

# Cloud-based Computing for LHAASO experiment at IHEP

Qiulan Huang, Weidong Li, Haibo Li, Yaodong  
Cheng, Tao Cui, Jingyan Shi, Qingbao Hu

[huangql@ihep.ac.cn](mailto:huangql@ihep.ac.cn)

Computing Center, IHEP, CAS

Grid2018 at Dubna

2018-09-10

# Outline

- **Overview of HEP computing in IHEP**
- **LHAASO Computing Requirements**
  - Computing, Storage and Network
- **Cloud-based Computing Solution for LHAASO**
  - Federate distributed resources
  - Job Scheduling
  - Remote Data Access
  - ...
- **Web-based Data Analysis for LHAASO**
- **Summary**



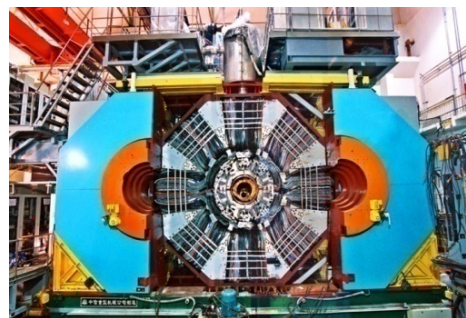


# Overview of HEP computing in IHEP

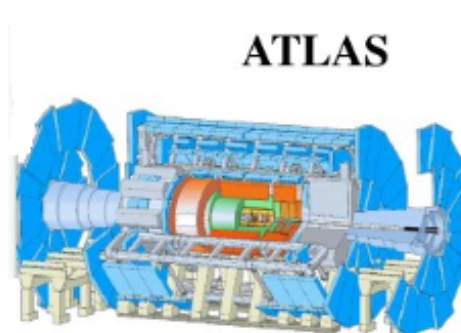


# HEP computing in IHEP

- BEPCII & BESIII
  - ~10PB data
- LHAASO
  - Start taking data in 2018, 6PB/year
- DayaBay
  - 200TB/year, >2PB data
- JUNO
  - Begin taking data in 2020, 2PB/year
- HXMT, CSNS, CMS, ATLAS experiments on LHC, HEPS



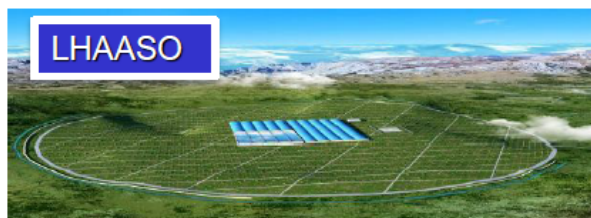
DYB (Daya Bay Reactor Neutrino Experiment)



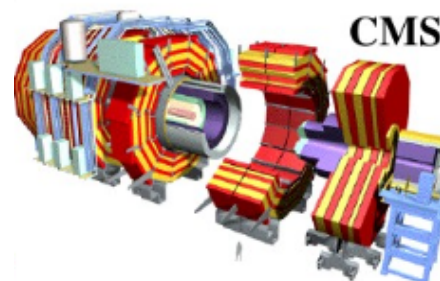
ATLAS



JUNO (Jiangmen Underground Neutrino Observatory)



LHAASO



CMS



HXMT



# Computing Environment in IHEP

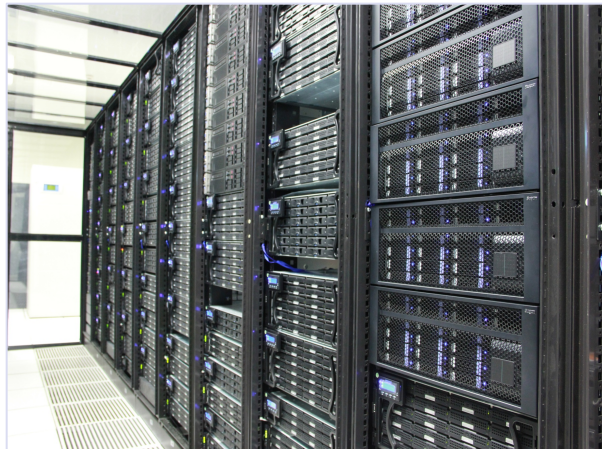


Power supply, cooling

CPU负责人 文朝朝007 (CPU-Servers)	91 OK 3 DOWN - 3 Unhandled	100% OK 3 WARNING - 3 Unhandled 2 on Problem Hosts 44 CRITICAL - 44 on Problem Hosts
存储服务器 (CRASS-Servers)	1 OK 1 OK	100% OK
集群服务器 (Cluster-Servers)	1 OK	100% OK
高性能计算节点 (HEP-GH)	100% OK	100% OK 11 WARNING - 11 Unhandled 14 CRITICAL - 4 Unhandled
IT网络计算节点 (ITWS-Servers)	100% OK	100% OK 12 WARNING - 12 Unhandled 40 CRITICAL - 40 Unhandled
高性能计算节点 (job-servers)	100% OK	100% OK 1 CRITICAL - 4 Unhandled
登录节点负责人 杜耀红 (Login-Servers)	100% OK	100% OK 12 WARNING - 12 Unhandled 1 CRITICAL - 1 Unhandled
Lustre-server (Lustre-servers)	100% OK	100% OK 12 WARNING - 12 Unhandled 1 CRITICAL - 1 Unhandled
Lustre-server-nds (Lustre-servers-nds)	100% OK	100% OK 12 WARNING - 12 Unhandled

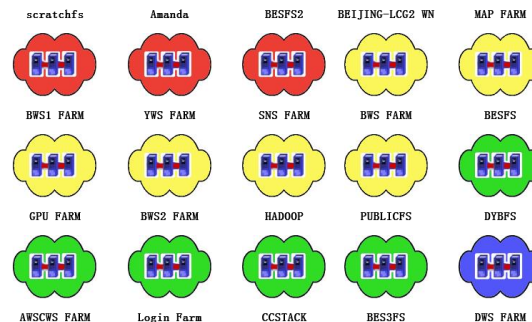


~15000 CPU cores  
(HTCondor/Slurm)



11PB disk space(Lustre/EOS)

## Monitoring



5PB tape library(Castor)

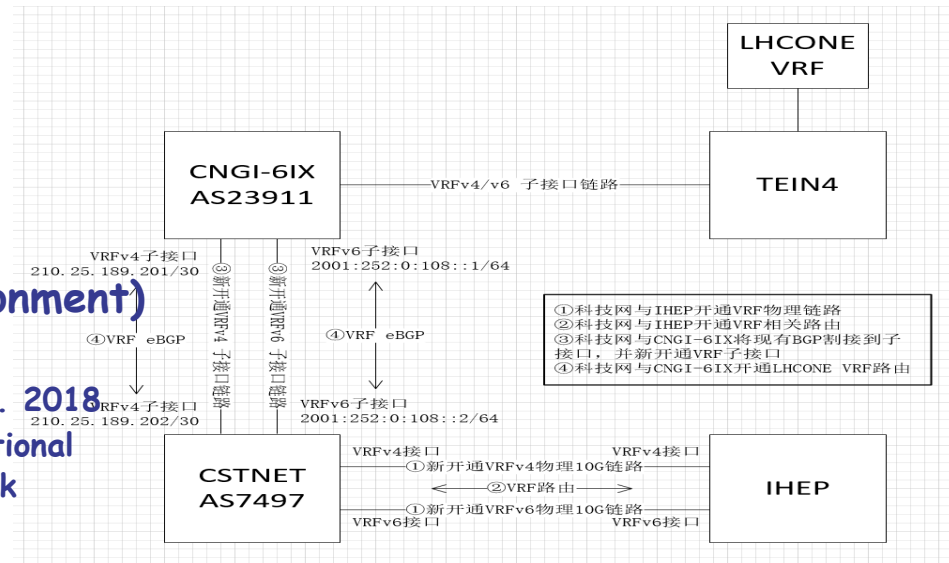
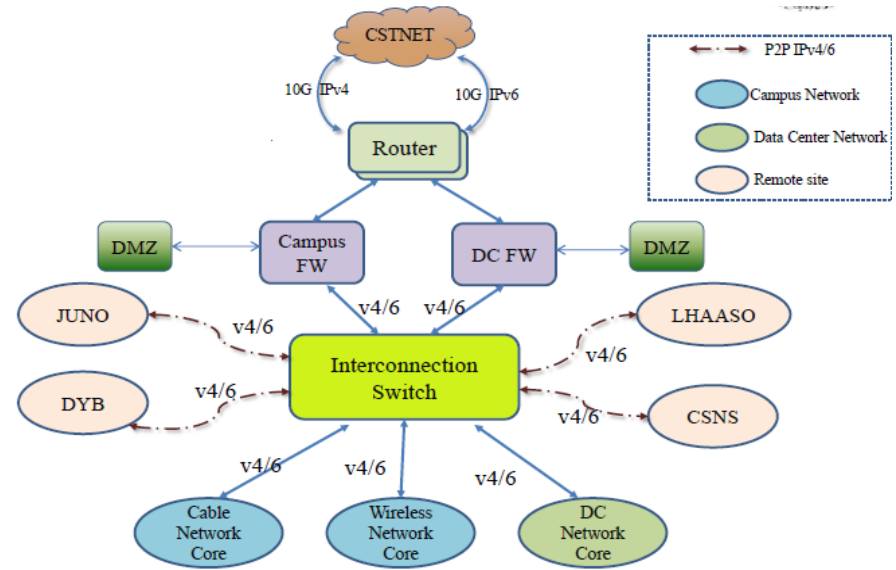




# Network at IHEP

## WAN

- **IHEP-USA(Internet2)**
  - IHEP-CSTNet-CERNET-USA
  - Bandwidth: **10Gbps**
- **IHEP-EUR(GEANT)**
  - IHEP-CSTNet-CERNET-LONDON-Euro
  - Bandwidth: **10Gbps**
- **IHEP-Asia**
  - IHEP-CSTNet-HONGKONG-Asia
  - Bandwidth: **2.5Gbps**
- **IHEP-\*.edu.cn**
  - IHEP-CSTNet-CERNET-University
  - Bandwidth: **10Gbps**
- **LHCONE(LHC Open Network Environment)**
  - A Virtual dedicated network
  - Became a member of LHCONE from Mar. 2018
  - Various route information of the international collaborators can be seen in this network



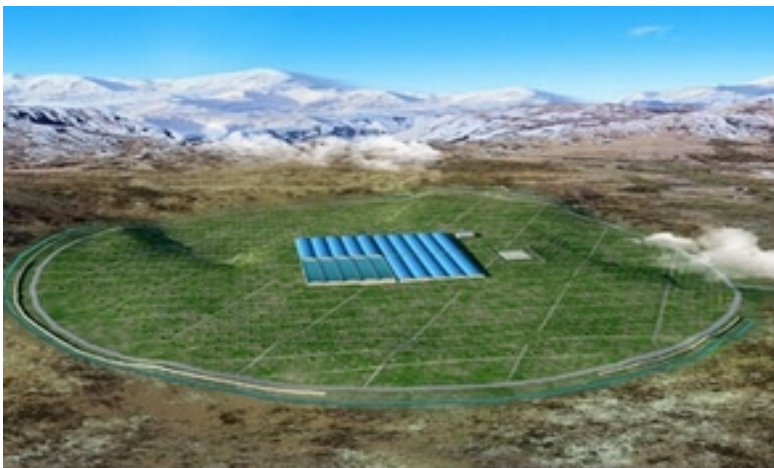


# LHAASO Computing Requirements



# LHAASO

- **Large High Altitude Air Shower Observatory**
  - Located in Daocheng, Sichuan province (at the altitude of 4410 m)
  - Expected to be the most sensitive project to study the problems in Galactic cosmic ray physics
  - Start to take data this year
  - 6PB/year \* 10 years
- **Computing Requirements**
  - 5120 CPU cores, 6 GPU servers
  - 4PB disk storage, 5PB Tape storage
  - Dedicated network between IHEP and DaoCheng: > 500Mbps, 2PB to be transferred annual





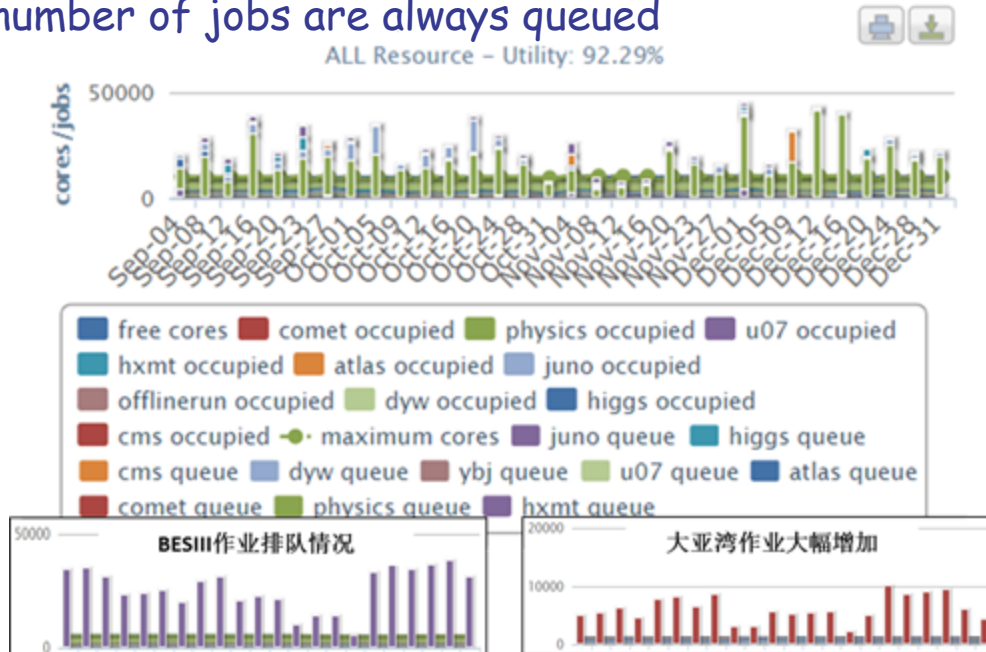


# Cloud-based Computing Solution for LHAASO



# Motivation & Challenge

- IHEP provides the computing service for BESIII, Dayabay, JUNO, LHAASO, CMS and Atlas experiment
  - ~15000 CPU cores, 11PB disk storage
- The computing resources of the existing single data are tight
  - A large number of jobs are always queued



- To meet the requirement of huge amount of storage and computing power, we need to integrate distributed heterogeneous resources to expand computing scale

# Motivation & Challenges

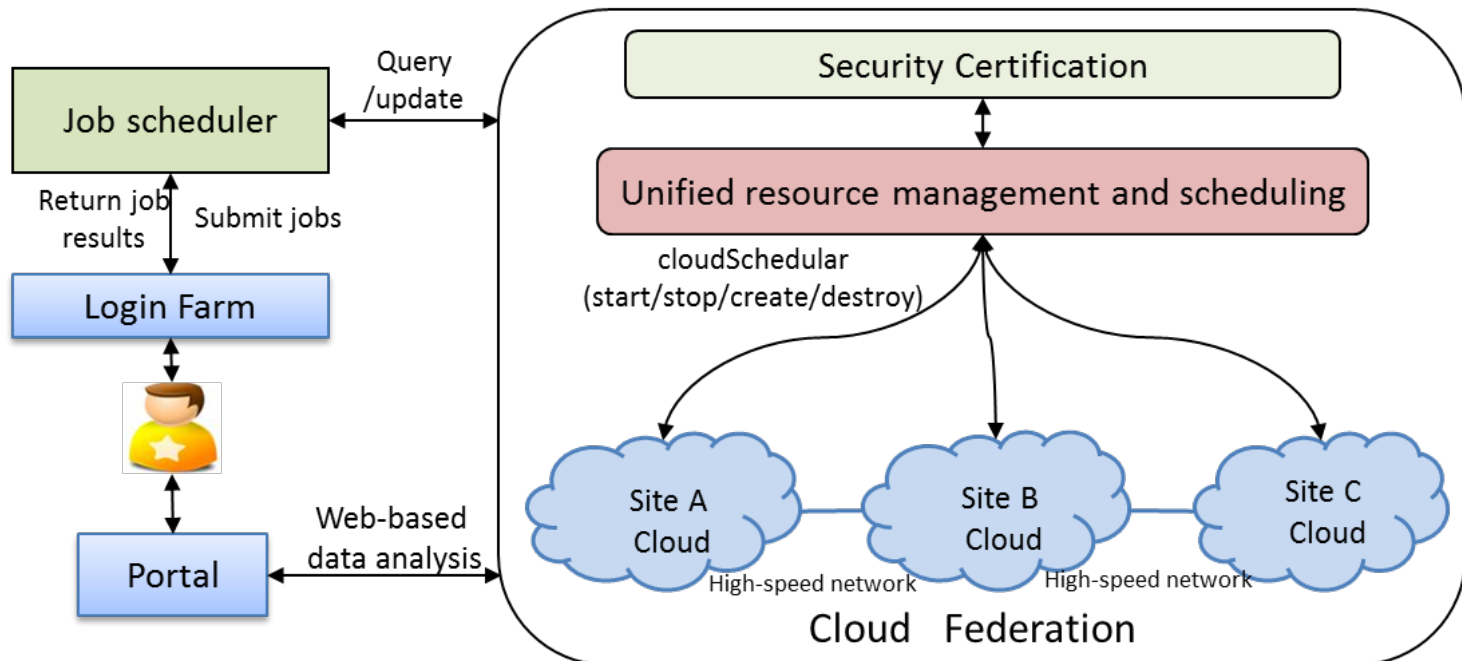
- Resource integration promotes resource sharing between experiment cooperation groups
- HEP experiments using cross-border resources are troubled with some issues
  - High operation and maintenance costs
  - Computing system instability of remote sites
  - Operation and maintenance ability is poor
  - Shortage of experienced administrators
- We introduce virtualization and cloud computing technology into LHAASO
  - Use virtualization technology to hide the underlying details
  - Make sure of the system availability and stability
  - Significantly reduce the maintenance cost



# Architecture

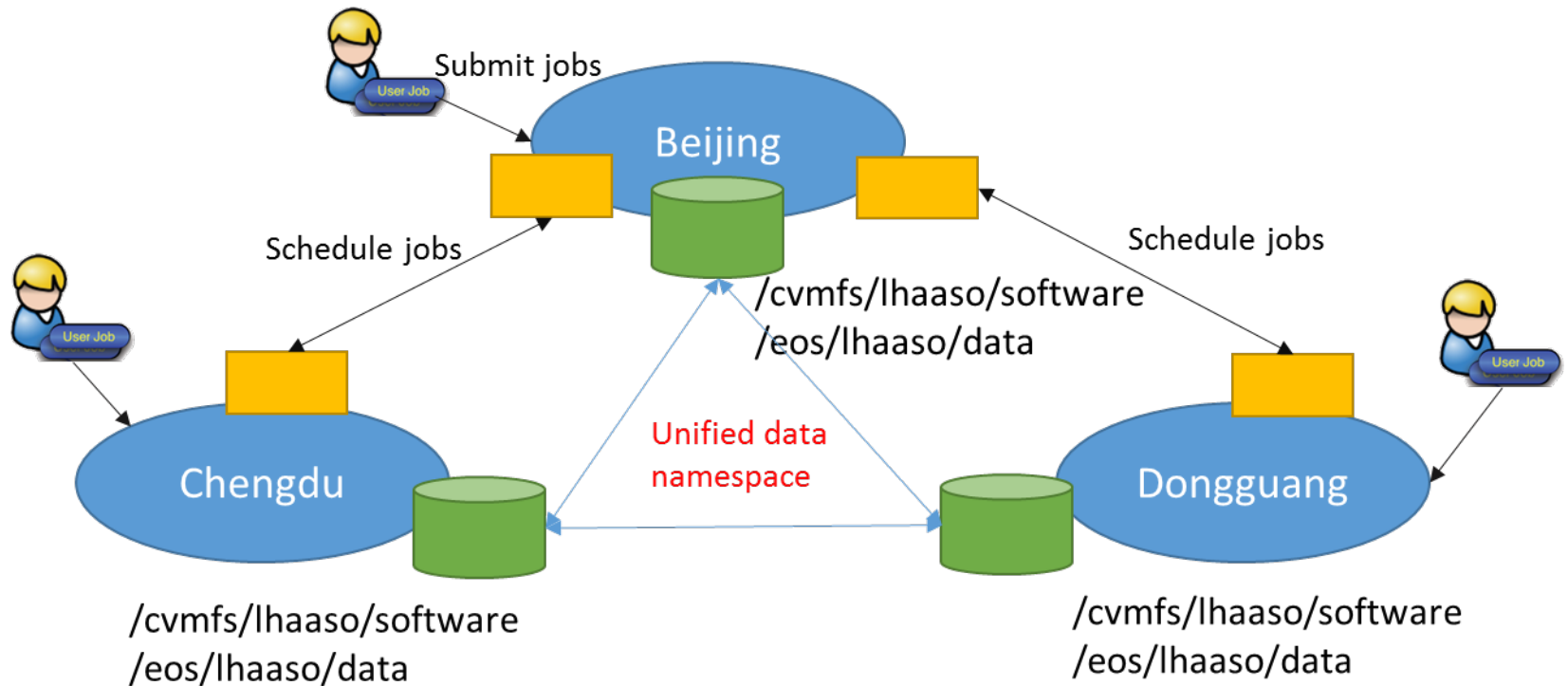
- **Key points**

- Unified distributed resource management
- To schedule jobs across regions transparently
- Dynamic resource provision to meet the peak demand
- Distributed monitoring and automated deployment
- Security certification



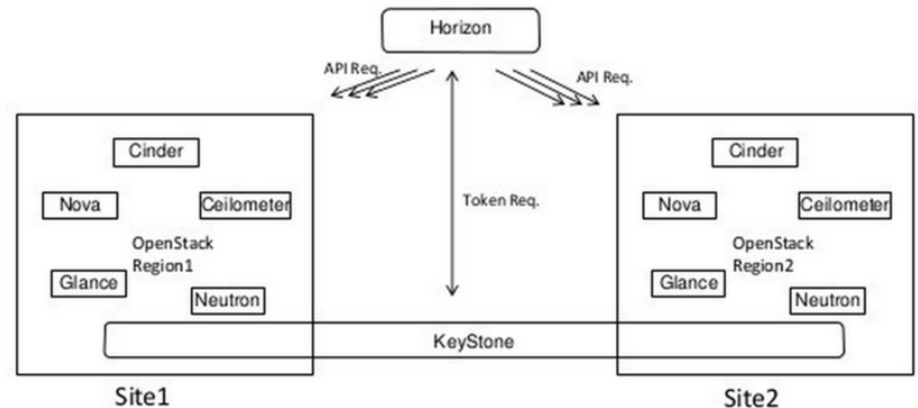
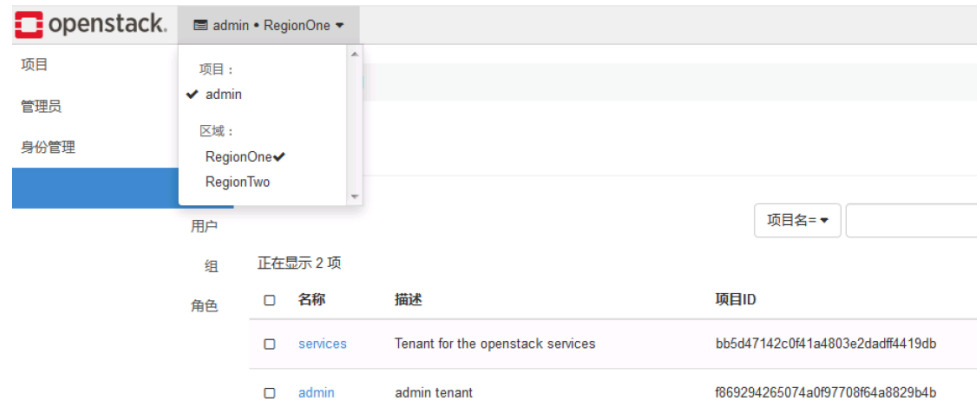
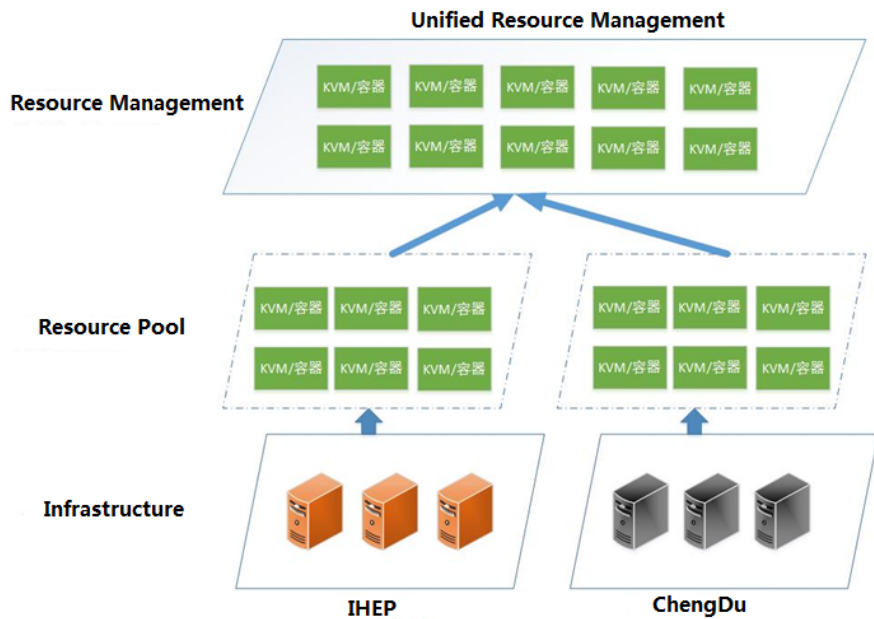
# Features

- Based cloud computing models to achieve unified management across regions
- Remote operation and maintenance
- Unified data namespace and support remote data access
- When the local site is busy, the jobs can be scheduled to remote sites **transparently**



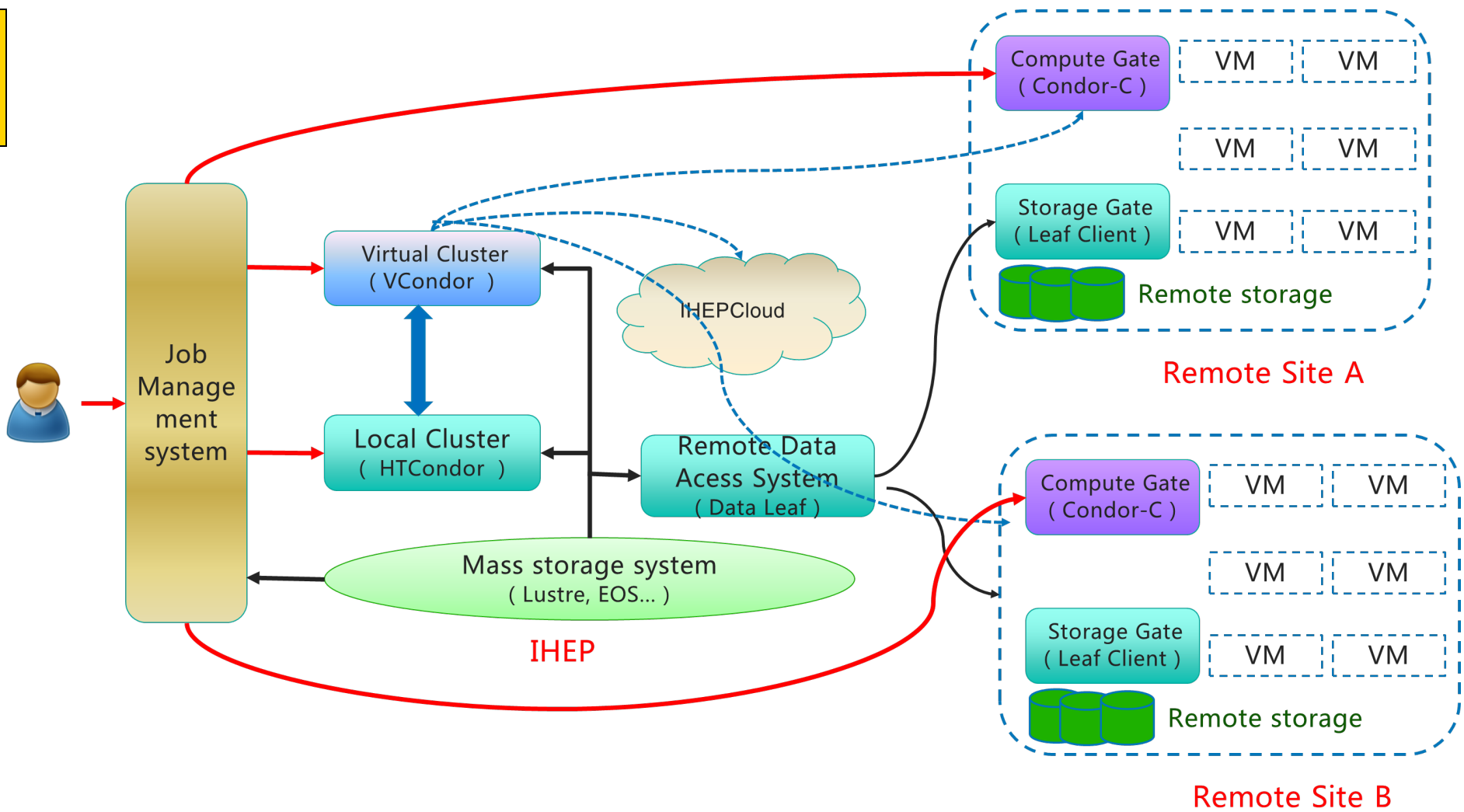
# Unified Resource Management

- Based on Openstack Queen and HTCondor
- Adopted Multi-region to manage the resources across domain
- A prototype is located in Beijing, Chengdu and Dongguang





# Job Scheduling(1)



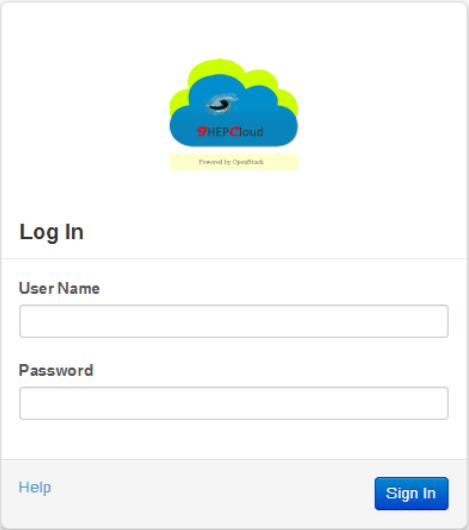
# Job Scheduling(2)

- **Job Management system**
  - Design the a toolkit(submit/query/delete) based on HTCondor
  - hep\_sub/hep\_rm/hep\_q
- **Schedule job to remote via "Condor-C" model**
  - Job in queue could be transferred to remote cluster via Condor-C
- **Dynamic resource provision**
  - Developed Vcondor
  - Remote control Vms dynamically
- **User management**
  - All the users uses AFS account managed by IHEP
  - Submit jobs from Login farm in IHEP
- **IHEP Virtual Cluster**
  - ~1000CPU cores
  - Resource provision dynamically to meet peak demand
  - Improve resource sharing between different experiments
  - Based on IHEPCloud(a private cloud)



# IHEPCloud

- Launched in May 2014, ~1670 CPU cores
- A private IaaS platform aiming to provide a self-service cloud platform for users and IHEP scientific computing
- SSO authentication: Any user who has IHEP email account (>1000 users, >87 active users)
- Three use scenario
  - User self-Service virtual machine platform
    - User register and destroy VM on-demand
  - Virtual Computing Cluster
    - Job will be allocated to virtual queue automatically when physical queue is busy
  - Distributed computing system
    - Work as a cloud site: Dirac call cloud interface to start or stop virtual work nodes



Log In

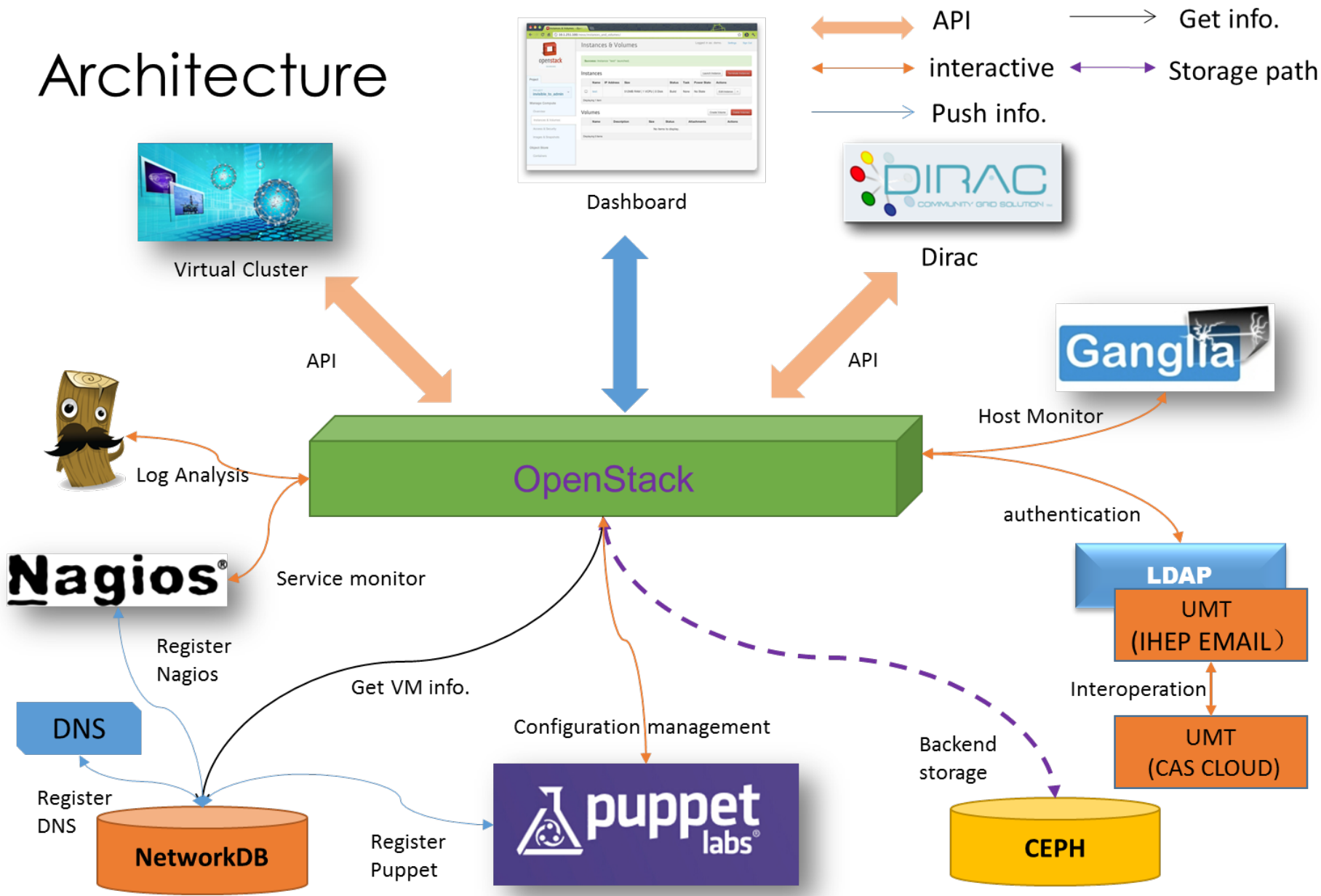
User Name

Password

[Help](#) [Sign In](#)

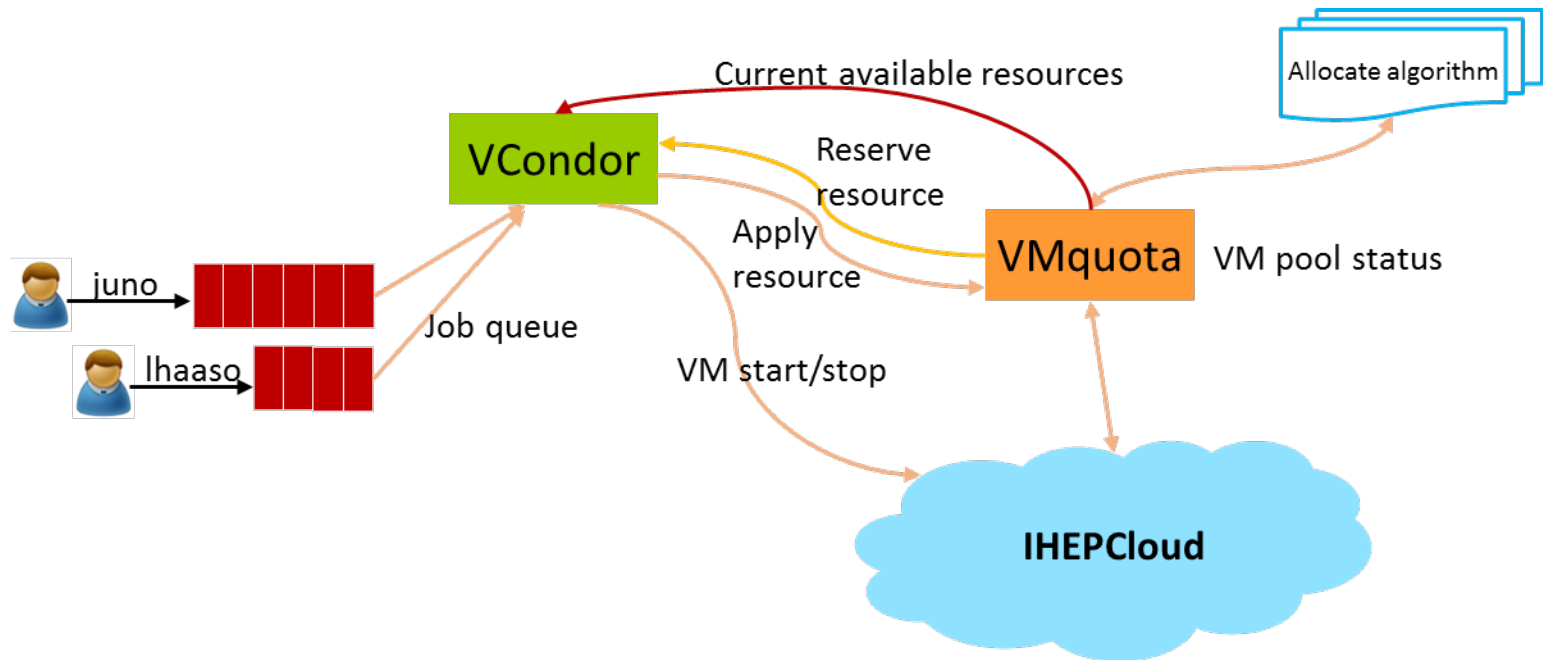


# Architecture

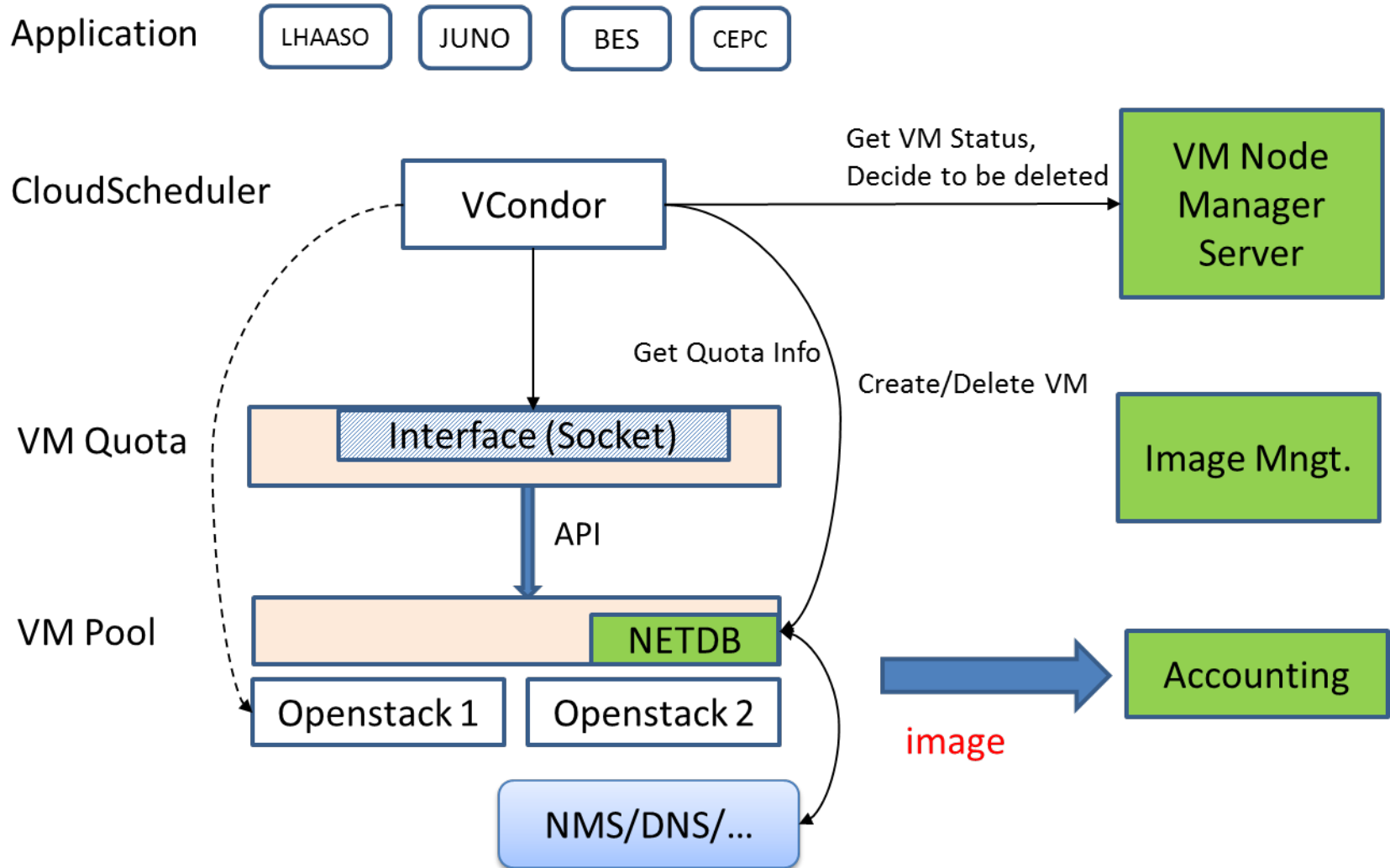


# CloudScheduler-Vcondor(1)

- VCondor is a cloud scheduler providing elastic resource allocation service based on HTCondor
- Take fine-grained resource allocation to schedule tasks instead of taking nodes.
- Design flexible allocating policy to provisioning VMs dynamically, considering job types, system load and cluster real-time status.



# CloudScheduler-Vcondor(2)

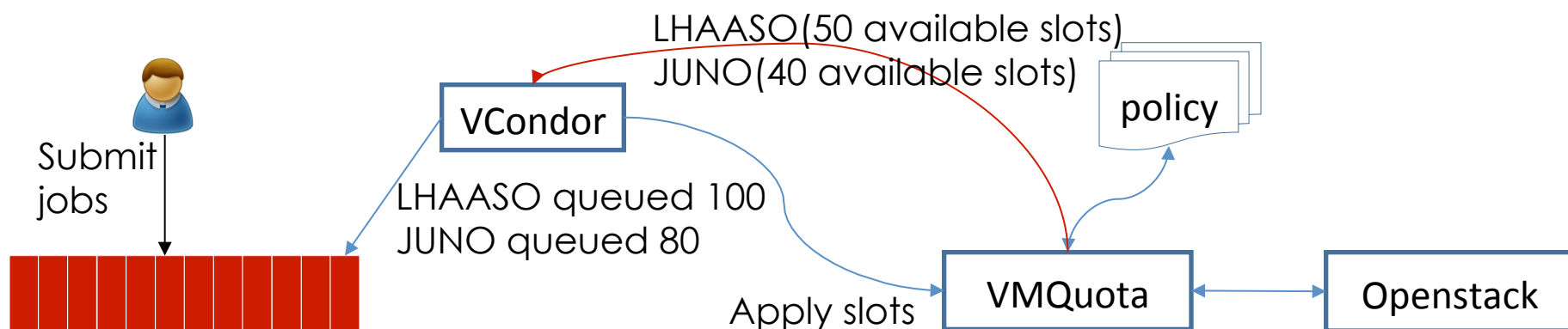




# VMQuota-Resource share management

- Set up virtual computing queues for each application, like LHAASO and JUNO
- Define the upper thresholds, lower thresholds, and resource reservation time for each queue

Queue	Lower thresholds	Upper thresholds	Available resource	Resource reservation time(seconds)
LHAASO	100	400	200	600
JUNO	100	300	200	600



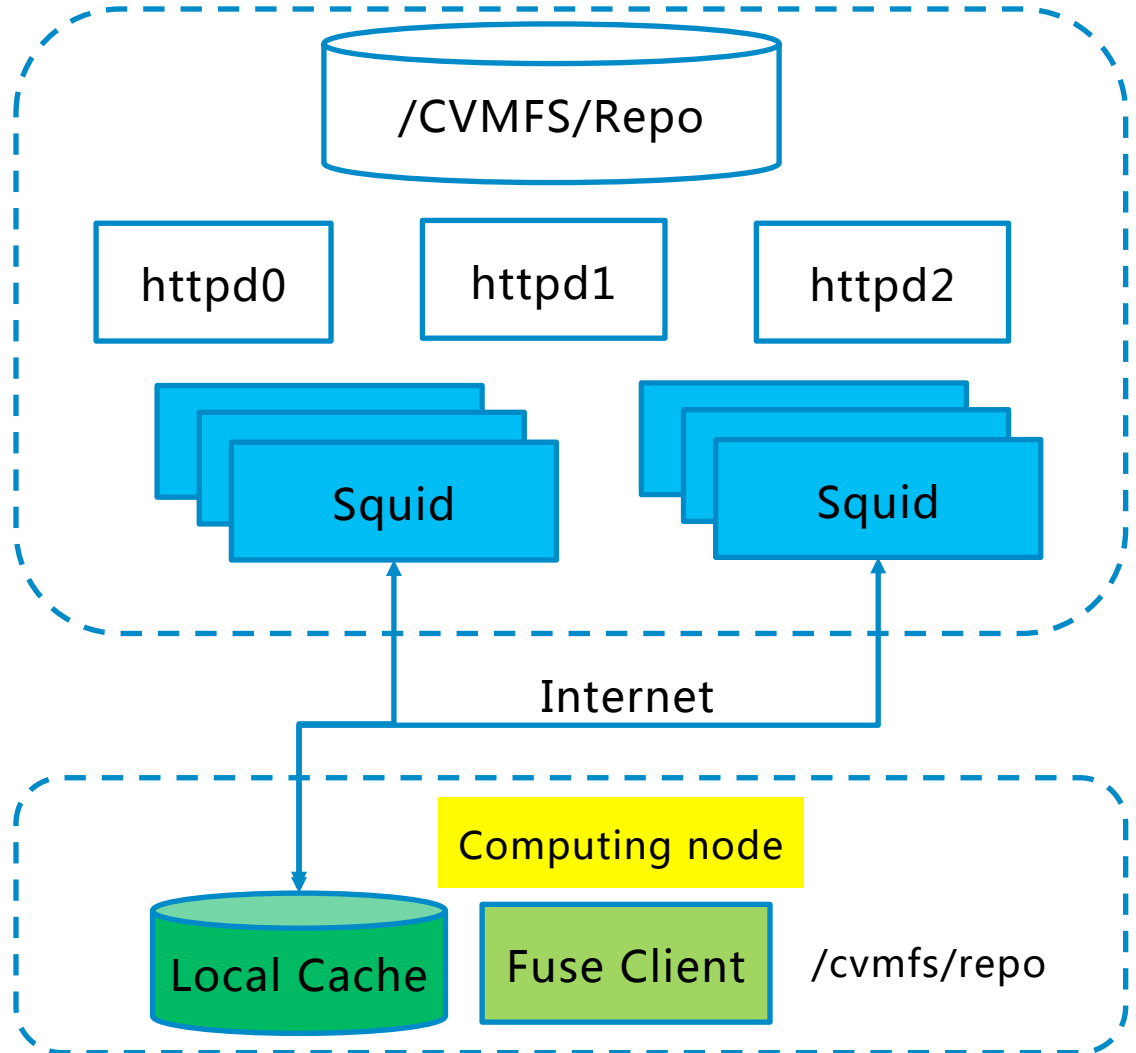
# Data Access(1)

- **Software**

- Physical library(Boss, Gaudi), common software(gcc),etc
- Version consistent among distributed sites

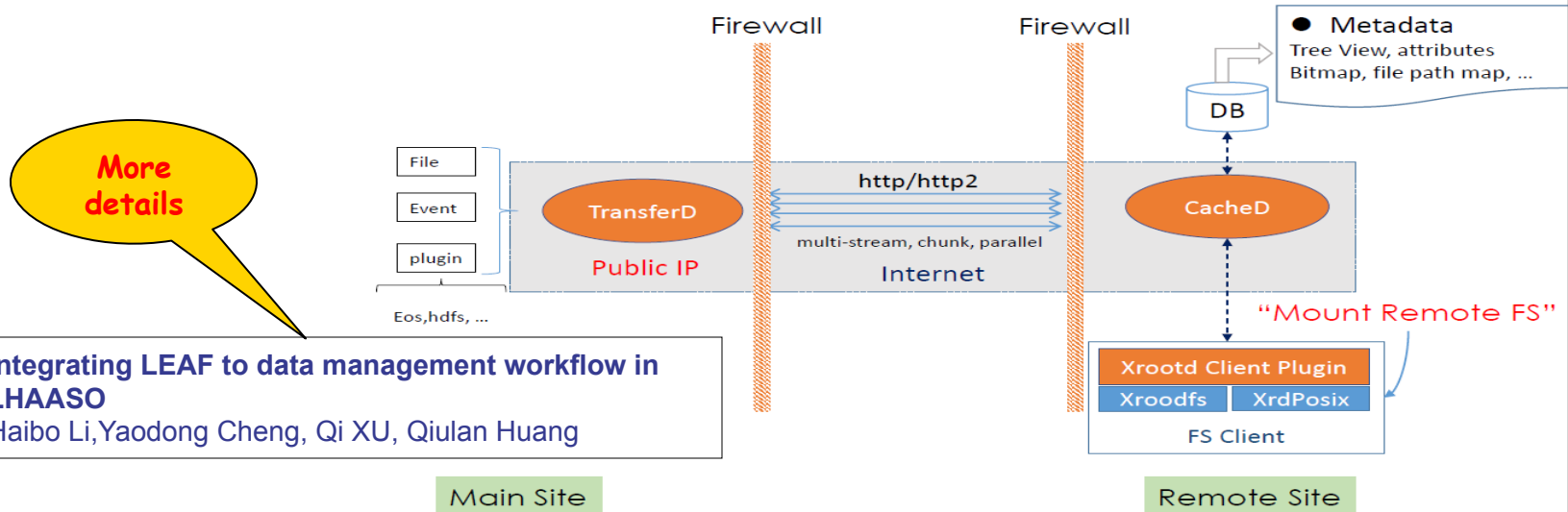
- **Solution**

- AFS
- CVMFS



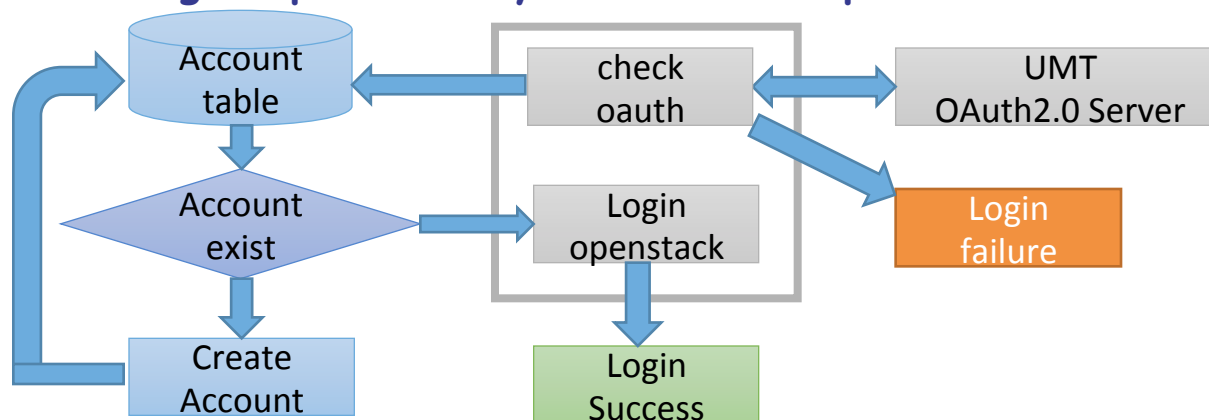
# Data Access(2)

- LEAF: a data cache and access system across remote sites
  - Full Metadata synchronization from main site periodically
  - Data transfer technologies: multi-stream, chunk, non-block, etc
  - Use HTTP protocol to go through firewall
  - Use Xrootd framework and fuse to mount file system



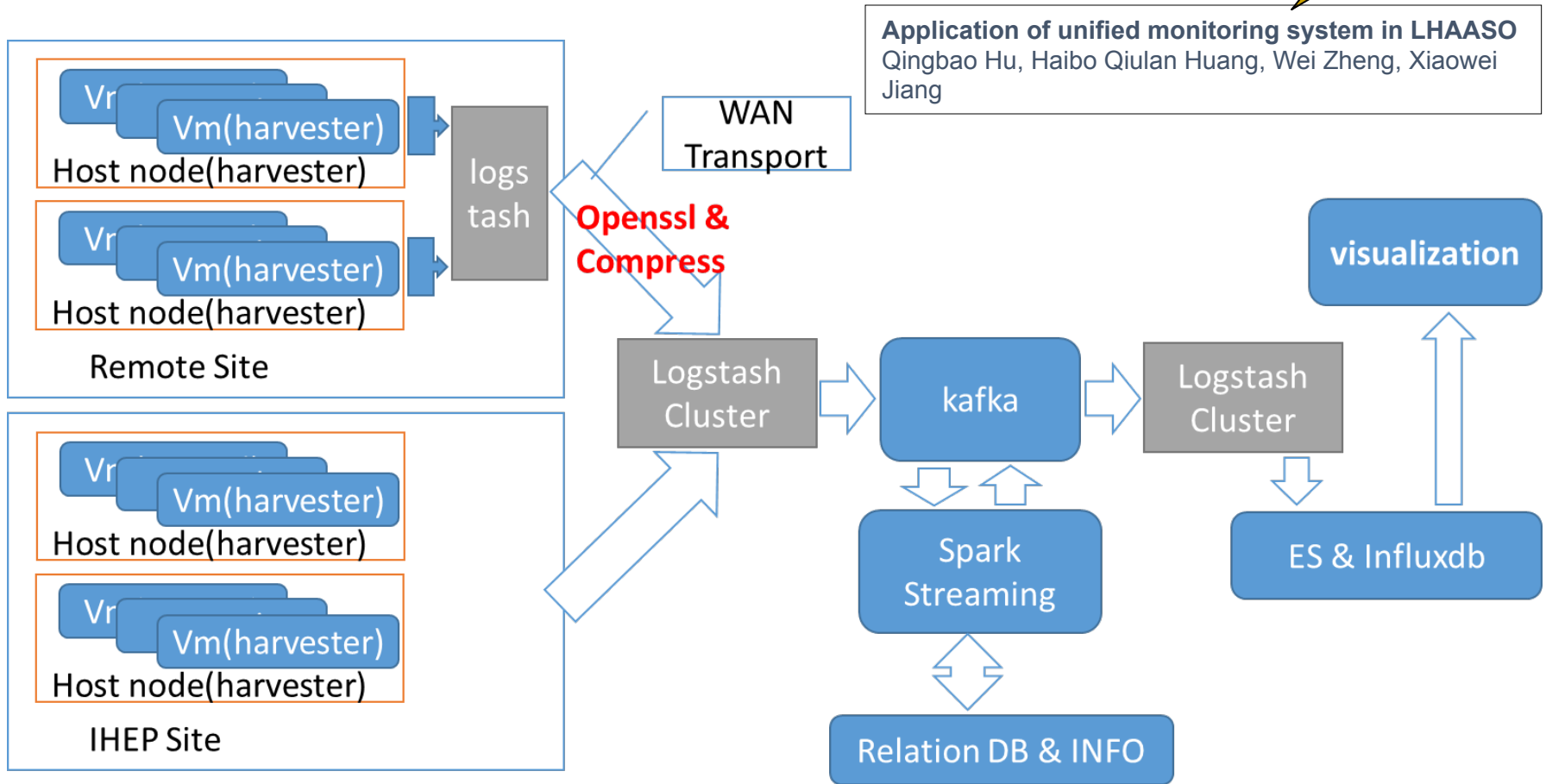
# Certification

- Single Sign On(SSO) in IHEP
- Motivation
  - Lack of the ability to achieve unified user management access control to the Openstack cloud.
  - IHEP UMT users can login openstack with their UMT password.
- A lightweight access control
  - A module is added to interact with the oauth2.0 server.
  - An table created to record all Openstack accounts and their passwords.
  - After user authorized by UMT, openstack will create a new account or find an exist account-password for the current user.
  - UMT users login Openstack by the recorded password automatically.



# Distributed Monitoring

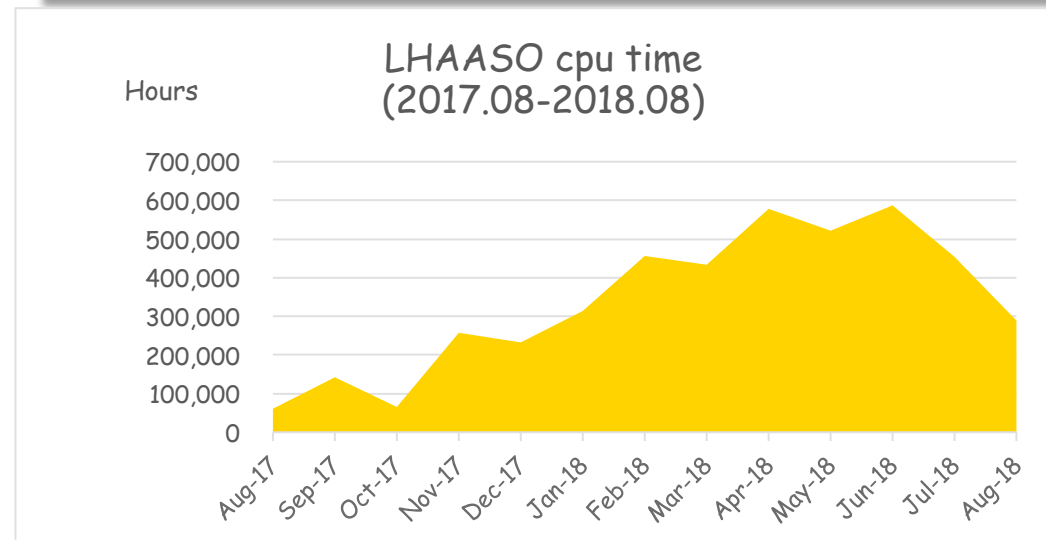
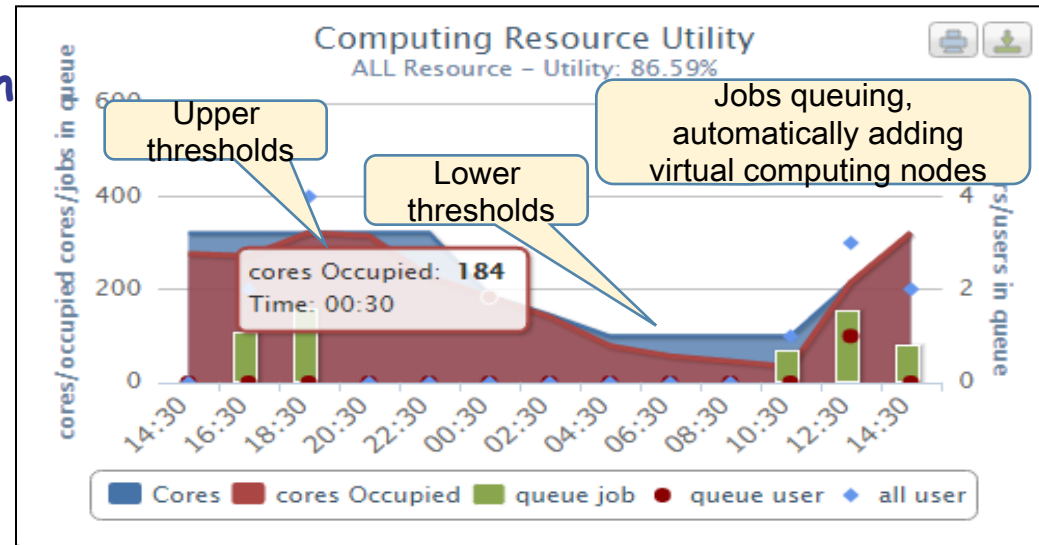
More details



- Remote site monitoring data compression and encryption transmission
- The virtual machine and host monitor can display together
- Real-time streaming data processing

# Running status

- The cloud-based computing system became in production in September, 2014
  - Local cloud cluster: ~1000 CPU cores.
  - ~30,000 jobs, 250,000 CPU hours in average one week
  - **1,509,424 jobs, 4,394,655 CPU hours (2017.8-2018.8)**
  - Resource utilization reaches up to 86.59%
- The resource of remote site is being integrated
  - ChengDu, Dongguang
  - Alibaba Cloud resource is also evaluated.

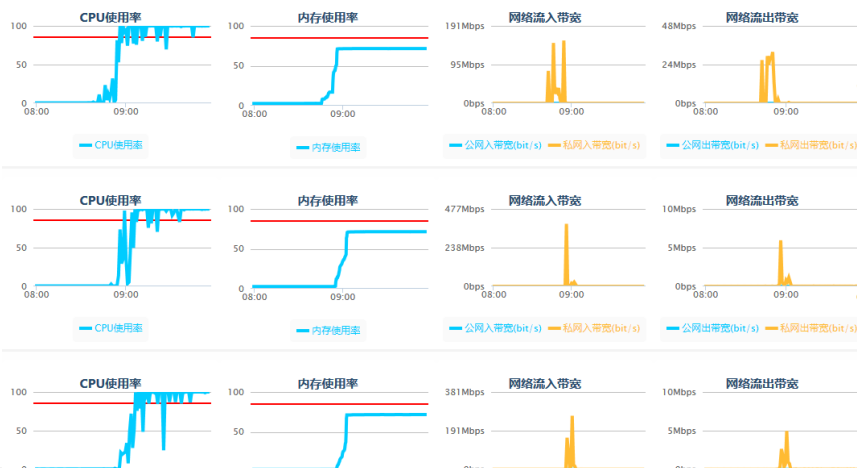




# Evaluation on Alibaba Cloud

- Expand the Vcondor to remote control Vms
  - Modular designed
  - Called the Alibaba Cloud API to create/start/stop/delete VMs
- CPU benchmark(The performance is optimistic)

	2 CPU 4GB	2CPU 8GB	4CPU 8G	4CPU 16GB	8CPU 16GB
Alibaba Cloud	32.84	51.96	96.95	96.26	110.45
Equal	4 CPU cores Intel Xeon E5420 @ 2.50GHz	4CPU cores Intel Xeon E5620 @ 2.40GHz	6CPU cores Intel Xeon X5650 @ 2.66GHz		



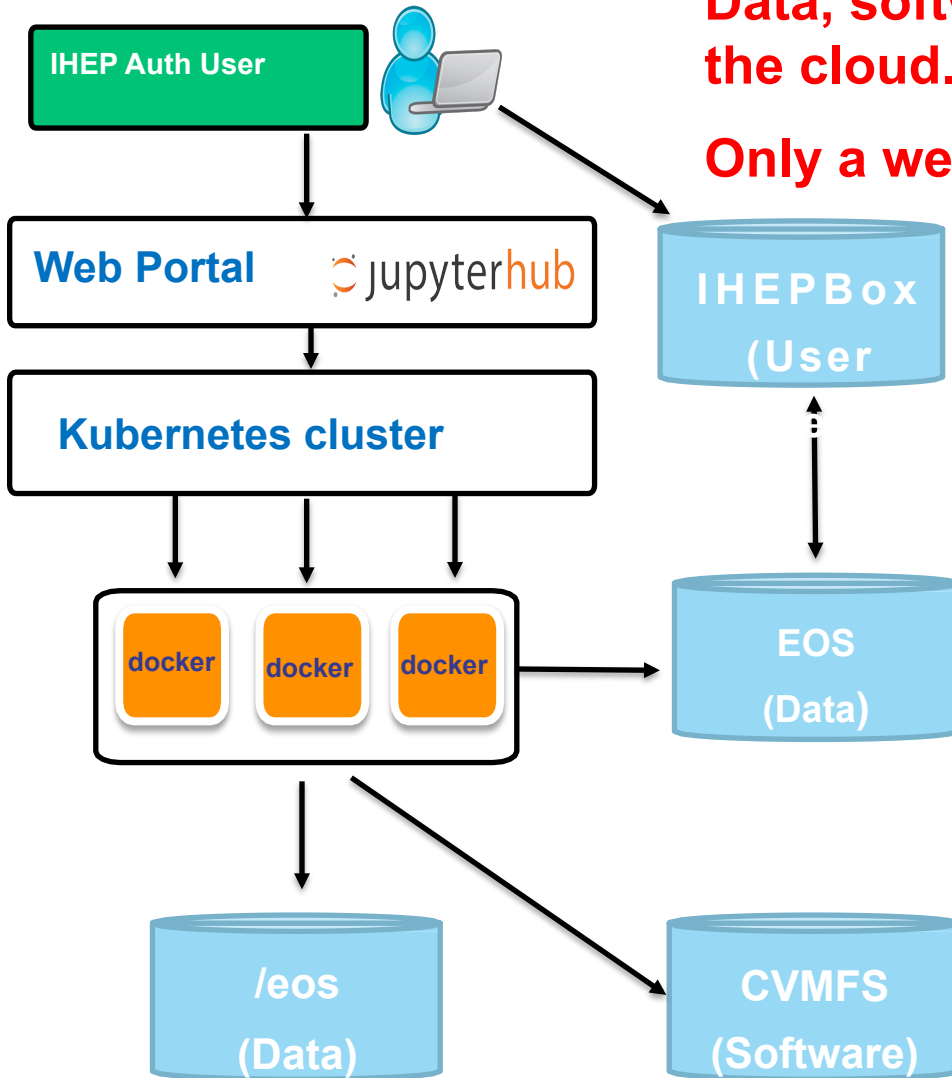
Test results shows the job efficiency running in Alibaba Cloud is equal to our local cluster

Some tests achieve better performance in Alibaba Cloud especially when the local farm is busy

# Web-based Data Analysis for LHAASO



# Web-based Data Analysis for LHAASO



**Data, software and computing resources in the cloud.**

**Only a web browser is needed!**

- **Authentication :** IHEP SSO
- **Infrastructure:**
  - JupyterHub+ IHEPBox +Kubernetes+ EOS
- **Software distribution:** CVMFS
- **Storage, in two flavours:**
  - DATA : EOS
  - User Files: IHEPBox
- **Jupyterhub-kubespawner**
  - enables JupyterHub to spawn single-user notebook servers on a Kubernetes cluster



# Data Analysis as a Service

- > **Analysis only with a web browser**
  - Available everywhere and at anytime
- > **Easy to use (but powerful)**
  - No local installation and configuration needed
- > **Create easily sharable scientific results: plots, data, code**
  - Storage is crucial: mass & synchronized
- > **Integration with existing resources**
  - Access software, user/experiments data, mass processing power
- > **Integration with other analysis ecosystems**
  - ROOT, R, Python, machine learning and deep learning...



# Summary

- **Virtualization and cloud computing technologies were adopted to support LHAASO experiment successfully**
  - > Using Openstack and HTCondor to integrate remote heterogeneous resources to expand computing scale
  - > Web based data analysis provides a portable, easy but powerful work mode
- **Easy to integrate more resources with this solution**
  - > More collaborated sites
  - > Commercial cloud
- **Many R&D activities are in progress**
  - > Enables JupyterHub to spawn users' notebook servers on a Kubernetes cluster
  - > Expand the function of job management toolkit
  - > LEAF
- > **Hope we could continue the collaboration with JINR on HEP computing**





Thank you  
Question?

