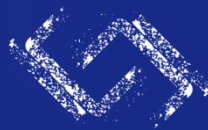




国家高能物理科学数据中心

National HEP Science Data Center



高能所计算中心

IHEP Computing Center

Integration of A Supercomputing Center in LHAASO Distributed Computing System

Qingbao HU, Xiaowei JIANG

IHEP Computing Center

2024-10-24

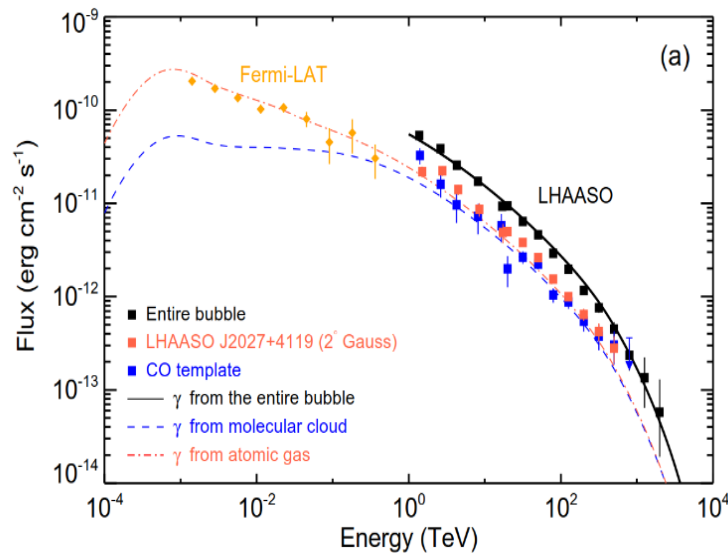


- 1 LHAASO Distributed Computing System
- 2 Integration of Supercomputing Center
- 3 Job and Data Test
- 4 Summary

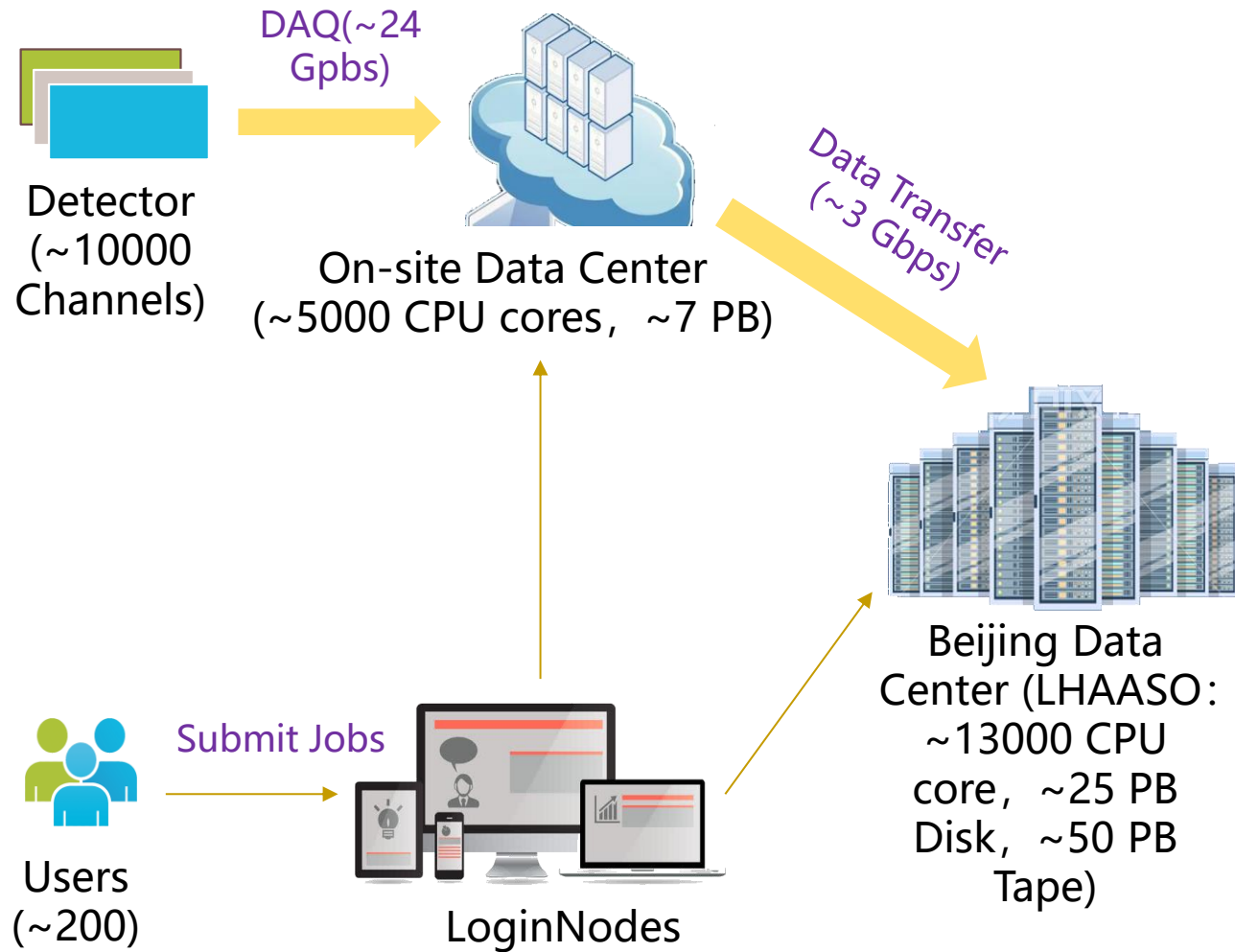
Large High Altitude Air Shower Observatory (LHAASO)



- World largest air shower array (with e, m, water Č detectors and Č telescope) for the high energy γ -astronomy and cosmic-ray physics
- Construction completed in 2023 and interesting results came out:
 - Highest γ -rays from the Milky Way: 2.5 PeV
 - 43 identified γ -rays sources up to ~ 1 PeV \rightarrow PeVatrons in the Milky Way
 - Energy spectrum of high energy γ -rays from the Crab Nebula as the standard candle
- International Collaboration : Countries/regions:5, members:~300



LHAASO Data Processing



• Resources

- 30 PB Disk
- 50 PB Tape
- 18000 CPU cores

• Services

- Computing, Disk storage, Tape storage, Network, Data Transfer

• ~11 PB increase up every year

Distributed Computing Model for LHAASO

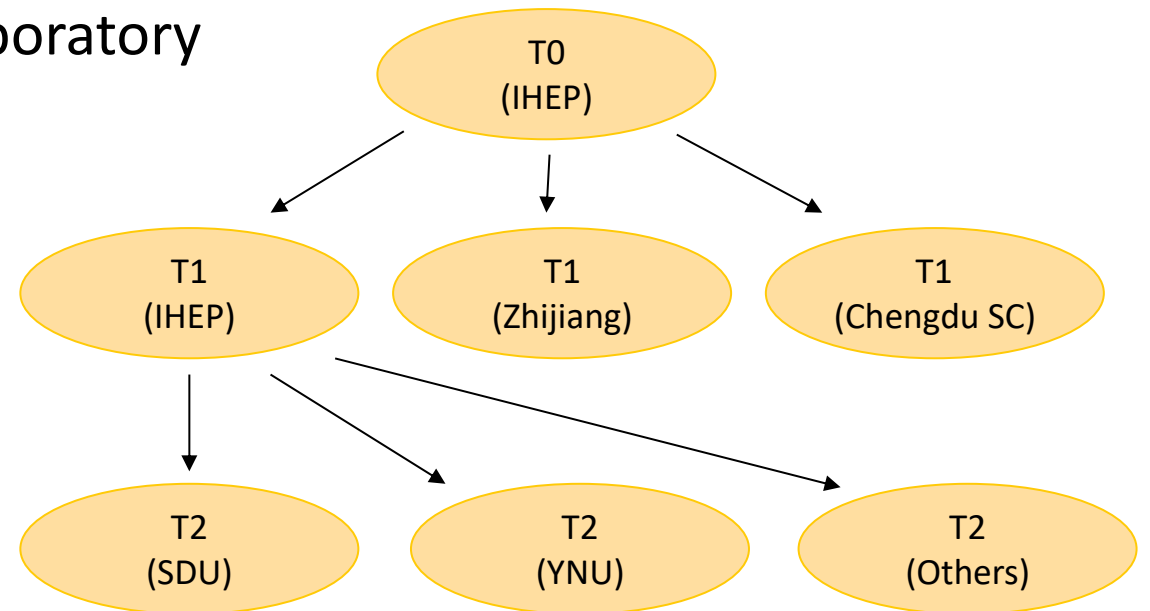


- Tier model (similar with WLCG's grid computing model)

- T0: IHEP
- T1: IHEP, Chengdu SC Center, Zhijiang Laboratory
- T2: collaboration sites

- Systems and Services

- Distributed computing system
- Distributed storage system
- Service and job environment
- Authentication and Authorization
- User interfaces (HepJob)



LHAASO Distributed Computing (1)



- Distributed Computing System

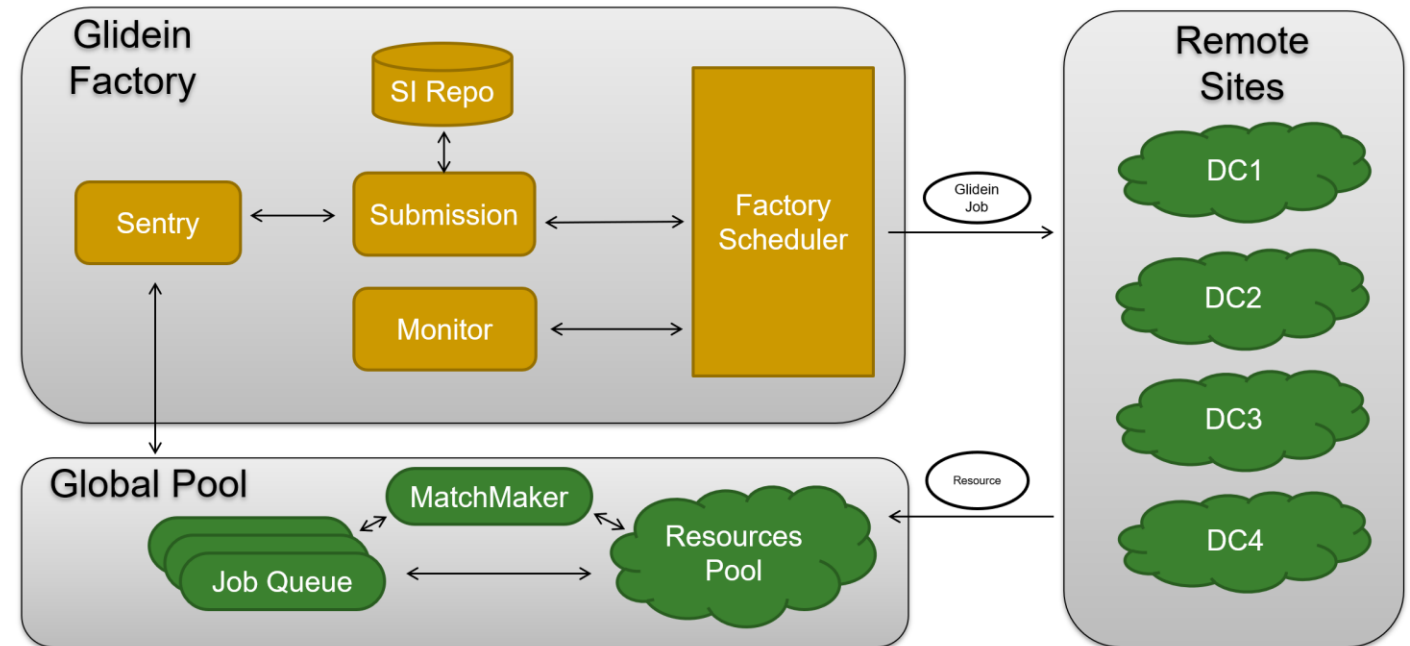
- Find and publish the remote computing resources to global side
- Dispatch the user job to remote worker node

- Components

- Global Pool
- Glidein Factory
- Glidein Client Tool

- HTCondor-C is also used

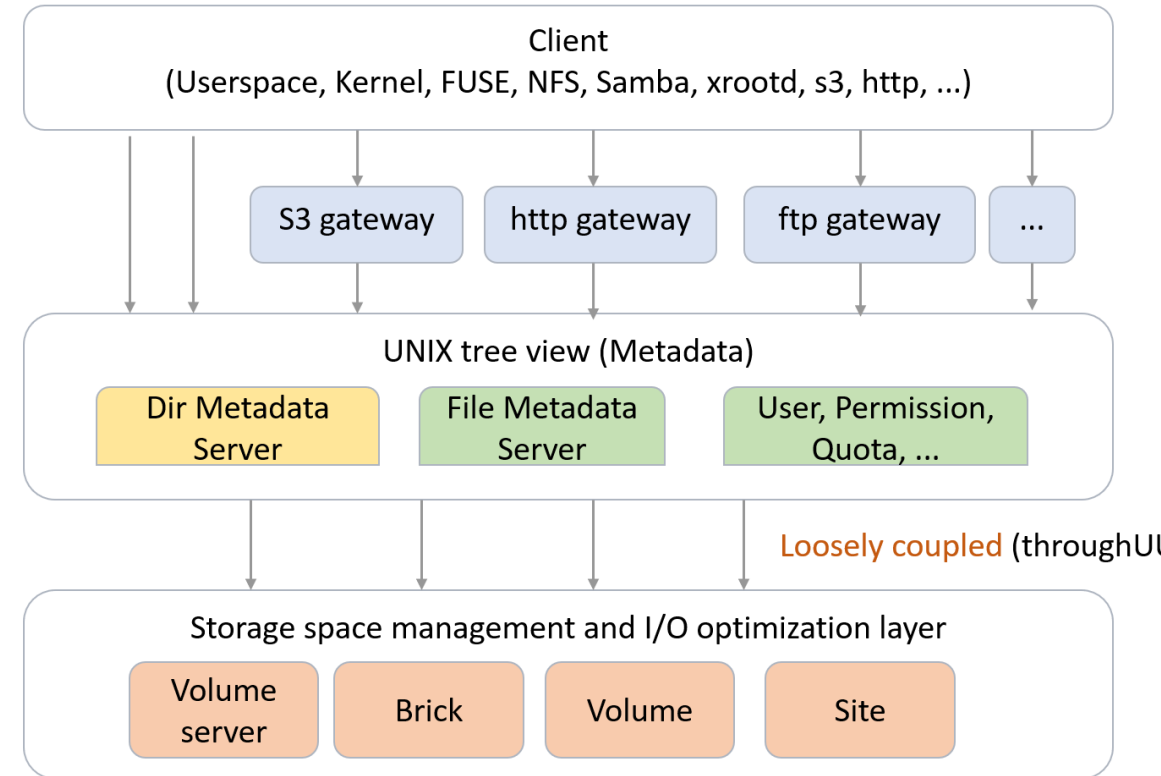
- Schedd server is different with glidein pool



LHAASO Distributed Computing (2)



- Distributed Storage System
- XRootD Proxy
 - A central way to read/write data from EOS
- Particlefs or Ocloud
 - A unified file system above multiple sites
- Rucio
 - A data management system to manage data
- First step is still using xrootd proxy
 - Particlefs/Ocloud or Rucio would be the solution in future





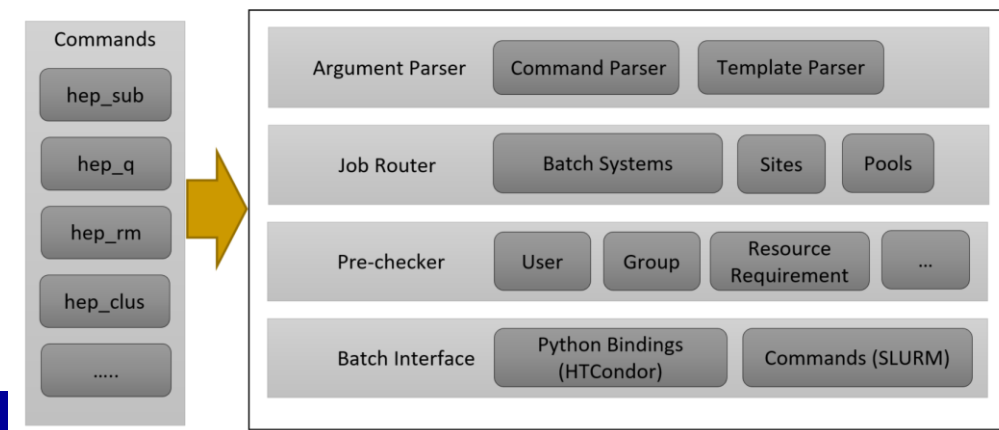
- Authentication and Authorization – tokens

- Kerberos token have been used in production
 - ◆ Local cluster (>10 years) and Dongguan data center

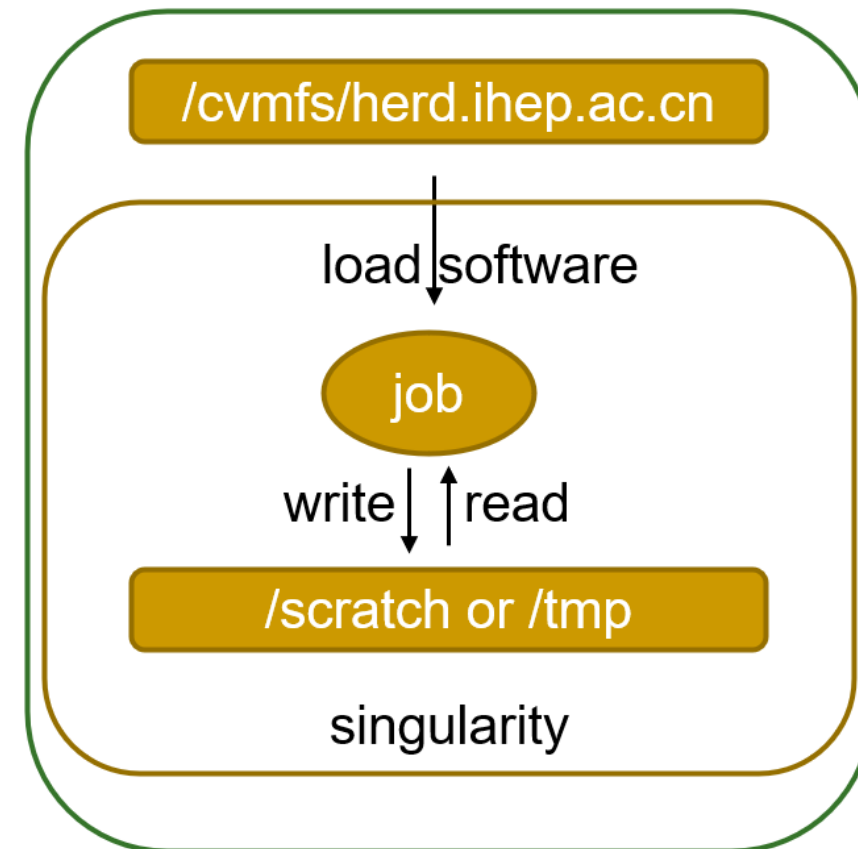
- Running Environment - Singularity

- Container

- User Interface - HepJob



Worker Node



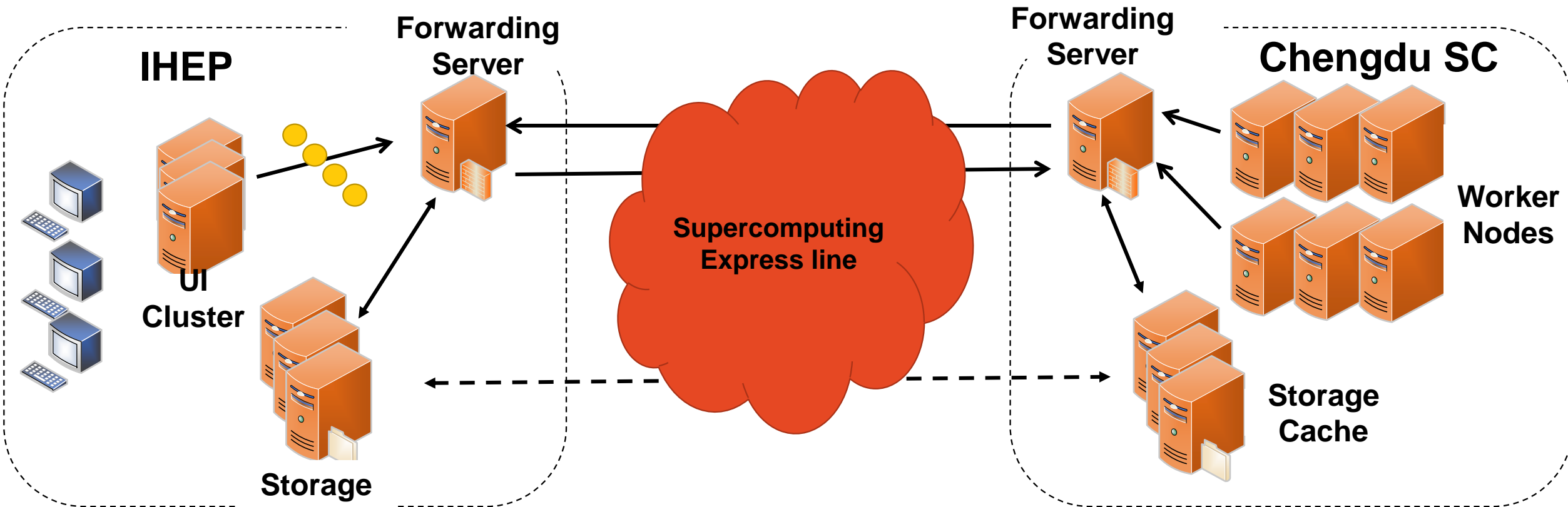


- Resources provided from Chengdu Supercomputing Center
 - >10,000 CPU cores and ~8PB Storage space
- The network link between Beijing and Chengdu was the big problem
 - Chengdu SC does not allow the inner servers to connect with outside
- A 10-Gbps network link was established
 - Cooperated with China Telecom Research Institute
 - Deployed the hardware VPNs on both of centers

Trying to Integrate Chengdu Supercomputing Center



- Supercomputing Express line is built by China Telecom Research Institute
 - Aim to connect the supercomputing centers in China
 - This is the first application

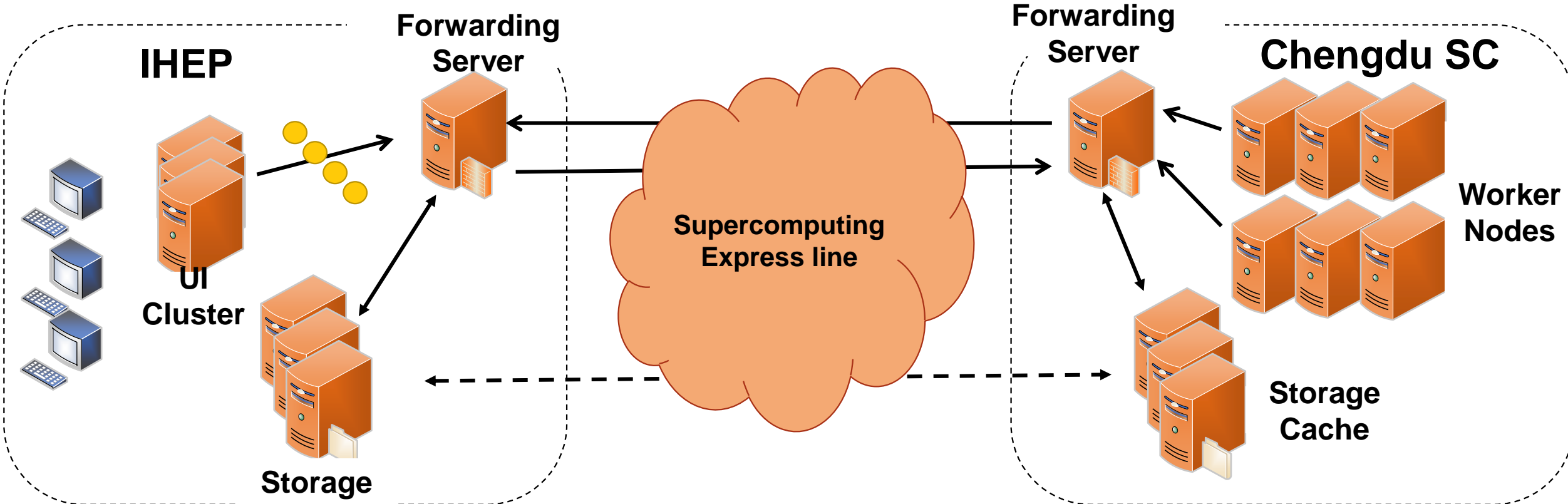


Trying to Integrate Chengdu Supercomputing Center



- All involved services are deployed on the forwarding servers

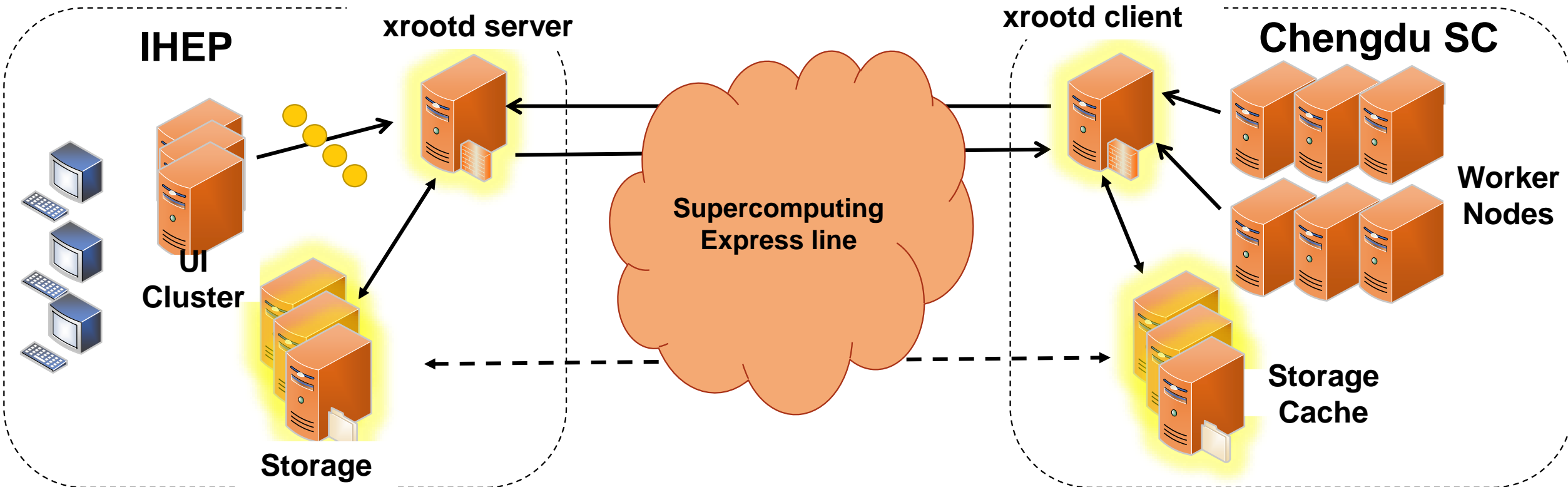
- Forwarding servers are visible with each other via supercomputing express line
- HTCondor for computing / Xrootd for data access



Data Access with XRootD



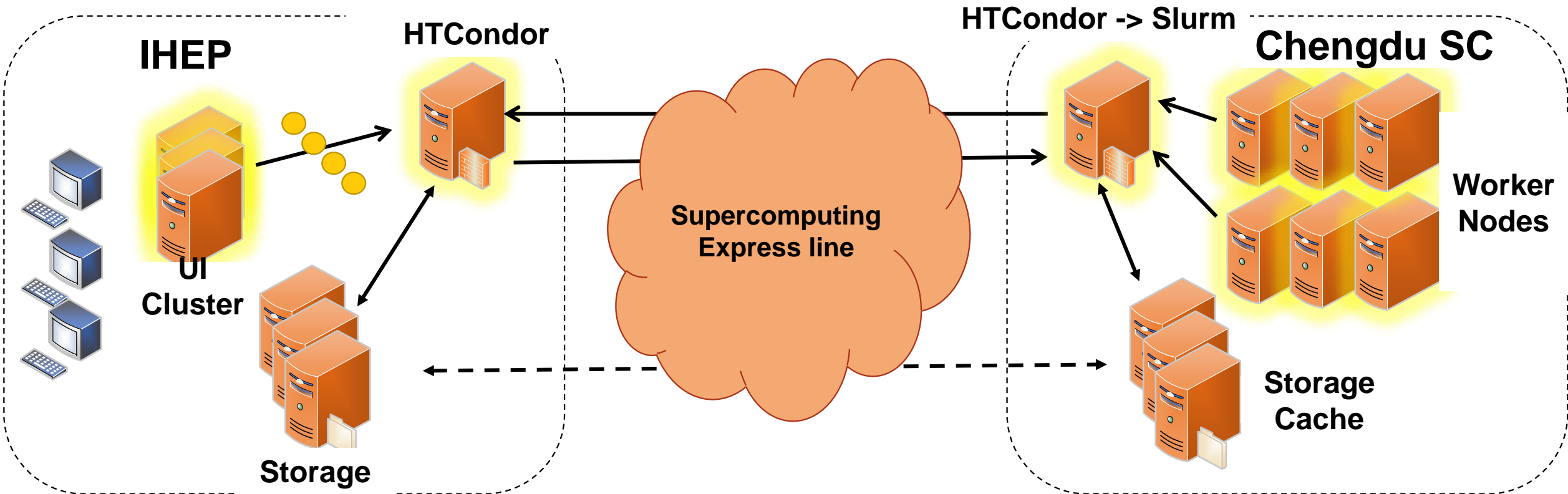
- The disk data storage is mounted on the xrootd server
 - Currently the data is manually transferred from IHEP to Chengdu SC via xrootd
- The file directories keep same structure on both sides



Computing Solution – HTCondor



- Job schedule: condor-c & condor-g
 - IHEP HTCondor -> Chengdu HTCondor -> Slurm

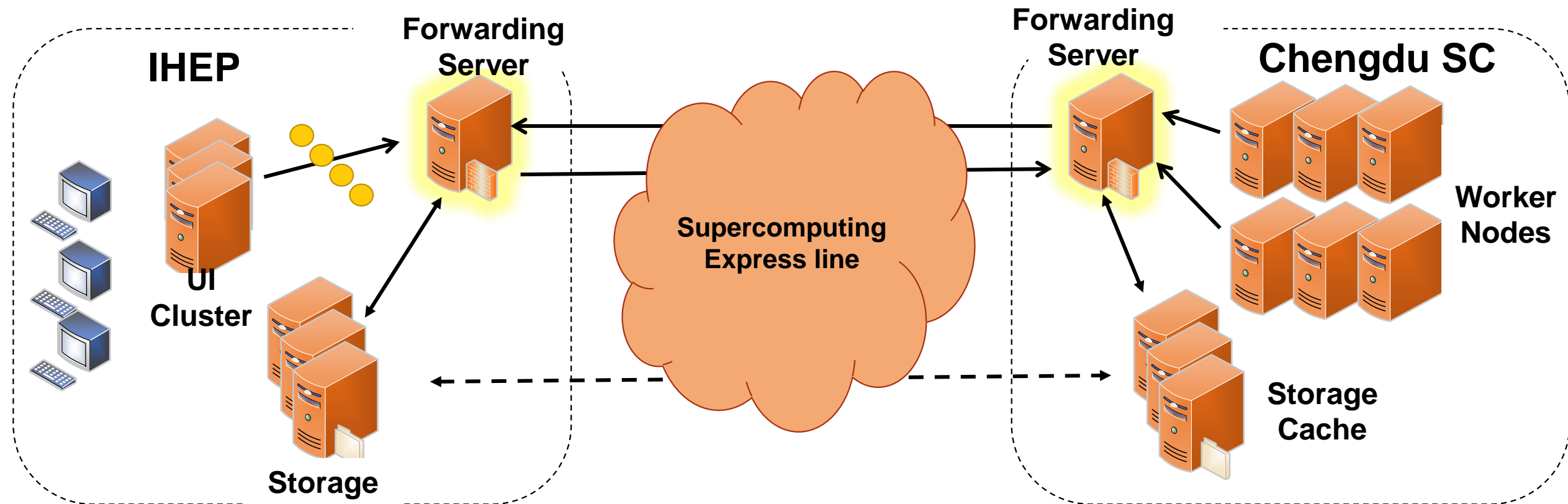


Software Synchronization



- Software is synchronized from IHEP side to Chengdu SC

- /cvmfs is mounted on forwarding server, where the Ihaaso software stores
- The software files are transferred to the local storage on Chengdu SC forwarding server
- The directories on Chengdu SC keep same with /cvmfs

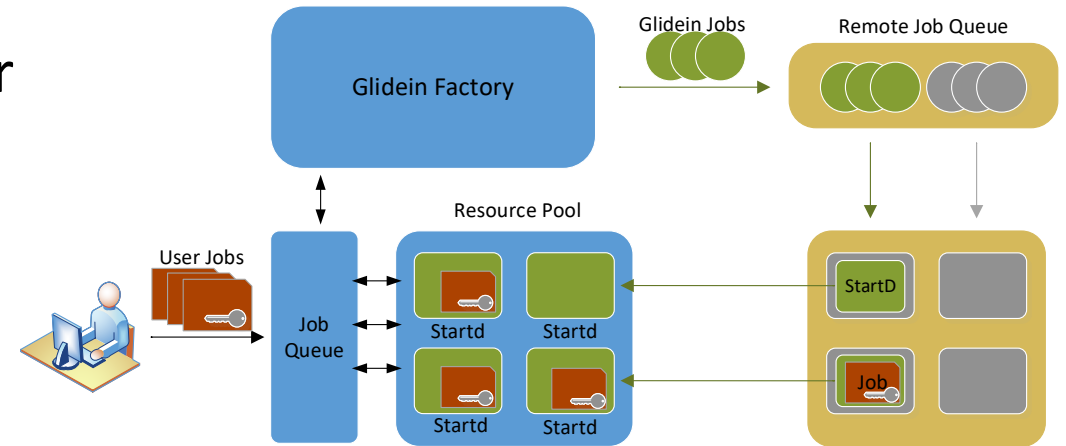


Cert&Auth with Kerberos Tokens



- Kerberos tokens for jobs and other services

- Kerberos token is using in LHAASO local cluster
- Access data from xrootd server with Kerberos
- Possible for other services



- Current solution is to transfer token credential as a normal input file

Submit Side

```
transfer_input_files = /tmp/krb5cc_10634  
+HepJob_KRB5CCNAME = "krb5cc_10634"
```



Execute Side

```
$ ls /var/lib/condor/execute/dir_6412/  
condor_exec.exe _condor_stderr _condor_stdout krb5cc_10634 tmp var
```

- Job schedule: condor-c & condor-g

- IHEP HTCondor -> Chengdu HTCondor -> Slurm

```
Every 2.0s: condor_q Thu Mar 7 22:58:37 2024

-- Schedd: scheduler@192-168-51-253.oscg.cn : <192.168.51.253:9431?... @ 03/07/24 22:58:37

OWNER BATCH_NAME SUBMITTED DONE RUN IDLE TOTAL JOB_IDS
shijy ID: 86 3/7 21:40 60 40 - 100 86.0-99
shijy ID: 87 3/7 21:41 58 42 - 100 87.0-97
shijy ID: 88 3/7 22:02 39 61 - 100 88.0-99
shijy ID: 89 3/7 22:02 27 73 - 100 89.0-99
shijy ID: 90 3/7 22:02 36 64 - 100 90.0-99
shijy ID: 91 3/7 22:54 - 56 44 - 100 91.0-99
shijy ID: 92 3/7 22:54 - - 100 100 92.0-99
shijy ID: 93 3/7 22:54 - - 100 100 93.0-99
shijy ID: 94 3/7 22:55 - - 100 100 94.0-99
shijy ID: 95 3/7 22:55 - - 100 100 95.0-99
shijy ID: 96 3/7 22:55 - - 100 100 96.0-99
shijy ID: 97 3/7 22:55 - - 100 100 97.0-99
shijy ID: 98 3/7 22:55 - - 100 100 98.0-99
shijy ID: 99 3/7 22:55 - - 100 100 99.0-99
shijy ID: 100 3/7 22:56 - - 100 100 100.0-99

Total for query: 1280 jobs; 0 completed, 0 removed, 944 idle, 336 running, 0 held, 0 suspended
Total for all users: 1280 jobs; 0 completed, 0 removed, 944 idle, 336 running, 0 held, 0 suspended
```



```
[lhc@log012 cn]$ condor_q

-- Schedd: scheduler@192-168-51-253.oscg.cn : <192.168.51.253:9431?... @ 03/07/24 22:59:15

JOBID PARTITION NAME USER ST TIME NODES NODELIST(REASON)
42221152 new bl_04370 shjep R 0:01 1 g00-00-00
42221151 new bl_T80V shjep R 0:01 1 g00-00-01
42221154 new bl_LynZL shjep R 0:01 1 g00-00-06
42221148 new bl_06945 shjep R 0:06 1 g00-00-06
42221149 new bl_H0485 shjep R 0:06 1 g00-00-06
42221150 new bl_07075 shjep R 0:06 1 g00-00-06
42221151 new bl_79007 shjep R 0:06 1 g00-00-06
42221145 new bl_07015 shjep R 0:11 1 g00-00-06
42221147 new bl_p0011 shjep R 0:11 1 g00-00-06
42221148 new bl_H0485 shjep R 0:15 1 g00-00-06
42221144 new bl_10000 shjep R 0:16 1 g00-00-06
42221140 new bl_05314 shjep R 0:22 1 g00-00-06
42221141 new bl_15000 shjep R 0:22 1 g00-00-06
42221142 new bl_15000 shjep R 0:22 1 g00-00-06
42221143 new bl_60351 shjep R 0:22 1 g00-00-06
42221138 new bl_14785 shjep R 0:27 1 g00-00-06
42221137 new bl_09581 shjep R 0:27 1 g00-00-06
42221138 new bl_04800 shjep R 0:27 1 g00-00-06
42221139 new bl_04490 shjep R 0:27 1 g00-00-06
42221132 new bl_04000 shjep R 0:32 1 g00-00-06
42221133 new bl_F3161 shjep R 0:32 1 g00-00-06
42221134 new bl_H0505 shjep R 0:32 1 g00-00-06
42221135 new bl_p0004 shjep R 0:32 1 g00-00-06
42221129 new bl_04950 shjep R 0:37 1 g00-00-06
42221130 new bl_04000 shjep R 0:37 1 g00-00-06
42221131 new bl_07800 shjep R 0:37 1 g00-00-06
42221128 new bl_g0000 shjep R 0:41 1 g00-00-06
42221127 new bl_04000 shjep R 0:43 1 g00-00-06
42221128 new bl_rj001 shjep R 0:44 1 g00-00-06
42221134 new bl_04512 shjep R 0:46 1 g00-00-06
42221131 new bl_04000 shjep R 0:46 1 g00-00-06
42221130 new bl_g_000 shjep R 0:51 1 g00-00-06
42221131 new bl_51100 shjep R 0:53 1 g00-00-06
42221132 new bl_H1200 shjep R 0:53 1 g00-00-06
```



```
[lhc@log012 cn]$ squeue

JOBID PARTITION NAME USER ST TIME NODES NODELIST(REASON)
42221152 new bl_04370 shjep R 0:01 1 g00-00-00
42221151 new bl_T80V shjep R 0:01 1 g00-00-01
42221154 new bl_LynZL shjep R 0:01 1 g00-00-06
42221148 new bl_06945 shjep R 0:06 1 g00-00-06
42221149 new bl_H0485 shjep R 0:06 1 g00-00-06
42221150 new bl_07075 shjep R 0:06 1 g00-00-06
42221151 new bl_79007 shjep R 0:06 1 g00-00-06
42221145 new bl_07015 shjep R 0:11 1 g00-00-06
42221147 new bl_p0011 shjep R 0:11 1 g00-00-06
42221148 new bl_H0485 shjep R 0:15 1 g00-00-06
42221144 new bl_10000 shjep R 0:16 1 g00-00-06
42221140 new bl_05314 shjep R 0:22 1 g00-00-06
42221141 new bl_15000 shjep R 0:22 1 g00-00-06
42221142 new bl_15000 shjep R 0:22 1 g00-00-06
42221143 new bl_60351 shjep R 0:22 1 g00-00-06
42221138 new bl_14785 shjep R 0:27 1 g00-00-06
42221137 new bl_09581 shjep R 0:27 1 g00-00-06
42221138 new bl_04800 shjep R 0:27 1 g00-00-06
42221139 new bl_04490 shjep R 0:27 1 g00-00-06
42221132 new bl_04000 shjep R 0:32 1 g00-00-06
42221133 new bl_F3161 shjep R 0:32 1 g00-00-06
42221134 new bl_H0505 shjep R 0:32 1 g00-00-06
42221135 new bl_p0004 shjep R 0:32 1 g00-00-06
42221129 new bl_04950 shjep R 0:37 1 g00-00-06
42221130 new bl_04000 shjep R 0:37 1 g00-00-06
42221131 new bl_07800 shjep R 0:37 1 g00-00-06
42221128 new bl_g0000 shjep R 0:41 1 g00-00-06
42221127 new bl_04000 shjep R 0:43 1 g00-00-06
42221128 new bl_rj001 shjep R 0:44 1 g00-00-06
42221134 new bl_04512 shjep R 0:46 1 g00-00-06
42221131 new bl_04000 shjep R 0:46 1 g00-00-06
42221130 new bl_g_000 shjep R 0:51 1 g00-00-06
42221131 new bl_51100 shjep R 0:53 1 g00-00-06
42221132 new bl_H1200 shjep R 0:53 1 g00-00-06
```




- Data Transfer: xrootd

- Transfer speed is above 1GB/s

```
guocq@ccopt.ihep.ac.cn:22
[1.398GB/2.441GB][ 57%][=====] ] [10.38MB/s]
[952MB/2.441GB][ 38%][=====] ] [6.899MB/s]
[1.82GB/2.441GB][ 74%][=====] ] [12.86MB/s] [1.422GB/2.441GB][ 58%][==
[1.539GB/2.441GB][ 63%][=====] ] [10.79MB/s] [1.062GB/2.441GB][ 43%][==
[1.578GB/2.441GB][ 64%][=====] ] [10.7MB/s] [992MB/2.441GB][ 39%][=====
[1.078GB/2.441GB][ 44%][=====] ] [6.9MB/s]
[1.539GB/2.441GB][ 63%][=====] ] [9.85MB/s]
[1.273GB/2.441GB][ 52%][=====] ] [8.15MB/s]
[root@ocloud ~]#
[root@ocloud ~]#
[976MB/2.441GB][ 39%][=====] ] [6.1MB/s] ] [root@ocloud ~]#
[1.383GB/2.441GB][ 56%][=====] ] [8.85MB/s]
[1.5GB/2.441GB][ 61%][=====] ] [9.54MB/s]
[1.328GB/2.441GB][ 54%][=====] ] [8.447MB/s]
[1.367GB/2.441GB][ 56%][=====] ] [8.696MB/s]
[1.57GB/2.441GB][ 64%][=====] ] [9.988MB/s]
[1.242GB/2.441GB][ 50%][=====] ] [7.901MB/s]
[1.703GB/2.441GB][ 69%][=====] ] [10.83MB/s]
[1.508GB/2.441GB][ 61%][=====] ] [9.59MB/s]
[1.75GB/2.441GB][ 71%][=====] ] [11.13MB/s]
[2.367GB/2.441GB][ 96%][=====] ] [15.06MB/s]
[1.305GB/2.441GB][ 53%][=====] ] [8.298MB/s] [root@ocloud ~]#
[1.344GB/2.441GB][ 55%][=====] ] [8.547MB/s] [root@ocloud ~]#
[1.57GB/2.441GB][ 64%][=====] ] [9.988MB/s]
[1.516GB/2.441GB][ 62%][=====] ] [9.64MB/s]
[1.375GB/2.441GB][ 56%][=====] ] [8.745MB/s]
[root@ocloud ~]#
[1.352GB/2.441GB][ 55%][=====] ] [8.596MB/s]
[1.32GB/2.441GB][ 54%][=====] ] [8.398MB/s]
[1.484GB/2.441GB][ 60%][=====] ] [9.441MB/s]
[1.547GB/2.441GB][ 63%][=====] ] [9.839MB/s]
[1.102GB/2.441GB][ 45%][=====] ] [7.006MB/s]
[root@ocloud ~]#
[1.281GB/2.441GB][ 52%][=====] ] [8.149MB/s]
[root@ocloud ~]#
[root@ocloud ~]#
[root@ocloud ~]#
[1.172GB/2.441GB][ 48%][=====] ] [7.453MB/s]
[984MB/2.441GB][ 39%][=====] ] [6.112MB/s]
[1.391GB/2.441GB][ 56%][=====] ] [8.845MB/s]
[1.109GB/2.441GB][ 45%][=====] ] [7.056MB/s]
[1.367GB/2.441GB][ 56%][=====] ] [8.537MB/s] [1.453GB/2.441GB][ 59%][==
[2.441GB/2.441GB][100%][=====] ] [14.79MB/s]
[1.414GB/2.441GB][ 57%][=====] ] [8.468MB/s]
```



- LHAASO is building the distributed computing system to handle the big scale of data processing
 - Supercomputing Center can contribute computing and storage
- LHAASO is trying to integrate the Chengdu Supercomputing Center to the distributed computing system
 - Network solution: Supercomputing Express line
 - Computing solution: HTCondor-C & HTCondor-G
 - Data solution: Xrootd
- Problem and Next Plan
 - To develop a real-time data transfer service and deploy it on Supercomputing side
 - Investigate a proxy way to use glidein to publish the worker nodes from SC to Central Glidein Pool
 - The network solution is too heavy now and needs a better one

Thanks!

Q&A