New Computing Environment for LHAASO offline data analysis

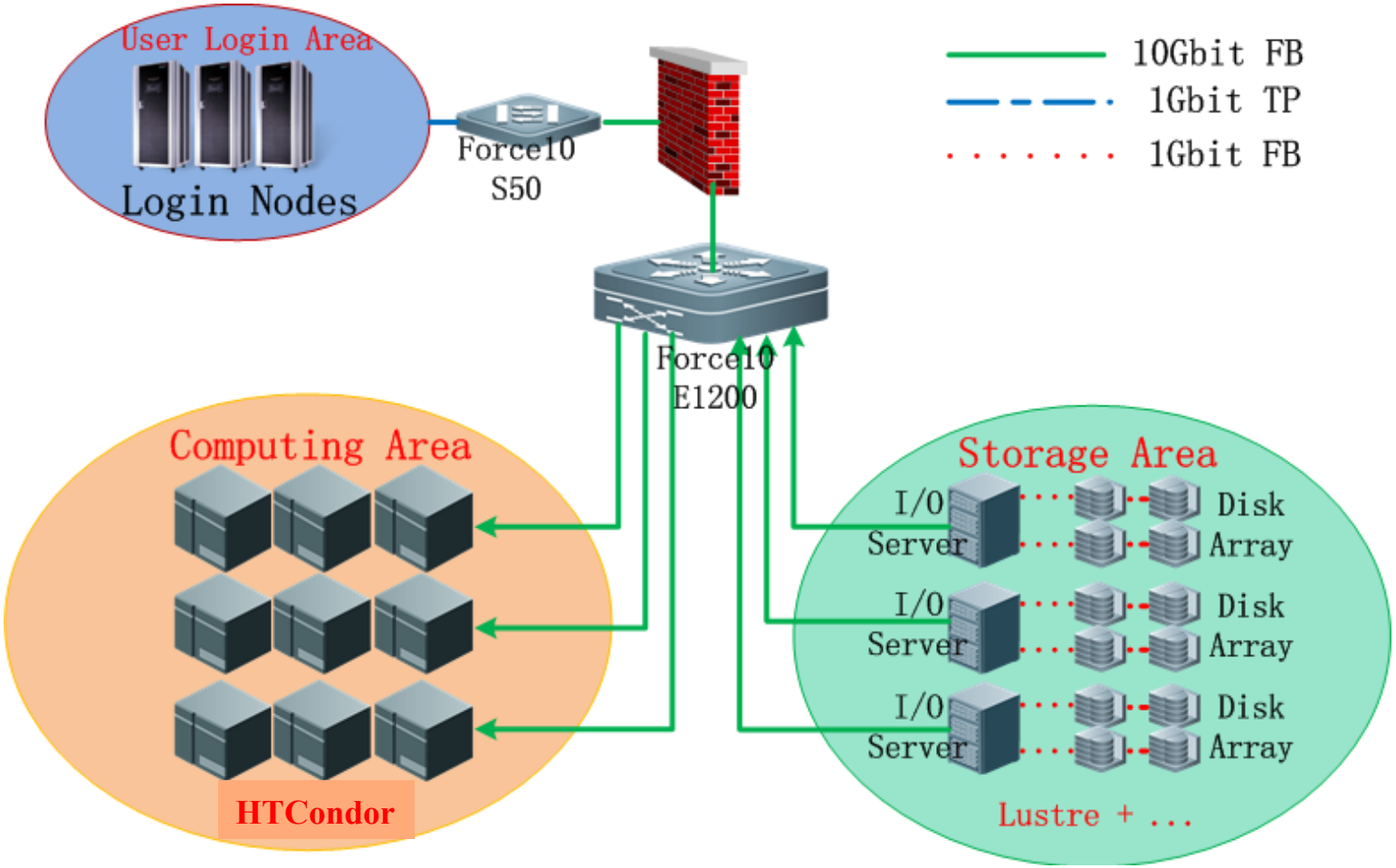
Qiulan Huang, Gongxing Sun, Zhanchen Wei, Qiao Yan
Institute of High Energy Physics, CAS

ISGC 2018
Mar 23, 2018

Outline

- Overview of the Traditional Computing System
- Problems & Challenges
- New Computing System with Hadoop
- Activities and Evaluation
- Hadoop Status in LHAASO
- Summary

The Current Computing System



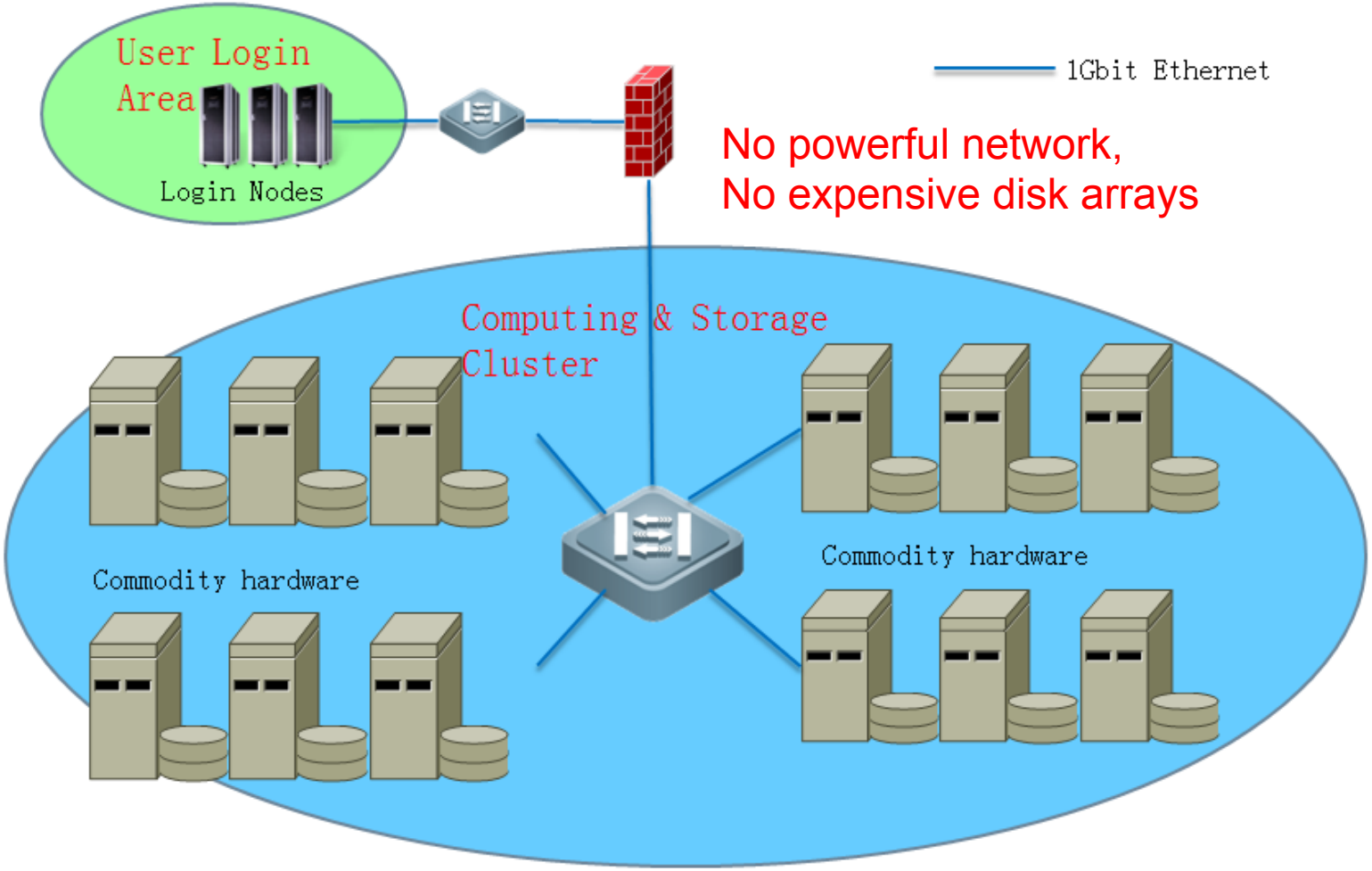
~15000CPU Cores

~11PB(Lustre/EOS)

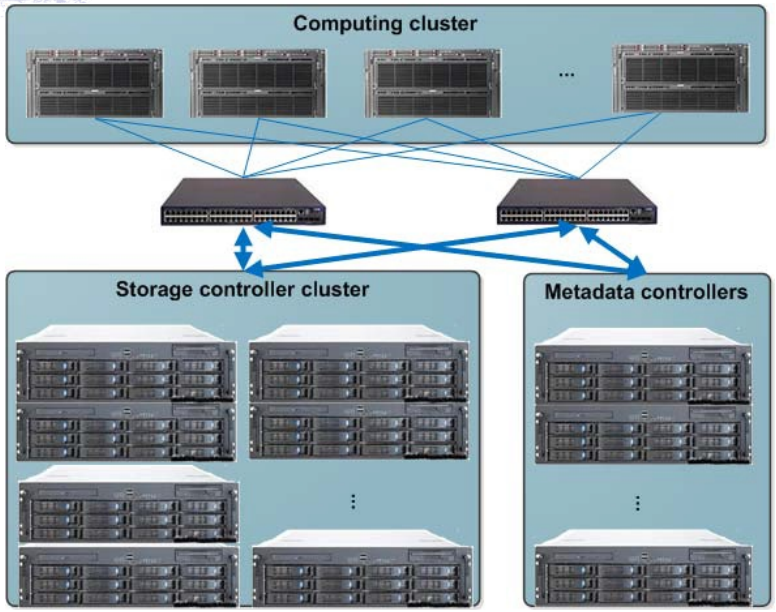
Problems & Challenges

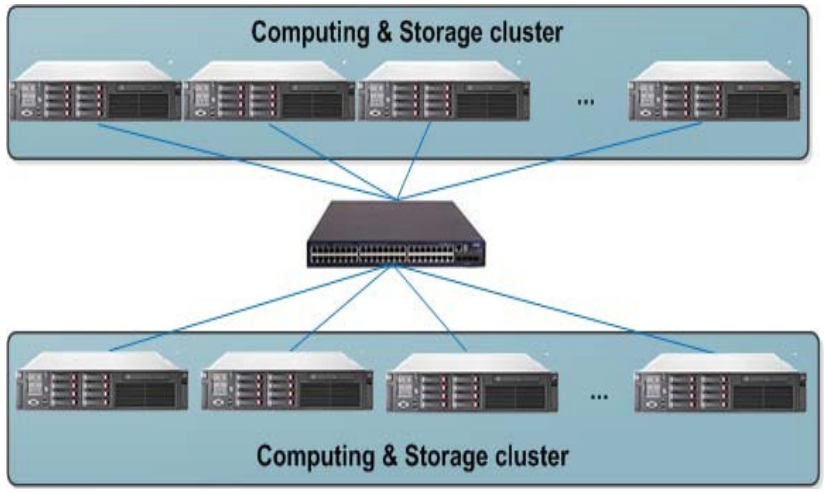
- Traditional computing architecture has certain limitations in scalability, fault tolerance and so on
  - Communication bottleneck: All data transmission pass through the central network switch
  - One IO server failure may cause storage system unavailable
- Network I/O becomes the bottleneck for data-intensive jobs
- More money should be devoted to purchase expensive facilities
- In the big data era, HEP experiments require new and intelligent computing technology

The New Architecture with Hadoop



Traditional architecture VS New architecture





Data to computation

Computation to data

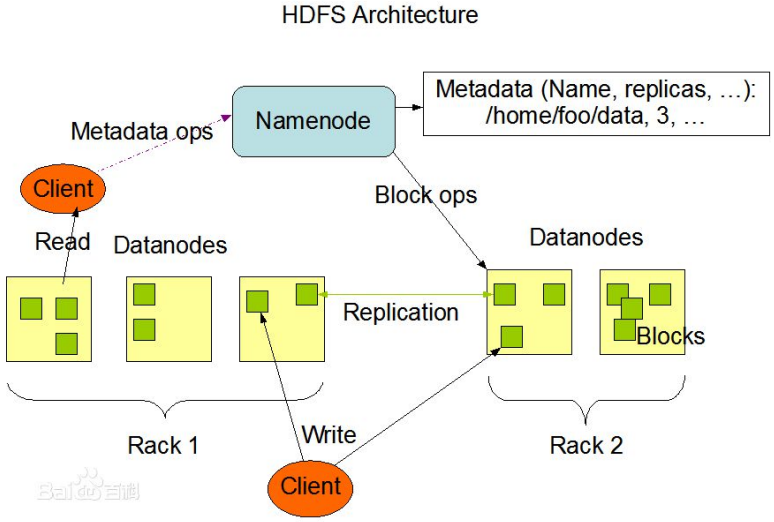
	Traditional architecture	New architecture
Network	10Gpbs	1Gpbs
Storage	Disk array	Local disk
Data access	Access through network, limited by network	Access local disk

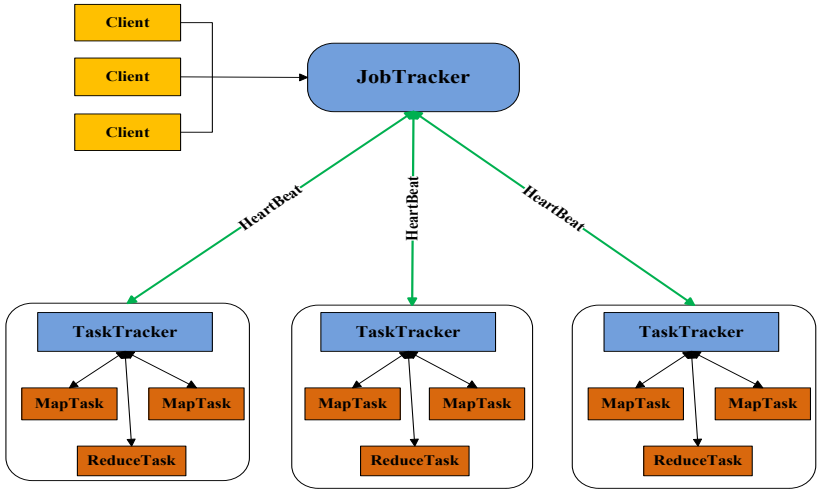
What is Hadoop?

Apache Hadoop

An open-source software framework for distributed storage and distributed processing huge amount of data sets

- A highly reliable distributed file system (HDFS)
- Parallel computing framework for large data sets(MapReduce)
- Some tools: HBase, Hive, Pig, Spark, etc
- Widely adopted in the Internet industry





Why Hadoop?

- ✓ High scalability
  - ✓ one master cluster can reach 4000 nodes
- ✓ IO intensive jobs achieve higher CPU efficiency
  - ✓ Local data read/write
- ✓ Lower cost
  - ✓ Without powerful network equipment
  - ✓ Without expensive disk arrays
- ✓ Some HEP experiments introduced Hadoop in scientific computing
- ✓ Widely used in industry, and commercial support is available from a number of companies
  - ✓ Three Hadoop software providers : Apache, Cloudera, Hortonworks
  - ✓ More than 150 companies are using

Challenges

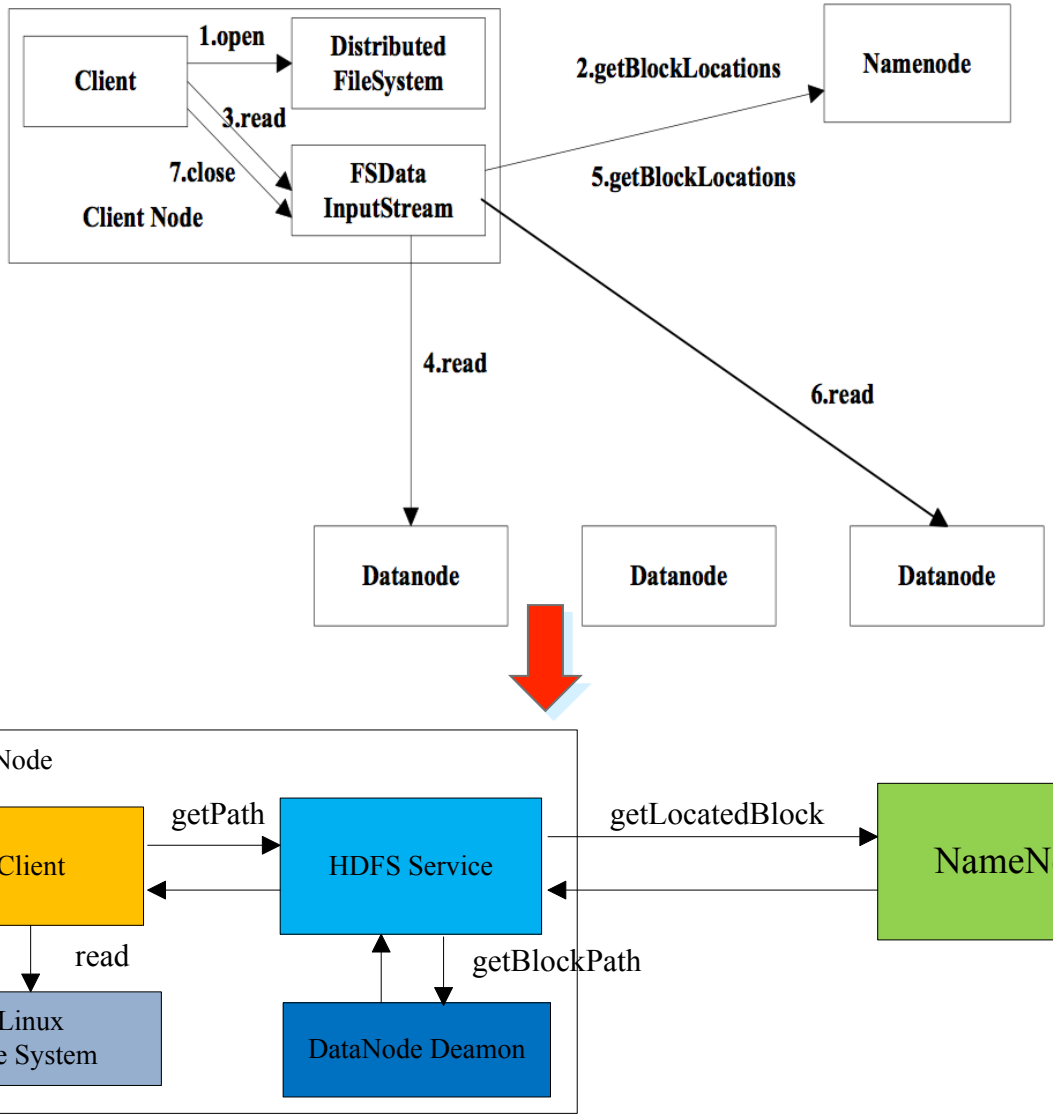
- Hadoop uses streaming access data, only support sequential write and append, not support random write
- Hadoop is written in Java, while C/C++ support is very limited
- HEP jobs read files via FUSE or other plugins
- HDFS fuse interface is not strong

Activities

- A new data access designed and implemented
  - Support random data access
  - Support files modification in HDFS
- Data migration system
  - Move data between HDFS and other storage systems
- User-friendly interface
  - Hide the underling details, to avoid learning mapreduce programming for users
  - Only one job option file needed according to the template

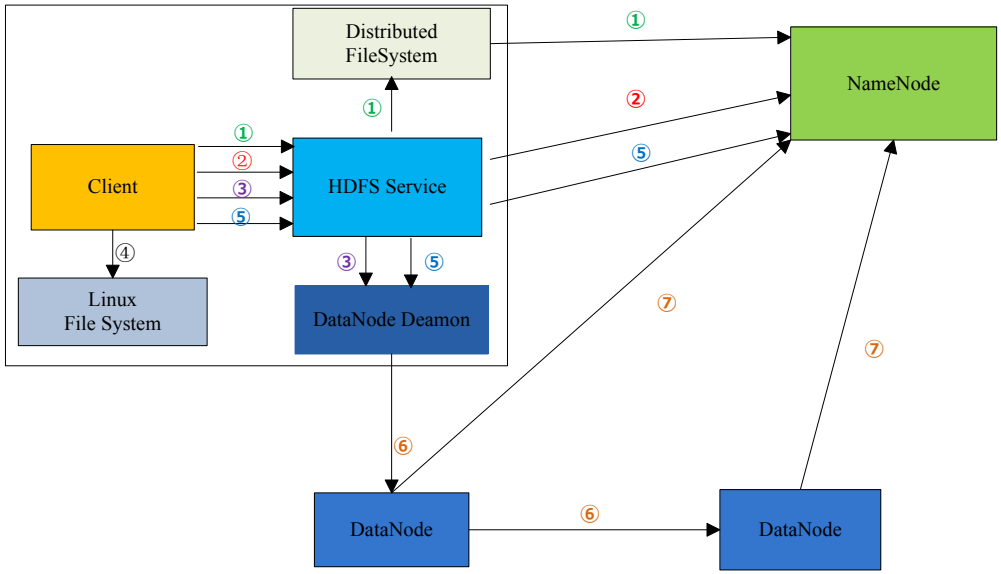
New data access: Read

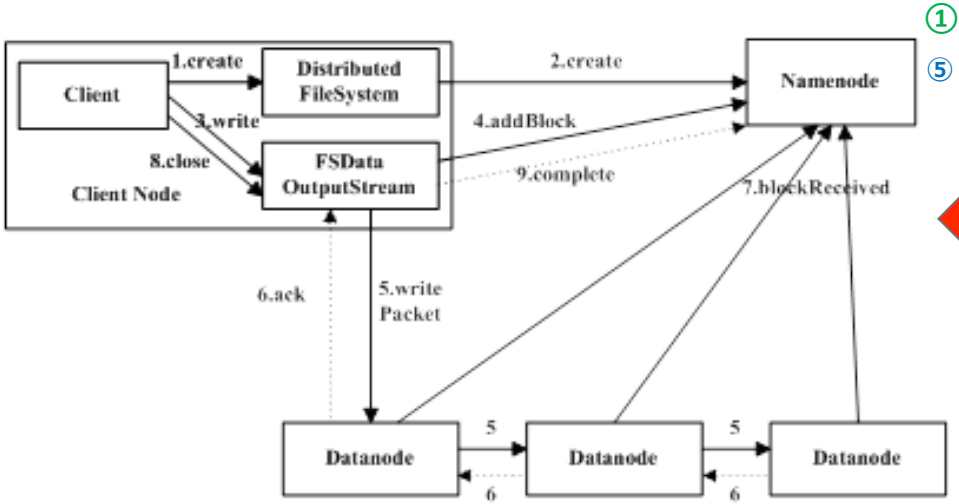
- One file one block
  - File in Root format cannot be divided
- Local read data(Job completely localized execution)
  - No data transmission
  - No network I/O
  - Low latency



New data access: Write

- ROOT API write to HDFS directly
- Local write if only have one replica
  - No data transmission
  - No network I/O
  - Low latency
- Support random write





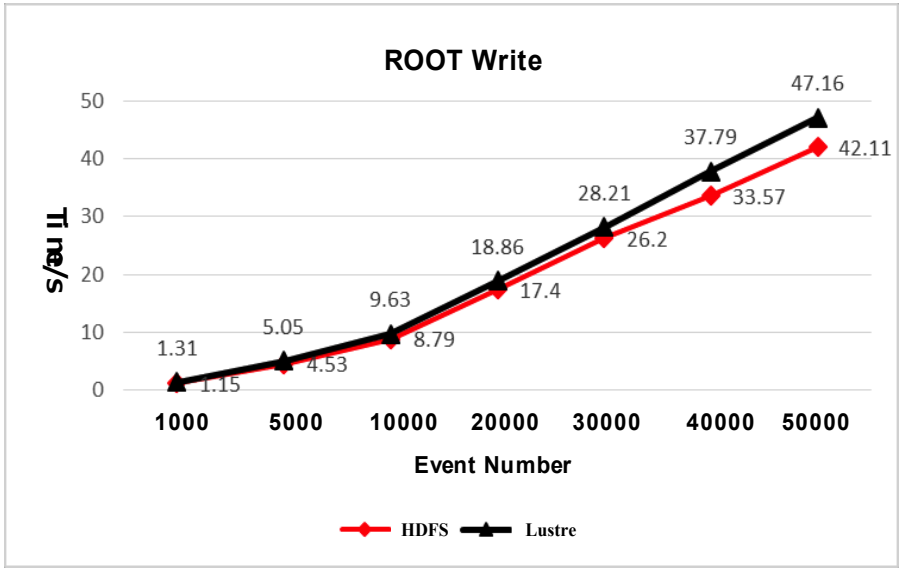
① createFile ② addBlock ③ getBlockPath ④ writeBlock
⑤ Complete ⑥ Copy ⑦ blockReceivedReport

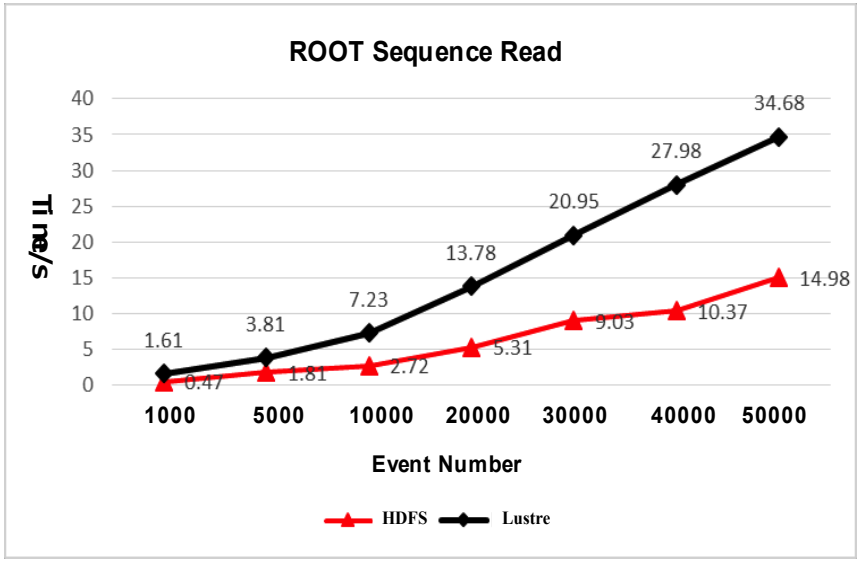
Test specification

- HDFS
  - ✓ 1 NameNode, 5 DataNode (6*6TBdisks, Raid5)
  - ✓ 1Gigabit Ethernet
- Lustre
  - ✓ 1 Metadata server, 5 OSS servers with 2 Disk Arrays(24*3TB,Raid6)
  - ✓ 10 Gigabit Ethernet

Performance Test

- ROOT tool
  - Root Write: $ROOTSYS/test/Event EventNumber 0 1 1
  - Root Read: $ROOTSYS/test/Event EventNumber 0 1 20

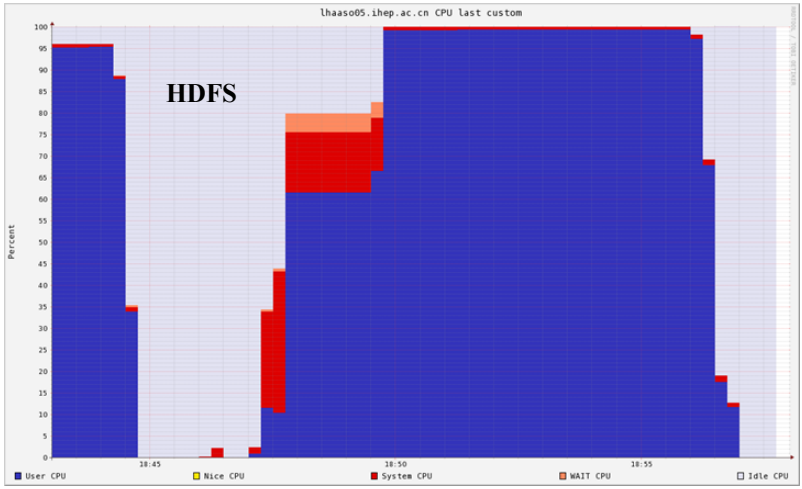


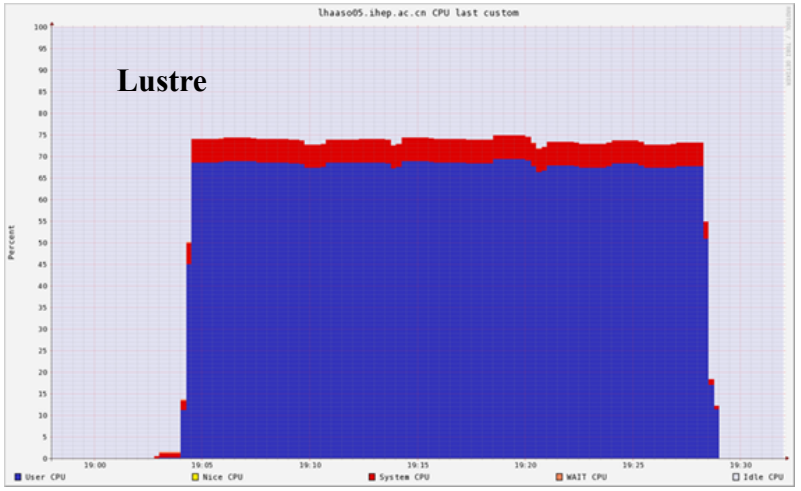


- Compared to Lustre, write event performance of HDFS improved 10% and read performance increased 2~3 times

Real job test

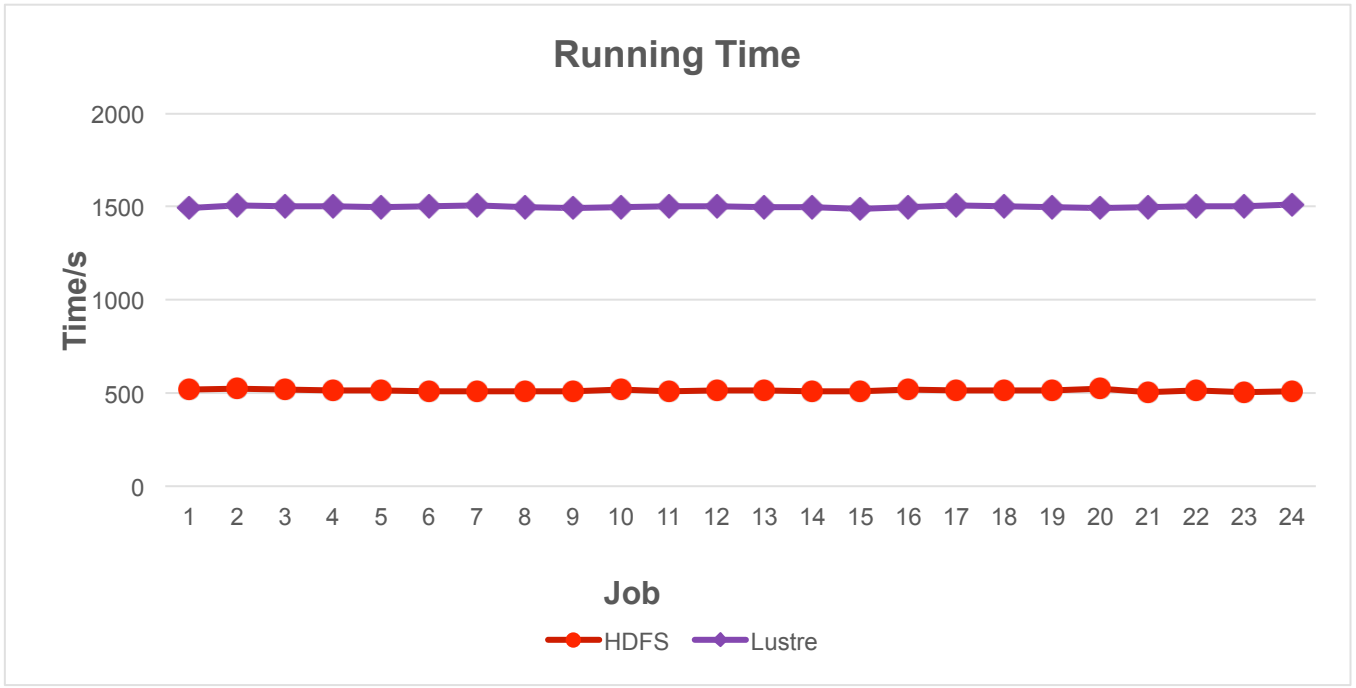
- Real job
  - Cosmic ray simulation job(corsika)
  - Detector simulation job(Geant4)
  - ARGO reconstruction job(medea++)
- Result and analysis
  - The CPU efficiency of CPU intensive job (corsika and Geant4) is up to 100%. The performance of HDFS and Lustre is comparable
  - The CPU efficiency of IO intensive job(medea++) is 100% with HDFS, while 67% with Lustre
  - IO intensive job needs large IO over network and the Lustre cient service consumes additional system overhead, which affect job execution





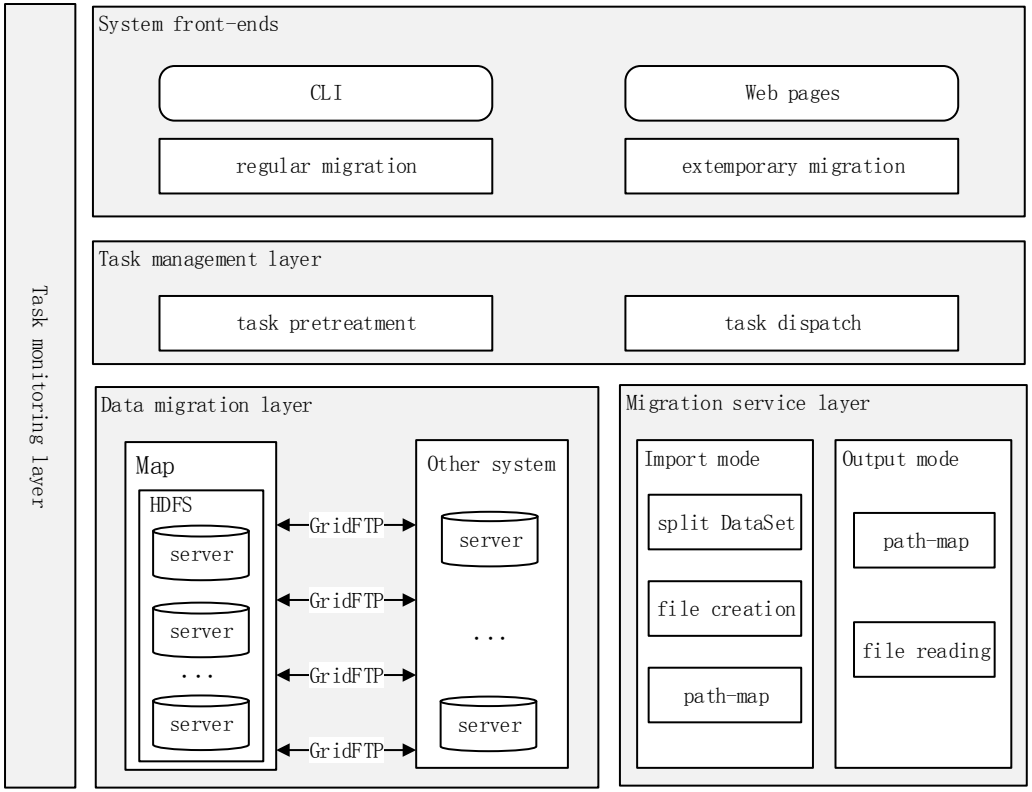
Real job test

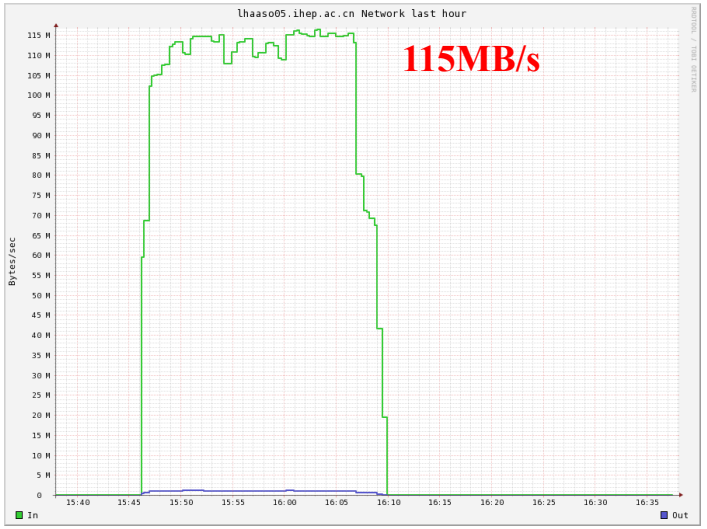
- Job execution time
  - Count the job execution time of medea++ job
  - Job running on HDFS is one third of Lustre



Data migration System

- A good supplement of Hadoop computing cluster in HEP
- provide import mode/output mode
- Move data between HDFS and other storage systems
- Mapreduce & GridFTP
  - High parallelism
  - The performance of data transfer is up to 115MB/s each datanode





User-friendly Interface

- Submit jobs

  hsub + queue + jobType+jobOptionFile + jobname

- Descriptions

  queue: queue name(ybj、default)

  jobTpye: MC(simulation job),REC(Reconstruction job), DA(Analysis job)

  jobOptionFile: Job option file

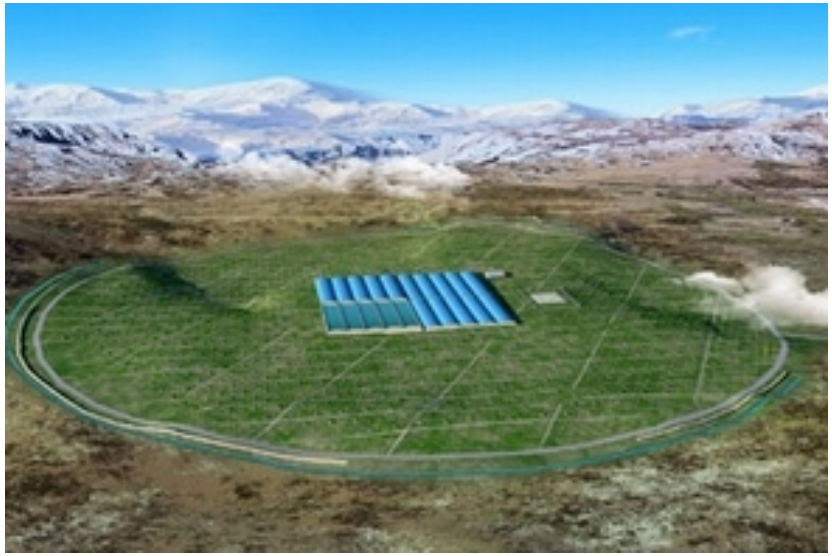
  jobname: job name

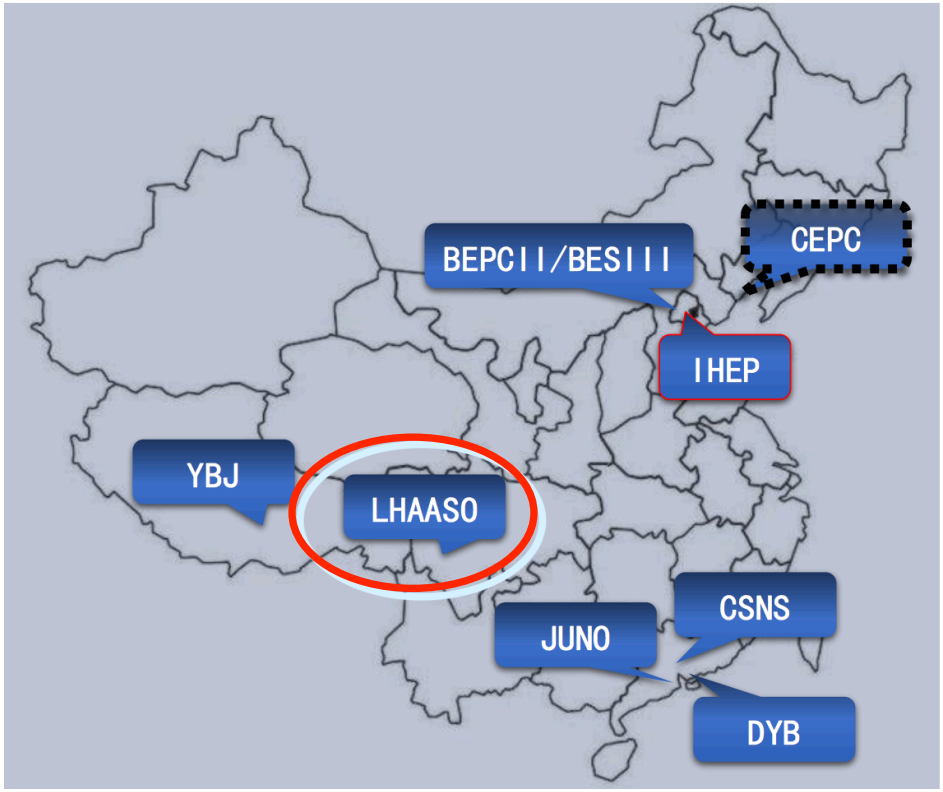
JobOptionFile Example

```
//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/
```

Successful Story in LHAASO

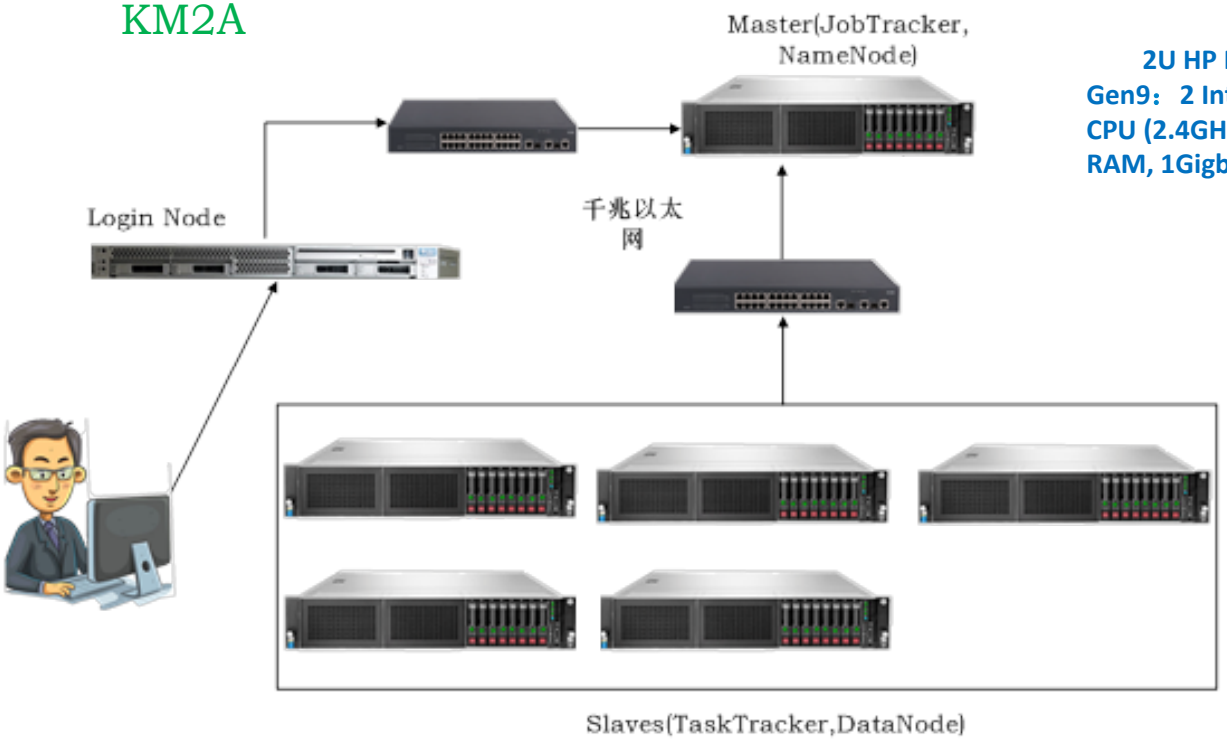
- LHAASO (Large High Altitude Air Shower Observatory)
  - Study the problems in Galactic cosmic ray physics
  - ~2PB raw data per year
  - Started to take data in 2018





Status in LHAASO

- Hadoop cluster
  - 5 Login nodes，1 Master node and 5 computing nodes，Link:1Gigbit
  - 120 CPU cores,140TB storage
  - Cosmic ray simulation(corsika), ARGO detector simulation(Geant4) and KM2A



2U HP ProLiant DL380 Gen9：2 Intel Xeon E5-2630 CPU (2.4GHz,8Cores)，64GB RAM, 1Gigbit

2U HP ProLiant DL380 Gen9：2 Intel Xeon E5-2680 CPU (2.5 GHz,12Cores), 64GB RAM，6*6TB disk, 1Gigbit

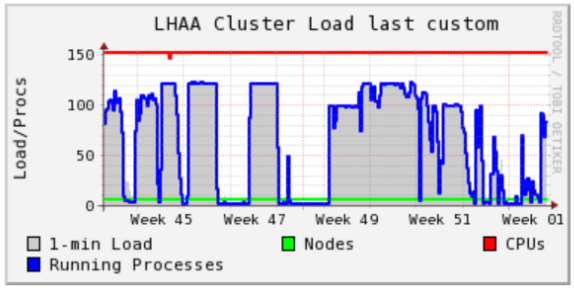
Status in LHAASO

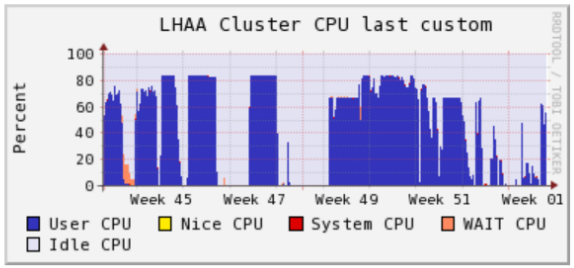
- Capacity
  - 119 TB used(88%)
- Job statistics (2017)
  - 20,225 jobs (502,341 tasks)
  - ~212,730 CPU hours

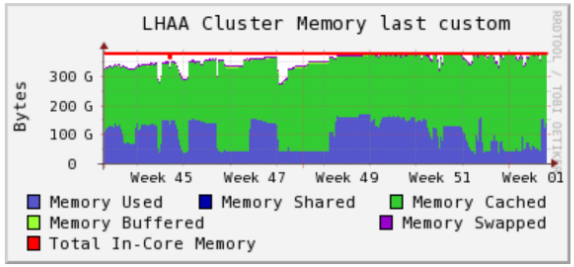
ExecHost	All Job Walltime(h)	All Job CPUTime(h)	Efficiency	All Job Sum		Failed Job WallTime(h)	Failed Job CpuTime(h)	Failed Job	Fail Rate	
ybjslc05.ihep.ac.cn	114064.083	212730.631	1.865	20225	Detail	0.000	0.000	3318	0.164	Detail
Tot/Ave	114064.083	212730.631		20225		0.000	0.000	3318	0.164	

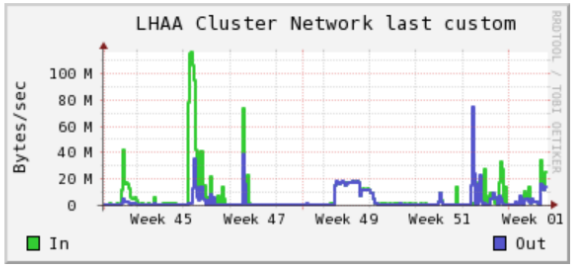
Cluster monitoring

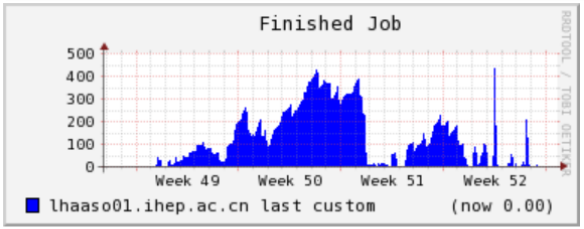
Overview of LHAA

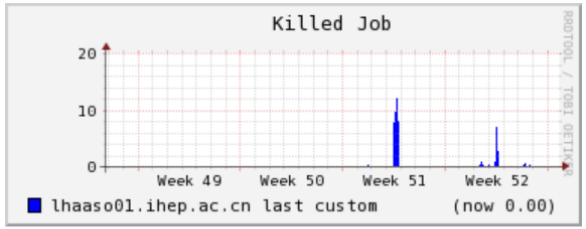


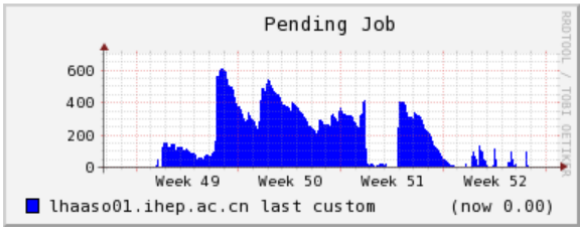


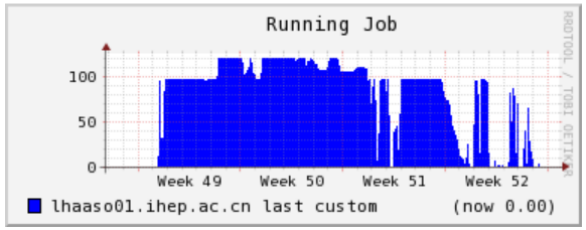






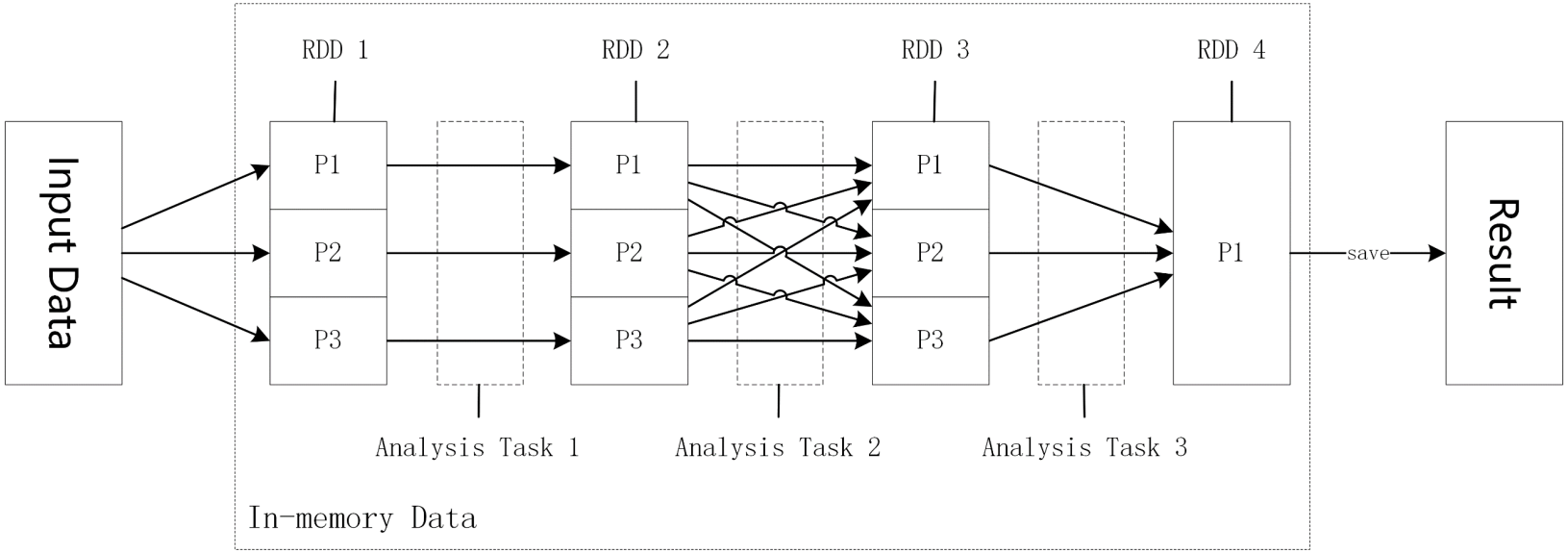






In memory computing in IHEP

- Start to introduce Spark into Partial Wave Analysis
- Study in-memory data-sharing mechanism based on Alluxio



Summary and Future

- Successfully applied in LHASSO experiment
- Reduce the cost of facilities
- Greatly improve the CPU efficiency of IO intensive jobs
- Data migration tool integrated into the exist Hadoop cluster for IHEP users
- Friendly interface are provided
- Plan to extend the solution to Ali experiment
- Plan to introduce Spark to HEP data analysis

Thank you!
Any questions?