuration

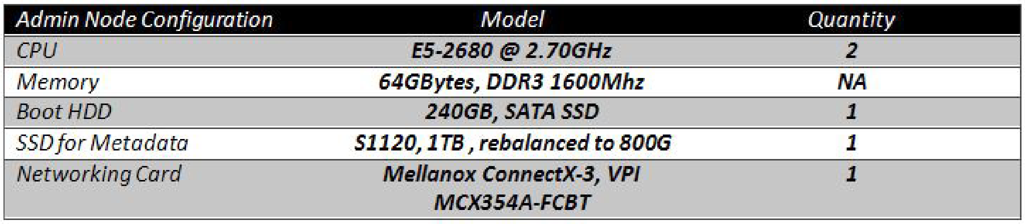
Hadoop Data Nodes:



Data Node Configuration	Model	Quantity
CPU	E5-2680 @ 2.70GHz	2
Memory	64GBytes, DDR3 1600Mhz	NA
Boot HDD	240GB, SATA SSD	1
HDFS Drives	1TByte, 7.2K RPM	4
Networking Card	Mellanox ConnectX-3, VPI MCX354A-FCBT	1

We used 3 nodes of the above configuration as Data/task nodes.

Hadoop Name Node:



Admin Node Configuration	Model	Quantity
CPU	E5-2680 @ 2.70GHz	2
Memory	64GBytes, DDR3 1600Mhz	NA
Boot HDD	240GB, SATA SSD	1
SSD for Metadata	S1120, 1TB , rebalanced to 800G	1
Networking Card	Mellanox ConnectX-3, VPI MCX354A-FCBT	1

We used one name node to control and manage the Hadoop cluster.

Cluster is connected using FDR 56Gbps Infiniband connectivity thorough Mellanox MSX6036, 36 ports, QSFP switch. Servers are connected to the switch using Mellanox FDR 56Gbps, copper cables. All servers are connected to the same switch and using the same subnet.

References

[1] Apache Hadoop: http://hadoop.apache.org/
[2] Ceph: http://ceph.com/
[3] https://www.usenix.org/legacy/publications/login/2010-08/openpdfs/maltzahn.pdf
[4] IBM GPFS w/ Hadoop: http://pic.dhe.ibm.com/infocenter/bigins/v2r1/index.jsp?topic=%2Fcom.ibm.swg.im.infosphere.biginsights.install.doc%2Fdoc%2Fgpfs_upgrade_hdfs_overview.html
[5] http://www.xyratex.com/sites/default/files/Xyratex_white_paper_MapReduce_1-4.pdf
[6] http://www.jeffshafer.com/publications/papers/yee-asbd12.pdf
[7] https://www.quantcast.com/engineering/qfs/
[8] http://hortonworks.com/products/hdp-1-3/
[9] http://ceph.com/papers/weil-rados-pdsw07.pdf
[10] https://github.com/openstack/cinder
[11] http://aws.amazon.com/s3/
[12] http://www.infinibandta.org/
[13] http://iperf.sourceforge.net/
[14] https://hadoop.apache.org/docs/current/api/org/apache/hadoop/examples/terasort/package-summary.html

Mellanox
TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com