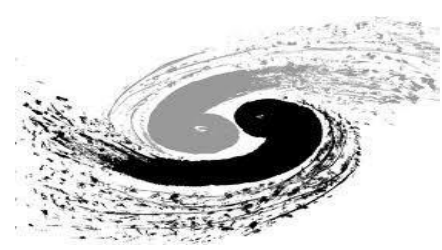# Expand Local Cluster to the Worker Node of Remote Site

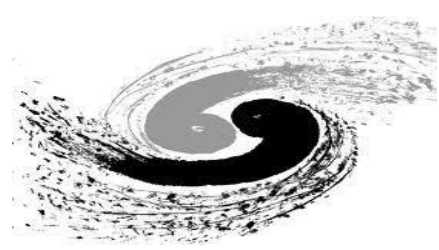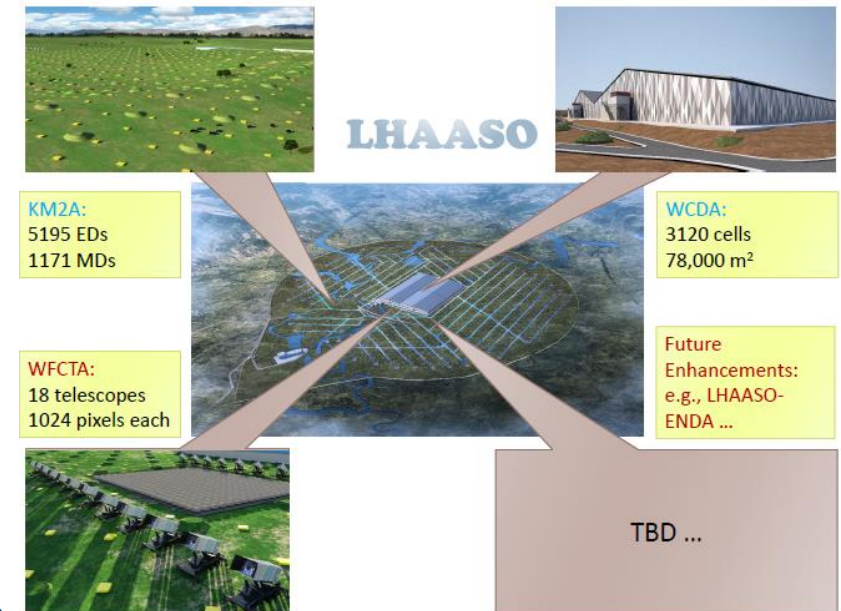Jingyan Shi, Xiaowei Jiang, Chaoqi Guo
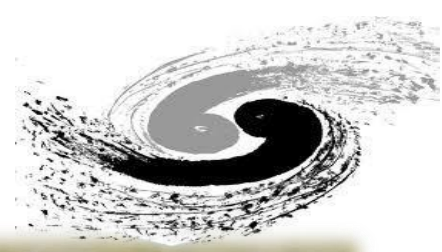IHEP – CC
shijy@ihep.ac.cn

Outline

Brief Introduction of LHAASO

- Large High Altitude Air Shower Observatory (LHAASO)
  - A new generation all-sky facility
    - Combined study of cosmic rays and gamma rays
    - Wide energy range of $10^{11} - 10^{17}$ ev
  - Located in Daocheng, Sichuan Province
    - Altitude: 4410 m
    - Coverage area: 1.3 km^2
  - Fully complete in Jun. 2021
    - Raw data per year: 13PB (7PB more than the plan)
    - Storage capacity: > 40 PB (20PB more than the plan)

KM2A:
5195 EDs
1171 MDs

WCDA:
3120 cells
78,000 m^2

WFCTA:
18 telescopes
1024 pixels each

Future
Enhancements:
e.g., LHAASO-
ENDA ...

TBD ...

LHAASO Data Processing

- **Computing issues**
  - No mature data management system developed
  - Most users are not sophisticated
- **Computing requirement**
  - LHAASO software is stored at /CVMFS
  - LHAASO data is stored at local EOS
  - Most are HTC job and running at HTCondor cluster at IHEP
  - User auth is based on kerberos

~2.5Gbps

The small on-site Data Center at Daocheng (altitude 4500m).
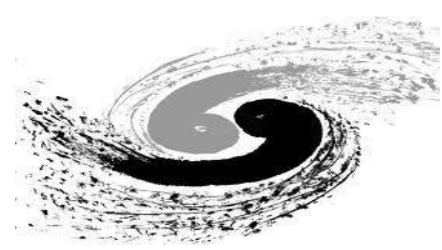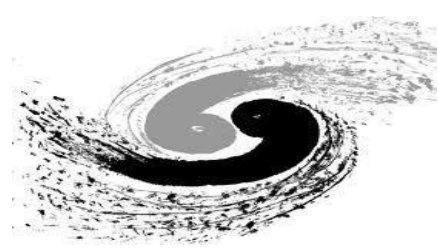
The large Offline Data Center at IHEP

- A simplified job management tool developed for users
  - For example: hep_sub -g lhaaso job.sh
- **Big gap** between the requirement and reality
  - Estimation: ~20k CPU cores and 40 PB disk storage are required
  - Reality: <11k CPU cores

Find More Resources for LHAASO

- IHEP local HTCondor cluster is the main place for LHAASO data processing
- IHEP Slurm cluster -- case 1
  - One partition (~1k cpu cores) are willing to accept LHAASO job when it is free
    - Know idle time period
    - Same user name space as IHEP HTCondor cluster; and IHEP EOS is accessible
- Big Data center located at Dongguan, Guangdong province– case 2
  - ~4k x86 CPU and 10k arm CPU resource with 10G network link
  - No permanent storage provided
  - Different user name space
- Small sites at domestic collaboration member organization -- case 3
  - Small resources with limited network connection
  - No mature technical support

LHAASO Local Cluster Extension

- To keep the same usage pattern for LHAASO data processing
  - IHEP LHAASO HTCondor cluster is the main cluster
  - Expand cluster to the remote resource
    - Add remote worker nodes as the IHEP cluster worker node
      - Submit glidein batch job to the remote site
      - Run IHEP Condor startd inside the glidein job
    - Schedule suitable jobs to run at remote job slots
    - User kerberos token is transferred with the user job to the remote worker node
    - No direct data access to IHEP EOS at remote site

Design of the LHAASO Cluster Extension
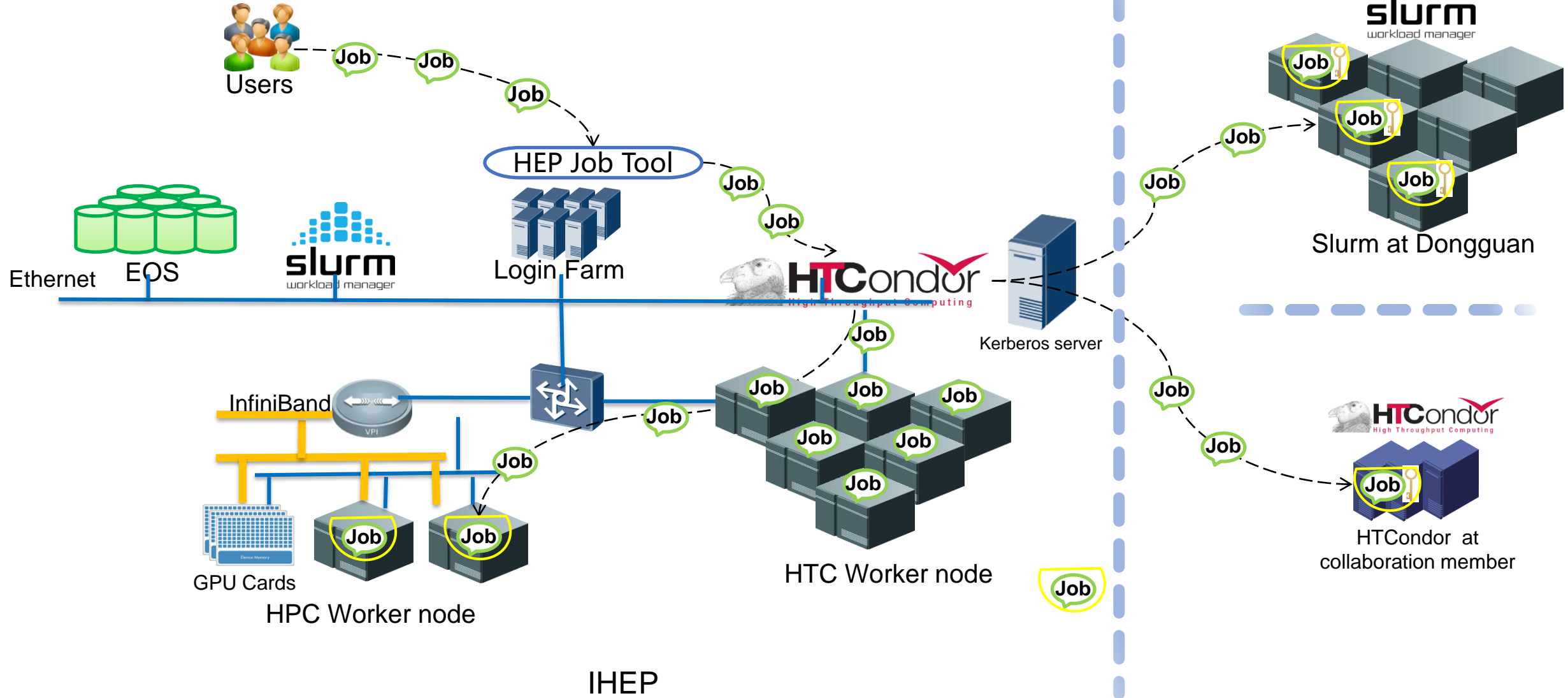
Classify LHAASO Job
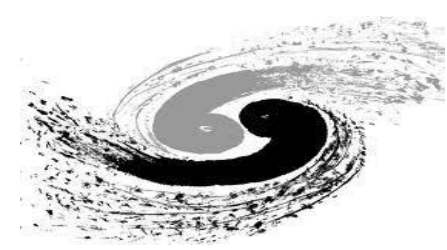
- **3 LHAASO detectors have their own simulation, reconstruction and analysis jobs**

  - Classify the jobs based on the computing time and IO access

  - Take one of the detector, WFCTA, as the example

Suitable to run at Dongguan

Suitable to run at remote site

Suitable to run at IHEP slurm

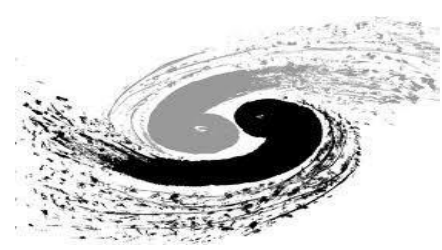| One Job	Input data	Output data	Cpu time
corsika	Little	a lot	too much
geant4	a lot	mid	too much
Corsika+geant4	little	mid	too much
reconstruction	mid	a little	a little
analysis	mid	little	A little

- **"jobtype" attribute is set by" hep job tool" when user submit the job**

User Authentication
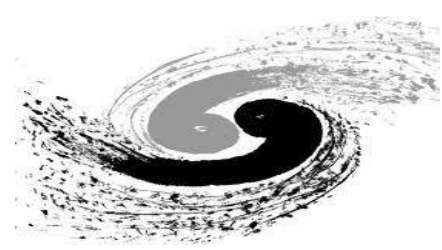
- User login to the IHEP cluster and kerberos token is generated
- Token is transferred to the worker node with job
  - Prolong token lifetime
    - Job is in the queue
      - User token is copied to the token dir by hep_job tool and a root deamon is responsible to prolong and clean the tokens
    - Job is running
      - Wrapper inside the glidein exports token path as the environment variable
        - Job access IHEP EOS from the remote site by the token
      - Wrapper start a process to prolong the token during the job lifetime
      - Token will be cleaned by startd of worker node when the job is finished

No Direct Data Access to IHEP EOS

- Provide WFCTA job script (saved at cvmfs)to the user.
- Both IHEP cluster and remote site use the same job script
  - Transfer the input data file to the local disk of worker node based on the authentication of job token
  - Job result is written to the local disk of worker node firstly
  - The result will be transferred back to the IHEP EOS based on the job token authentication
  - Clean the data in the job directory at worker node

Case 1: Running at IHEP Slurm Cluster
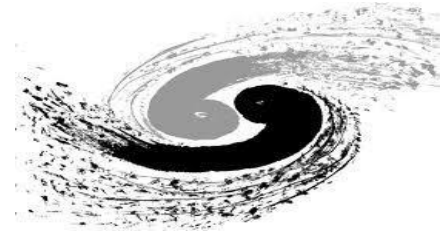
- User name space and EOS file system are same as that of IHEP HTCondor cluster
  - Submit glidein jobs to the Slurm worker nodes during the idle period as the root right
    - Glidein jobs run as user "condor" which is same as the owner of "startd" daemon running at the local HTCondor cluster
    - LHAASO jobs run inside startd

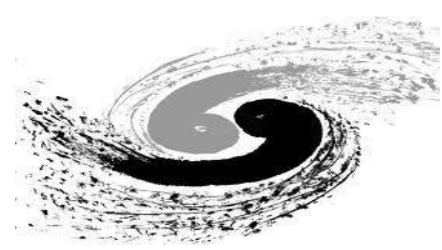# Case 2 and Case 3: Running at Remote Resource
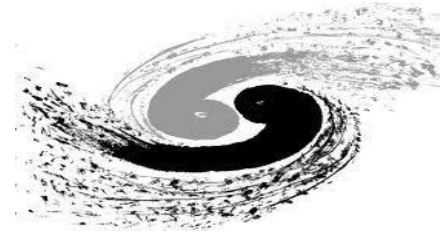
- Submit glidein slurm/htcondor jobs from login node of the remote cluster
    - Set the job slot only accept dedicated job type job (corsika, geant4 etc.)
- Corsika jobs and geant4 jobs are submitted to IHEP cluster by user
- The job will be scheduled to the glidein job slots at remote site
   - The last step of the job is to transfer result back to IHEP EOS

Others
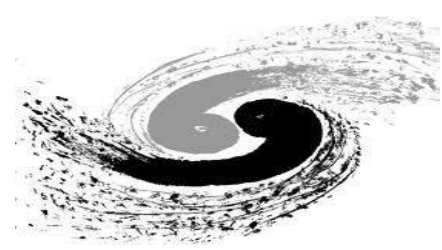
- ARM machine support – testing
  - Compile LHAASO software on ARM architecture
    - Physical Result evaluation is under going
  - Compile HTCondor on ARM architecture
    - ARM HTCondor worker node is ready

Next Plan

- The extension of IHEP cluster has provided 2.4M cpu hours and generated 80TB simulation data for LHAASO

- Next Plan
  - ARM machine will be OK next month
  - Glidein factory is under going
  - Try some more efficient scheduling algorithm

Summary
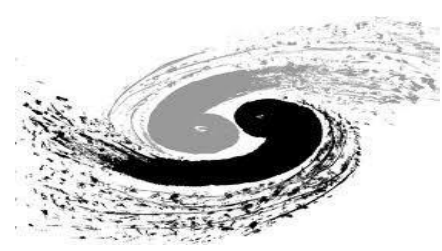
- LHAASO needs more computing resource
- A light dHTC designed and deployed for LHAASO
  - expand IHEP local cluster to the remote site
  - Keep the user cluster usage pattern
  - Have integrated remote resource from several sites
- More work need to be done