Introduction of BESIII Computing Platform

On Behalf of IHEP-CC

Jingyan Shi

shijy@ihep.ac.cn

第八届 **BESIII R** 值与 **QCD** 强子结构研讨会

Outline

1 Brief overview of HEP Computing

2 BESIII Computing Platform

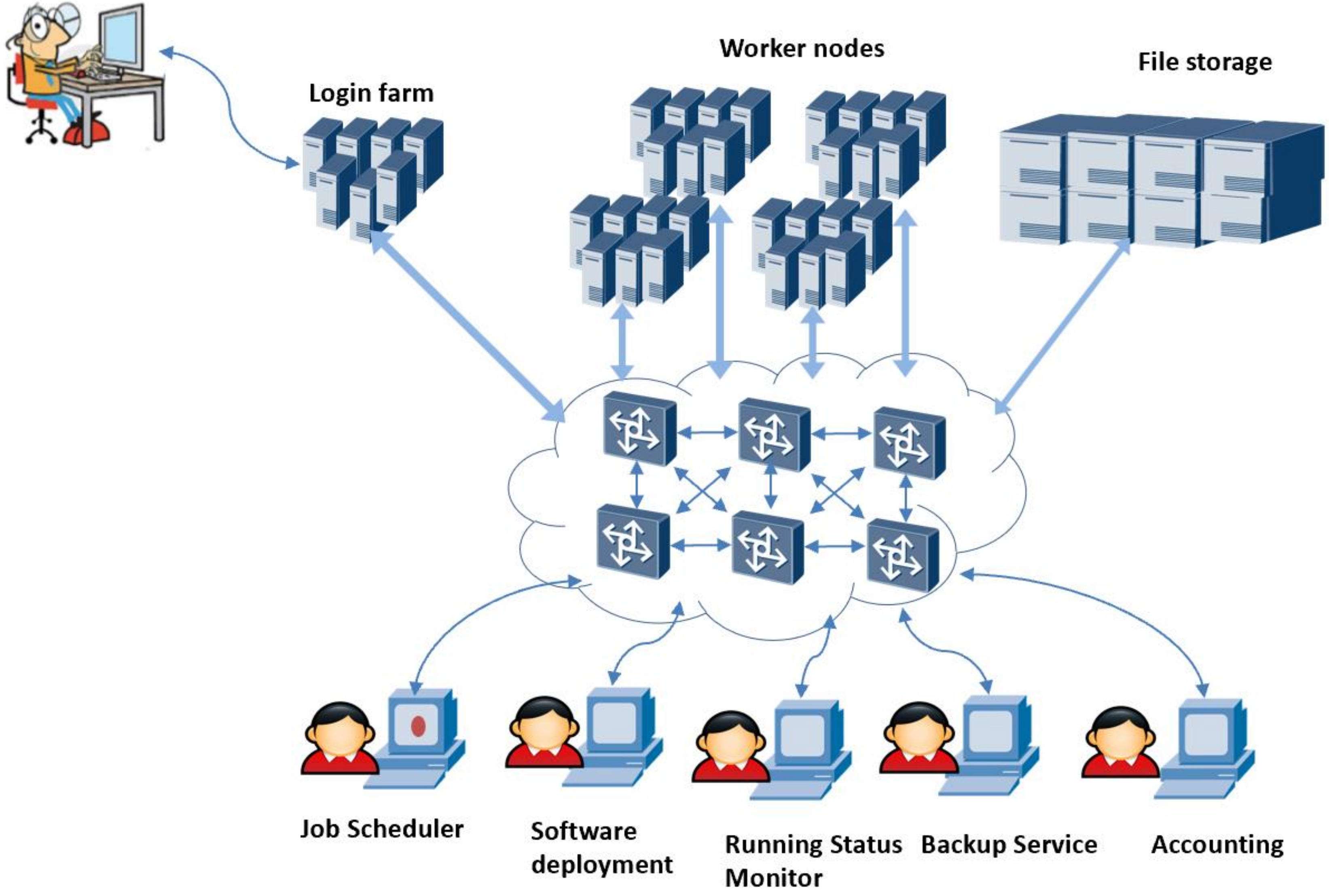
3 Use BESIII Computing Platform Efficiently

4 Summary

Quick View of HEP Computing

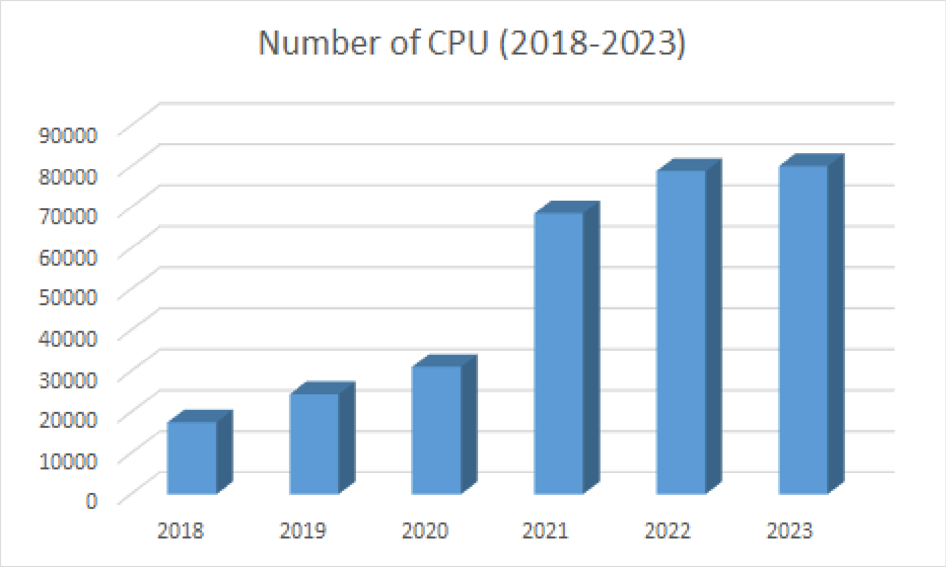
- The success of HEP Research depends on the development of computing technology
  - Big Data, HPC, HTC, AI4Science 。。。
- A powerful computing platform is critical for HEP Experiment
  - From online data acquisition to offline data analysis
- Different task needs different computing model
  - Computing-intensive, data intensive
- The development of HEP Computing platform has been driven by experimental needs
  - International collaboration → Grid computing
  - Large computing jobs volumes→ High throughput computing
  - Heavy IO → distributed file system (EOS)
  - .。。。。。。

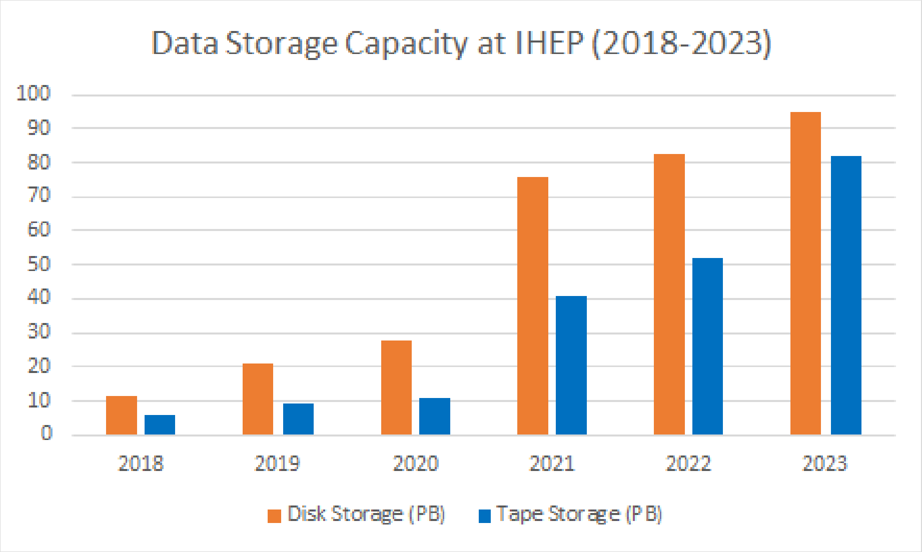
A Typical HEP Offline Computing Platform



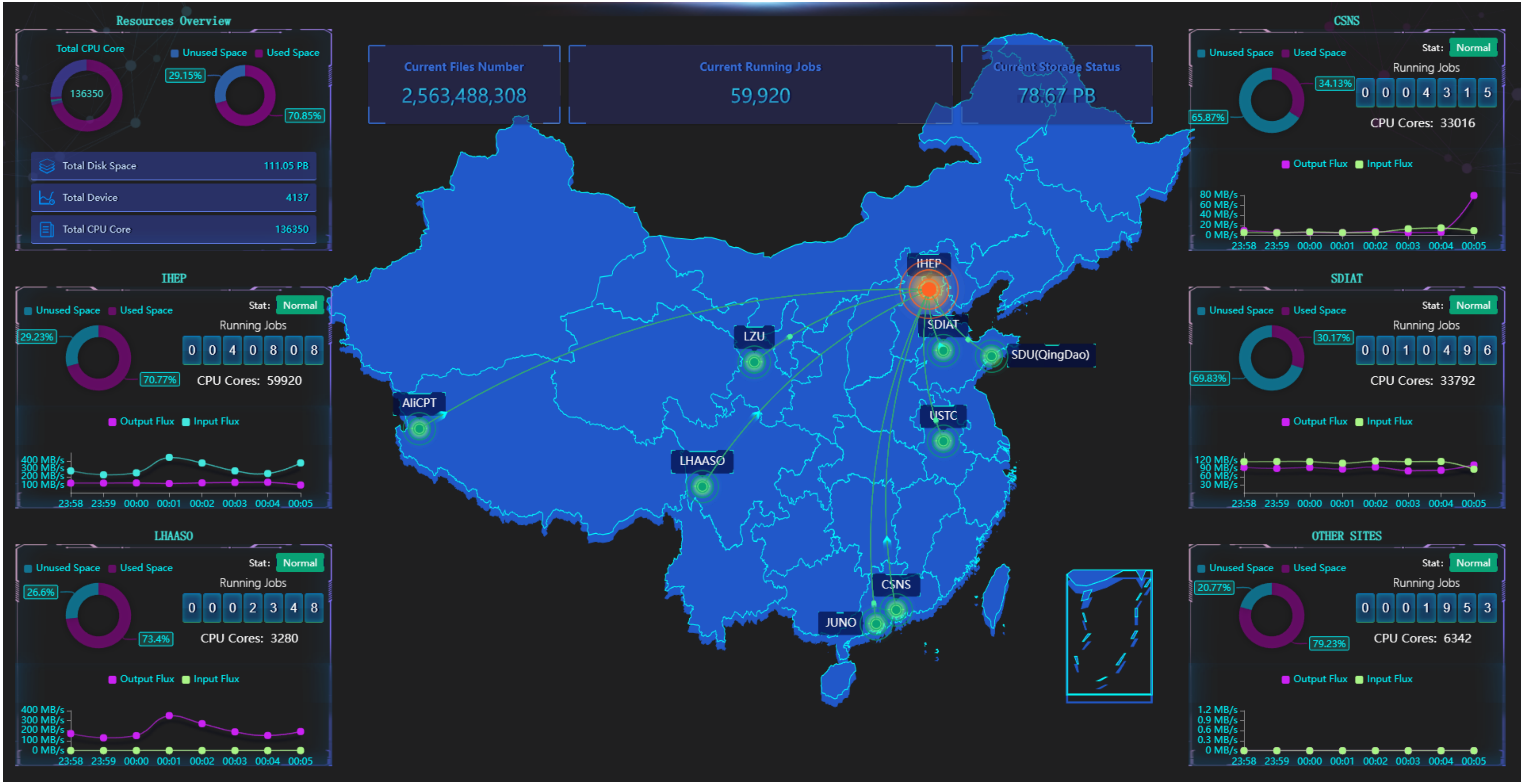
Computing Center of IHEP

- Distributed centers
  - Beijing, Huairou, Dongguan,
  - Daocheng, Jiangmen, …
- Provides and Supports:
  - HTC, HPC and Grid for 28 experiments / project
  - Data archive and sharing for HEP projects of China
- Quantity of resources grew exponentially
  - ~100K CPU cores
  - ~100 PB Disk Storage
  - ~137 PB Tape Storage





HEP Computing Platform – Multi Exp. and Multi Sites



Resources Overview

Total CPU Core 136350

🟦 Unused Space 🟪 Used Space

29.15%

70.85%

Total Disk Space 111.05 PB

Total Device 4137

Total CPU Core 136350

Current Files Number 2,563,488,308

Current Running Jobs 59,920

Current Storage Status 78.67 PB

IHEP

🟦 Unused Space 🟪 Used Space

Stat: Normal

Running Jobs 0 0 4 0 8 0 8

29.23%

70.77%

CPU Cores: 59920

🟪 Output Flux 🟦 Input Flux

LHAASO

🟦 Unused Space 🟪 Used Space

Stat: Normal

Running Jobs 0 0 0 2 3 4 8

26.6%

73.4%

CPU Cores: 3280

🟪 Output Flux 🟩 Input Flux

CSNS

🟦 Unused Space 🟪 Used Space

Stat: Normal

Running Jobs 0 0 0 4 3 1 5

34.13%

65.87%

CPU Cores: 33016

🟪 Output Flux 🟩 Input Flux

SDIAT

🟦 Unused Space 🟪 Used Space

Stat: Normal

Running Jobs 0 0 1 0 4 9 6

30.17%

69.83%

CPU Cores: 33792

🟪 Output Flux 🟩 Input Flux

OTHER SITES

🟦 Unused Space 🟪 Used Space

Stat: Normal

Running Jobs 0 0 0 1 9 5 3

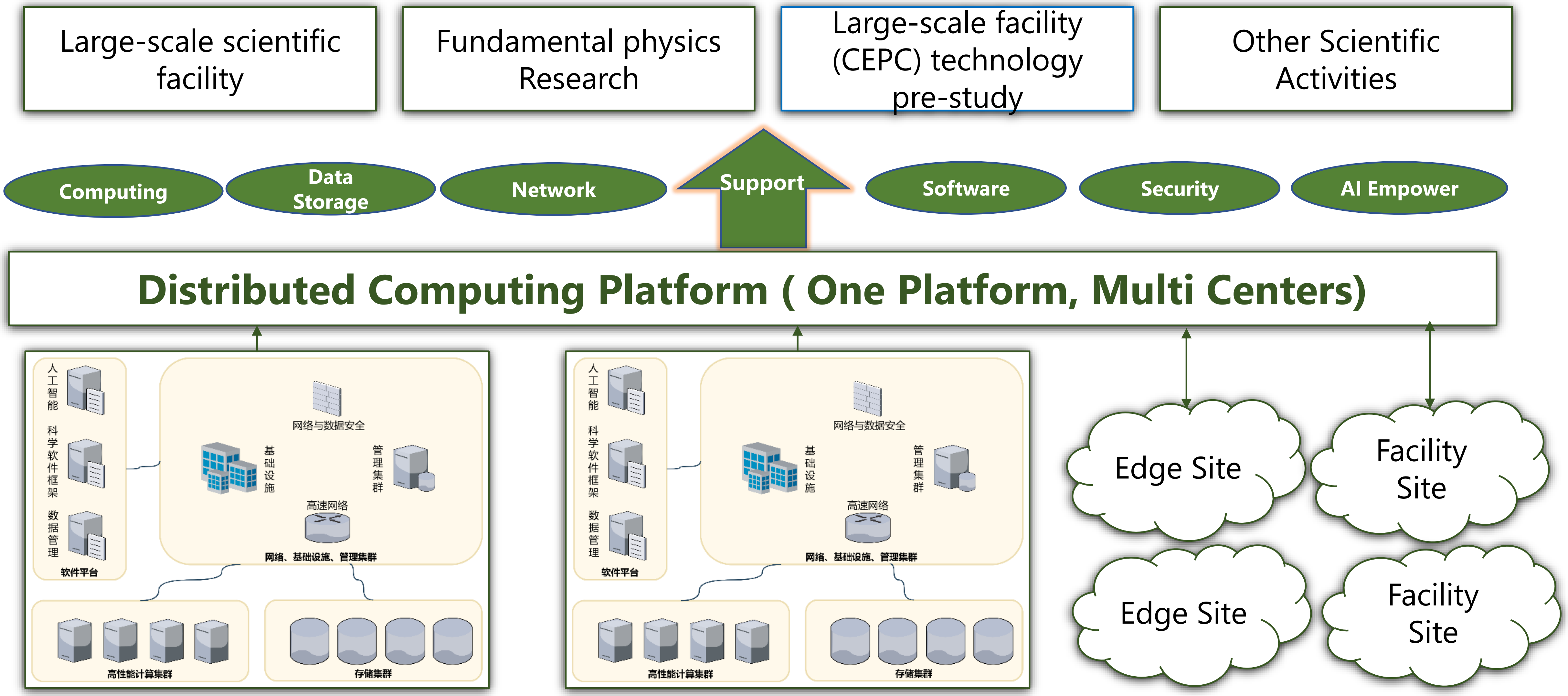
20.77%

79.23%

CPU Cores: 6342

🟪 Output Flux 🟩 Input Flux

Computing Platform Supports Science Research



Large-scale scientific facility

Fundamental physics Research

Large-scale facility (CEPC) technology pre-study

Other Scientific Activities

Computing

Data Storage

Network

Support

Software

Security

AI Empower

Distributed Computing Platform (One Platform, Multi Centers)

Edge Site

Facility Site

Edge Site

Facility Site

Data Center of CSNS at Dongguan

Computing Center of IHEP at Beijing

Outline

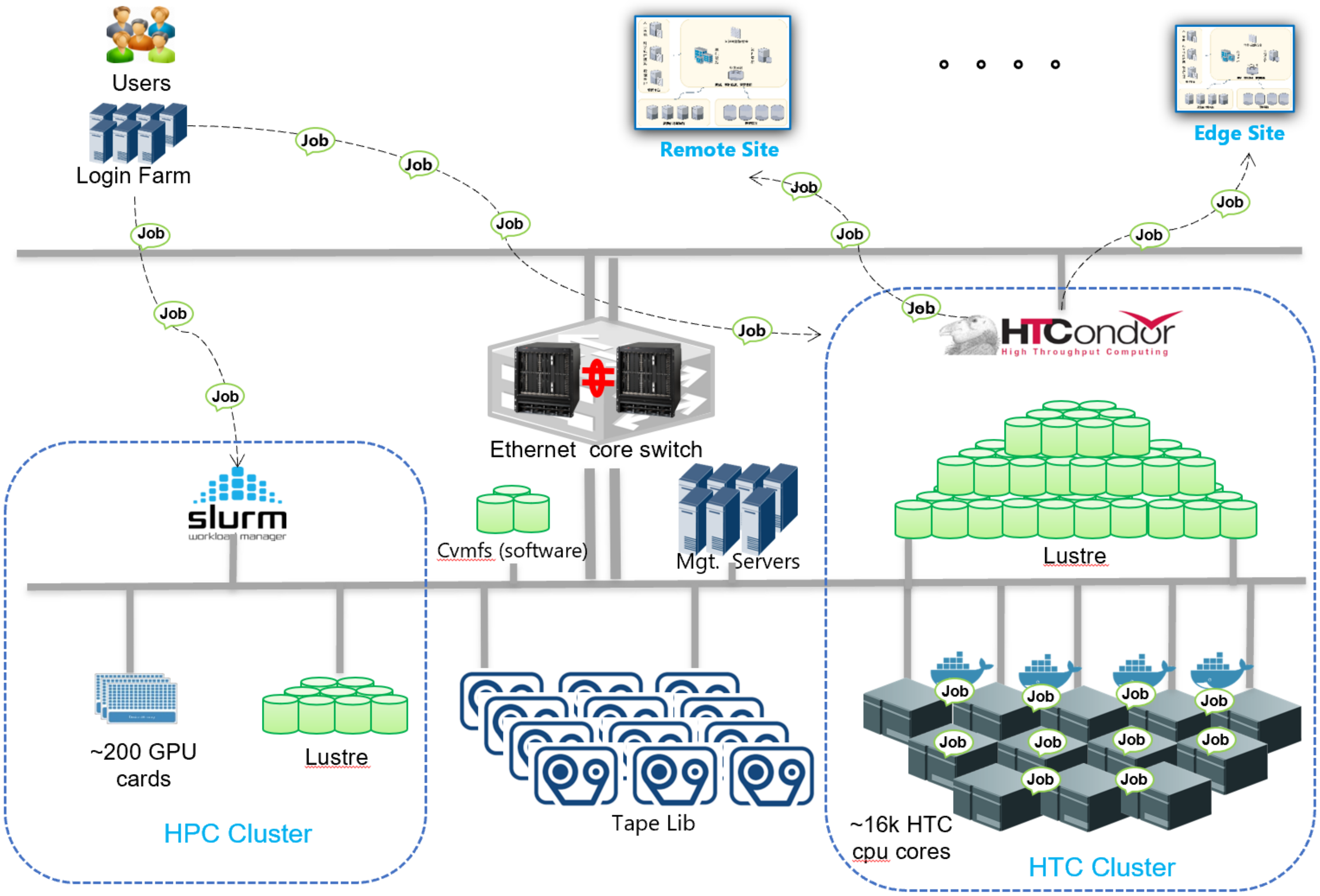
1 Brief overview of HEP Computing

2 **BESIII Computing Platform**

3 Use BESIII Computing Platform Efficiently

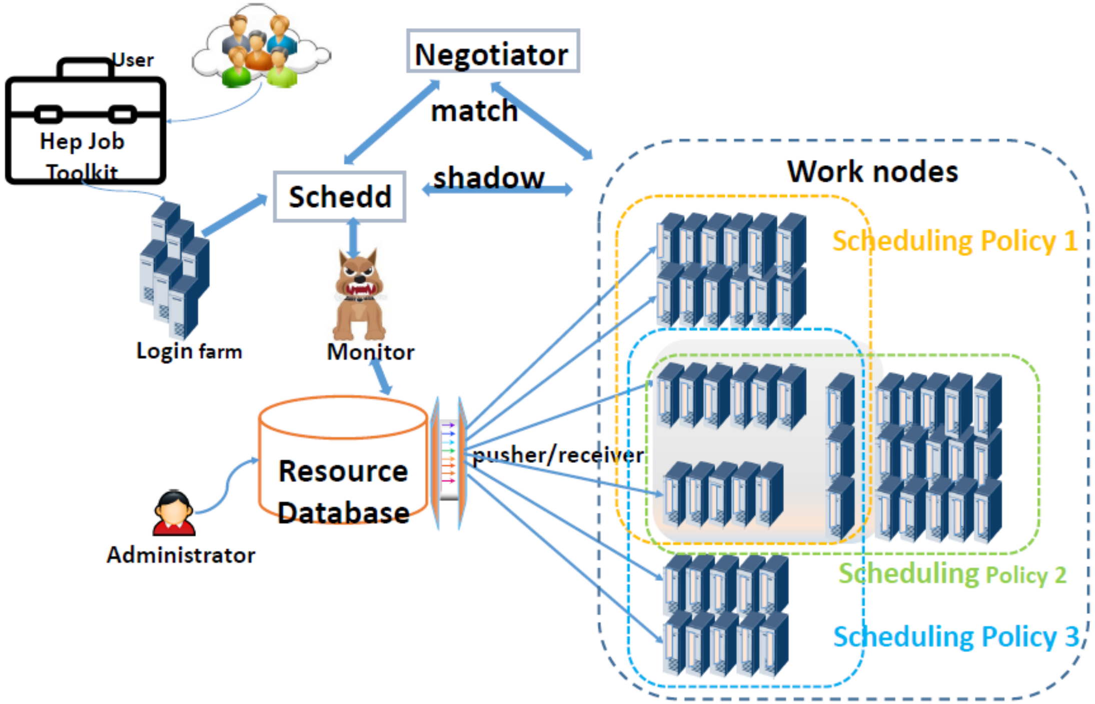
4 Summary

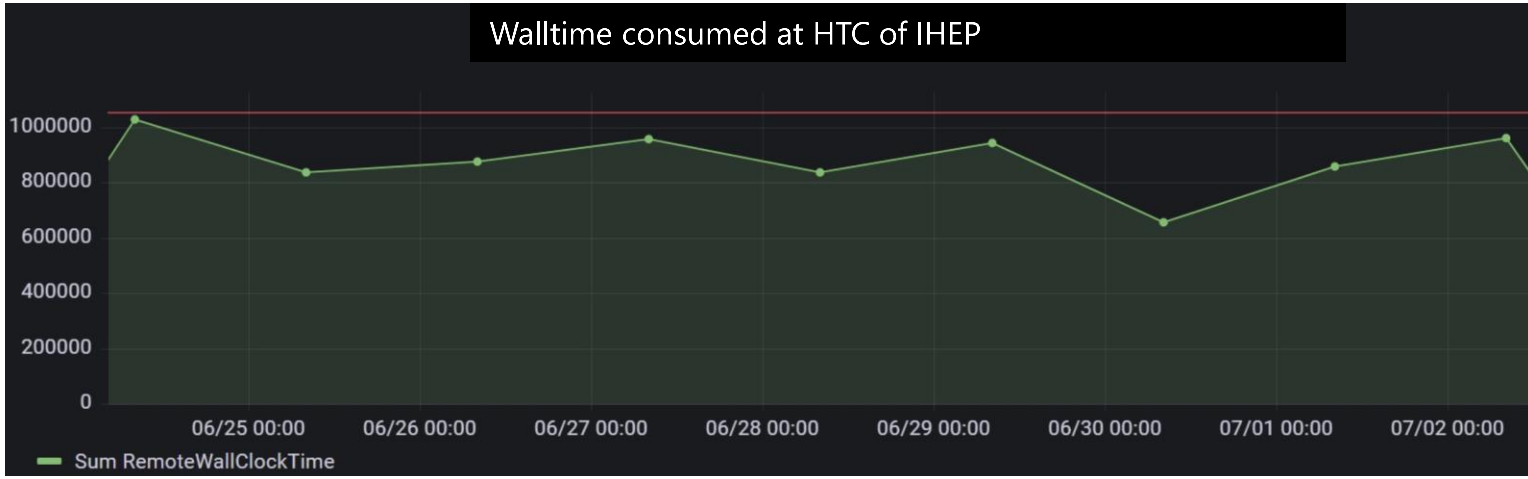
BESIII Computing Platform



Provides More CPU Times Inside HTC Cluster

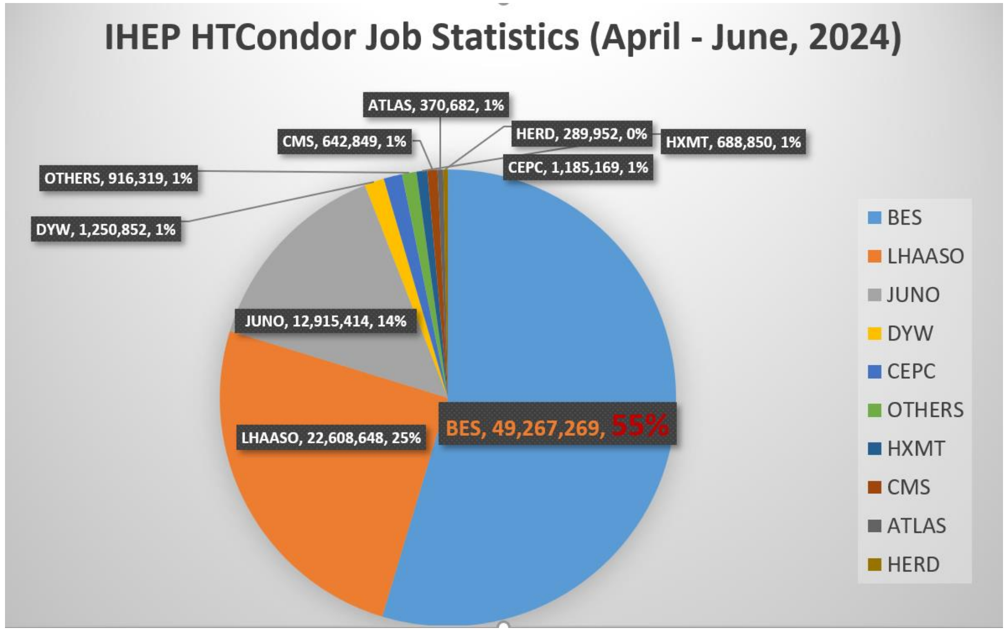
- “Resource Sharing Pool” at local HTC cluster: ~40k CPU cores
  - CPUs contributed by all Exp.
  - Let the jobs from busy Exp. run on the job slots of unbusy Exp.
  - Fairshare policy guarantee the higher priority for the unbusy Exp. jobs
  - Monitor tool developed guarantees the quick error reaction
- >85% CPU utilization and stable worker nodes

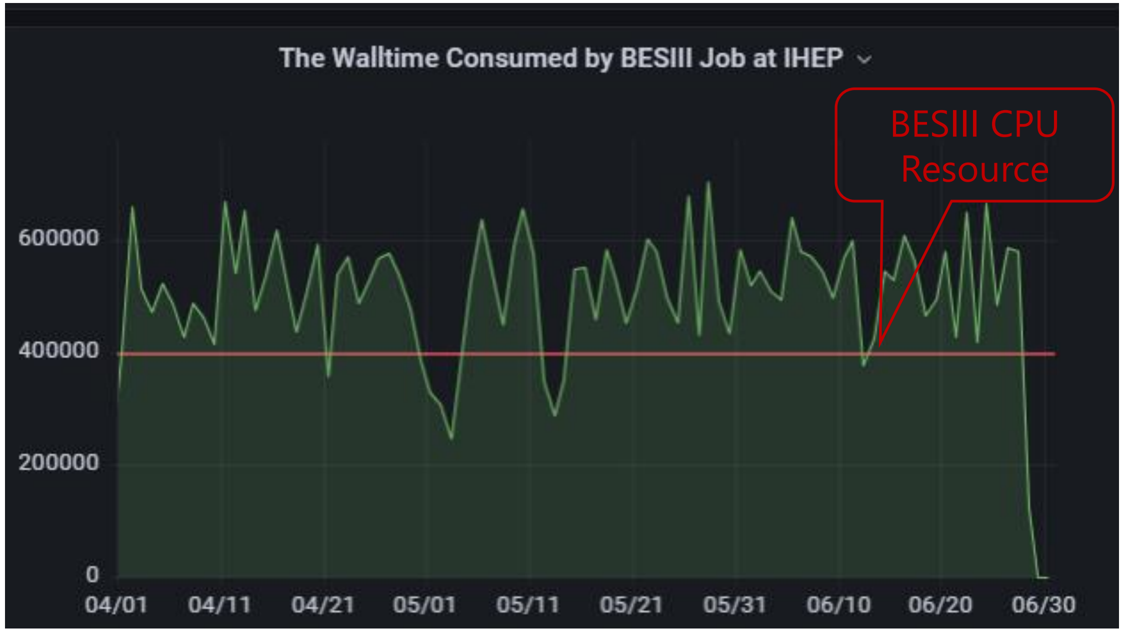




Job Statistics of HTCondor Cluster

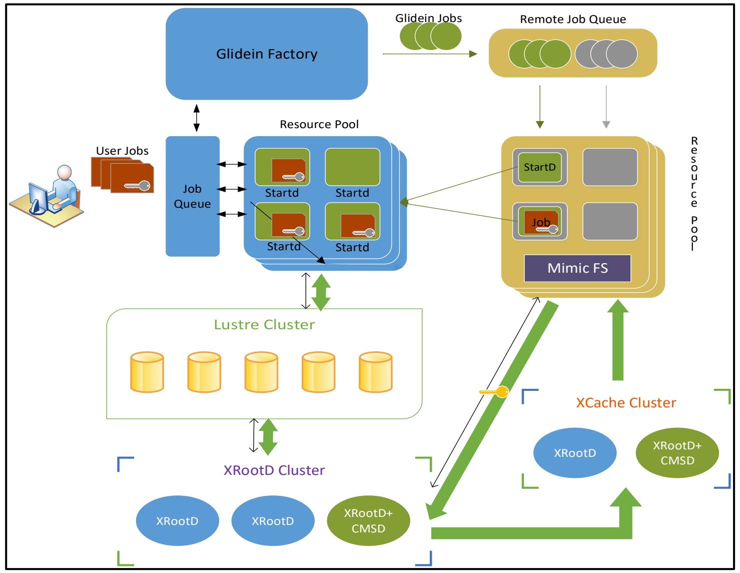
- IHEP HTCondor Local cluster serves 18 Experiments and Applications
  - The amount of CPU cores: 39,972
    - BESIII contributed 16,568 CPU cores, 41.4%
- Job Statistics of IHEP HTCondor from April to June, 2024
  - Shared pool provides BESIII extra 36% CPU time





Distributed Computing -- Local Cluster Expansion

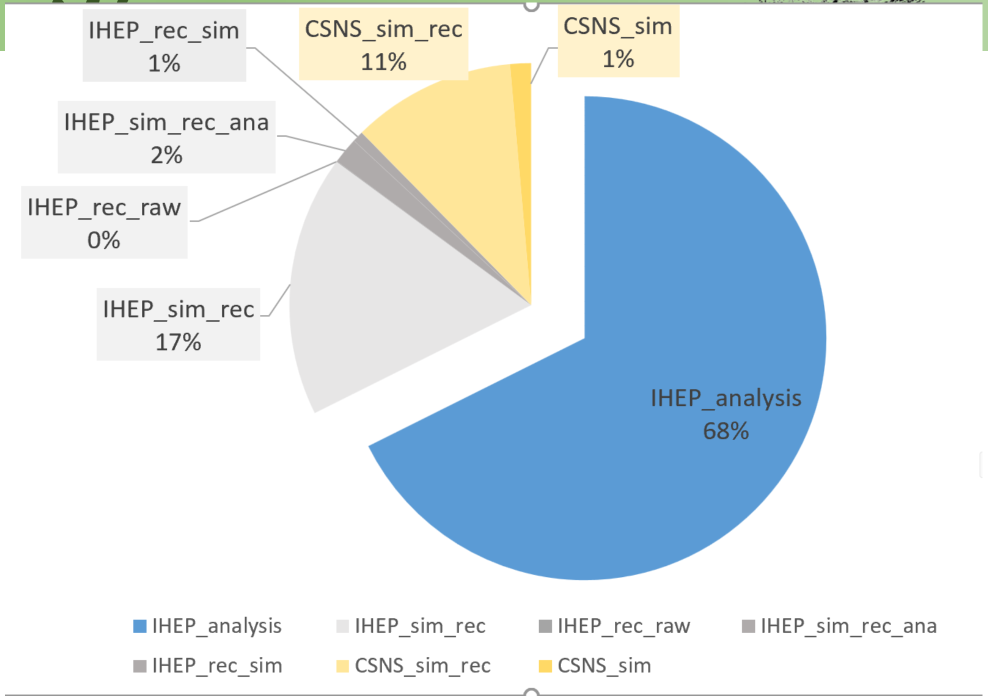
- Local cluster is the main place of the data processing for some Exp.
- Local cluster expansion
  - IHEP-centered, and Computing resource extension on-demand
  - Classification to jobs and sites
    - Dispatch the suitable jobs to the suitable remote site
  - Transparent data access / transfer
    - Token-based user authentication
- Keep the original user cluster way



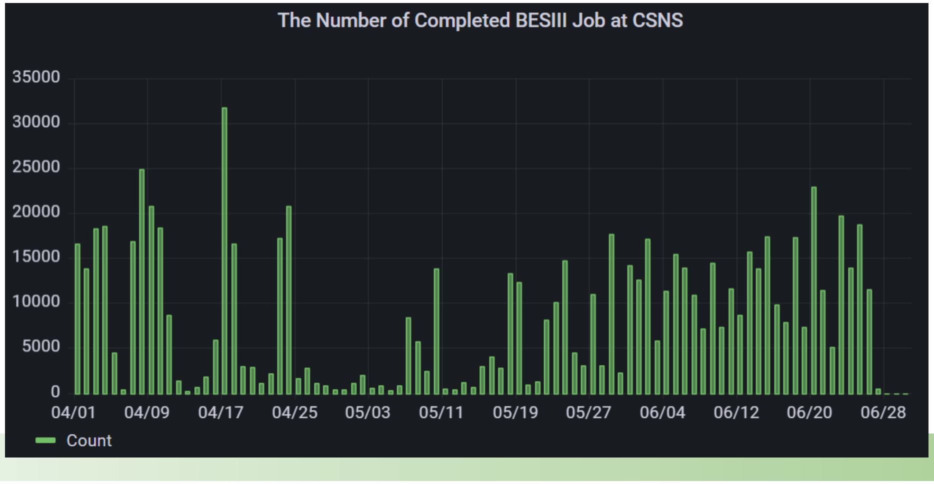
Design of Local Cluster Expansion

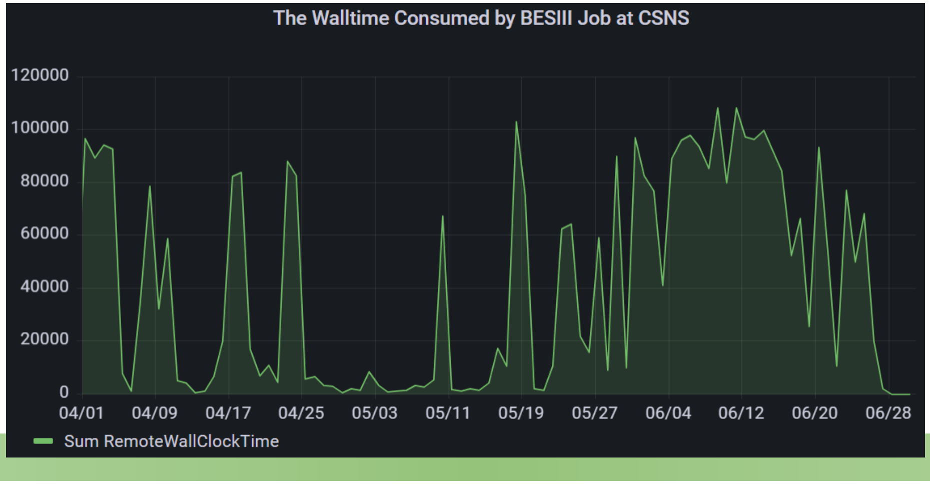
BESIII Cluster Expansion (Ongoing)

- Aim: Dispatch more jobs run on more remote resource
- Study and development keeps going
  - Dispatch 20%-30% BES jobs to the remote resource transparently
    - Simulation and reconstruction
- Sim and small part of rec jobs have been run remotely.
  - Statistics of BESIII jobs submitted last three months
    - Completed dHTC Jobs: 2,495,129
    - Consumed CPU Hours: 14,161,868 7.6% of CPU time of BESIII jobs at IHEP cluster
- More study focus on:
  - Random trigger access
  - How to use the resource inside close network
  - Performance optimization



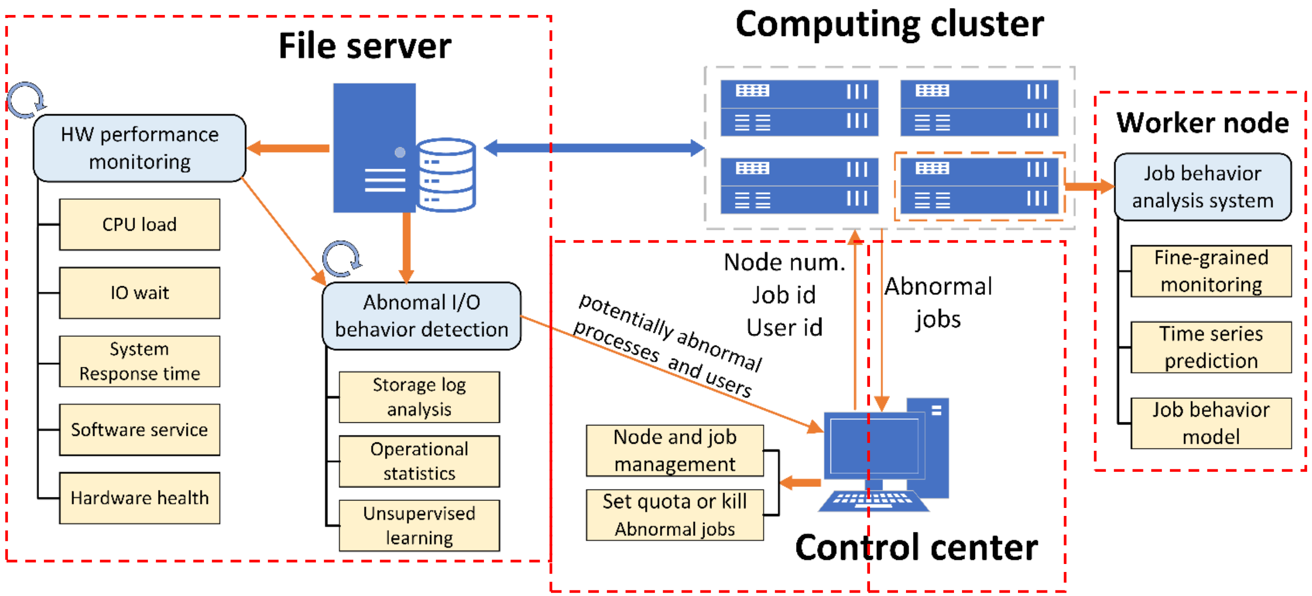
Statistics of BESIII jobs in 2024

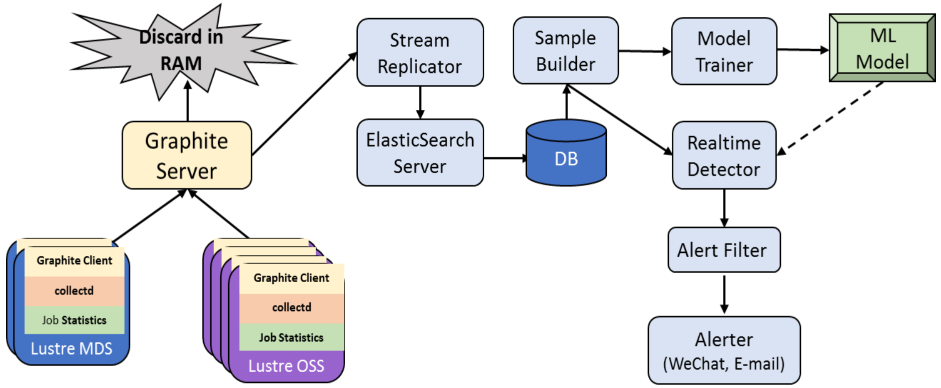

The Number of Completed BESIII Job at CSNS


The Walltime Consumed by BESIII Job at CSNS

The Intelligent Operation System (ongoing)

- Detect abnormal I/O behavior of user processes from the file system
  - Near real-time statistics of the file system resources consumed by user processes
  - Score and rank the user's I/O behaviors
    - Using unsupervised machine learning algorithm
  - Identify the potentially abnormal worker nodes
- Analysis job behavior from the worker node
  - Identify the abnormal job inside potentially abnormal worker nodes
- Adjust the available resource scale for each user dynamically
  - Limit resource usage of abnormal user





Outline

1 Brief overview of HEP Computing

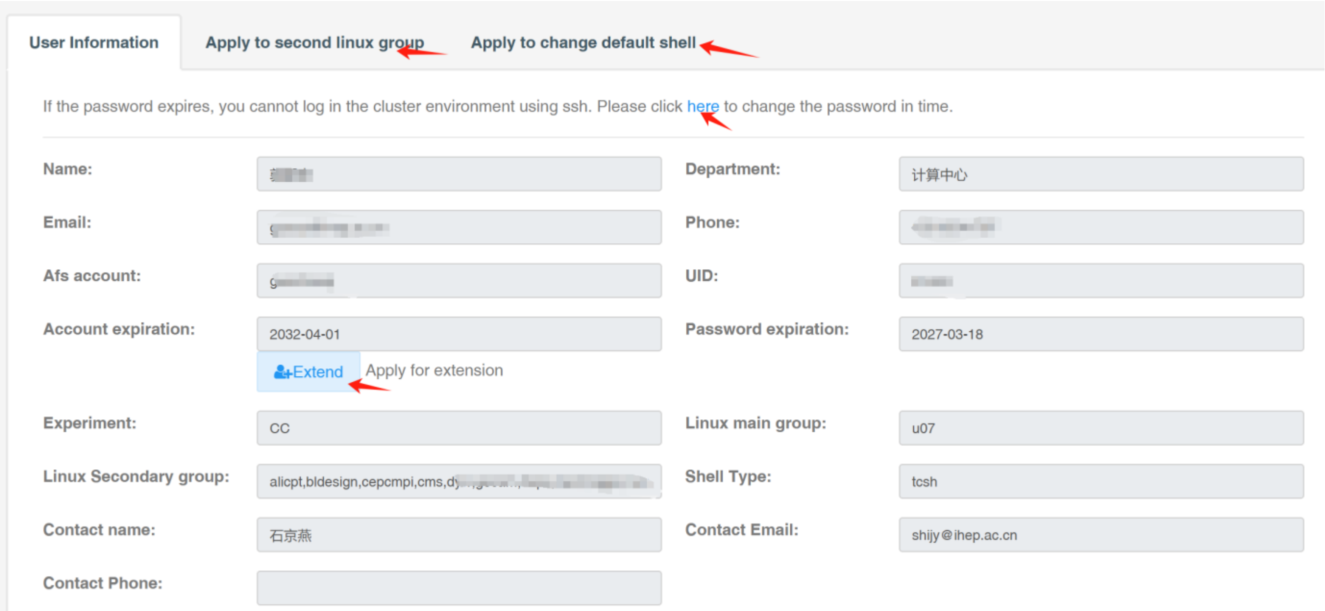
2 BESIII Computing Platform

3 **Use BESIII Computing Platform Efficiently**

4 Summary

Self Services Provided

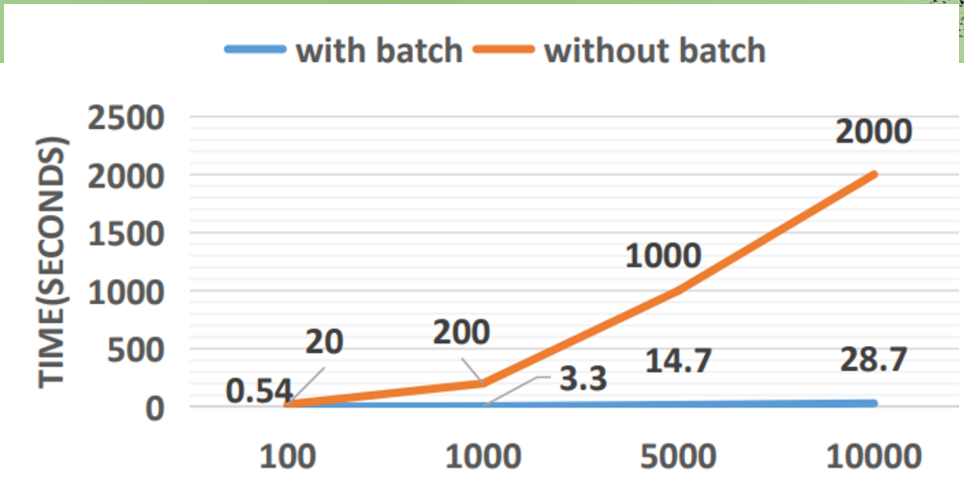
- Password is same as the one of IHEP SSO
- User dashboard:
    - http://ccsinfo.ihep.ac.cn
        - Job and storage statistics
        - Self services
            - Account extension
            - Default Bash change
            - Secondary group apply
- Helpdesk: helpdesk@ihep.ac.cn



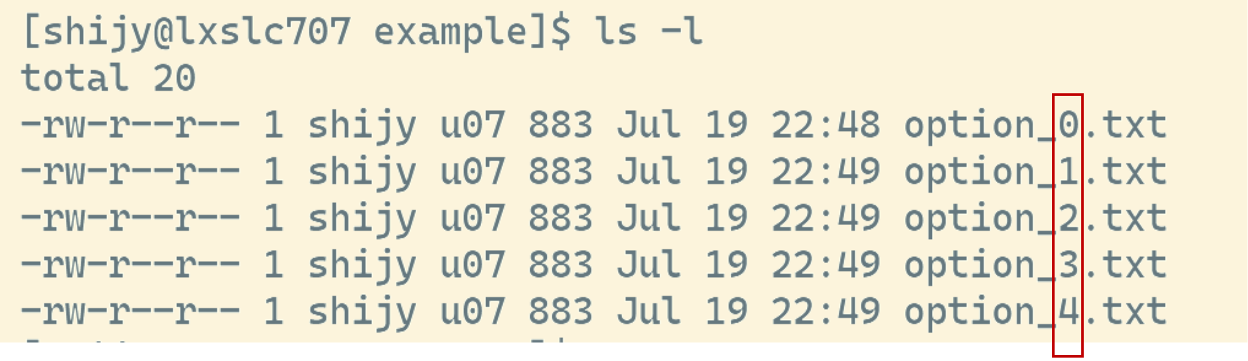
If the password expires, you cannot log in the cluster environment using ssh. Please click here to change the password in time.

Name:	[redacted]	Department:	计算中心
Email:	[redacted]	Phone:	[redacted]
Afs account:	[redacted]	UID:	[redacted]
Account expiration:	2032-04-01 Extend Apply for extension	Password expiration:	2027-03-18
Experiment:	CC	Linux main group:	u07
Linux Secondary group:	alicpt,bldesign,cepcmpi,cms,d[redacted]	Shell Type:	tcsh
Contact name:	石京燕	Contact Email:	shijy@ihep.ac.cn
Contact Phone:			

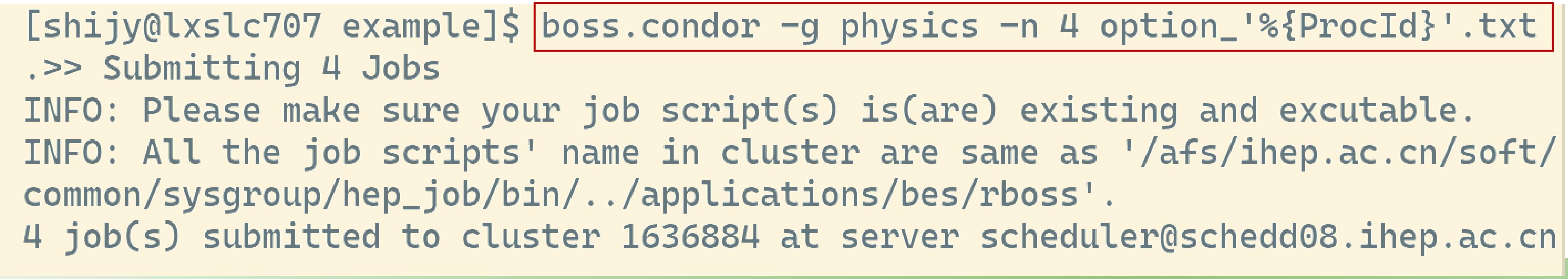
批量作业提交 (1 / 2)



- 非常推荐批量作业提交
  - 大部分BESIII作业可以“批量作业”
  - 可以极大减少用户作业提交时间
  - 可以极大减轻调度器的负载压力
- 简单的批量作业提交示例
  - Option文件名字格式相同:
  - %{ProcId}用于替换升序数字



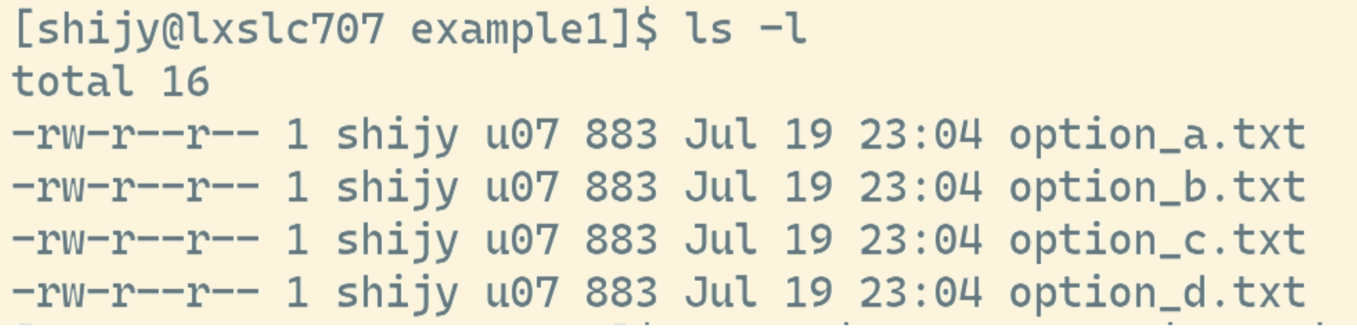
```
[shijy@lxslc707 example]$ ls -l
total 20
-rw-r--r-- 1 shijy u07 883 Jul 19 22:48 option_0.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 22:49 option_1.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 22:49 option_2.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 22:49 option_3.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 22:49 option_4.txt
```



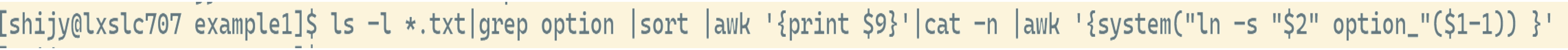
```
[shijy@lxslc707 example]$ boss.condor -g physics -n 4 option_'%{ProcId}'.txt
.>> Submitting 4 Jobs
INFO: Please make sure your job script(s) is(are) existing and excutable.
INFO: All the job scripts' name in cluster are same as '/afs/ihep.ac.cn/soft/
common/sysgroup/hep_job/bin/../applications/bes/rboss'.
4 job(s) submitted to cluster 1636884 at server scheduler@schedd08.ihep.ac.cn
```

批量作业提交 (1 / 2)

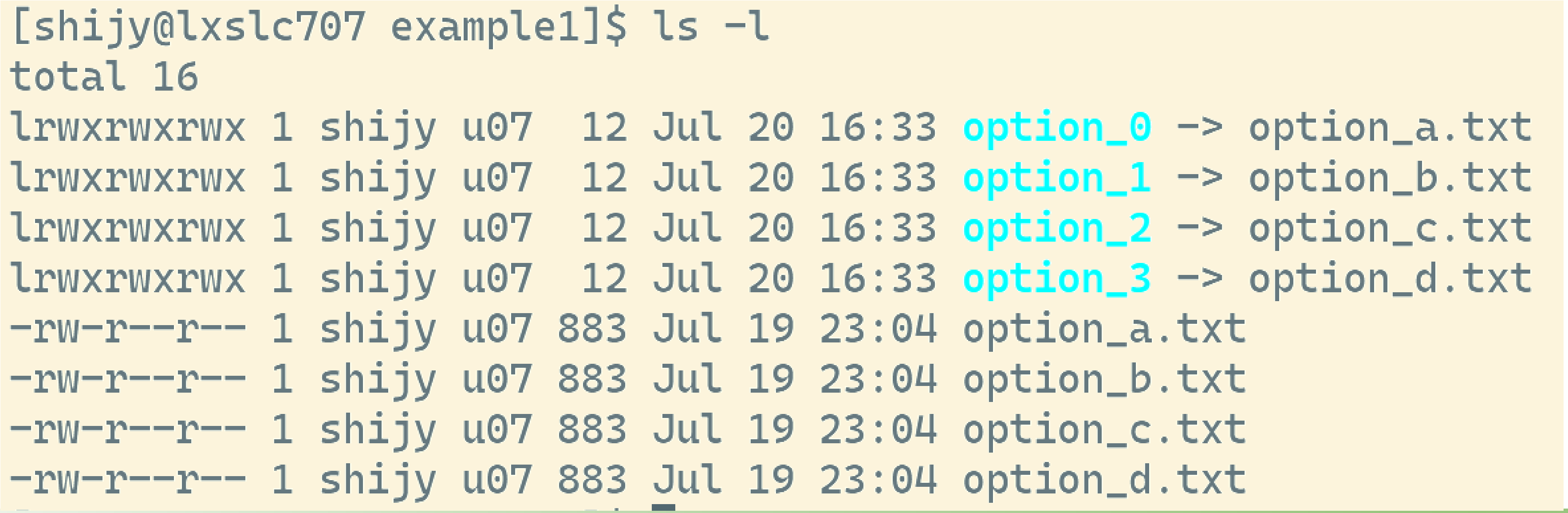
- 同格式，非数字的option文件，可以增加有数字的链接文件后再提交



```
[shijy@lxslc707 example1]$ ls -l
total 16
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_a.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_b.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_c.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_d.txt
```



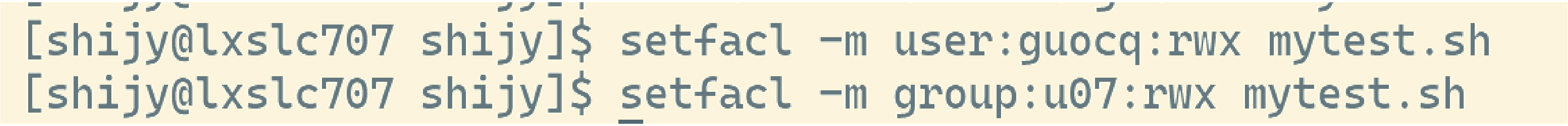
```
[shijy@lxslc707 example1]$ ls -l *.txt|grep option |sort |awk '{print $9}'|cat -n |awk '{system("ln -s "$2" option_"($1-1)) }'
```



```
[shijy@lxslc707 example1]$ ls -l
total 16
lrwxrwxrwx 1 shijy u07  12 Jul 20 16:33 option_0 -> option_a.txt
lrwxrwxrwx 1 shijy u07  12 Jul 20 16:33 option_1 -> option_b.txt
lrwxrwxrwx 1 shijy u07  12 Jul 20 16:33 option_2 -> option_c.txt
lrwxrwxrwx 1 shijy u07  12 Jul 20 16:33 option_3 -> option_d.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_a.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_b.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_c.txt
-rw-r--r-- 1 shijy u07 883 Jul 19 23:04 option_d.txt
```

文件存储

- 用户可以自己授权，向指定人员开放目录/文件的访问
  - http://afsapply.ihep.ac.cn/cchelp/zh/local-cluster/storage/Lustre/#%E8%AE%BE%E7%BD%AE%E7%9B%AE%E5%BD%95%E7%9A%84acl%EF%BC%9A



```
[shijy@lxslc707 shijy]$ setfacl -m user:guocq:rwx mytest.sh
[shijy@lxslc707 shijy]$ setfacl -m group:u07:rwx mytest.sh
```

- 尽量避免在单个目录下存放过多的数据文件
  - 建议单目录下文件数量控制在3000以内
  - 如果避免不了，可以生成一个文件列表，之后的数据处理直接访问该列表，而不要直接使用ls *, rm *等命令
  - 文件数量很大的目录，用ls –color=never， 响应会更快
- 不要把文件系统当成消息通信管道
  - 会给Lustre带来额外的负载
  - 考虑MPI等数据通信和同步协议
- 程序中打开了文件一定要关闭

IHEP School of Computing 2024 is coming!

- IHEP School of computing 2024 will be held in Yanqing, Beijing from the 21th to the 24th of August 2024
- 2.5 days, 21 lectures, and 4 hours of hands-on
- Indico: https://indico.ihep.ac.cn/event/22917/
- The course covers
  - Data processing in the field of high-energy physics,
  - AI technology for high-energy physics,
  - Computing technology for high-energy physics
  - Hands-on practice on computational platform

高能物理计算暑期学校
IHEP School of Computing

高能物理 第五届
IHEP School of Computing 2024
计算暑期学校

2024.8.21-8.24 北京·延庆

» 针对人群

暑期学校主要针对从事高能物理及其他学科的高年级本科生、研究生、博士后和科研工作者

» 培训内容

主要授课内容涉及主要授课内容涉及高能物理领域数据处理、高能物理AI技术、高能物理计算技术、计算平台上机实践等。

» 会议网址

https://indico.ihep.ac.cn/e/isc2024

请扫描上方二维码注册

» 联系方式

010-88236853；13811502631（晋老师）
010-88236883；13811338941（李老师）
Email: ComputingSchool@ihep.ac.cn

Summary

- BESIII computing platform is a important components of the Experiment
- The scale of BESIII computing platform continue to grow, and demand for data processing is also becoming diverse
- Try the efficient way to run jobs and access files on BESIII computing platform

Thank you!

Question?