Running Hadoop Map Reduce Jobs on Lustre*

Zhiqi Tao and Gabriele Paciucci

April 08, 2014

** Some name and brands may be claimed as the property of others.*

Agenda

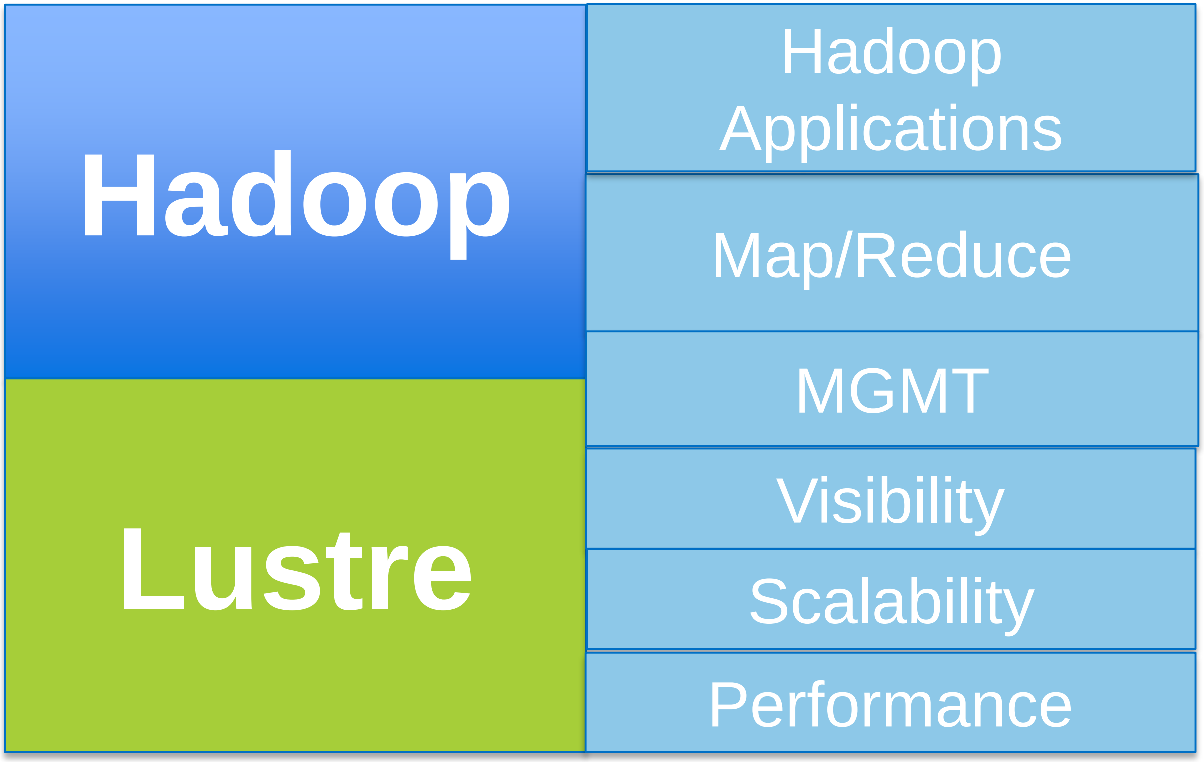
- Overview
- How to configure Hadoop with Lustre*
- Benchmarks results
- Future works

*other names and brands may be claimed by others

Why runs Hadoop Map Reduce Jobs on Lustre*?

Effective Data Processing

High Performance Storage



Hadoop

Lustre

Hadoop Applications

Map/Reduce

MGMT

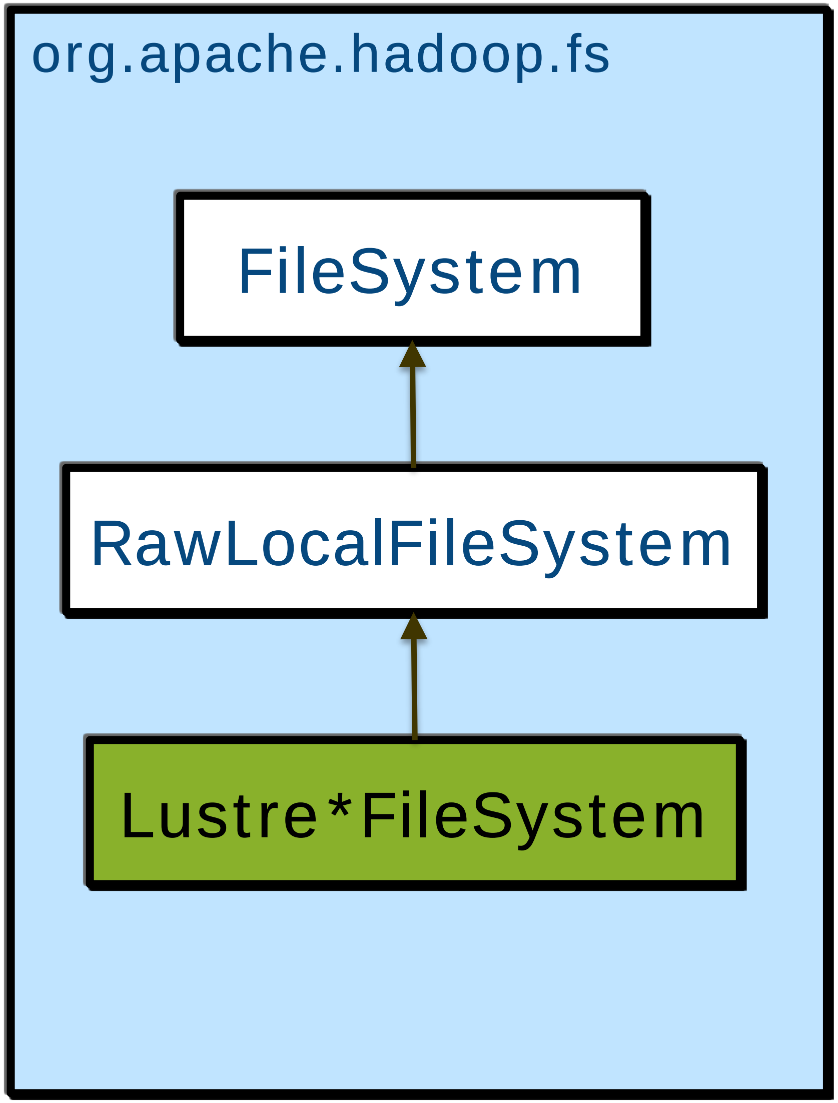
Visibility

Scalability

Performance

*other names and brands may be claimed by others

Recall Omkar's talk at LUG'13 ?



- Used Hadoop's built-in LocalFileSystem class to add the Lustre file system support
- Defined new URL scheme for Lustre, i.e. `lustre:///`
- Optimized the shuffle phase
- Demonstrated huge performance improvement

*other names and brands may be claimed by others

Setup Overview

- Install and Setup Lustre*
- Mount Lustre
- Install and Setup Hadoop
- Direct Hadoop IOs to Lustre instead of HDFS

I'm here to just talk about the approach I know. There would certainly be more than one way that leads to Rome. ☺

*other names and brands may be claimed by others

Preparation

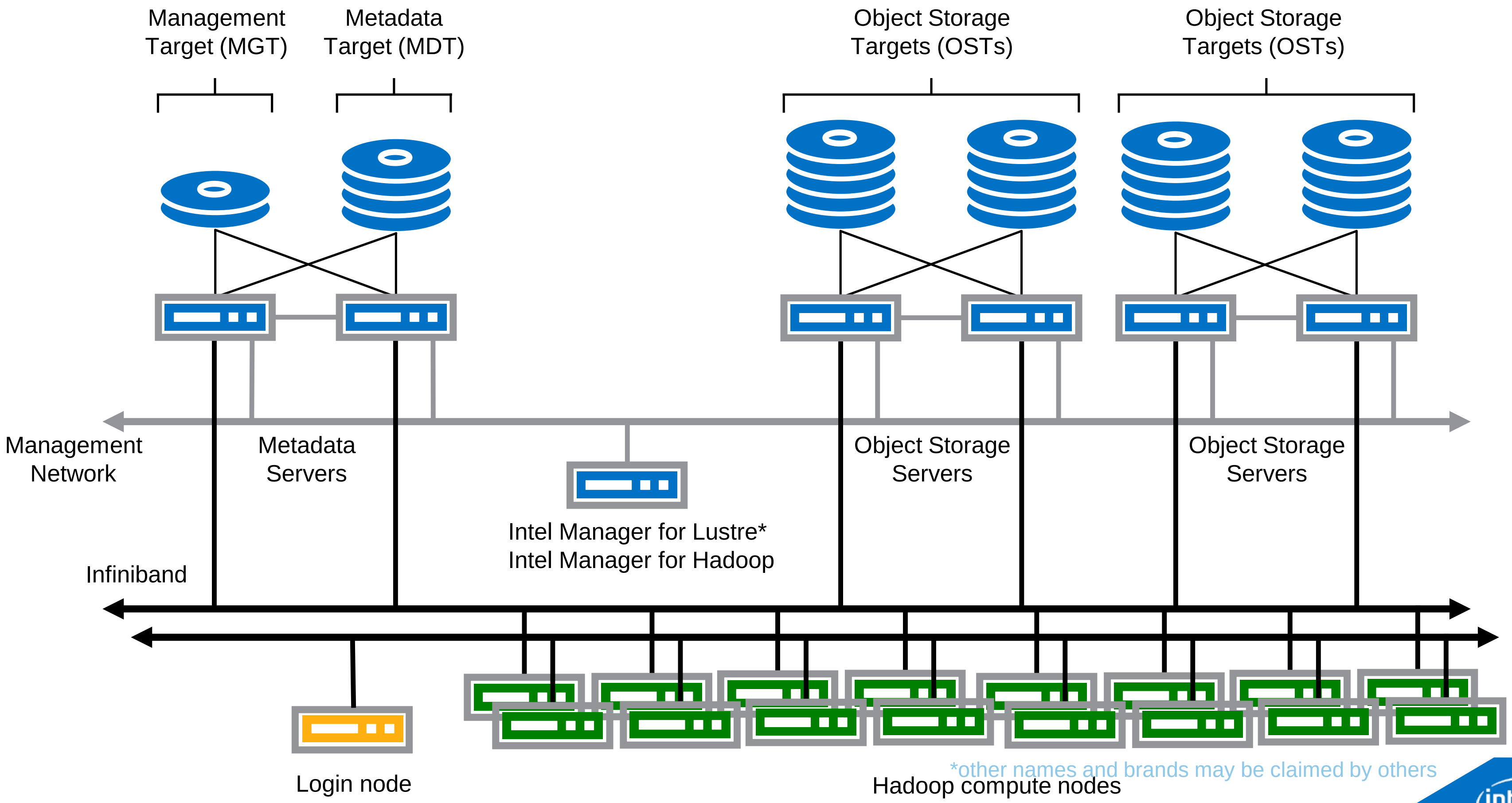
- Consistent UID and GID, especially for the Hadoop users
  - The best way is to setup the global naming server and connect Lustre* server and Hadoop server there.
  - For a small test system, try this script.

```
VALUE=10000;
for i in hive hbase hdfs mapred yarn;
do
    VALUE=$(expr $VALUE + 1);
    groupadd -g $VALUE $i;
    adduser -u $VALUE -g $VALUE $i;
done;
groupadd -g 10006 hadoop;
groupmems -g hadoop -a yarn;
groupmems -g hadoop -a mapred;
groupmems -g hadoop -a hdfs;
usermod -d /var/lib/hive -s /sbin/nologin hive;
usermod -d /var/run/hbase -s /sbin/nologin hbase;
usermod -d /var/lib/hadoop-yarn -s /sbin/nologin yarn;
usermod -d /var/lib/hadoop-mapreduce -s /sbin/nologin mapred;
usermod -d /var/lib/hadoop-hdfs -s /bin/bash hdfs
```

*other names and brands may be claimed by others

Preparation

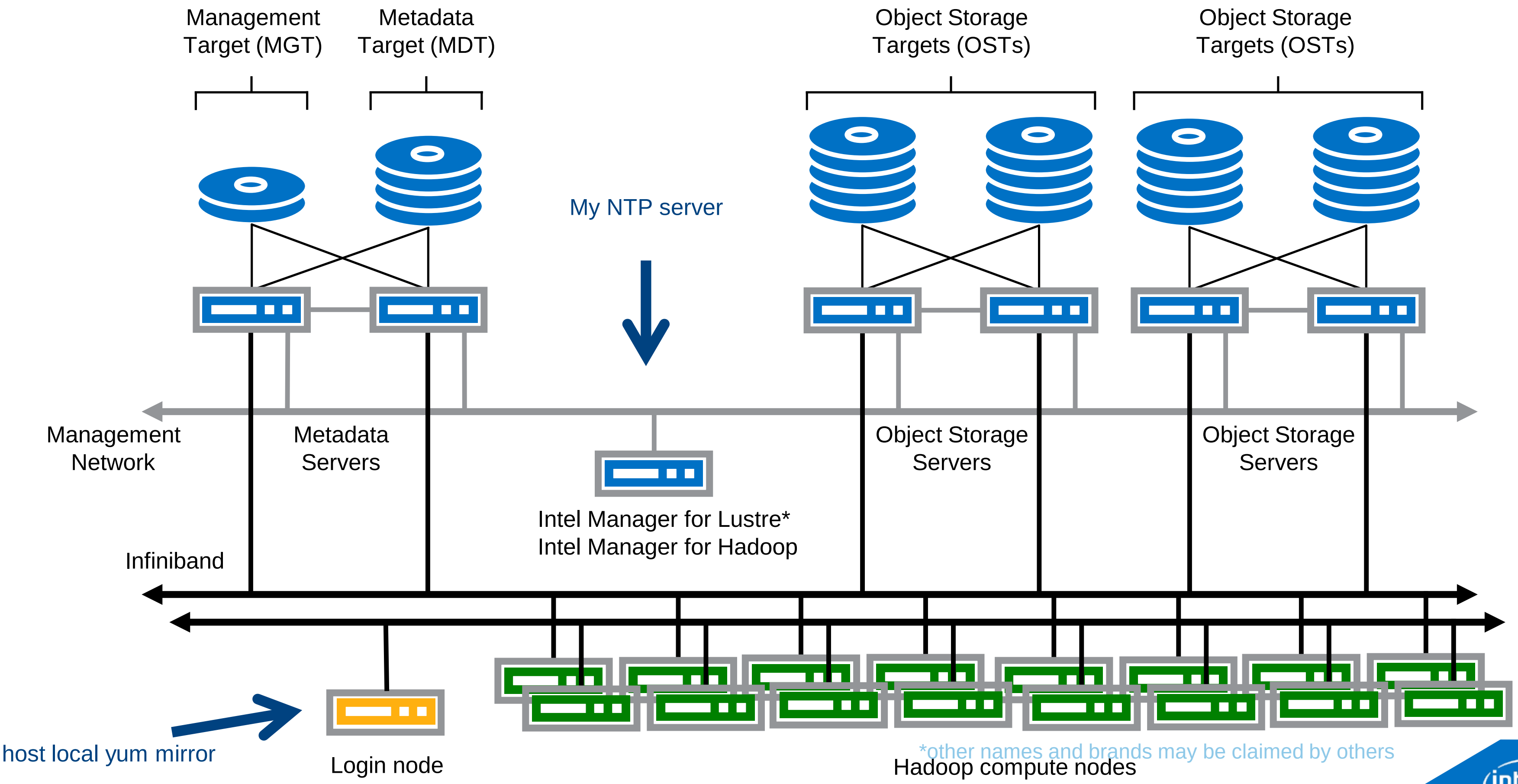
- Setup a reasonable size test system. My setup has
  - 2x MDS and 4x OSS with shared storage
  - 1x MDT SSD based, 1x MGT, 16x OST
  - 16x Hadoop nodes.



*other names and brands may be claimed by others

Preparation

- Consistent Clock
  - Setup a local NTP server
- Local yum repositories if no good connection to public network



*other names and brands may be claimed by others

Setup and Mount Lustre*

- On all of Hadoop nodes, mount the same Lustre file system before installing any Hadoop software.

```
# mkdir /mnt/lustrefs
# mount -t lustre 10.99.0.21@tcp1:/lustrefs /mnt/lustrefs
#df -h
Filesystem             Size  Used Avail Use% Mounted on
/dev/mapper/myvg-rootvol
                       219G  3.8G  204G   2% /
tmpfs                   63G     0   63G   0% /dev/shm
/dev/sda1              194M   35M  149M  19% /boot
10.99.0.21@tcp1:/lustrefs
                       175T   65G  166T   1% /mnt/lustrefs
```

I did a quick “dd” to make sure that I can indeed write data to the Lustre file system. Always good to be cautious.

Install Hadoop

- Make sure that the yum repositories are configured properly.

- I’d remove any pre-install JRE environment. Avoid the conflicts later on.

- Run the install script.

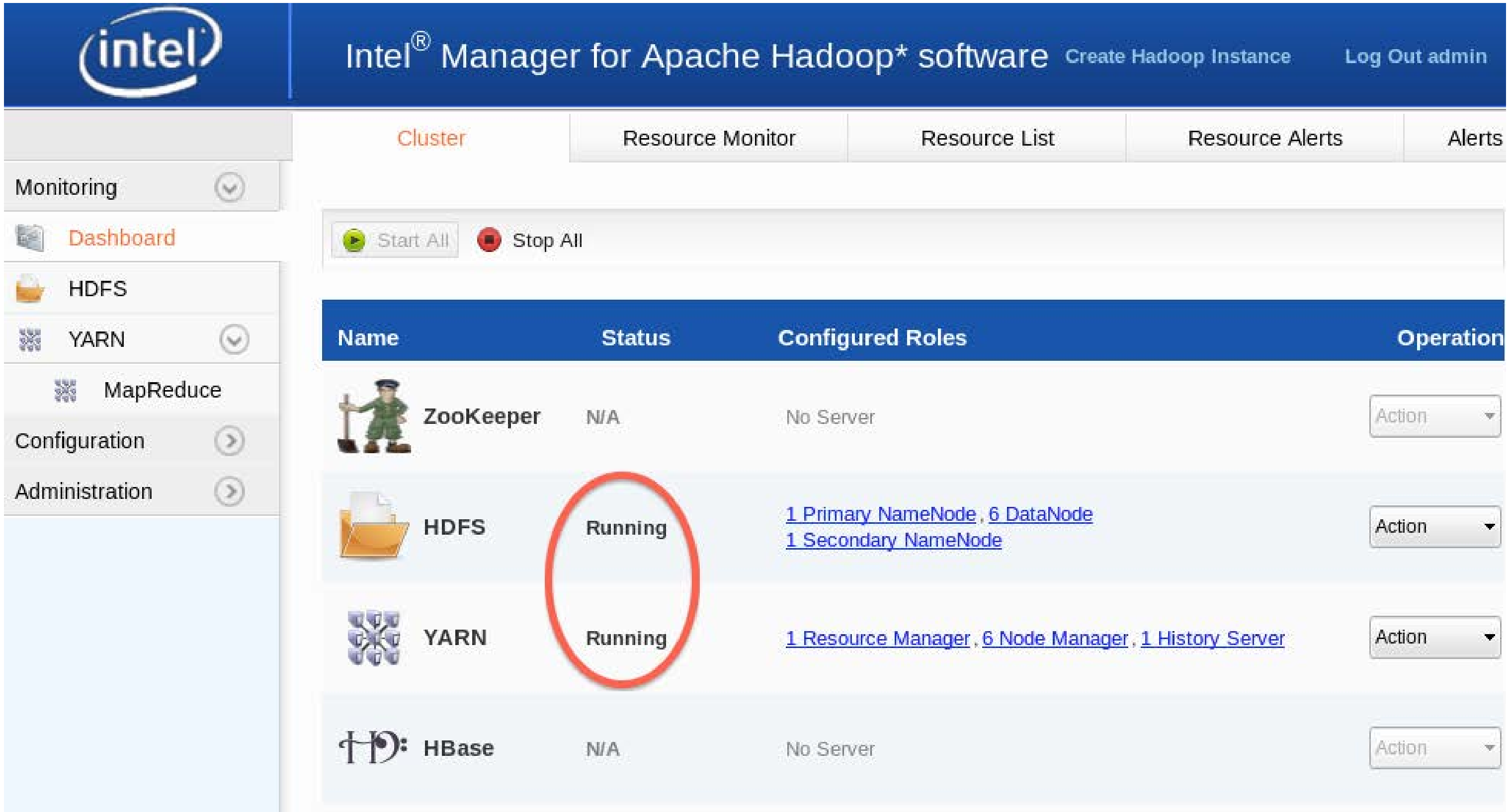
Install Hadoop

- Make sure that the yum repositories are configured properly.
- I'd remove any pre-install JRE environment. Avoid the conflicts later on.
- Run the install script.
- I did not need to install Lustre* adapter separately as the adapter is shipped along with Intel Distribution for Hadoop.

*other names and brands may be claimed by others

Setup Hadoop

- Configure a Hadoop cluster with the conventional HDFS firstly.
  - Not really a necessary step. I just like to build things step by step. If I can run Map Reduce jobs with HDFS, I know my Hadoop part was setup correctly.



Name	Status	Configured Roles	Operation
ZooKeeper	N/A	No Server	Action
HDFS	Running	1 Primary NameNode, 6 DataNode 1 Secondary NameNode	Action
YARN	Running	1 Resource Manager, 6 Node Manager, 1 History Server	Action
HBase	N/A	No Server	Action

Setup Hadoop

- Run a sample Map Reduce job

```
# yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 4 1000

Number of Maps  = 4
Samples per Map = 1000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3

...

Job Finished in 13.52 seconds
Estimated value of Pi is 3.14000000000000000000
```

Direct Hadoop IOs to Lustre*

- On all of client nodes, edit the /etc/sudoers by using visudo

```
#
# Disable "ssh hostname sudo <cmd>", because it will show the password in
#         You have to run "ssh -t hostname sudo <cmd>".
#
# Defaults    requiretty
Defaults:%hadoop !requiretty

...

## Same thing without a password
# %wheel        ALL=(ALL)       NOPASSWD: ALL
%hadoop         ALL=(ALL)       NOPASSWD: ALL
```

Please note that the location of these configuration syntax does matter. When editing the sudoers list, find the related words, e.g. "Defaults", "wheel" and add a new line right below and comment out the "Default requiretty" line.

*other names and brands may be claimed by others

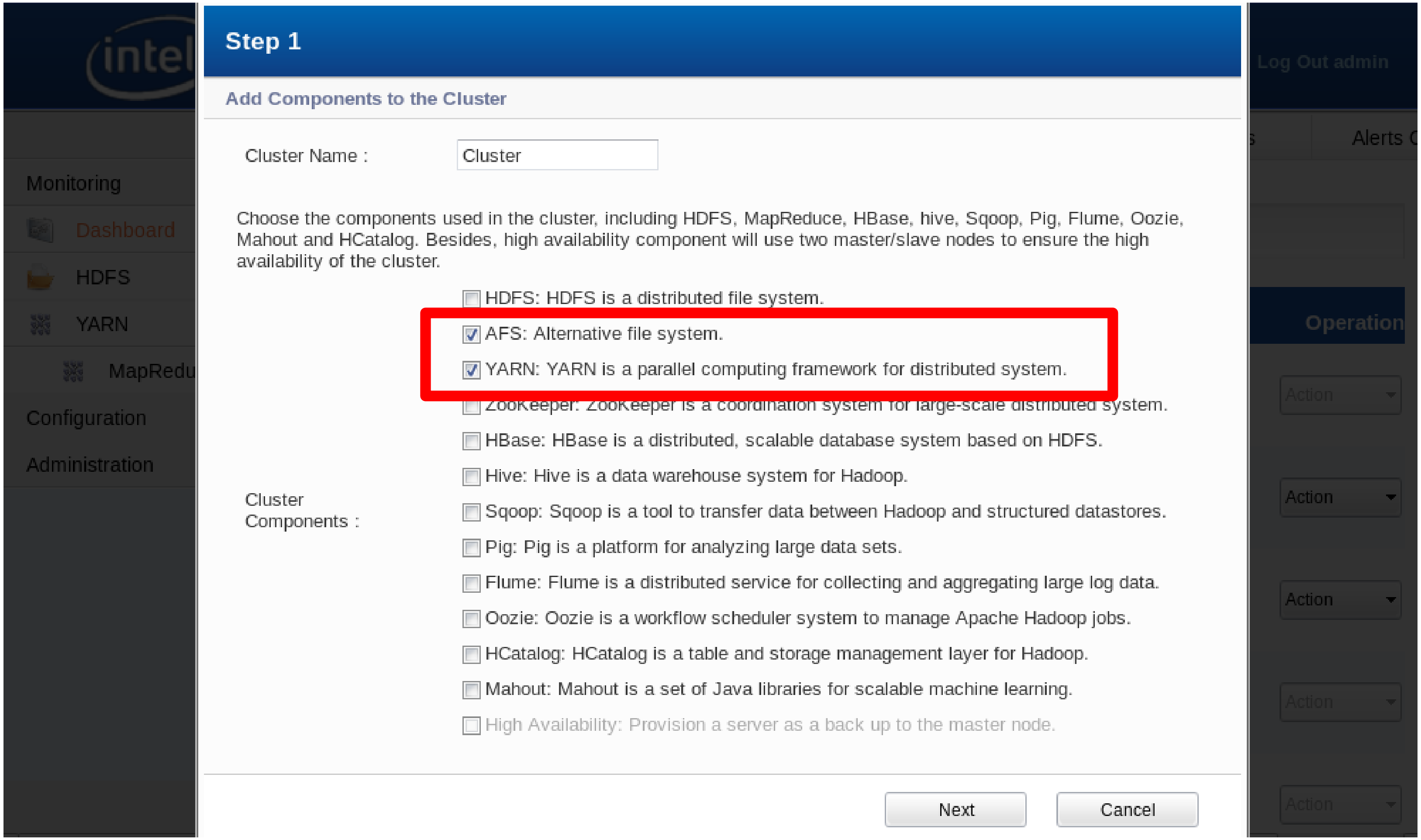
Direct Hadoop IOs to Lustre*

- Stop both HDFS and Yarn services
- Make sure the Lustre adapter modules are loaded in Hadoop

```
On the Intel Manager for Hadoop
# echo "export USE_LUSTRE=true" \
>> /usr/lib/deploy/puppet/modules/hadoop/templates/hadoop

# mkdir /mnt/lustre/hadoop
# chmod 777 /mnt/lustre/hadoop
```

Direct Hadoop IOs to Lustre*

- Create a new Hadoop instance with AFS (Alternative File System) and YARN



Step 1

Add Components to the Cluster

Cluster Name : Cluster

Choose the components used in the cluster, including HDFS, MapReduce, HBase, hive, Sqoop, Pig, Flume, Oozie, Mahout and HCatalog. Besides, high availability component will use two master/slave nodes to ensure the high availability of the cluster.

Cluster Components :

- □ HDFS: HDFS is a distributed file system.
- ☑ AFS: Alternative file system.
- ☑ YARN: YARN is a parallel computing framework for distributed system.
- □ ZooKeeper: ZooKeeper is a coordination system for large-scale distributed system.
- □ HBase: HBase is a distributed, scalable database system based on HDFS.
- □ Hive: Hive is a data warehouse system for Hadoop.
- □ Sqoop: Sqoop is a tool to transfer data between Hadoop and structured datastores.
- □ Pig: Pig is a platform for analyzing large data sets.
- □ Flume: Flume is a distributed service for collecting and aggregating large log data.
- □ Oozie: Oozie is a workflow scheduler system to manage Apache Hadoop jobs.
- □ HCatalog: HCatalog is a table and storage management layer for Hadoop.
- □ Mahout: Mahout is a set of Java libraries for scalable machine learning.
- □ High Availability: Provision a server as a back up to the master node.

*other names and brands may be claimed by others

Direct Hadoop IOs to Lustre*

- Add parameters for AFS
  - Edit the "`fs.defaultFS`" property to “`lustre:///`”

Property	Value	Description
fs.root.dir	/mnt/lustre/hadoop	Root directory on Lustre for Hadoop operations.
hadoop.tmp.dir	${fs.root.dir}/tmp/${user.name}	A base for other temporary directories
fs.lustre.impl	org.apache.hadoop.fs.LustreFileSystem	

Please make sure that these configuration changes are saved and also replicated to all of Hadoop nodes. IDH Manager provides the graphical wizard for these editing and replication.

Direct Hadoop IOs to Lustre*

- Edit the Map Reduce property

Property	Value
mapreduce.job.map.output.collector.class	org.apache.hadoop.mapred.SharedFsPlugins $MapOutputBuffer
mapreduce.job.reduce.shuffle.consumer.plugin.class	org.apache.hadoop.mapred.SharedFsPlugins $Shuffle

Please make sure that these configuration changes are saved and also replicated to all of Hadoop nodes. IDH Manager provides the graphical wizard for these editing and replication.

*other names and brands may be claimed by others

Direct Hadoop IOs to Lustre*

- Start the YARN service.
  - No HDFS necessary. AFS uses Lustre in this case.

- Common errors
  - No consistent UID and GID
  - Permission errors to Lustre file system
  - The Lustre Hadoop specific parameters changes were not replicated to all of Hadoop nodes. The IDH manager has a button for replication. Navigate to "`Configuration`" → "`Nodes`" and click the "`Provisioning Service Properties`" button

*other names and brands may be claimed by others

Check the configuration and Run a sample test

- Check if the Lustre* file system is recognized by Hadoop.

```
# hadoop fs -df -h
Filesystem      Size     Used  Available  Use%
lustre:///   174.5 T  64.5 G    174.5 T    0%
```

- Let's do another fun exercise - see what word Jane Austen used the most frequently in the "Pride and Prejudice"

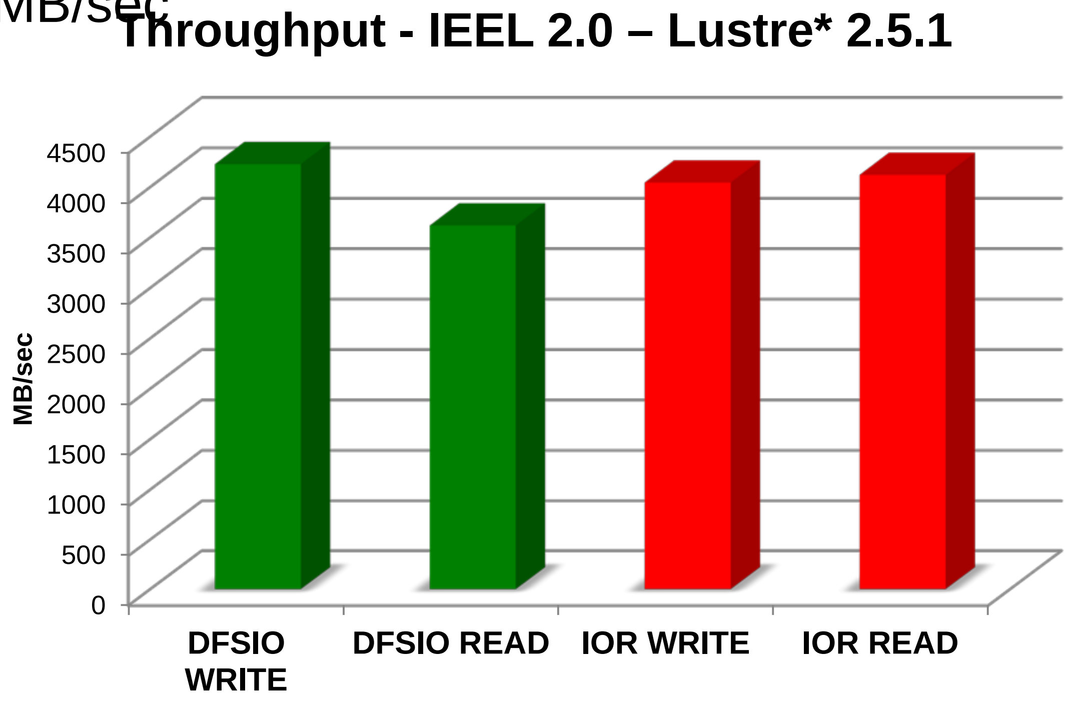
```
# wget http://www.gutenberg.org/cache/epub/1342/pg1342.txt

# yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar \
 wordcount /mnt/lustrefs/hadoop/wordcount/pg1342.txt \
/mnt/lustrefs/hadoop/wordcount/result

# cat /mnt/lustrefs/hadoop/wordcount/result/part-r-00000 | sort -k2 -n
...
of      3660
to      4121
the     4205
```

*other names and brands may be claimed by others

Benchmark

- IOR Baseline
  - WRITE: 4043 MB/sec
  - READ: 4119 MB/sec

- DFSIO from Hibench - 120 files each 10GB
  - Write throughput: 4225 MB/sec
  - Read throughput: 3617 MB/sec



*other names and brands may be claimed by others

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance

Future Work
And Call for collaboration

- A big scale testing. Looking for the 100+ node compute clusters with 50GB/s+ Lustre* file system

- Real life applications
  - Process many many files
  - Process a large amount of data
  - Need the result as quickly as possible
  - Have some nice eye-candies. Let us put up a show at SC'14.

Acknowledgements
Server and storage infrastructure, networking and hardware support was supplied by Intel's High Performance computing Labs in Swindon (UK). Special thanks to Jamie Wilcox, Adam Roe and Stephen Anderson.

*other names and brands may be claimed by others

intel®