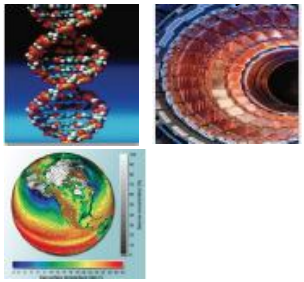
Experience with Apache Hadoop in Spider Filesystem

Seung-Hwan Lim
Computational Data Analytics Group
Computational Sciences and Engineering Division
ORNL

Knowledge-discovery lifecycle



Data Generation

Data Processing & Organization

Mining, Discovery, Predictive Modelng

Drive insights, Refine, Plan, Execute

Tasks in data processing & organization: they might run in parallel with data generation.

- Data Management
- Data Reduction/Query
- Data Visualization
- Data Sharing

From Synergistic Challenges in Data-Intensive Science and Exascale Computing, DOE ASCAC Data Subcommittee Report, March 2013.

An option for tasks in data processing & organization

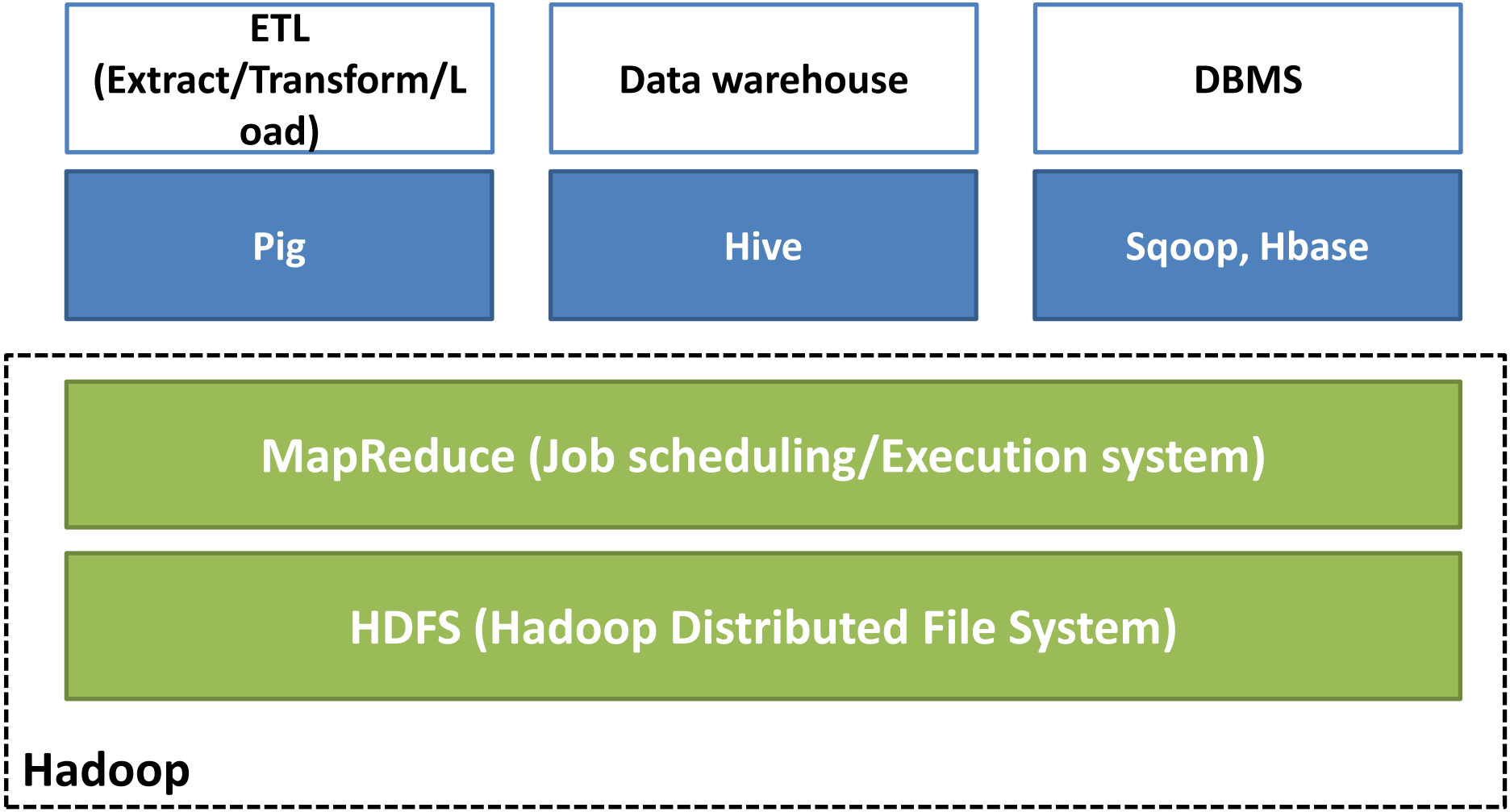
Data Management	Data Reduction/Query	Data Visualization	Data Sharing

Use tools for businesses such as database systems or statistical packages:

- They are for layman and general purpose tools.
  - you don’t need to be a computer scientist/ a software developer.
- They have evolved to process TB/PB scale data.

A major driver behind this trend is:

Apache Hadoop & it's ecosystem



Apache Hadoop

Creation.

- Dean and Ghemawat from Google announce MapReduce on Dec, 2004 at OSDI'04

Started to be spreaded across the universe.

- Cloudera distributes Hadoop, in 2009
- MapR gets a $9 million Fund , in 2009
- Facebook introduces Hive, at VLDB'09

Became a mainstream.

- Intel plans their own Hadoop distribution, in 2013

I want to use Apache Hadoop because of the free ecosystem for my data processing and organization for big-data.

But, I am not sure whether

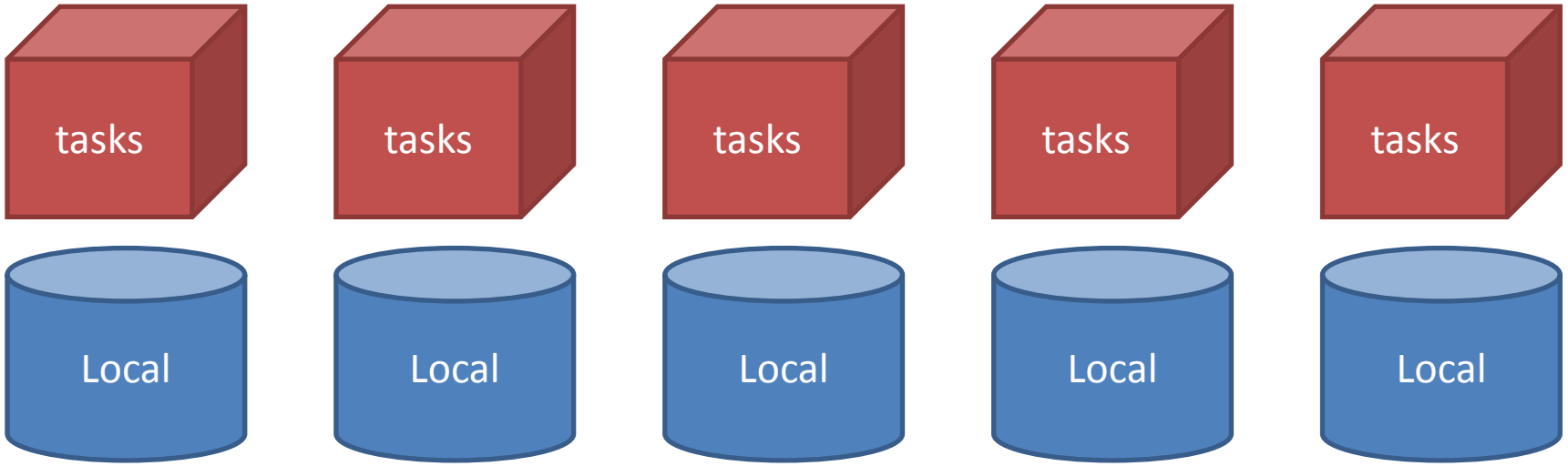
- I can cooperate with currently using programs and facilities (such as HPC applications and OLCF machines.)
- I can purchase a new, large enough, hardware dedicated to Hadoop.

In order to co-operate with existing HPC programs and infrastructure,

We can run Apache Hadoop on top of

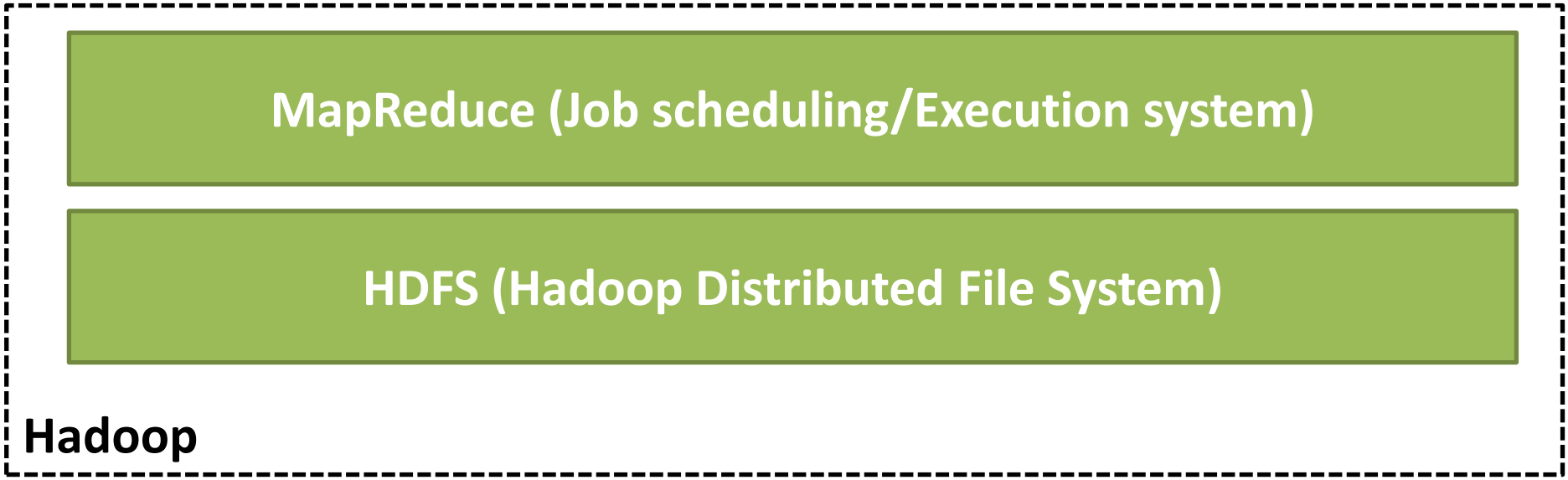
It sounds against my belief.

I believe that Hadoop runs where the data locates, that is, the local storage, instead of remote networked storage.



Also, I believe

Apache Hadoop consists of MapReduce and HDFS.



They may not be true.

People run Hadoop on networked shared storage or in virtualized cloud environment.

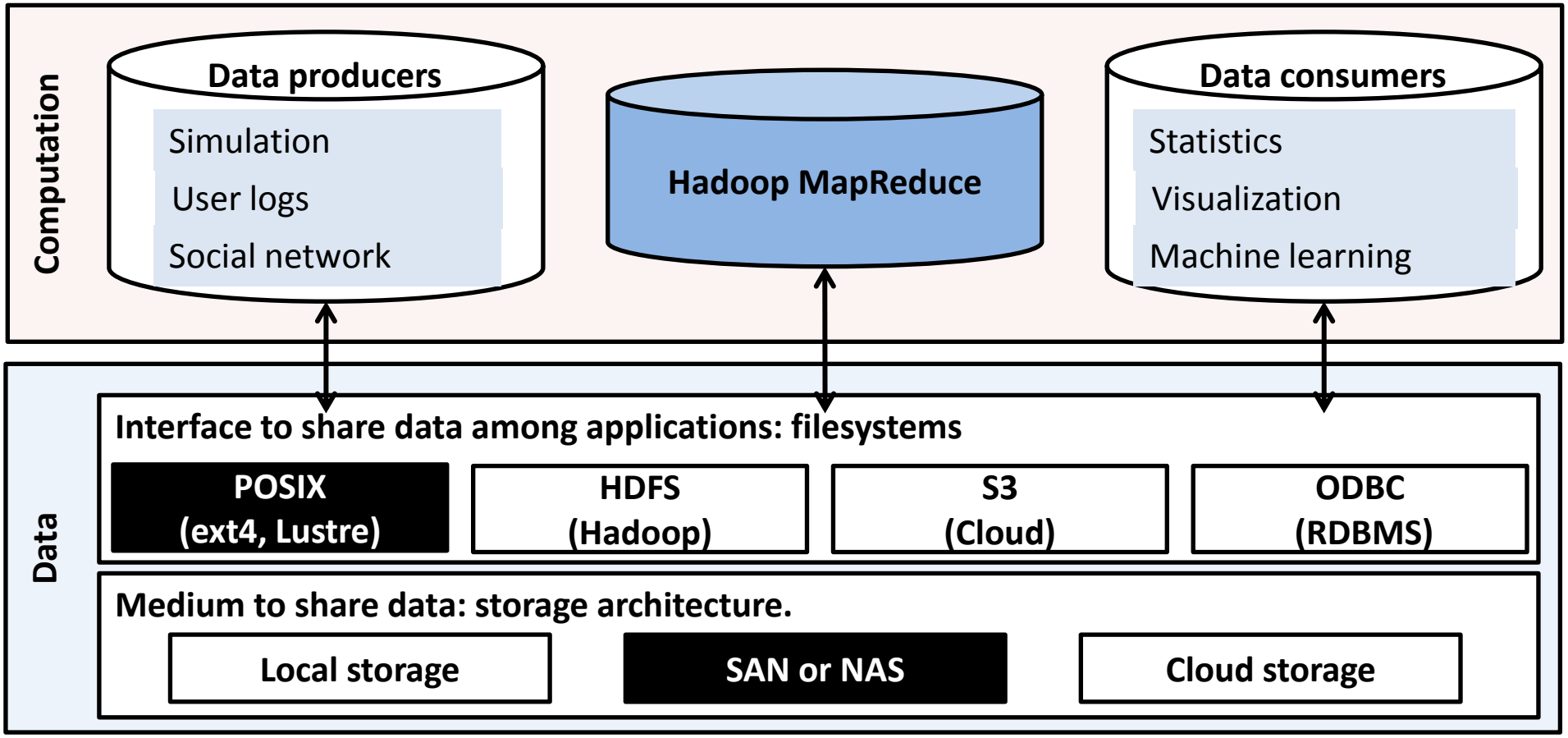
- Hadoop On EMC Isilon scale-out NAS (EMC whitepaper, 2012):
http://www.slideshare.net/emcacademics/h10528-hadoop-onemcisilonnas

- Hadoop on virtual machines (Hadoop summit 2012):
http://www.slideshare.net/rjmcdougall/hadoop-on-virtual-machines

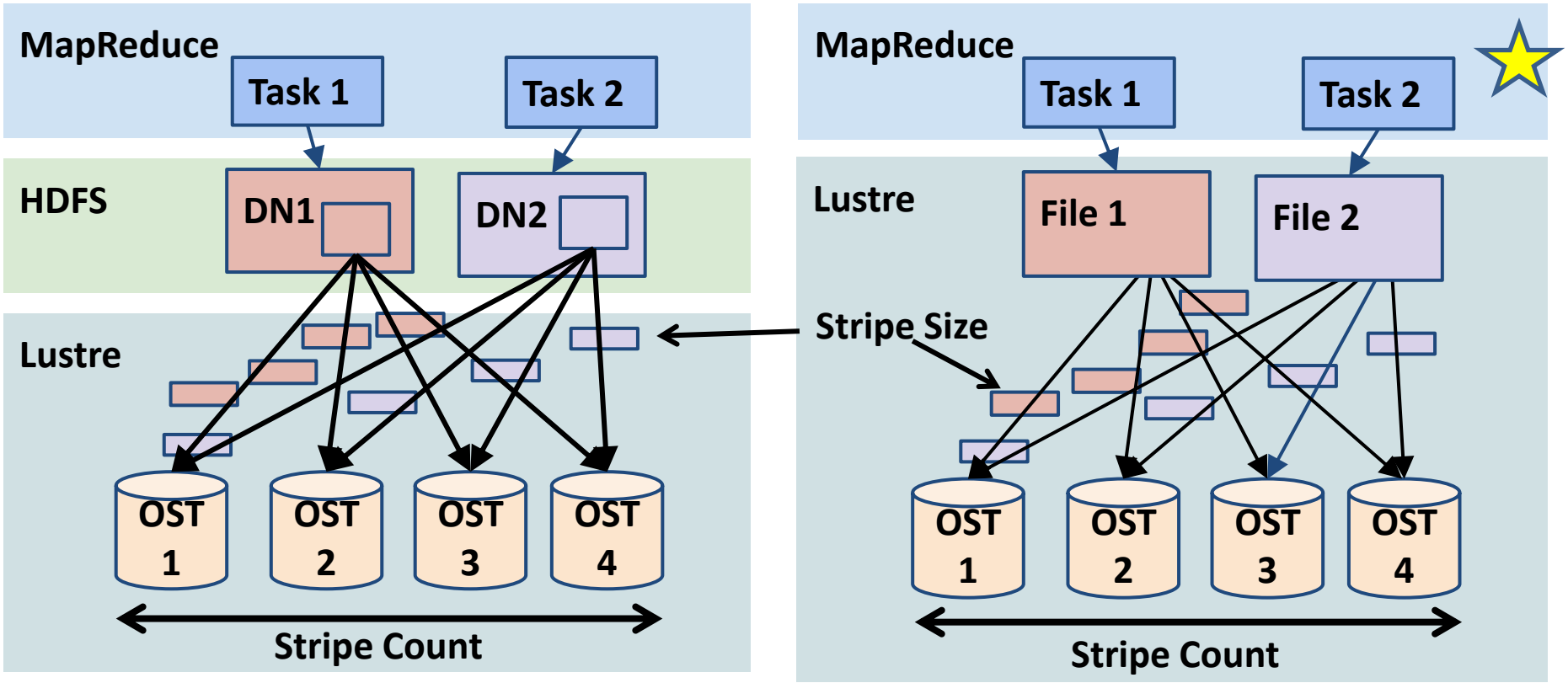
- For Amazon Web Services have Elastic MapReduce, both data input and output are over the Internet (S3 cloud storage).

Hadoop can use a variety of interfaces and medium to access data,

as it has become a part of data analytic systems.



Architectural Overview



Challenges

1. No dedicated hardware for Hadoop: we reserve compute nodes through PBS.
2. Language: Apache Hadoop is written in Java.

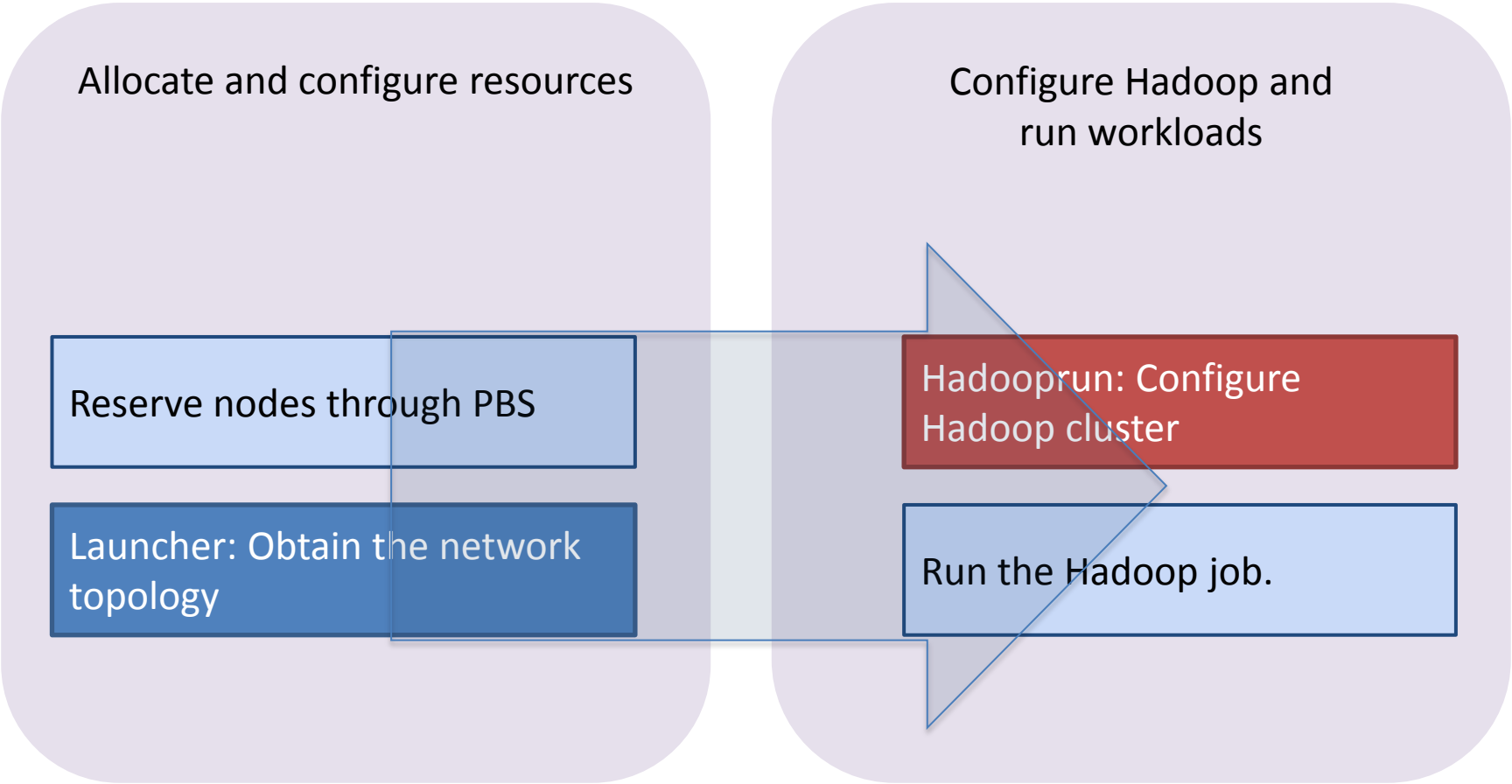
Node reservation (PBS)

- Done, open-sourced at github.com.

Java (can use Cray Gemini or Infiniband network?)

- Future, we have tested on infiniband cluster, Smoky.

How does it work through PBS?



Overview of Spot-Hadoop

#PBS –l nodes=${N}:ppn=16 –l walltime=02:00:00 –A <ProjectID>
mpirun –np ${N} launcher

Launcher (A simple MPI program)
- Obtains IP addresses of compute nodes for each MPI task.
- Store the pairs of (Rank, Hostname, IP)

Hadooprun.py

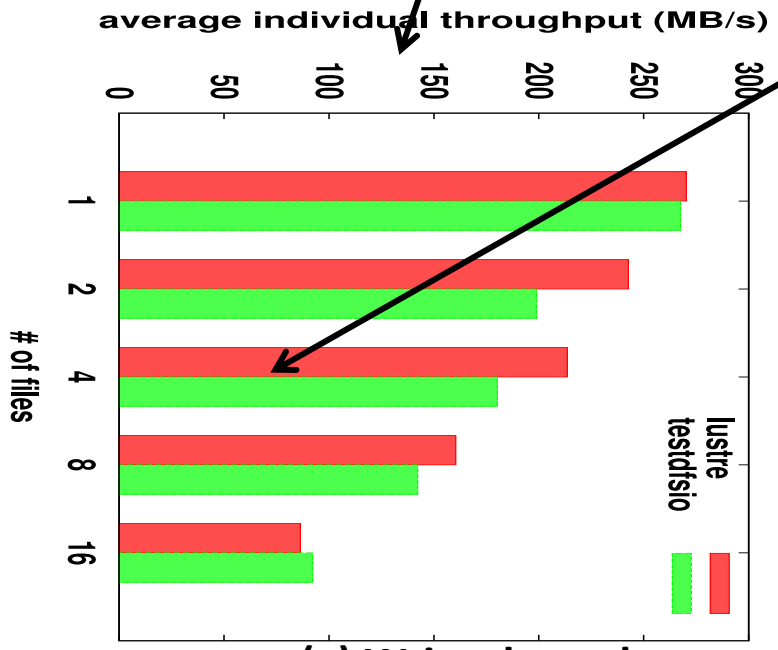
Hadooprun.py (A python program)
- Generate Hadoop configuration files and set proper environment variables.
- Network information (Master IP, Slave IPs) are from Launcher.
- Static parts (PATH of hadoop binary, etc) are from templates.
- Dynamic part of the configuration files (# of total map/reduce) are generated by heuristics.

${HADOOP_HOME}/bin/hadoop –jar <myjar> <parameters> : Run hadoop jobs

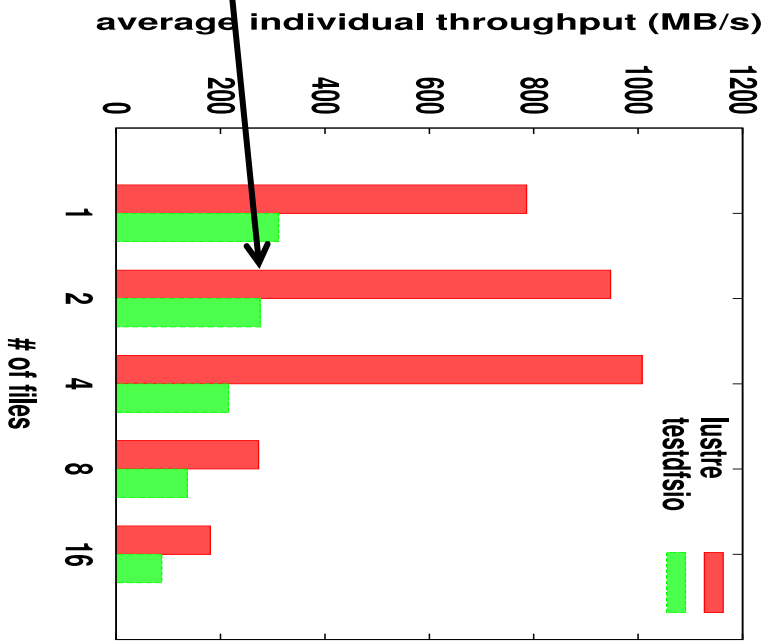
Performance
(accessing *n* files from one node)

Less than 18% of deviation from raw lustre.

One file cases represent the overhead from additional software stacks.



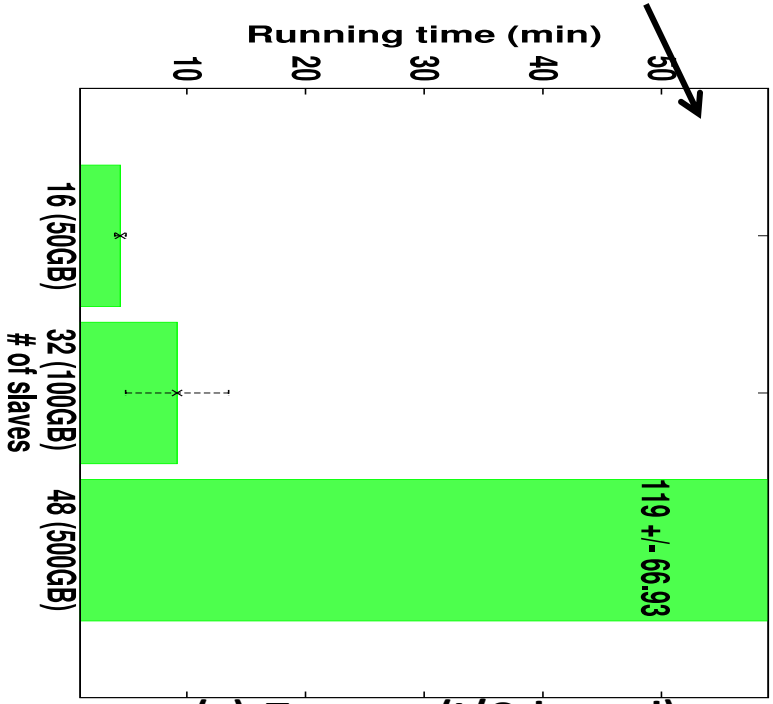
(a) Write throughput



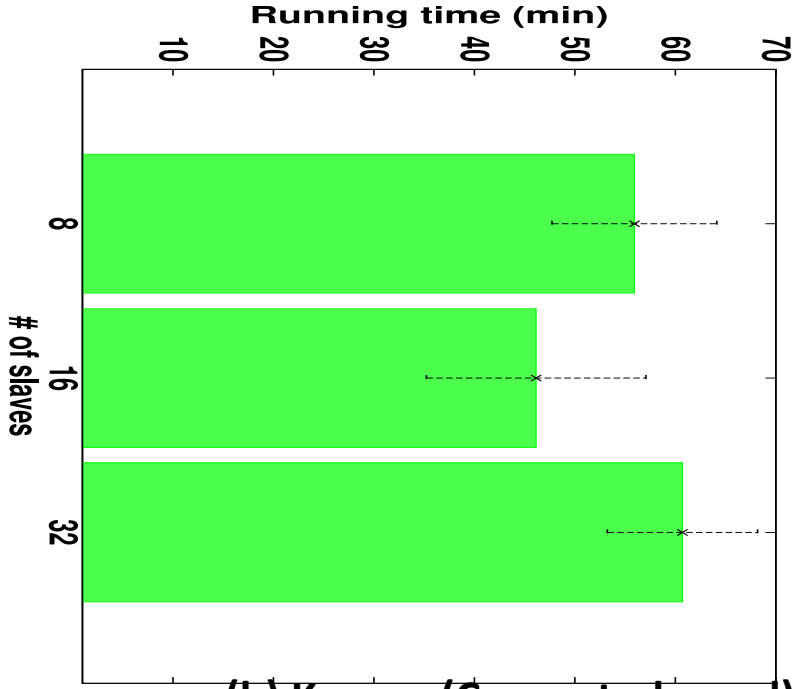
(b) Read throughput

Performance of sample applications (from 16~48 nodes)

95% confidence interval is almost a half of the average.

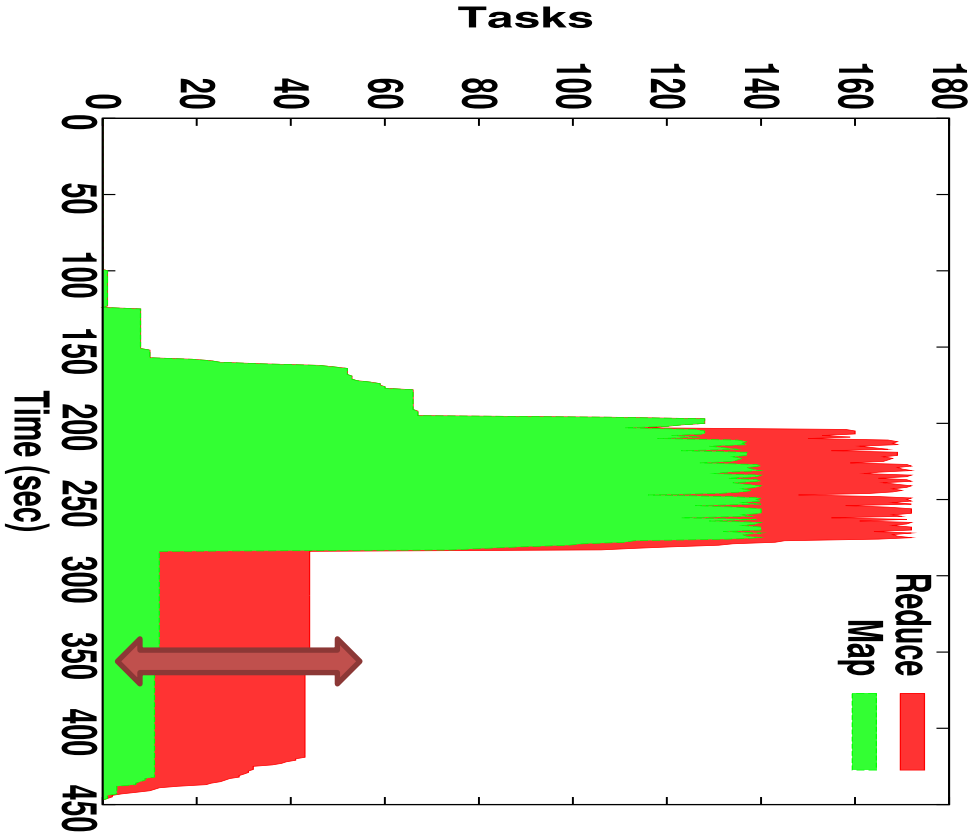


(a) Terasort (I/O-bound)
(1 record=100B)



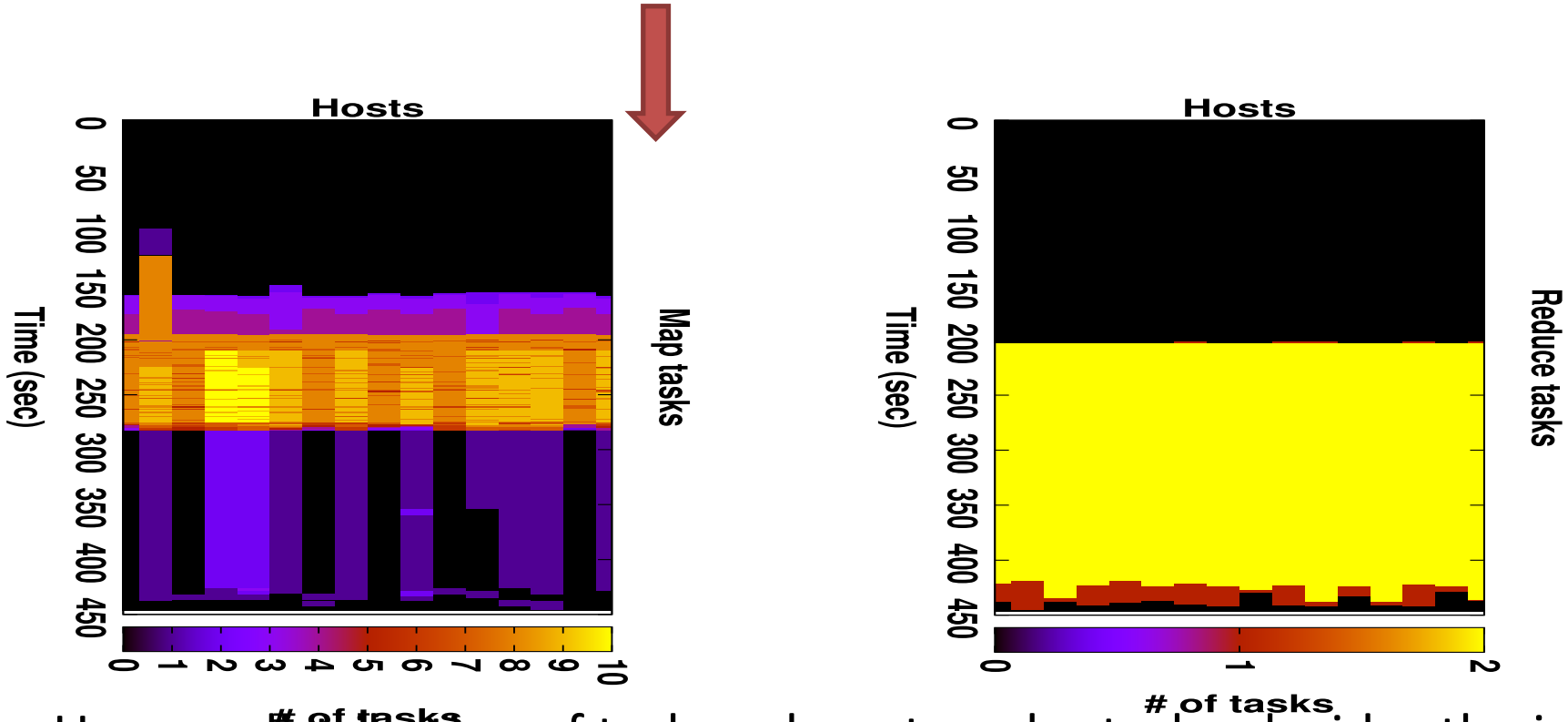
(b) Kmean (Compute-bound)
(sample=3million; 500MB)

Hadoop-specific challenges (1)



Task set up time can take ¼ of the actual running time of the job. (Data from Sorting 50GB using 16 slave nodes.)

Hadoop-specific challenges (2)



Uneven distribution of tasks, a.k.a. straggler tasks, decides the job running time (data from sorting 50GB with 16 slaves.)

Conclusion

We can dynamically construct a Hadoop cluster for each user Job through PBS.

Promising performance of Hadoop over Spider,

- Almost 100% write throughput of the filesystem
- But, 1/5 read throughput.

Additional Hadoop-specific optimization is on-going.

Toward deployment on Titan, we need to figure out networking.

Spot-Hadoop is open-sourced. https://github.com/jhorey/SpotHadoop

Seung-Hwan Lim
lims1@ornl.gov
Computational data analytics group
Oak Ridge National Laboratory, Oak Ridge, TN