Hadoop* on Lustre*

Liu Ying (emoly.liu@intel.com)
High Performance Data Division, Intel ® Corporation

Agenda

- Overview
- HAM and HAL
- Hadoop* Ecosystem with Lustre*
- Benchmark results
- Conclusion and future work

Agenda

- ***Overview***
- HAM and HAL
- Hadoop* Ecosystem with Lustre*
- Benchmark results
- Conclusion and future work

*Other names and brands may be claimed as the property of others.

Overview

Scientific Computing
- performance
- scalability

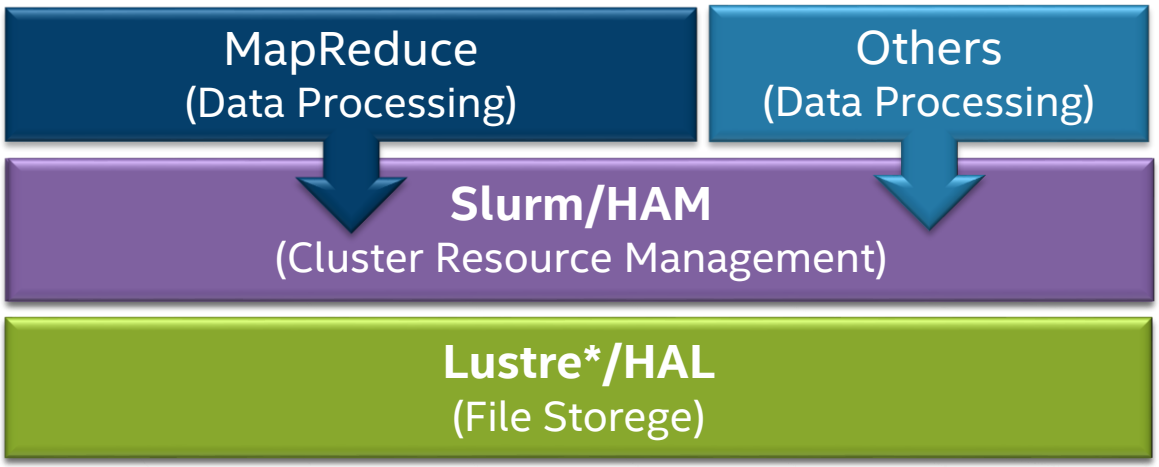
Commercial Computing
- application
- data processing

*Other names and brands may be claimed as the property of others.

Agenda

- Overview
- ***HAM and HAL***
  - HPC Adapter for Mapreduce/Yarn
  - Hadoop* Adaptor for Lustre*
- Hadoop* Ecosystem with Lustre*
- Benchmark results
- Conclusion and future work

HAM and HAL



HPC Adapter for Mapreduce/Yarn

- Replace YARN Job scheduler with Slurm
- Plugin for Apache Hadoop 2.3 and CDH5
- No changes to applications needed
- Allow Hadoop environments to migrate to a more sophisticated scheduler

Hadoop* Adapter with Lustre*

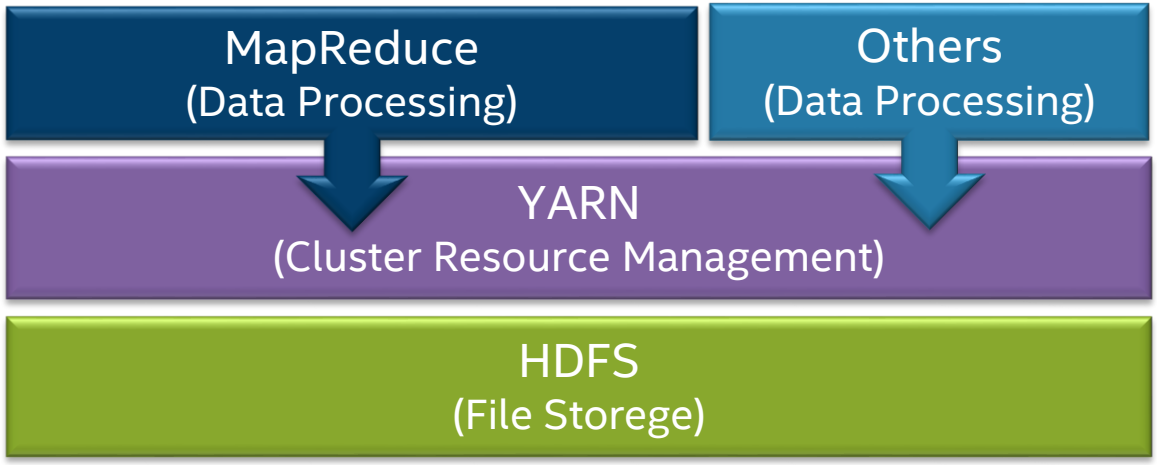
- Replace HDFS with Lustre
- Plugin for Apache Hadoop 2.3 and CDH5
- No changes to Lustre needed
- Allow Hadoop environments to migrate to a general purpose file system

HAM(HPC Adapter for Mapreduce)

- Why Slurm (Simple Linux Utility for Resource Management)
  - Widely used open source RM
  - Provides reference implementation for other RMs to model
- Objectives
  - No modifications to Hadoop* or its APIs
  - Enable all Hadoop applications to execute without modification
  - Maintain license separation
  - Fully and transparently share HPC resources
  - Improve performance

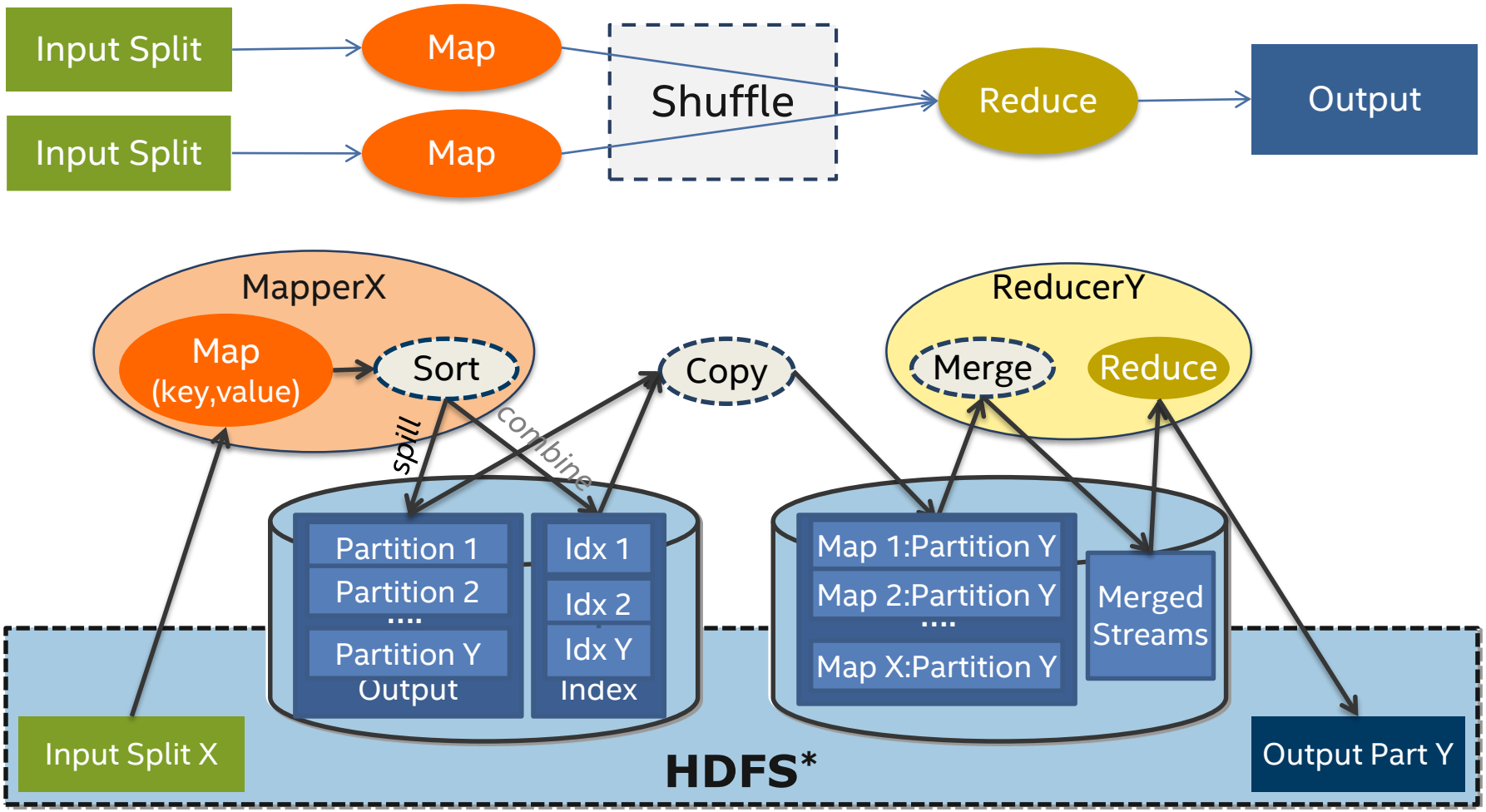
*Other names and brands may be claimed as the property of others.

HAL(Hadoop* Adaptor for Lustre*)

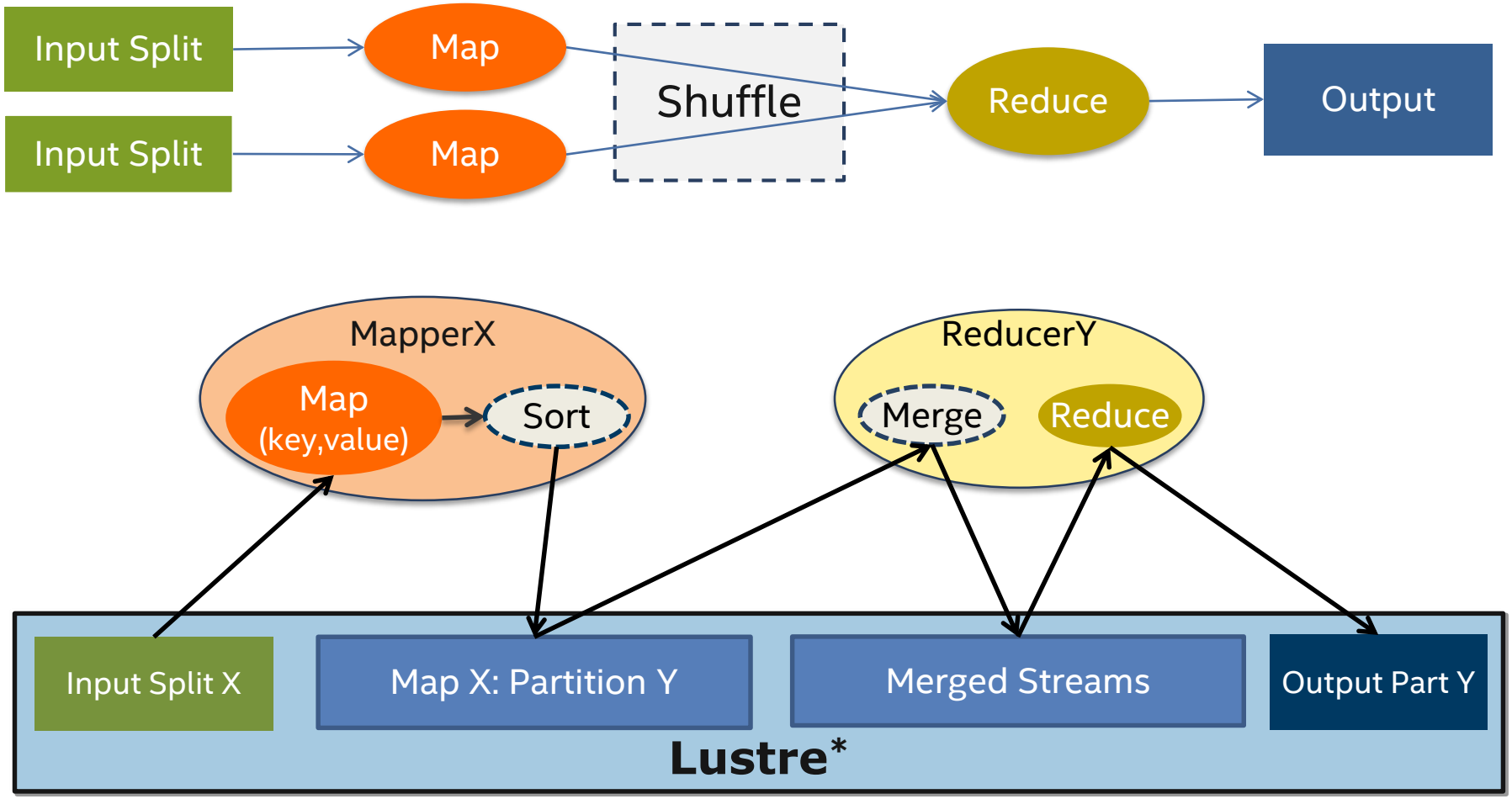


*Other names and brands may be claimed as the property of others.

The Anatomy of MapReduce



Optimizing for Lustre*: Eliminating Shuffle



*Other names and brands may be claimed as the property of others.

HAL

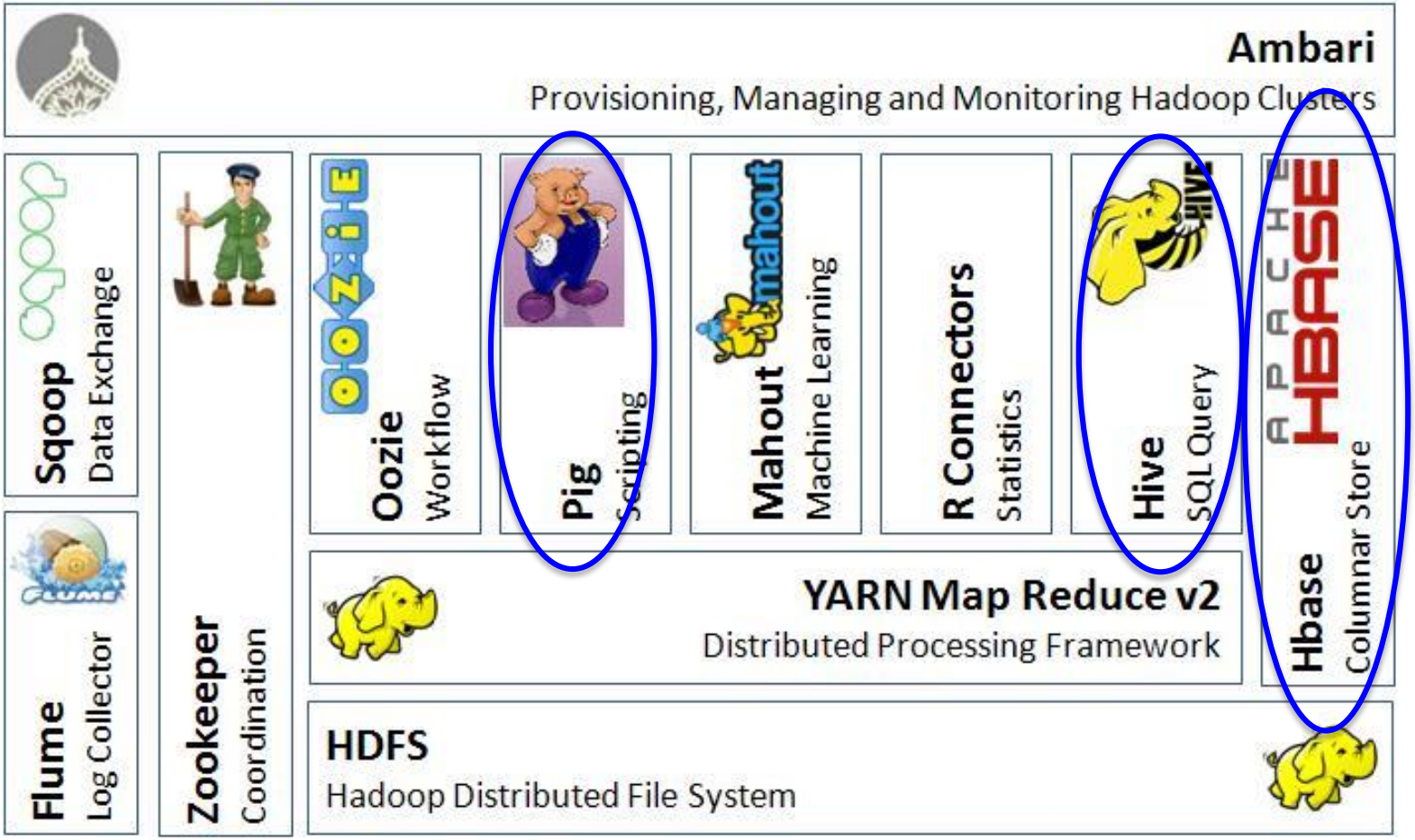
- Based on the new Hadoop* architecture
- Packaged as a single Java* library (JAR)
  - Classes for accessing data on Lustre* in a Hadoop* compliant manner. Users can configure Lustre Striping.
  - Classes for "Null Shuffle", i.e., shuffle with zero-copy
- Easily deployable with minimal changes in Hadoop* configuration
- No change in the way jobs are submitted
- Part of IEEL

Agenda

- Overview
- HAM and HAL
- ***Hadoop* Ecosystem with Lustre****
- Benchmark results
- Conclusion and future work

*Other names and brands may be claimed as the property of others.

Apache Hadoop Ecosystem

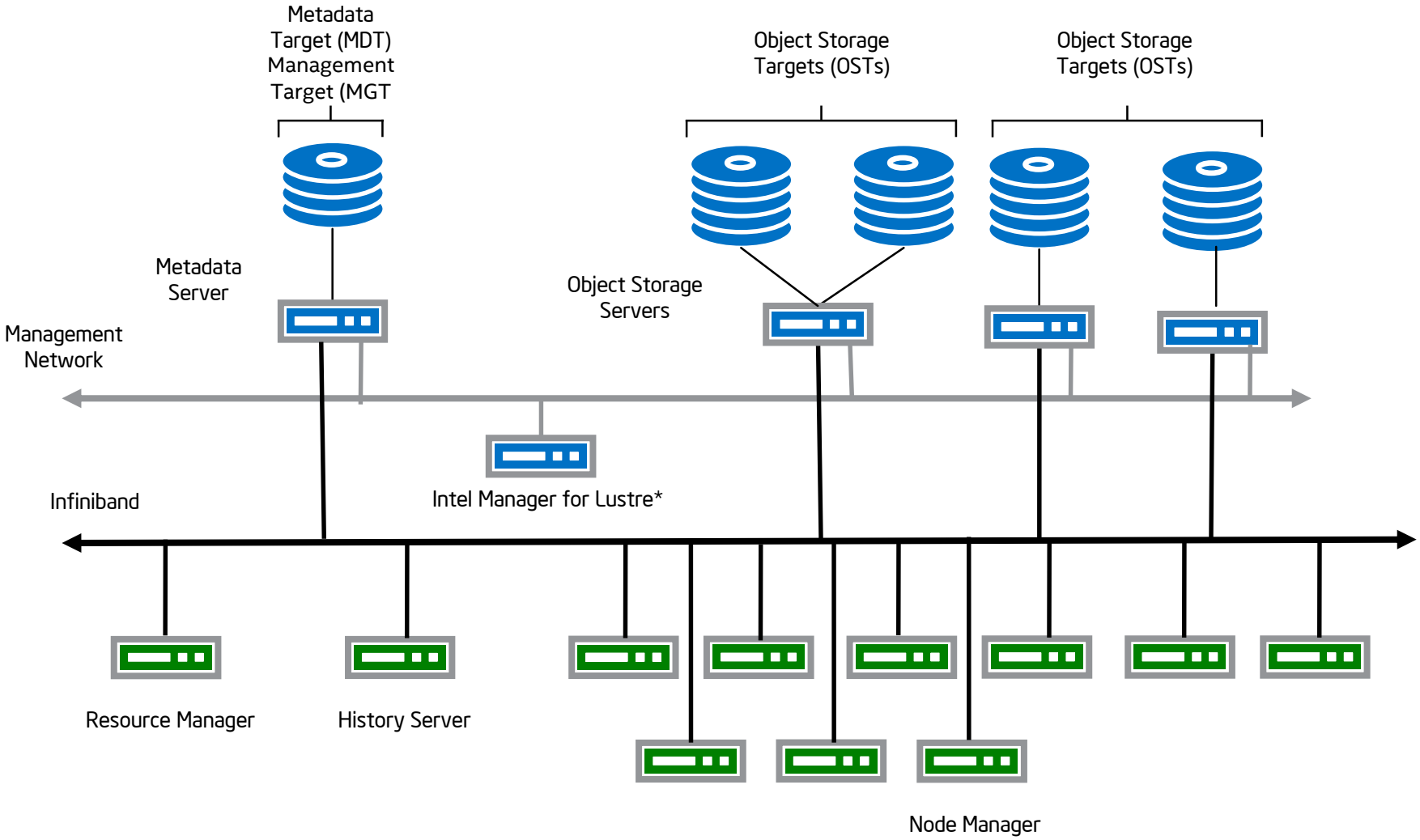


Agenda

- Overview
- HAM and HAL
- ***Hadoop* Ecosystem with Lustre****
  - Setup Hadoop*/HBase/Hive cluster with HAL
- Benchmark results
- Conclusion and future work

*Other names and brands may be claimed as the property of others.

Example: CSCS Lab



*Other names and brands may be claimed as the property of others.

Steps to install Hadoop* on Lustre*

- Prerequisite
  - Lustre* cluster, hadoop user
- Install HAL on all Hadoop* nodes, e.g.
  - *# cp ./ieel-2.x/hadoop/hadoop-lustre-plugin-2.3.0.jar $HADOOP_HOME/share/hadoop/common/lib*
- Prepare Lustre* directory for Hadoop*, e.g.
  - *# chmod 0777 /mnt/lustre/hadoop*
  - *# setfacl -R -m group:hadoop:rwx /mnt/lsutre/hadoop*
  - *# setfacl -R -d -m group:hadoop:rwx /mnt/lustre/hadoop*
- Configure Hadoop* for Lustre*
- Start YARN RM, NM and JobHistory servers
- Run MR job

Hadoop* configuration for Lustre*

- core-site.xml

Property name	Value	Description
fs.defaultFS	lustre:///	Configure Hadoop to use Lustre as the default file system.
fs.root.dir	/mnt/lustre/hadoop	Hadoop root directory on Lustre mount point.
fs.lustre.impl	org.apache.hadoop.fs.LustreFileSystem	Configure Hadoop to use Lustre Filesystem
fs.AbstractFileSystem.lustre.impl	org.apache.hadoop.fs.LustreFileSystem$LustreFs	Configure Hadoop to use Lustre class

Hadoop* configuration for Lustre*(cont.)

- mapred-site.xml

Property name	Value	Description
mapreduce.map.speculative	*false*	Turn off map tasks speculative execution (this is incompatible with Lustre currently)
mapreduce.reduce.speculative	*false*	Turn off reduce tasks speculative execution (this is incompatible with Lustre currently)
mapreduce.job.map.output.collector.class	org.apache.hadoop.mapred.SharedFsPlugins$MapOutputBuffer	Defines the MapOutputCollector implementation to use, specifically for Lustre, for shuffle phase
mapreduce.job.reduce.shuffle.consumer.plugin.class	org.apache.hadoop.mapred.SharedFsPlugins$Shuffle	Name of the class whose instance will be used to send shuffle requests by reduce tasks of this job

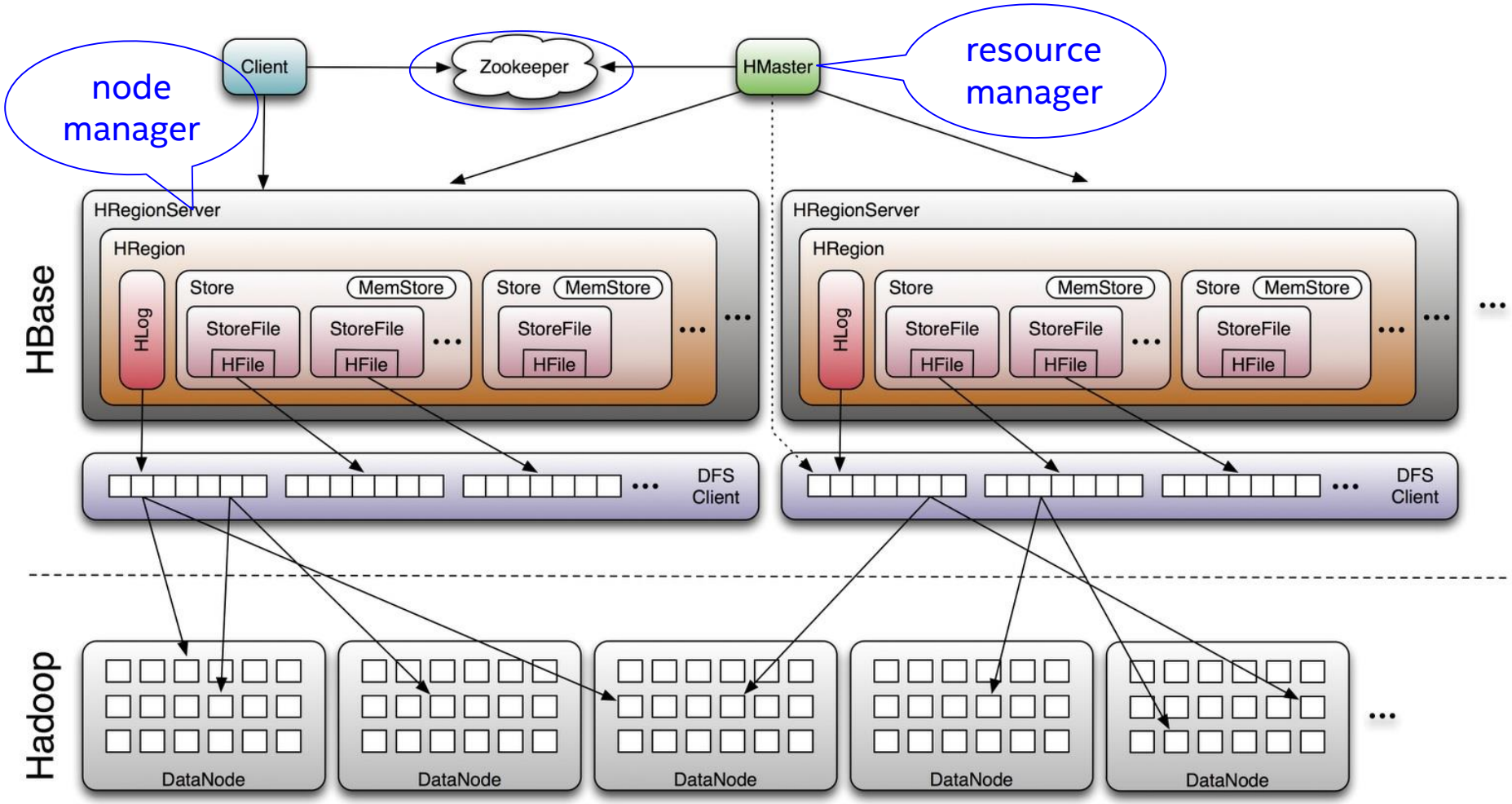
Start and run Hadoop* on Lustre*

- Start Hadoop*
  - start difference services in order on different nodes
    - *yarn-daemon.sh start resourcemanager*
    - *yarn-daemon.sh start nodemanager*
    - *mr-jobhistory-daemon.sh start historyserver*
- Run Hadoop*

#hadoop jar $HADOOP_HOME/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
Starting Job
...
Job Finished in 17.308 seconds
Estimated value of Pi is 3.14000000000000000000
```

HBase


node manager
Client
Zookeeper
HMaster
resource manager
HBase
HRegionServer
HRegion
HLog
Store
MemStore
StoreFile
HFile
DFS Client
Hadoop
DataNode

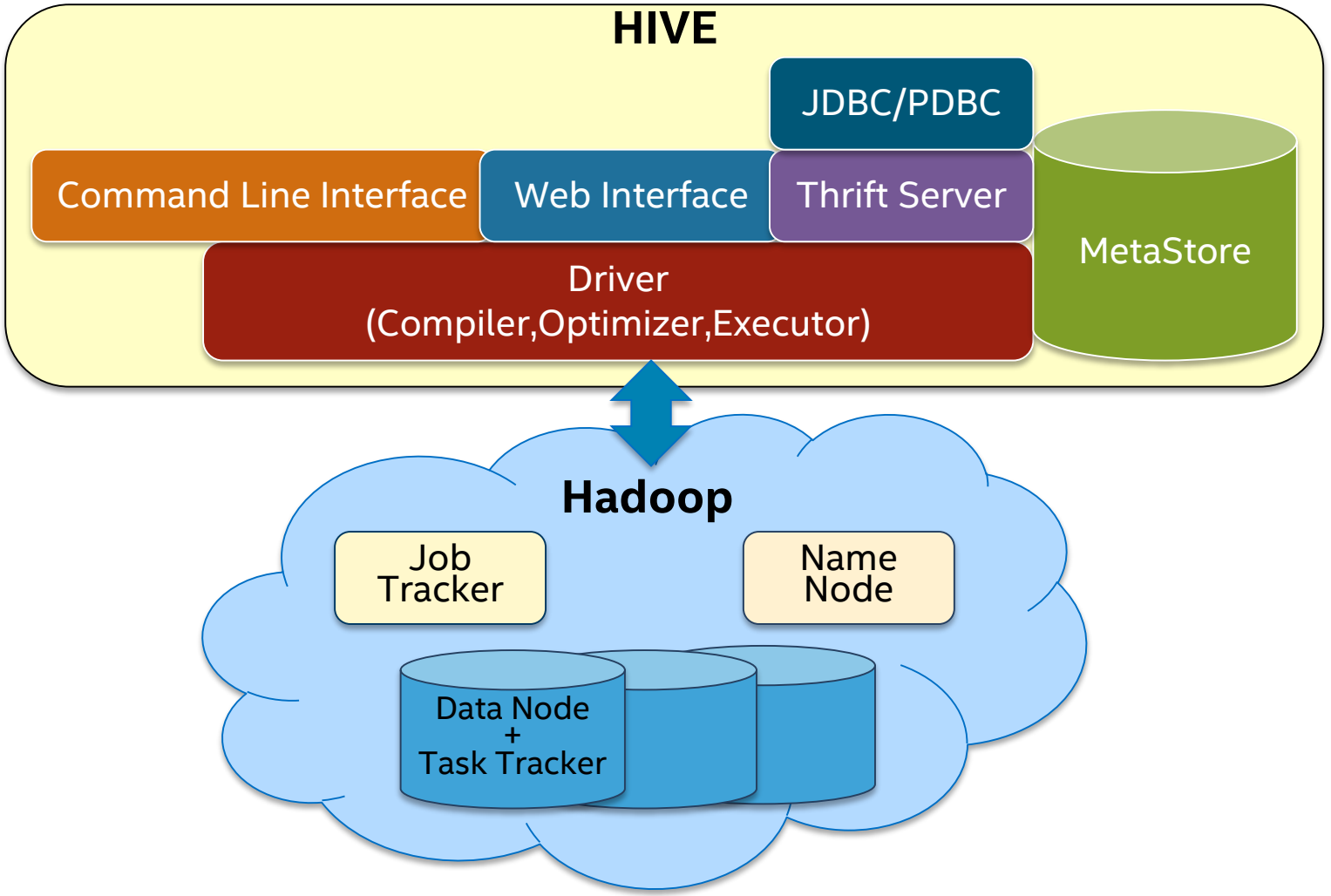
HBase configuration for Lustre*

- Include HAL to HBase classpath
- hbase-site.xml

Property name	Value	Description
hbase.rootdir	lustre:///hbase	The directory shared by region servers and into which HBase persists.
fs.defaultFS	lustre:///	Configure Hadoop to use Lustre as the default file system.
fs.lustre.impl	org.apache.hadoop.fs.LustreFileSystem	Configure Hadoop to use Lustre Filesystem
fs.AbstractFileSystem.lustre.impl	org.apache.hadoop.fs.LustreFileSystem$LustreFs	Configure Hadoop to use Lustre class
fs.root.dir	/scratch/hadoop	Hadoop root directory on Lustre mount point.

*Other names and brands may be claimed as the property of others.

HIVE


HIVE
JDBC/PDBC
Command Line Interface
Web Interface
Thrift Server
MetaStore
Driver
(Compiler,Optimizer,Executor)
Hadoop
Job
Tracker
Name
Node
Data Node
+
Task Tracker

Hive configuration for Lustre*

- hive-site.xml

Property name	Value	Description
hive.metastore.warehouse.dir	lustre:///hive/warehouse	Location of default database for the warehouse
Aux Plugin Jars (in classpath) for HBase integration: hbase-common-xxx.jar hbase-protocol-xxx.jar hbase-client-xxx.jar hbase-server-xxx.jar hbase-hadoop-compat-xxx.jar htrace-core-xxx.jar		

*Other names and brands may be claimed as the property of others.

Agenda

- Overview
- HAM and HAL
- Hadoop* Ecosystem with Lustre*
- ***Benchmark results***
- Conclusion and future work

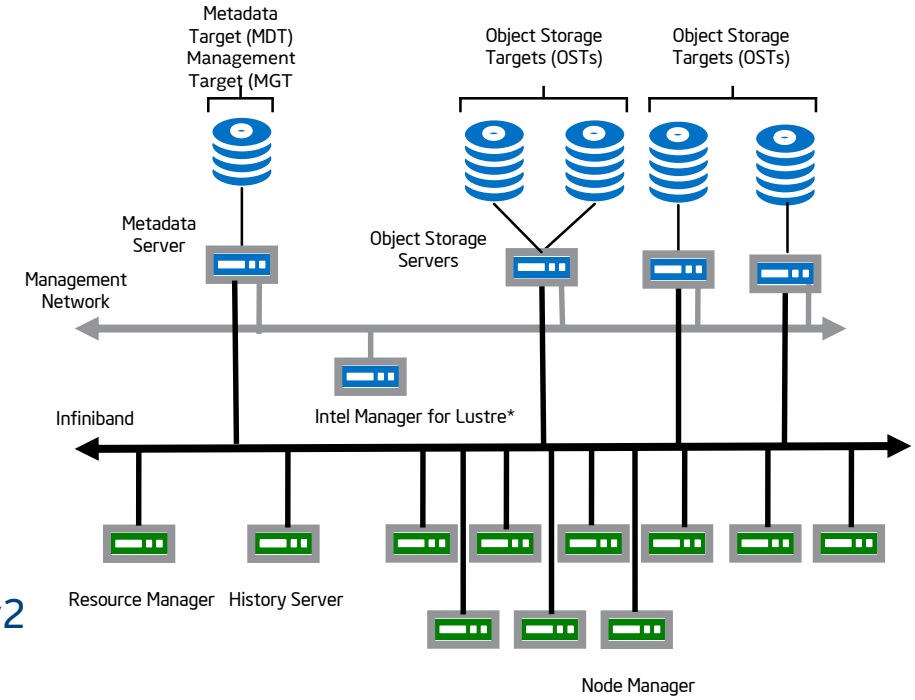
*Other names and brands may be claimed as the property of others.

Experiments

- Swiss National Supercomputing Centre(CSCS)
  - Read/write performance evaluation for Hadoop on Lustre*
  - Benchmark tools
    - HPC: iozone
    - Hadoop*: DFSIO and Terasort
- Intel BigData Lab in Swindon (UK)
  - Performance comparison of Lustre* and HDFS for MR
  - Benchmark tool: A query of Audit Trail System part of FINRA security specifications
    - Query average execution time

Experiment 1: CSCS Lab

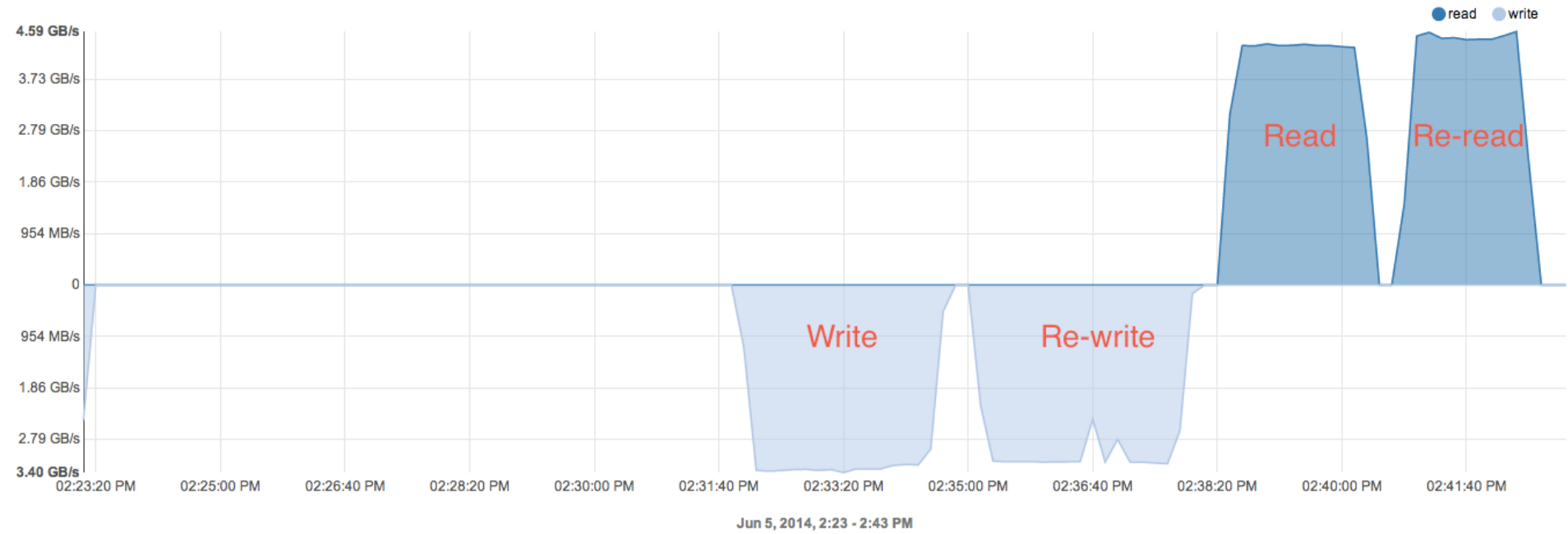
- Lustre*
  - 1x MDS
  - 3x OSS (4x OST)
- Hadoop*
  - 1x Resource Manager
  - 1x History Server
  - 9x Node Manager
    - 2x Intel(R) Xeon(R) CPU E5-2670 v2
    - 64GB RAM
    - Mellanox FDR RAMSAN-620 Texas Memory



*Other names and brands may be claimed as the property of others.

Iozone: baseline

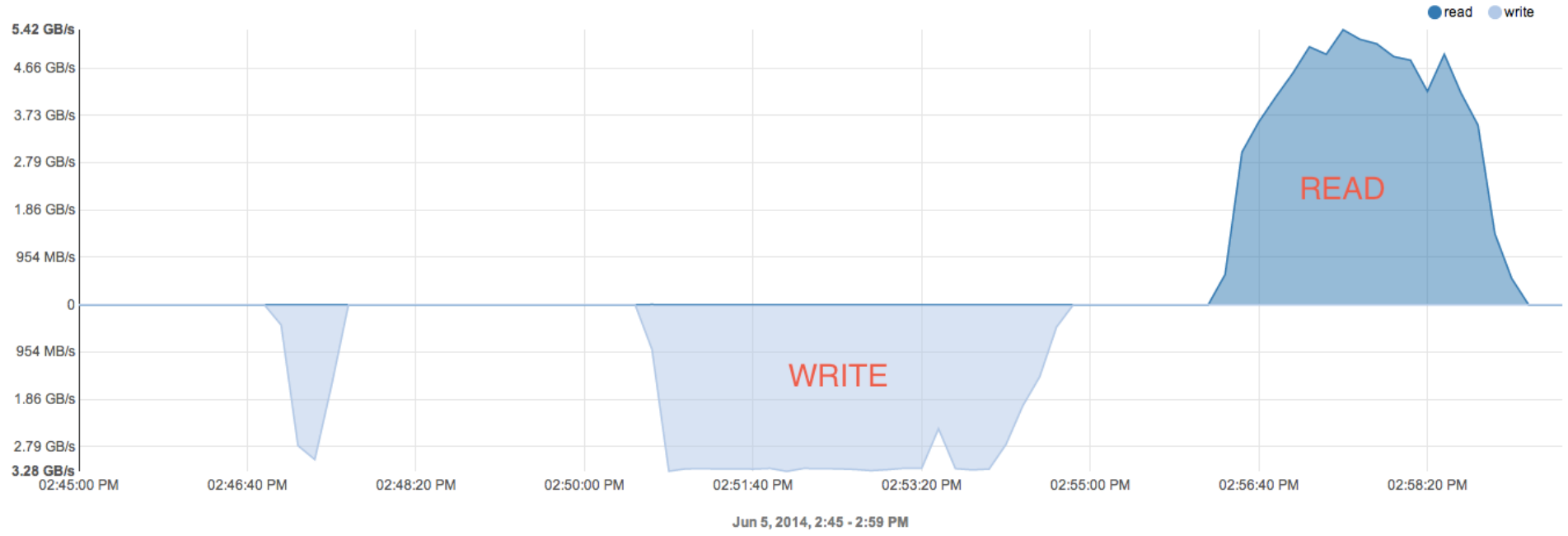
- Baseline: peak performance of *3.4GB/sec* writing and *4.59GB/sec* reading
- Our goal: achieve the same performance using Hadoop on Lustre*.



*Other names and brands may be claimed as the property of others.

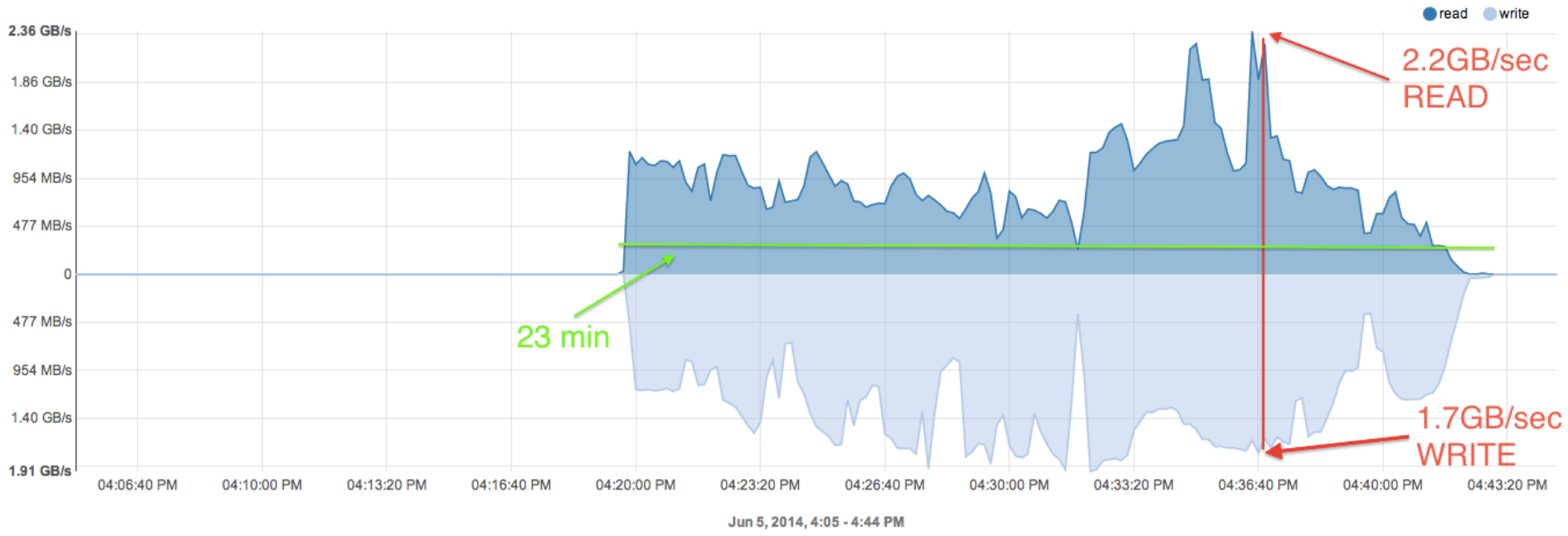
DFSIO

- 72 map tasks, 8 map tasks on each node manager, and 10GB data each map task
- Peak performance: *3.28GB/sec* writing and *5.42GB/sec* reading



Terasort

- 72 map tasks,144 reduce tasks and 500GB data size
- Peak performance: all throughput 3.9GB/sec (2.2GB/sec reading and 1.7GB/sec writing)

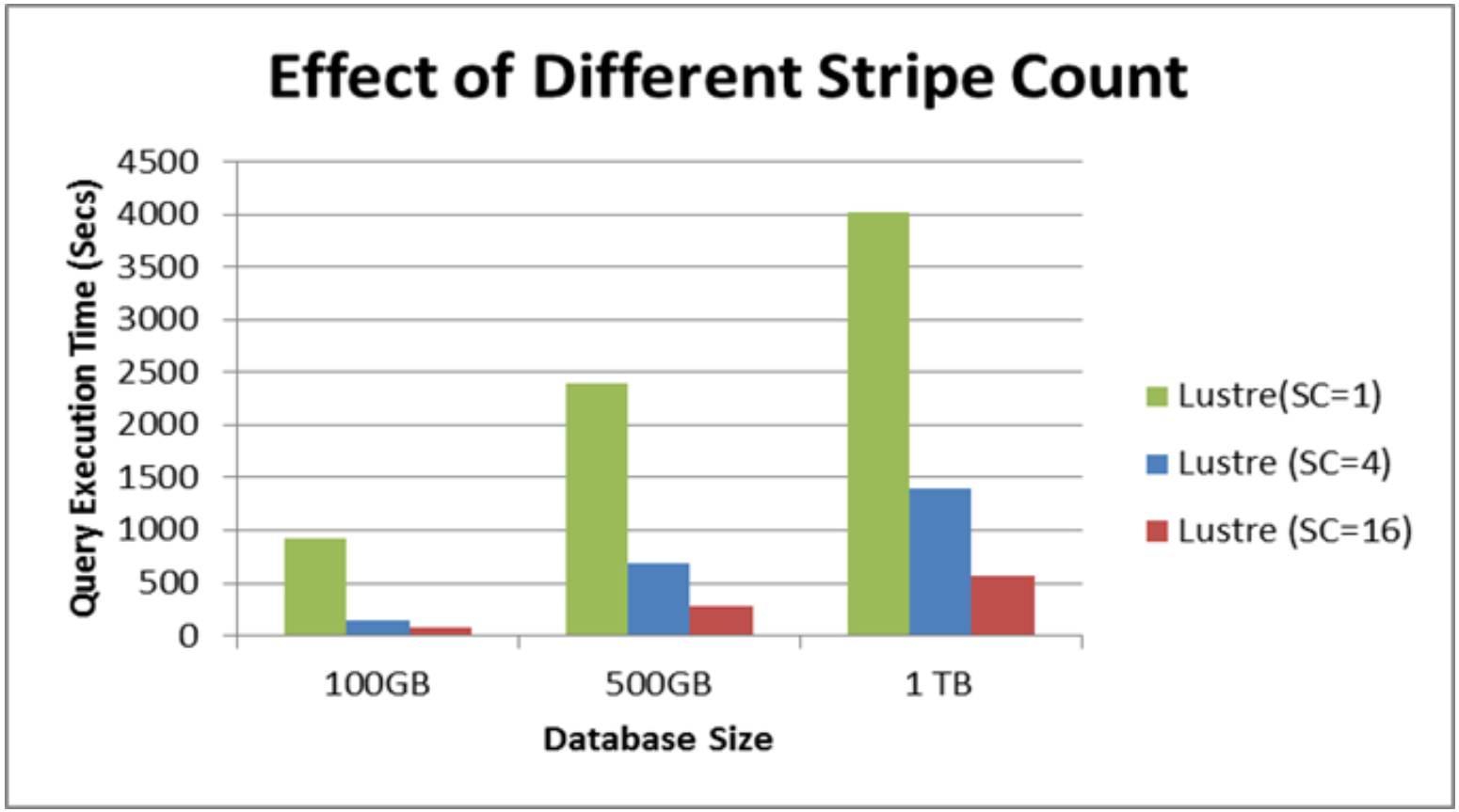


Experiment 2: Intel BigData Lab

- HDFS
  - 1x Resource Manager + 8x Node manager
    - Intel(R) Xeon(R) CPU E5-2695 v2 @ 2.40GHz, 320GB cluster RAM, 1 TB SATA 7200 RPM, 27 TB of usable cluster storage
- Lustre*
  - 1x MDS + 4x OSS + 16x OST
    - CPU- Intel(R) Xeon(R) CPU E5-2637 v2 @ 3.50GHz , Memory - 128GB DDr3 1600mhz, 1 TB SATA 7200 RPM, 165 TB of usable cluster storage
  - 1x Resource Manager + 1x History Server + 8x Node Manager
    - Intel(R) Xeon(R) CPU E5-2695 v2 @ 2.40GHz, 320GB cluster RAM, 1 TB SATA 7200 RPM
  - Stripe size = 4MB

(Redhat 6.5, CDH 5.0.2, IEEL*2.0+HAL, 10Gbps Network)

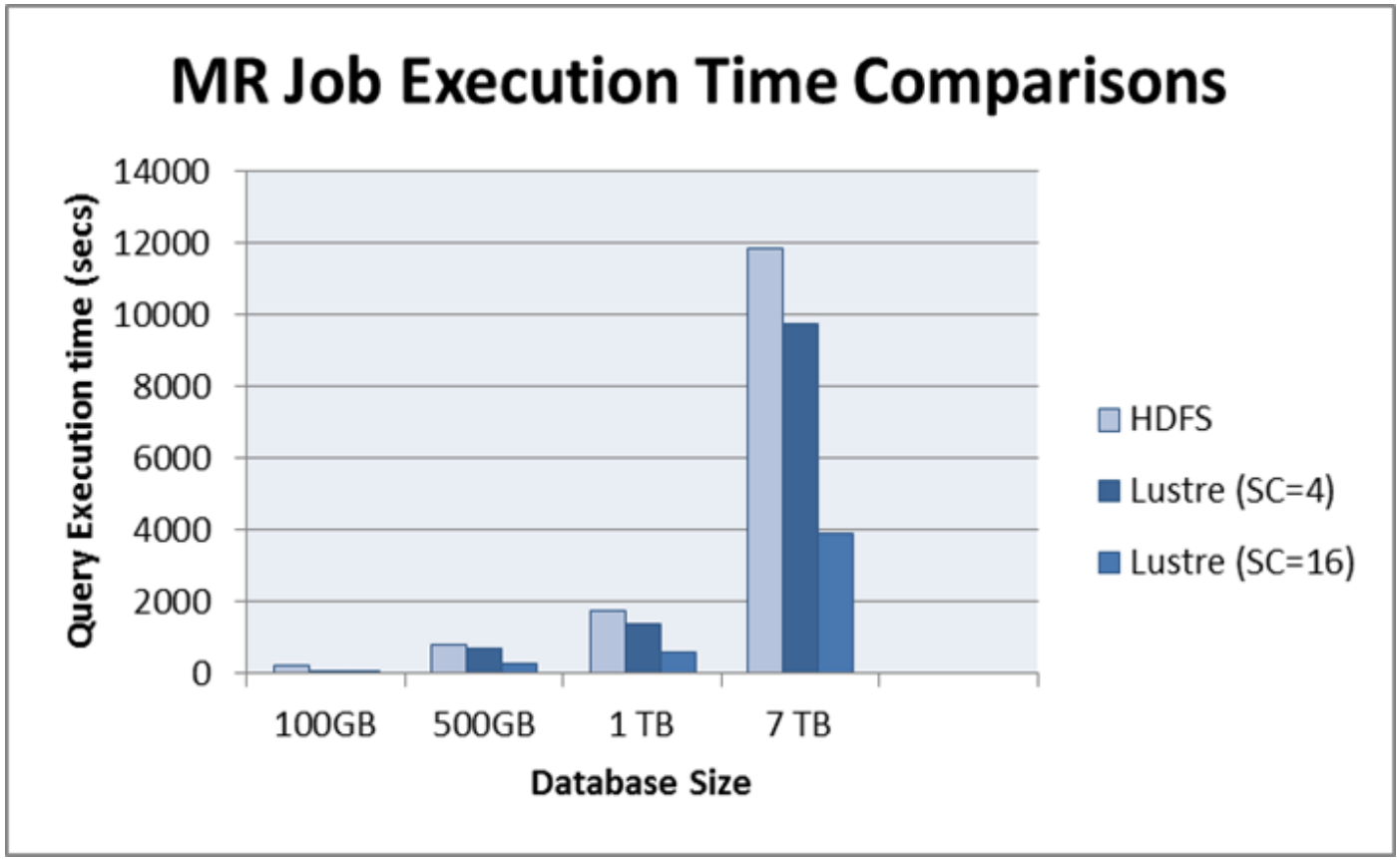
Results



Lustre* performs better on larger stripe count

*Other names and brands may be claimed as the property of others.

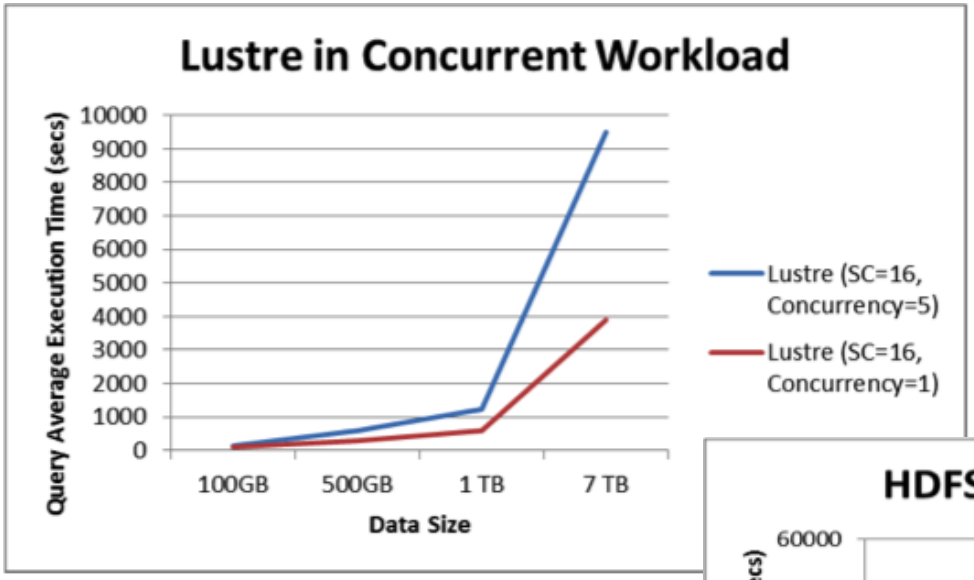
Results



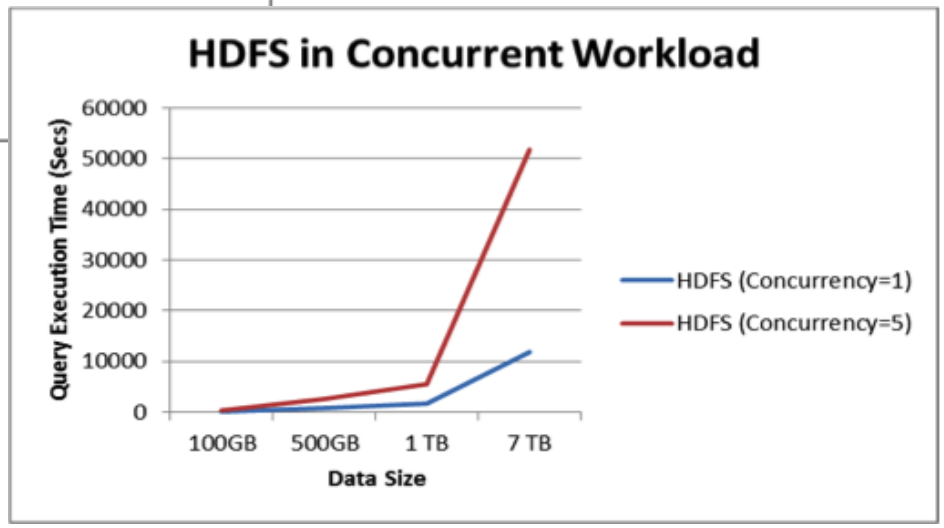
Lustre* = 3 X HDFS for optimal SC settings

*Other names and brands may be claimed as the property of others.

Results



$Saturation_{lustre}$ = 0.5



$Saturation_{HDFS}$ = 0.9

Agenda

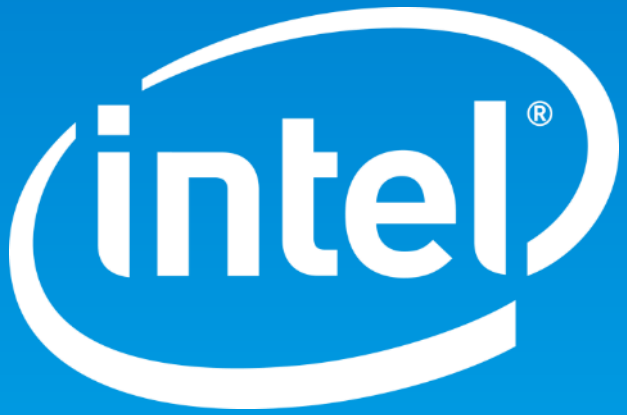
- Overview
- HAM and HAL
- Hadoop* Ecosystem with Lustre*
- Benchmark results
- ***Conclusion and future work***

*Other names and brands may be claimed as the property of others.

Conclusion and future work

- Intel is working to enable leveraging of existing HPC resources for Hadoop*.
- Hadoop* on Lustre* shows better performance than HDFS by increasing stripe count number.
- Full support for Hadoop
  - Cloudera cetification (in progress)
- Optimization and large scale performance testing
- Real life applications from different industries.

*Other names and brands may be claimed as the property of others.


intel

Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.
A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.
Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.
The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.
Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the third quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company’s expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; customer acceptance of Intel’s and competitors’ products; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel’s response to such actions; and Intel’s ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel’s results could be affected by the timing of closing of acquisitions and divestitures. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel’s ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel’s results is included in Intel’s SEC filings, including the company’s most recent reports on Form 10-Q, Form 10-K and earnings release.