The New Computing Environment for LHAASO

August 15 2016

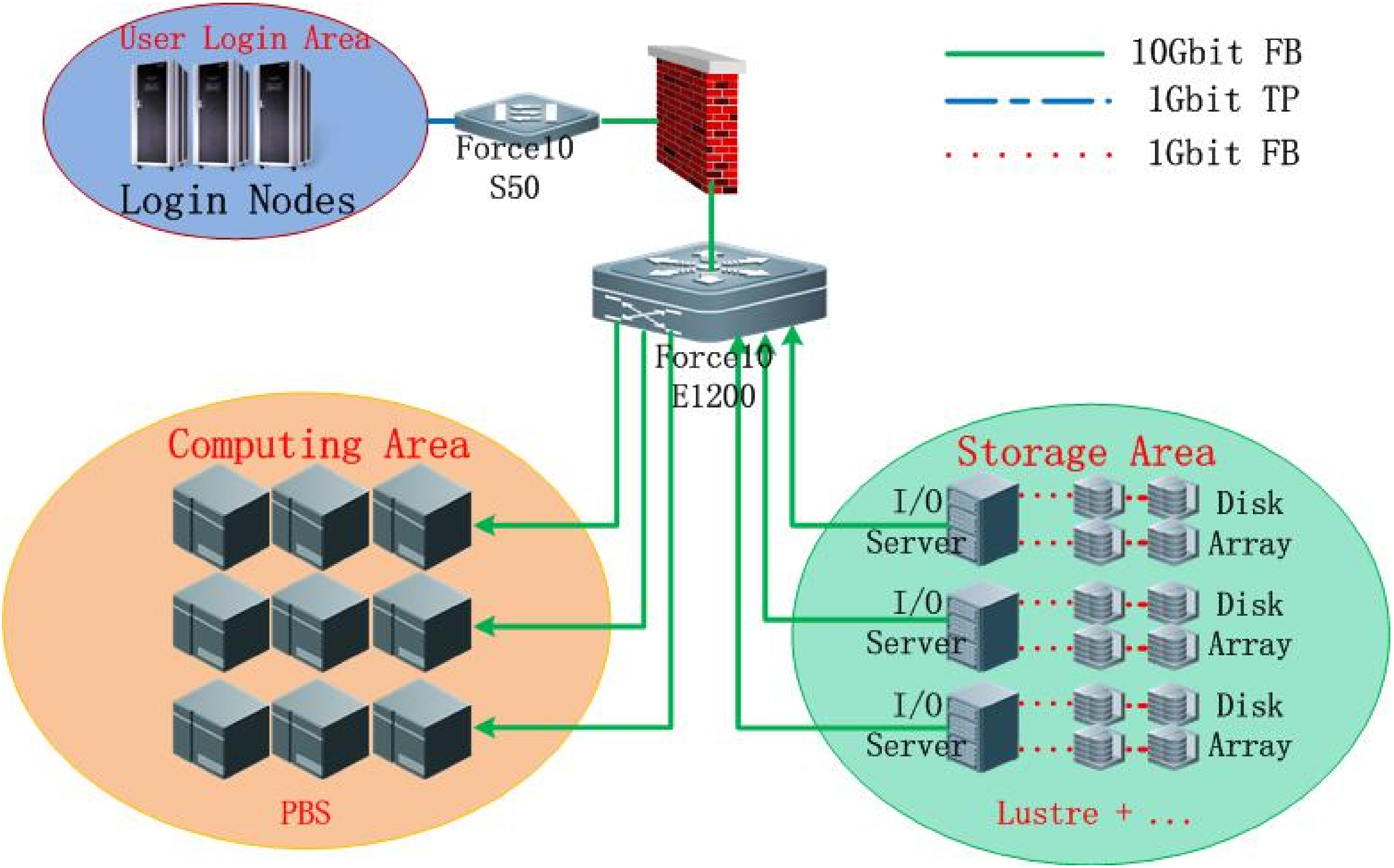
Li Qiang, Huang Qiulan, Sun Gongxing
IHEP-CC

Supported by the National Natural Science Fund

Outline

- The Current Computing System
- What is Hadoop?
- Why Hadoop?
- The New Computing System with Hadoop
- Test Results
- Summary

The Current Computing System



~10000CPU Cores

~5PB

Problems

- Network I/O becomes the bottleneck for data-intensive jobs
- More money should be invested to buy better network equipment and storage devices
- More data taking in the future, and new techniques should be explored

What is Hadoop?

Apache Hadoop

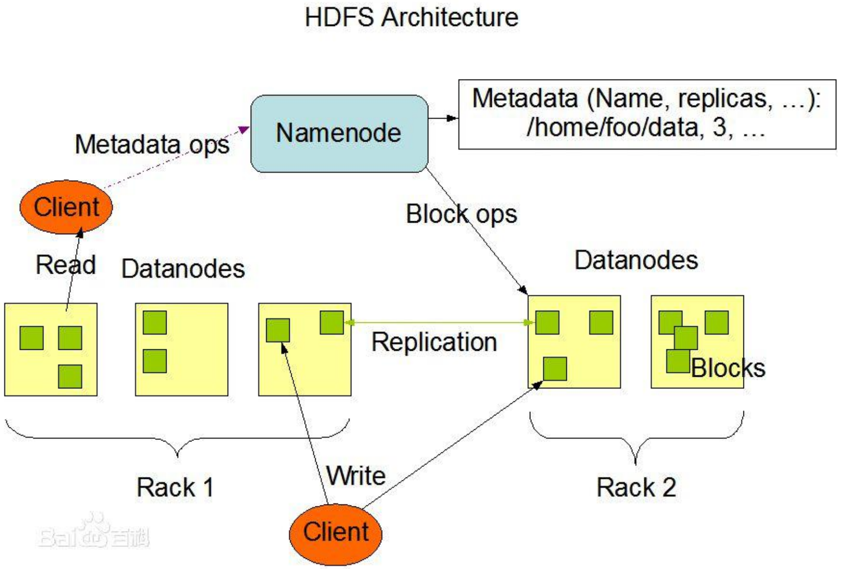
An open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

- Provides a highly reliable distributed file system (HDFS).
- Provides a parallel computing framework for large data sets(MapReduce).
- And also other tools: HBase, Hive, Pig, Spark, etc.
- Widely adopted througout the internet industry.

HDFS(Hadoop Distributed File System)

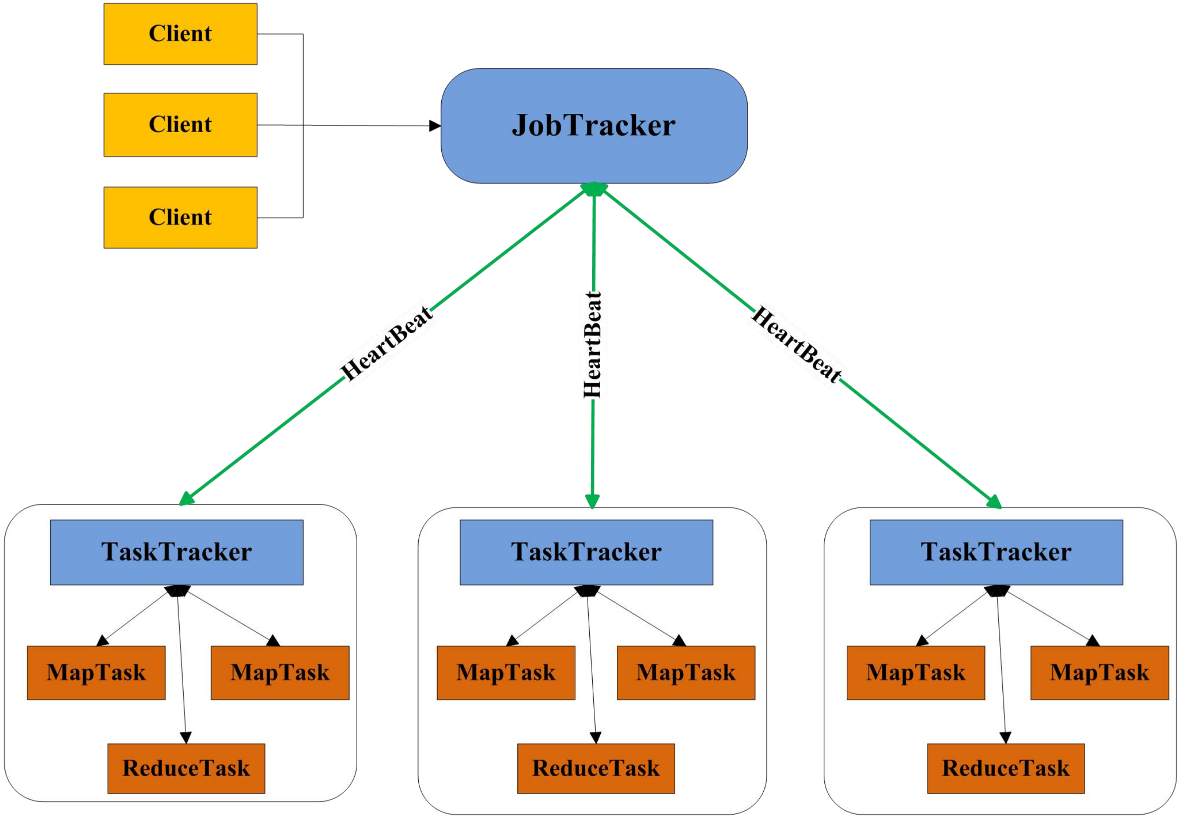
A distributed file-system that stores data on commodity machines.

- High fault-tolerant
- High throughput
- High reliability
- High scalability
- Suit large data set



MapReduce

- A distributed programing model by Google
- A job execution framework



Why Hadoop?

- ✓ Typical HEP physics analyses fit well to MapReduce paradigm
- ✓ Actively maintained and developed by the industry, and Commercial support is available from a number of companies

  --Three Hadoop software provider : Apache, Cloudera, Hortonworks

  -- More than 150 companies are using
- ✓ **Much cheaper**

  -- No need 10G optical network card

  -- No need powerful network equipment

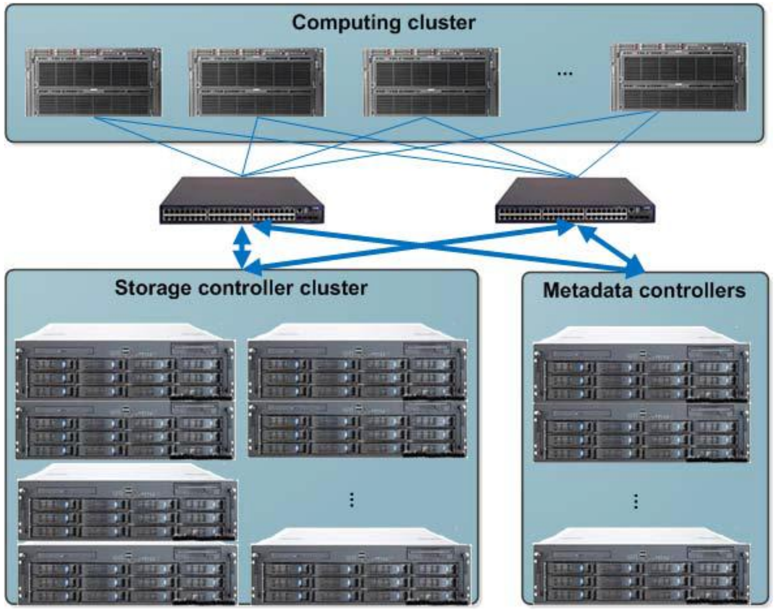
  -- Use local disks(more powerful), no need expensive disk arrays
- ✓ Very easy to scale out & up

  -- one master cluster can reach 4000 nodes

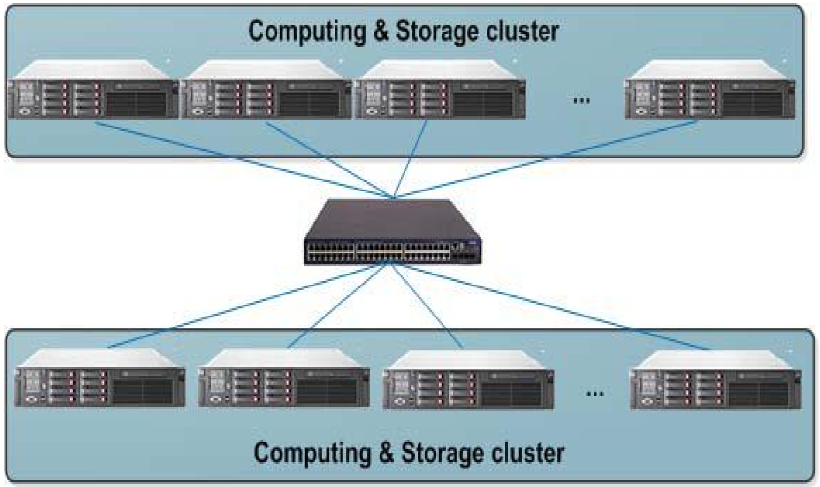
Hadoop & Hep

- CMS experiments from 2009 began to study the use of Hadoop framework for dealing with CMS experimental data, several sites in the United States have begun to use HDFS as a grid system OSG storage unit.
- CERN's Maria Lassnig and Vincent Garonne proposed use HBase to management data file/data set at CHEP (2012) .
- 2013, Fabian Glaser of the University of Iceland proposed use MapReduce substituted PROOF (Parallel ROOT Facility) to do parallel analysis.
- Dubna Joint Institute for Nuclear Research study using Hadoop to do Physical Analysis in 2015.

The New Computing System with Hadoop

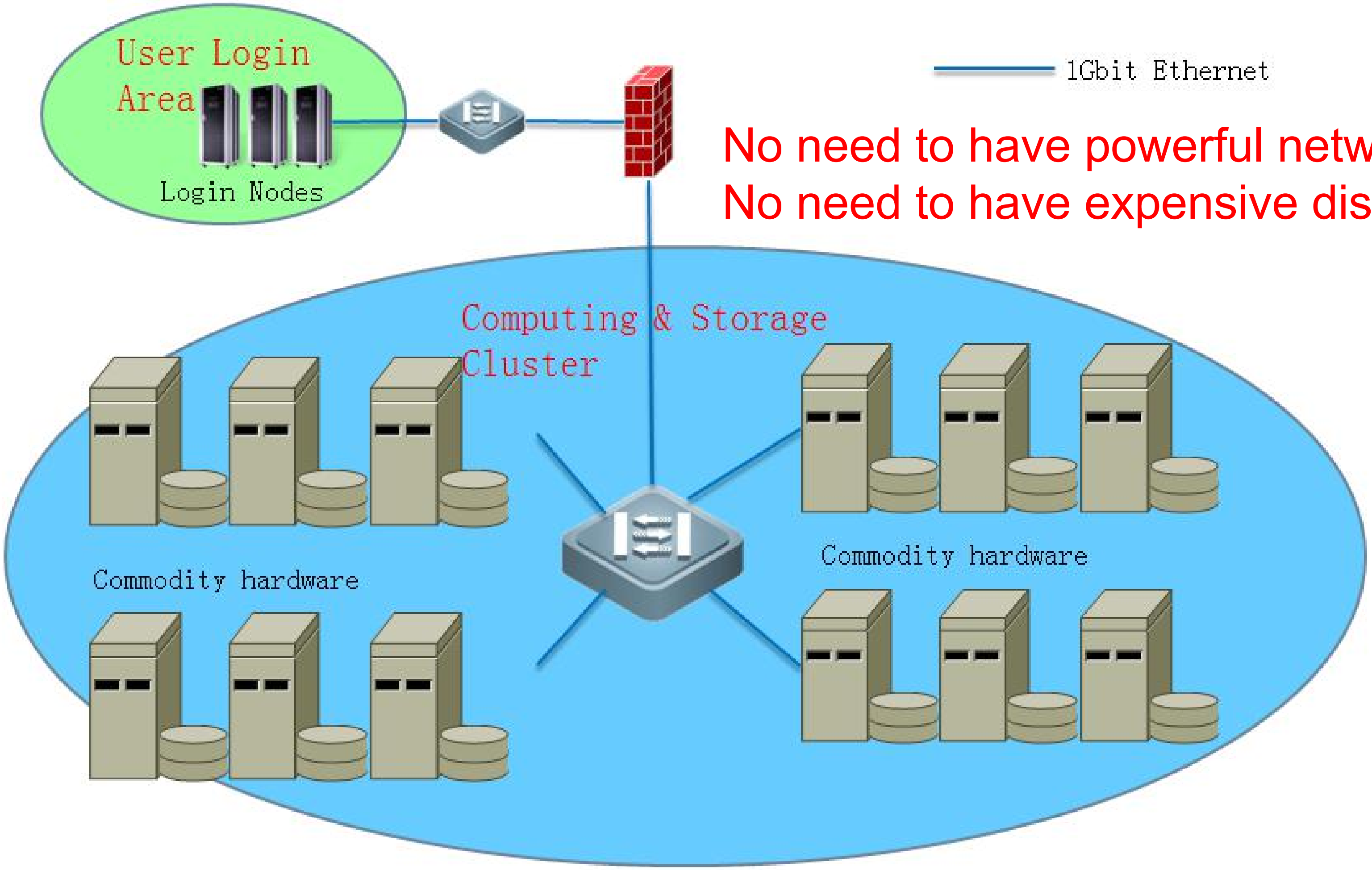


Moving data to computation



Moving computation to data

The New Architecture



No need to have powerful network,
No need to have expensive disk arrays

Challenges

- Hadoop use streaming access data, only support sequential write and append, not support random write.
- Hadoop is written in Java. C/C++ support are very limited.
- ROOT read HDFS files via FUSE or other plugins
- ROOT writes to HDFS via Temporary local files

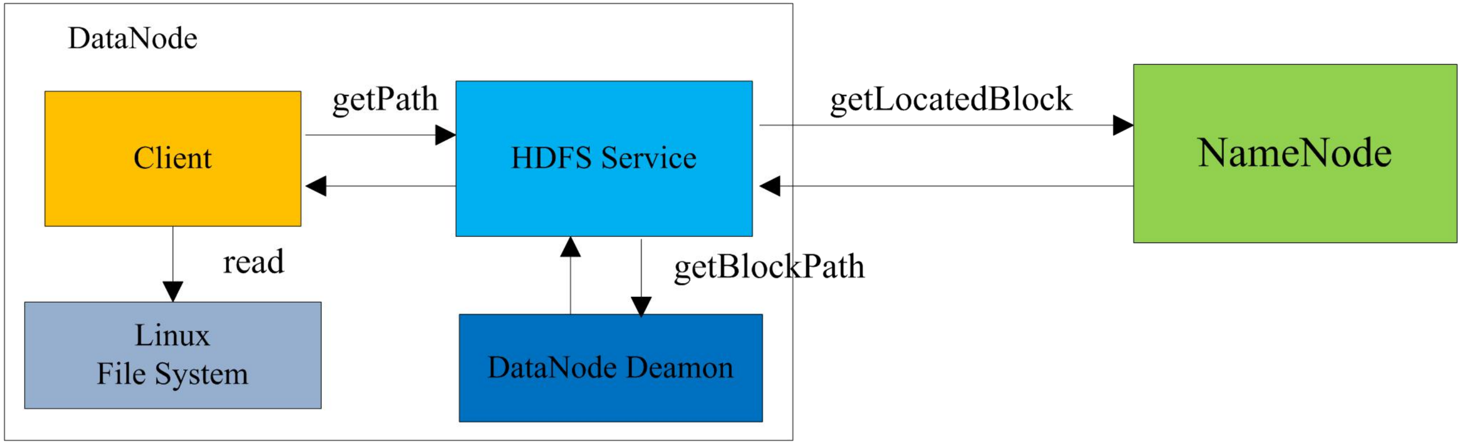
New Data Access Mechanism

- Local read data(Job completely localized execution)

  - No data transmission

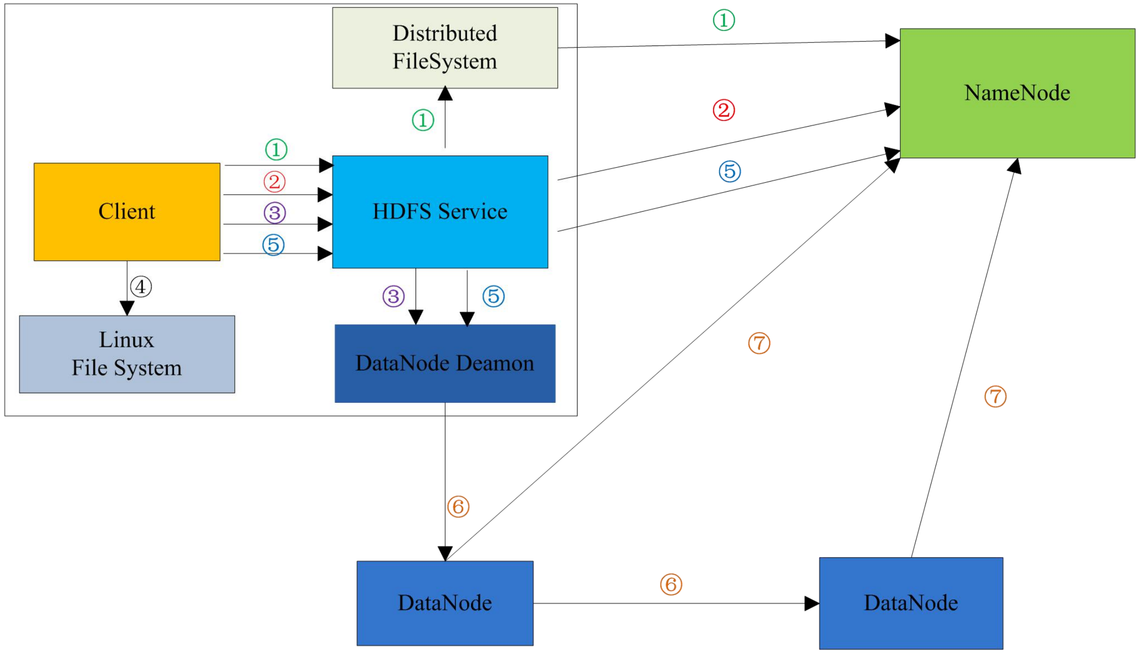
  - No network I/O

  - Low latency



New Data Access Mechanism

- ROOT write to HDFS directly
- Local write if only have one replica
  - No data transmission
  - No network I/O
  - Low latency
- Random write
- Modify file



① createFile ② addBlock ③ getBlockPath ④ writeBlock

⑤ Complete ⑥ Copy ⑦ blockReceivedReport

HDFS Blocks

- Define a block is a file
- The Block(Replic) has the same attributes as the file

  -- Original block's owner is hdfs(HDFS superuser)

  -- Now block's owner is the user who create it
- NameNode transfer block attributes to DataNode by HTTP
- DataNode set/change block attributes through JNI

IO Performance Test

- HDFS

  --1 NameNode,5 DataNode (6*6TBdisks, Raid5)

  -- Gigabit Ethernet

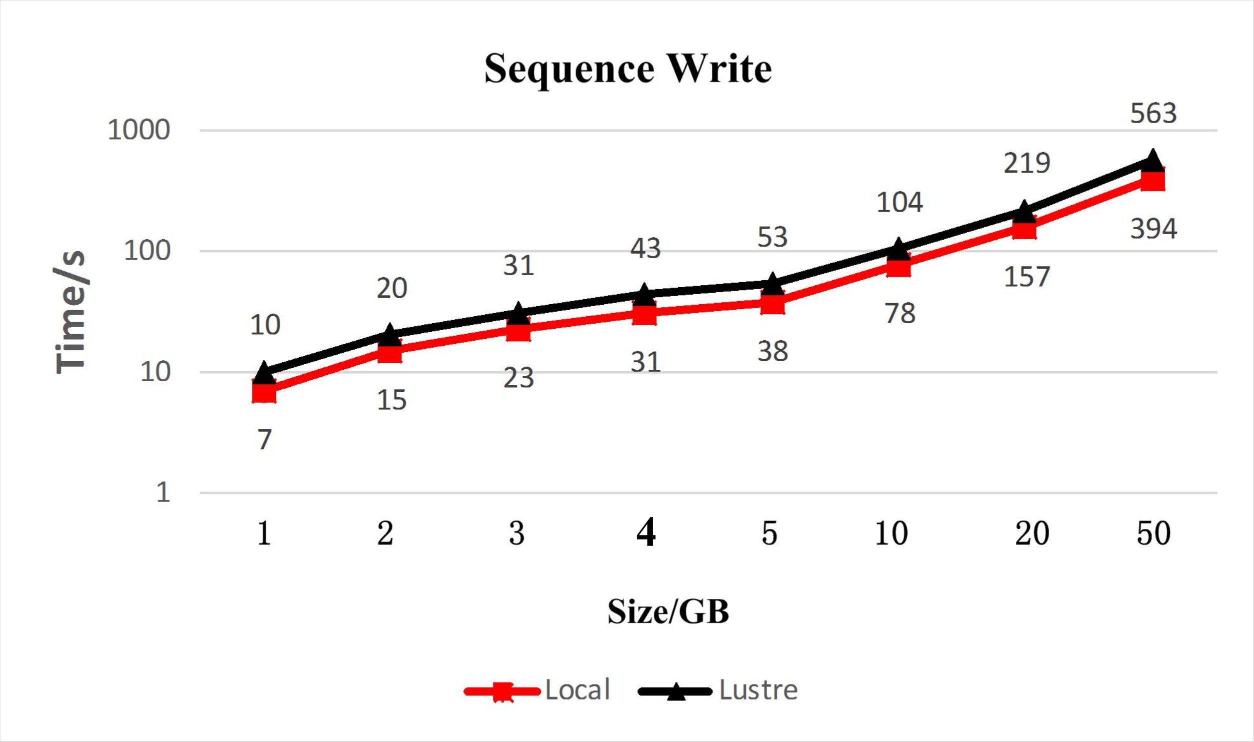
- Lustre

  -- 2*Disk Array (24*3TB,Raid6)

  -- 10 Gigabit Ethernet

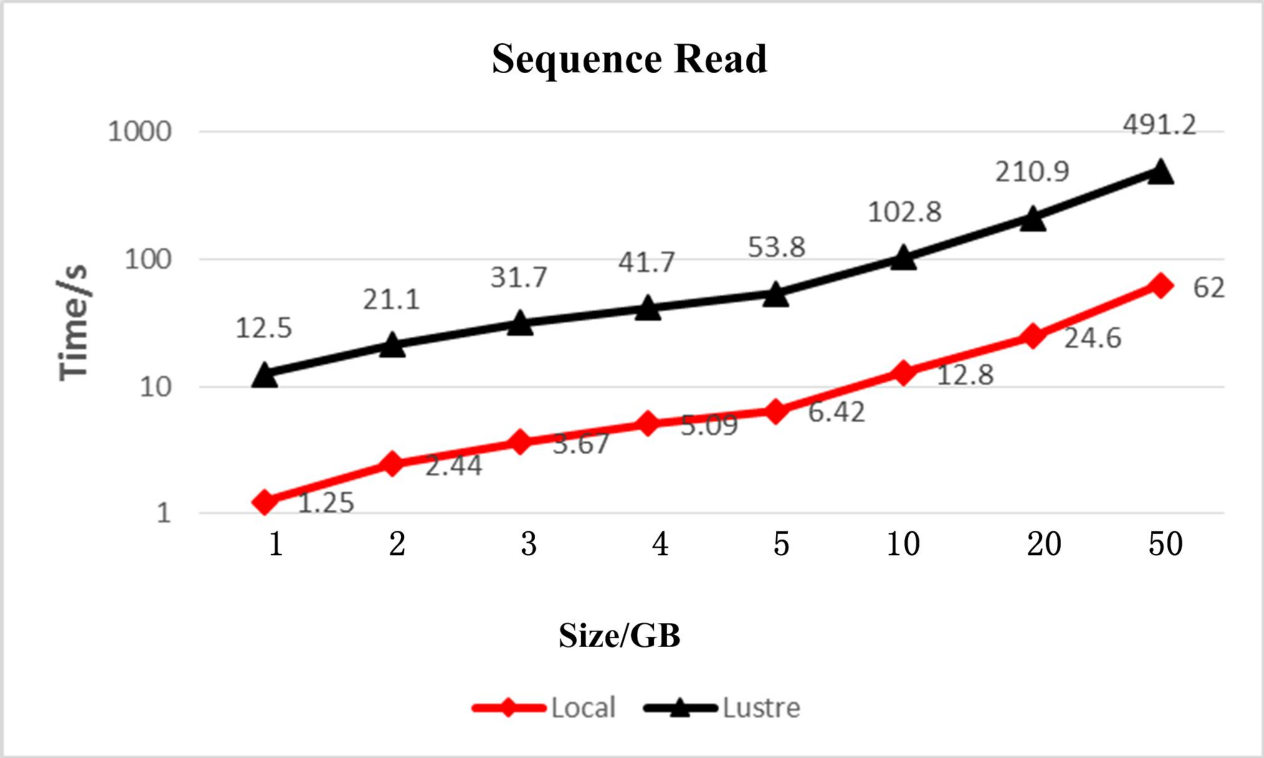
Sequence Write Test

- ✓ Write by bytes, buffer size is 64KB.
- ✓ Single process, Local write is 30% faster than Lustre.



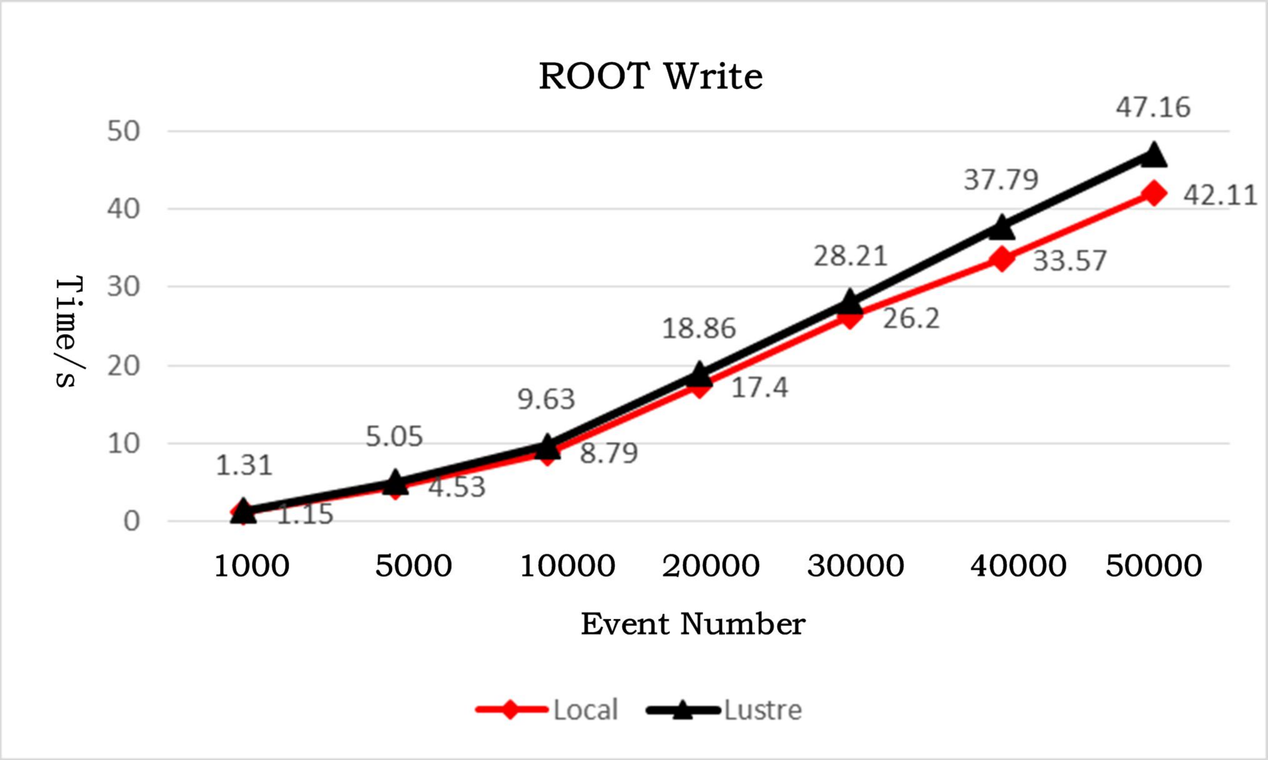
Sequence Read Test

- ✓ Read by bytes, buffer size is 64KB
- ✓ Single process, Local read is about 8~10 times faster than Lustre.



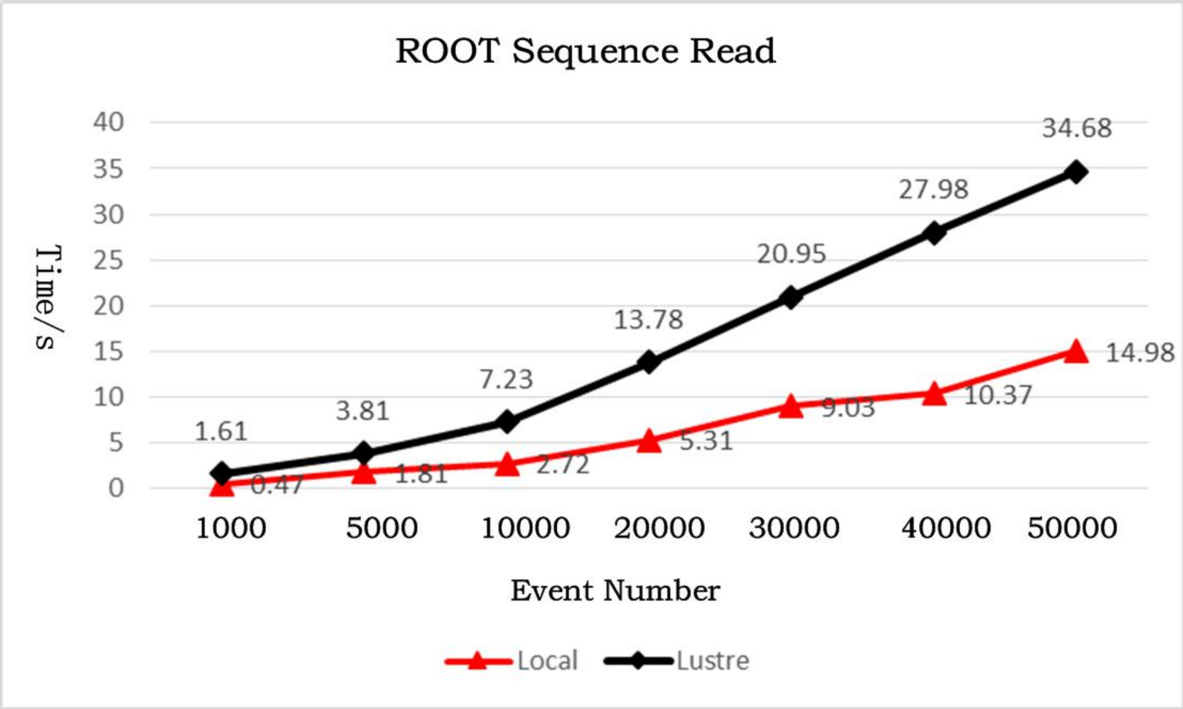
ROOT Tool Test

- ✓ Command :“$ROOTSYS/test/Event EventNumber 0 1 1”
- ✓ Single process, Local write is 10% faster than Lustre FS.



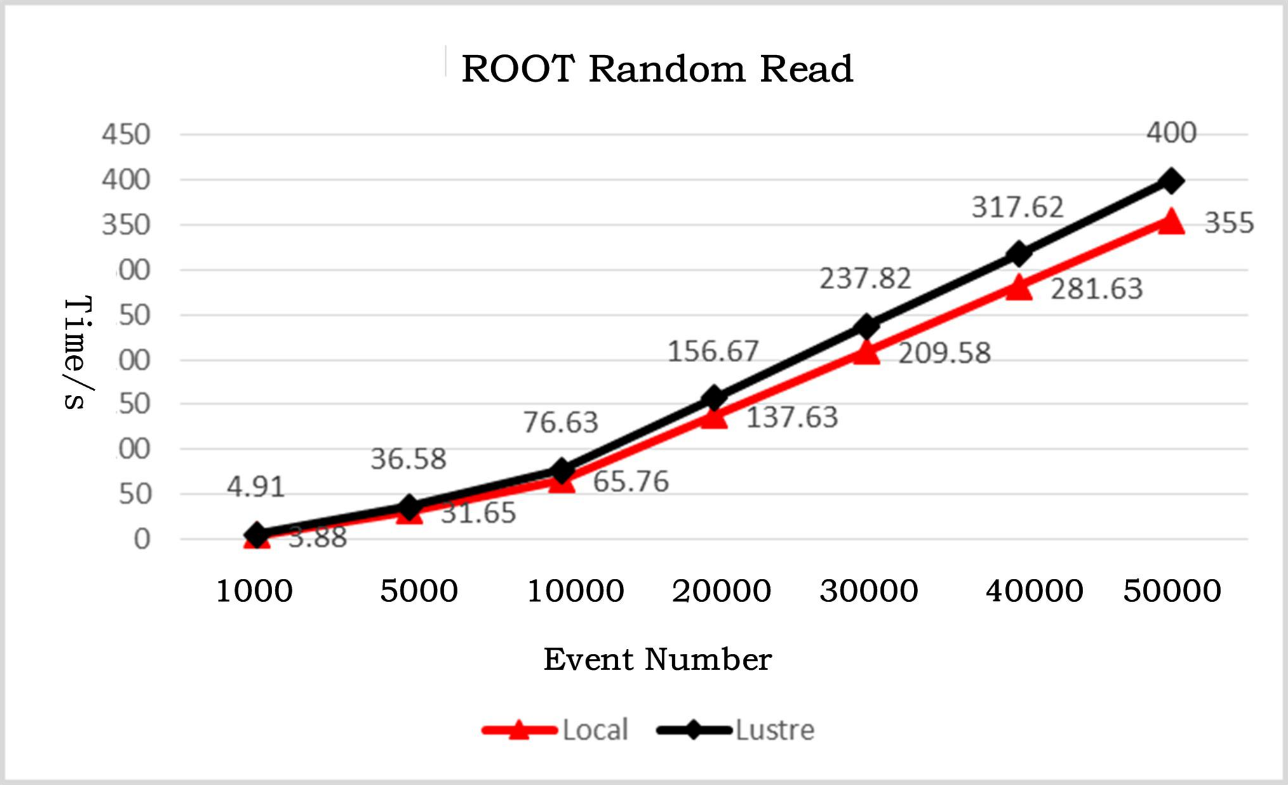
ROOT Sequence Read Test

- ✓ Command :“$ROOTSYS/test/Event EventNumber 0 1 20”.
- ✓ Single process, Local sequence read is 2~3 times faster than Lustre FS.



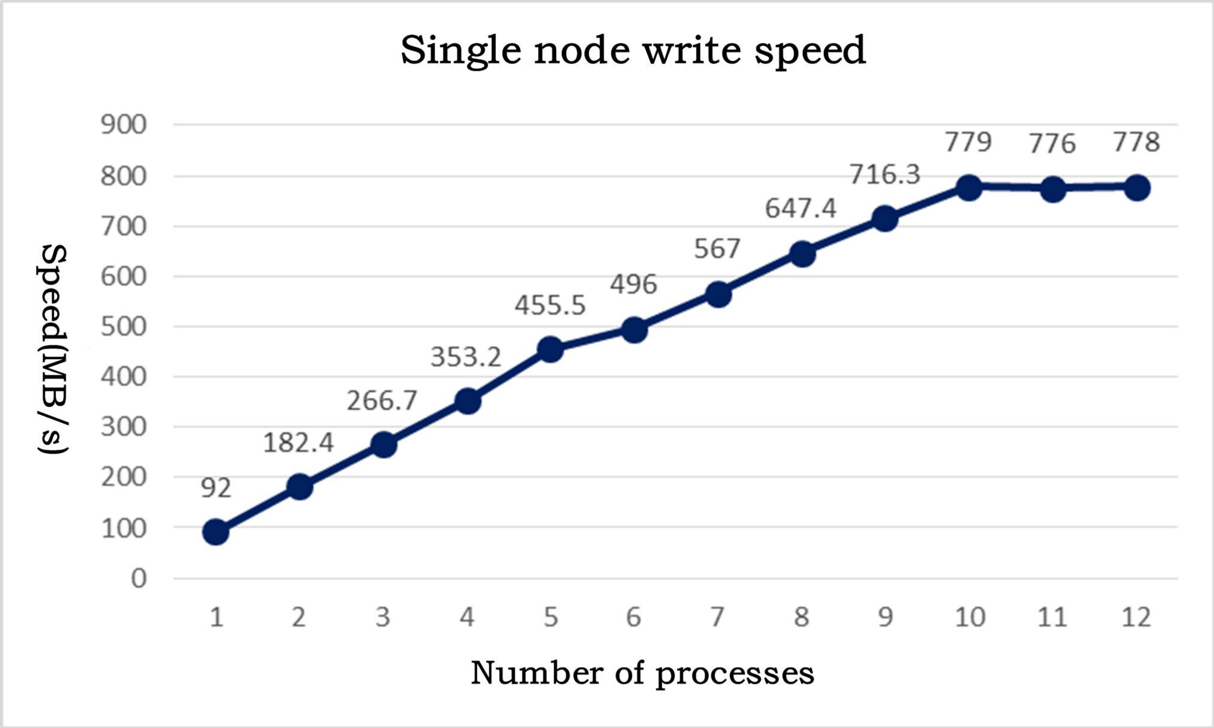
ROOT Random Read Test

- ✓ Command :“$ROOTSYS/test/Event EventNumber 0 1 25”.
- ✓ Single process, Local random read is 10% faster than Lustre FS.



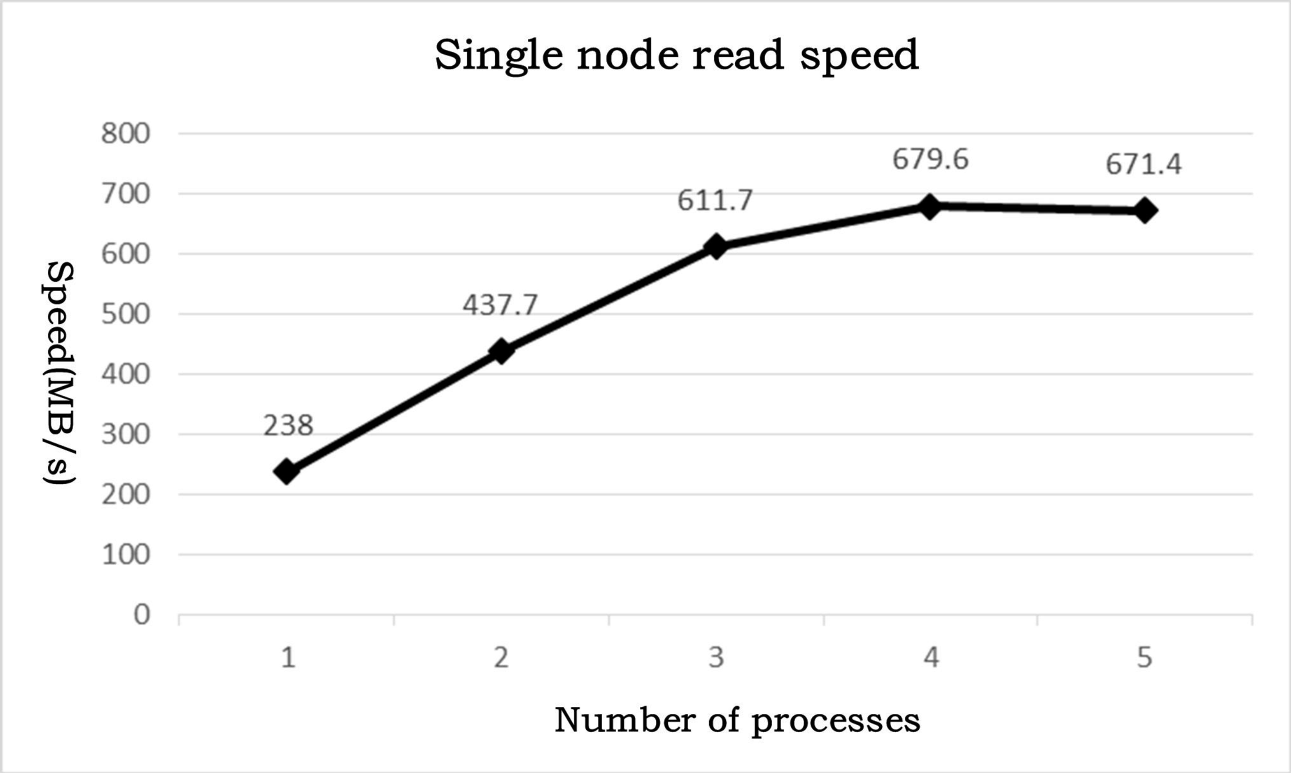
Concurrent Test

- ✓ Single node , local write speed maximum can reach to 779 MB/s.
- ✓ In HDFS, write speed maximum can reach to 779MB/s*datanodes number.
- ✓ Lustre maximum write speed depending on network.



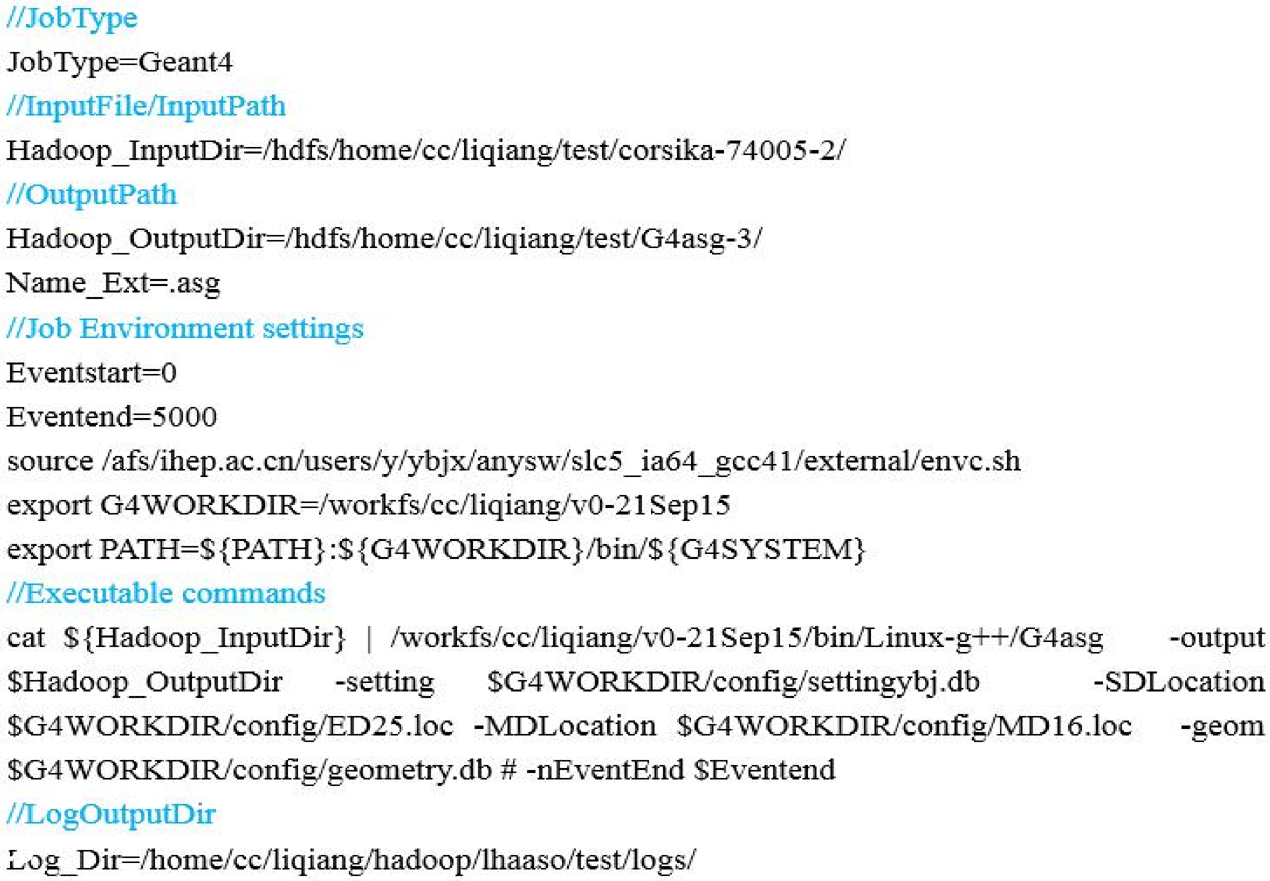
Read Concurrent Test

- ✓ Single node , local read speed maximum can reach to 680 MB/s.
- ✓ read speed maximum can reach to 680MB/s*datanodes number for HDFS.



Job Submit and Execute

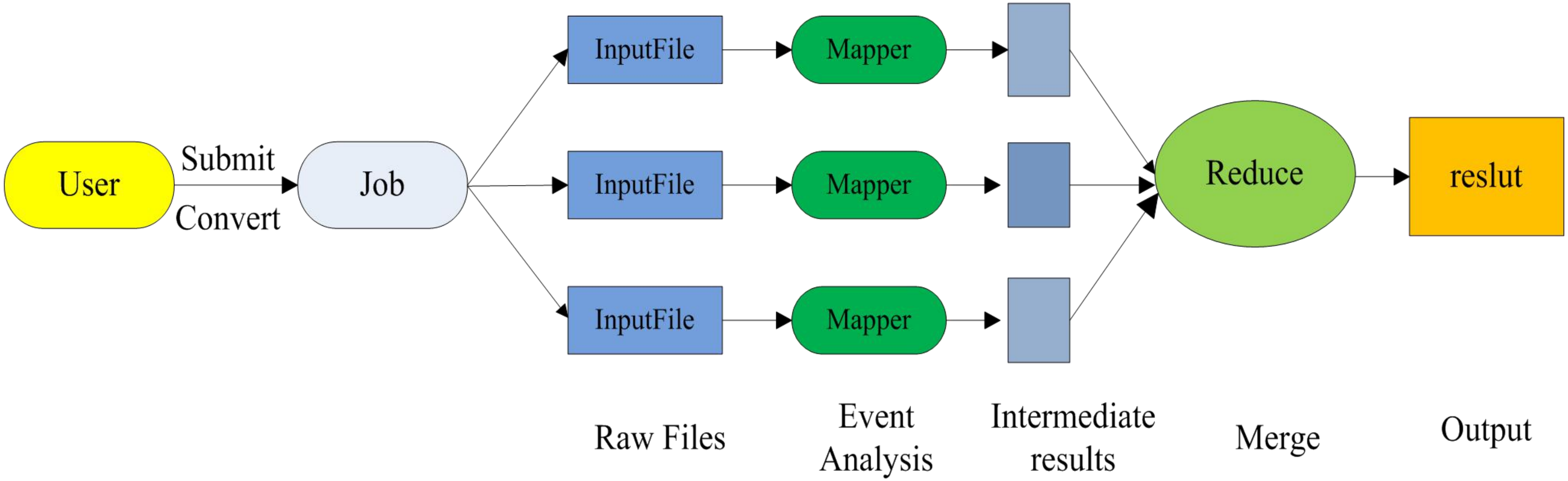
- Submit one job to process a large number of files(i.e 1000)
  - hsub + queue + jobOptionFile + JobName
- jobOptionFile can divide into Six parts: jobType, InputFile/InputPath, OutputPath, Job Environment settings, Executable commands, LogOutputDir.



```
//JobType
JobType=Geant4
//InputFile/InputPath
Hadoop_InputDir=/hdfs/home/cc/liqiang/test/corsika-74005-2/
//OutputPath
Hadoop_OutputDir=/hdfs/home/cc/liqiang/test/G4asg-3/
Name_Ext=.asg
//Job Environment settings
Eventstart=0
Eventend=5000
source /afs/ihep.ac.cn/users/y/ybjx/anysw/slc5_ia64_gcc41/external/envc.sh
export G4WORKDIR=/workfs/cc/liqiang/v0-21Sep15
export PATH=${PATH}:${G4WORKDIR}/bin/${G4SYSTEM}
//Executable commands
cat  ${Hadoop_InputDir} | /workfs/cc/liqiang/v0-21Sep15/bin/Linux-g++/G4asg     -output
$Hadoop_OutputDir    -setting    $G4WORKDIR/config/settingybj.db         -SDLocation
$G4WORKDIR/config/ED25.loc  -MDLocation  $G4WORKDIR/config/MD16.loc    -geom
$G4WORKDIR/config/geometry.db # -nEventEnd $Eventend
//LogOutputDir
Log_Dir=/home/cc/liqiang/hadoop/lhaaso/test/logs/
```

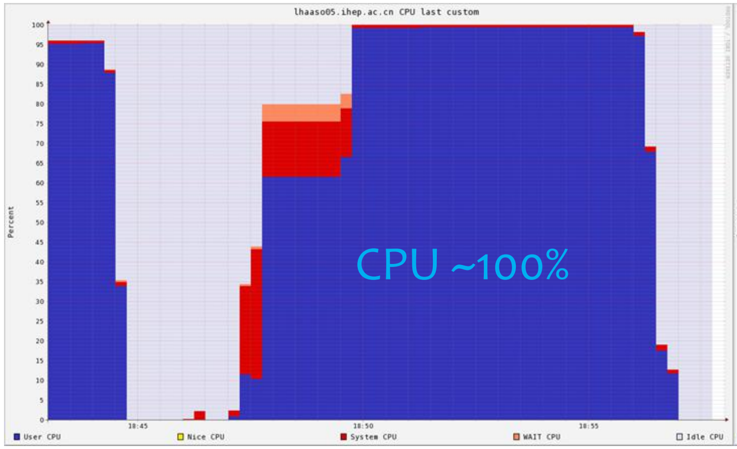
Job Submit and Execute

- The job is automatically split into multiple tasks, each task deal with one file
- Jobs are scheduled according to files' locations

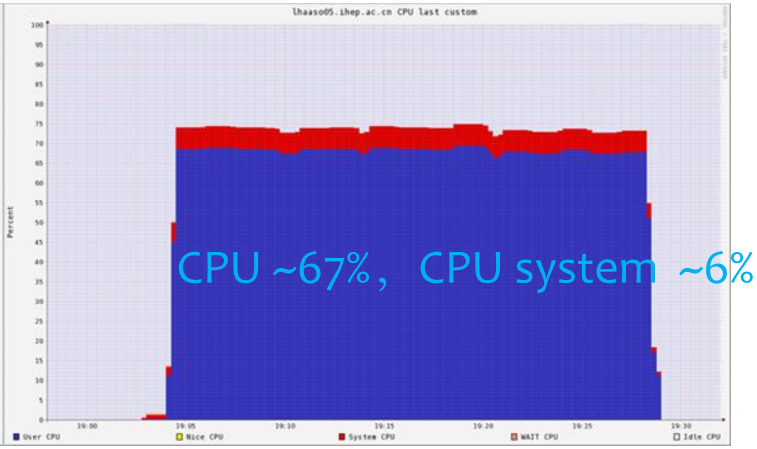


Application Test

- ✓ HDFS & Gluster performance on same node
- ✓ Jobs read data from Gluster/Lustre FS with enough bandwidth
- ✓ Test program: medea++3.10.02
- ✓ Gluster’s CPU utilization much lower than HDFS



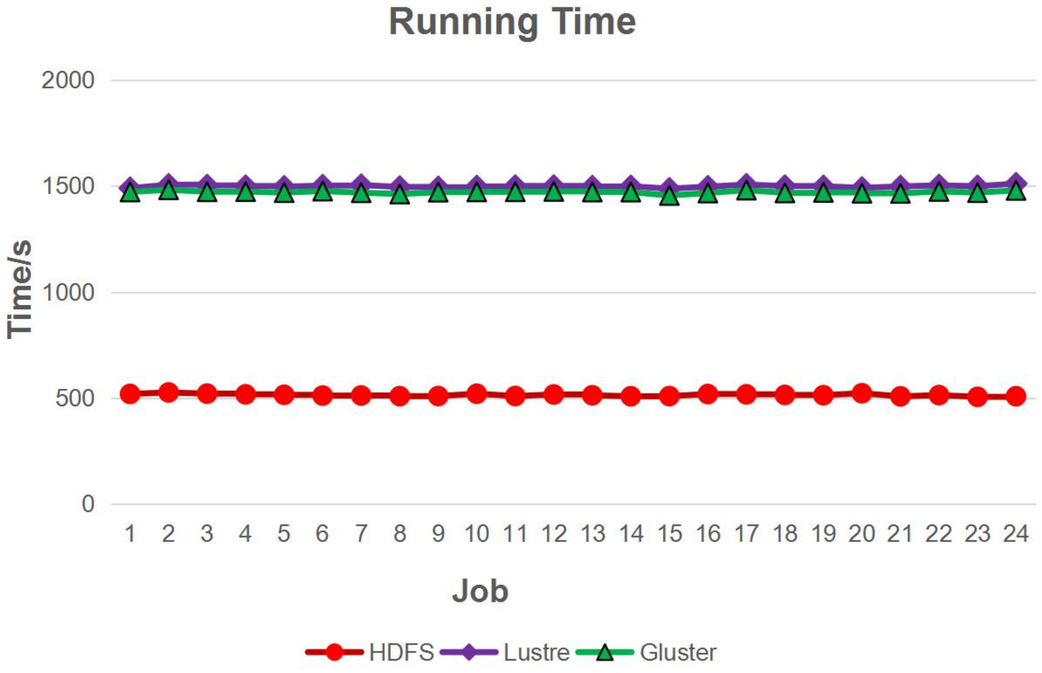
HDFS



Gluster/Lustre

Running Time

- ✓ HDFS job running time is about one-third of Gluster/Lustre FS.



Summary

The new Computing Environment:

-- Can reduce the cost to construct system

-- Improves the job execution time for IO-intensive jobs

-- IO speed not limited by network, Disk IO resources can be fully utilized.

THANKS
FOR
LISTENING!