Grid and Cloud Computing at IHEP in China

Weidong Li

IHEP, CAS, Beijing

Grid2016 at Dubna

2016-07-04

Contents

- HEP experiments in China
- Computing environment at IHEP
  - Computing and storage
  - Network and data transfer
- WLCG Tier 2 in Beijing
- BESIII grid computing
- Cloud computing
- High performance computing
- Summary

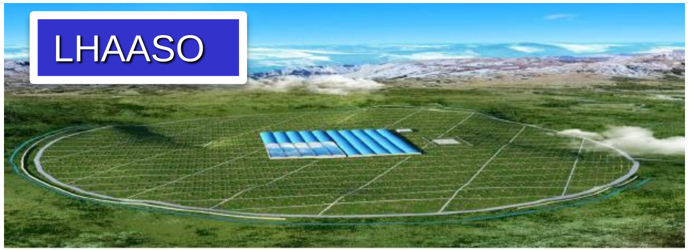
Experiments at IHEP

BESIII (Beijing Spectrometer III at BEPCII)

DYB (Daya Bay Reactor Neutrino Experiment)

JUNO (Jiangmen Underground Neutrino Observatory)

YBJ (Tibet-ASgamma ARGO-YBJ Experiments)



Large High Altitude Air Shower Observatory

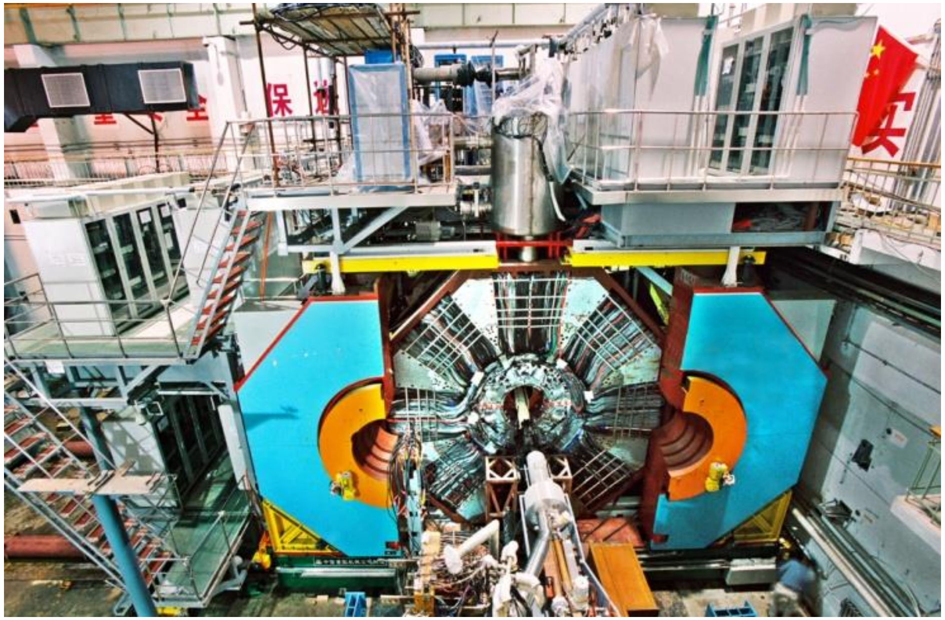
Hard X-Ray Moderate Telescope

BEPCII/BESIII

BEPC II: Beijing Electron-Positron Collider II

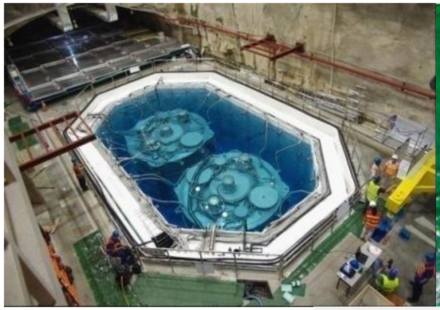
BES III: BEijing Spectrometer II, general-purpose detector on BEPC II

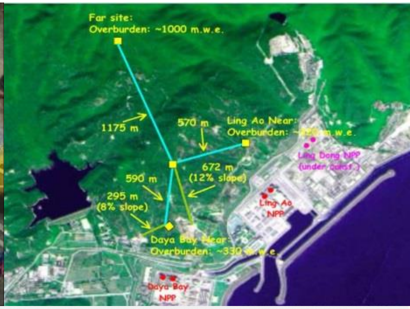
- Studying tau-charm physics
- Upgrade: BEPCII/BESIII, operational in 2008
- 2.0 ~ 4.6 GeV/C
- (3~10)×10^{32} $cm^{-2}s^{-1}$
- Produce ~100 TB/year raw data
- ~ 5000 CPU cores for data process and physics analysis



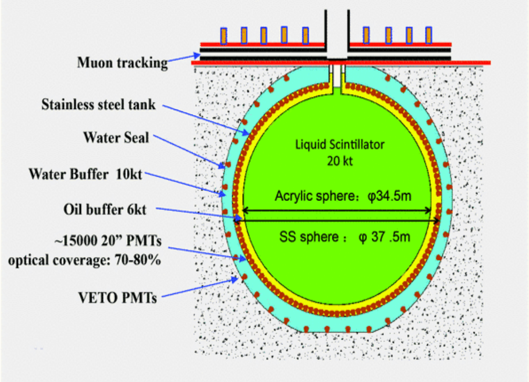
Neutrino experiments

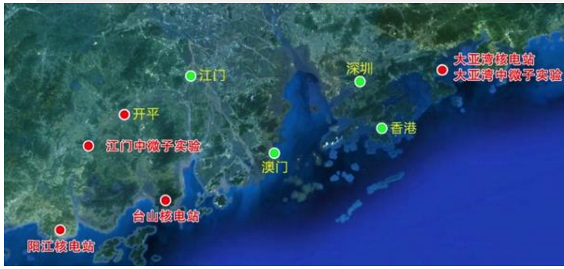
- Daya Bay Reactor Neutrino Experiment
  - To measure the mixing angle θ_{13}
  - 300 collaborators from 38 institutions
  - Produces ~200 TB/year (2011-2018)





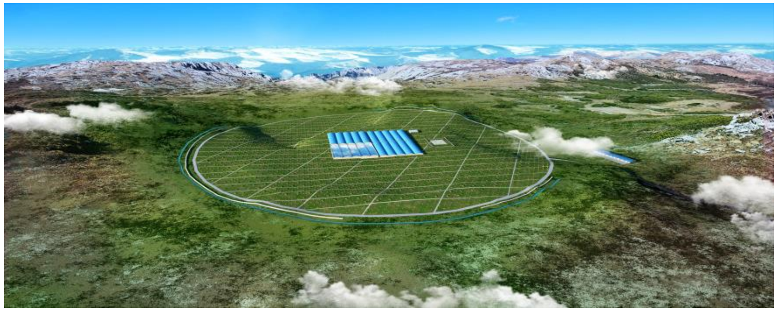
- JUNO - Jiangmen Underground Neutrino Observatory
  - Start to build in 2014, operational in 2020
  - 20 kt LS detector, 3% energy resolution
  - To determine the neutrino mass hierarchy using reactor antineutrino oscillations
  - Estimated to produce 2 PB data/year for 20 years





LHAASO

- Large High Altitude Air Shower Observatory, located on the border of Sichuan and Yunnan Province
- Multipurpose project with a complex detector array for high energy gamma ray and cosmic ray detection
- Expected to be operational in 2019
- ~1.2 PB data/year * 10 years
- On-site storage and computing resources. Data will be filtered and compressed and transferred back to IHEP.

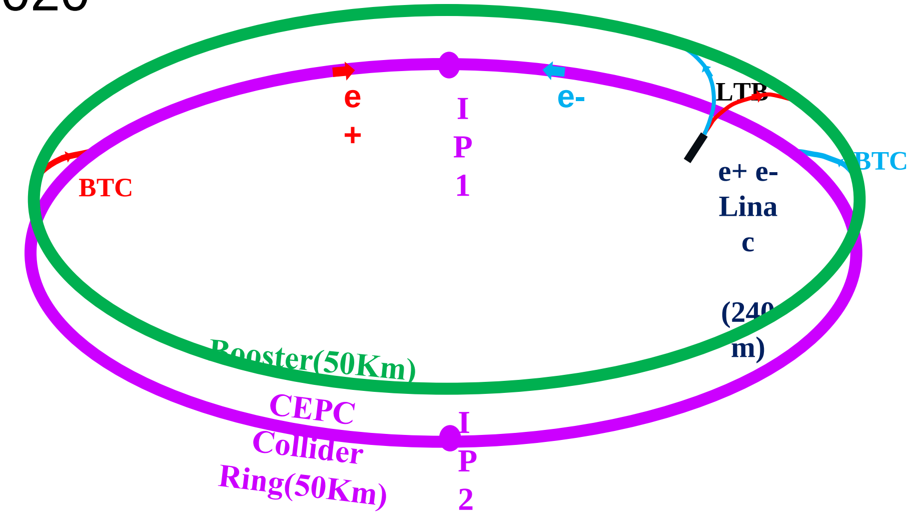


CEPC (1)

- Next Generation Accelerator in China after BEPCII which will complete its mission about 2021
- Two phases
  - CEPC (Circular Electron Positron Collider, e+e- ~ Higgs/Z factory)
    - Precision measurement of the Higgs/Z boson, about 12 years
    - Beam energy ~120 GeV
    - Estimated to produce 200TB/year raw data for Higgs factory and >100PB/year for Z factory
  - SPPC(Super Proton Proton Collider, pp ~ A discovery machine)
    - Discover new physics
    - Beam energy ~50 TeV
    - Estimated to produce 100PB/year

CEPC (2)

- CEPC collider is planed to build with the 50/100 km ring
- CEPC timetable
  - Pre-study, R&D and preparation work
    - pre-study: 2013-2015
    - R&D: 2016-2020
    - Engineering Design: 2015-2020
  - Construction: 2021-2027
  - Data taking: 2028-2035



Computing resources

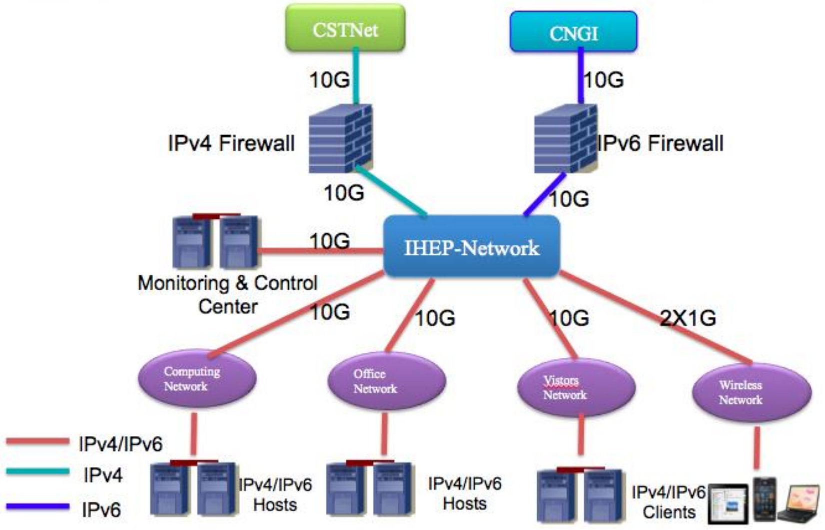
- Local clusters
  - ~13,500 CPU cores
  - 300 GPU cards
  - Scheduler:
    - PBS-2.5.5 with Maui-3.3.1
    - HTCondor 8.2.5
- Grid site (WLCG)
  - 1,200 CPU Cores
  - CreamCE (PBS-2.5.5 with Maui-3.3.4)
- The BESIII DIRAC-based distributed computing system
  - ~ 2,000 CPU cores
- IHEPCloud based on Openstack
  - ~ 720 CPU cores

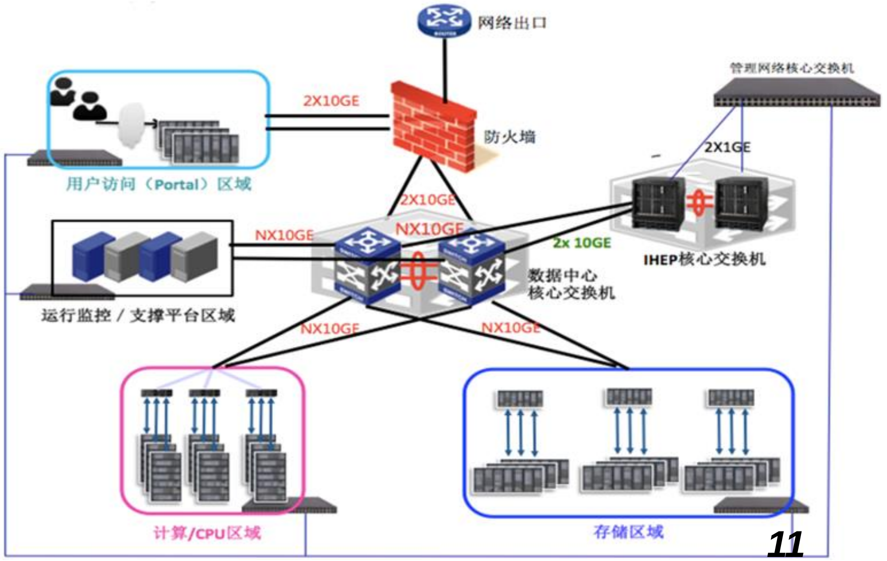
Storage

- **Lustre as main disk storage**
  - Capacity: 5.7 PB storage
- **Gluster system**
  - 734TB storage with replica feature
- **DPM & dCache**
  - 940TB, With SRM interface
- **HSM, with modified CASTOR**
  - 2 tape libraries + 2 robots, 26 drives
  - Capacity: 5 PB

Network at IHEP

- For office users
  - The largest campus network and bandwidth among all CAS institutes
    - 10G backbone
    - IPv4/IPv6 dual-stack
    - Wireless covered at (>250 APs)
  - Email/web/ services
  - >3000 end users
- For the data center at computing center
  - 160 Gbps (4X40Gbps) for 2-layer switches
  - 2X10 Gbps for storage nodes

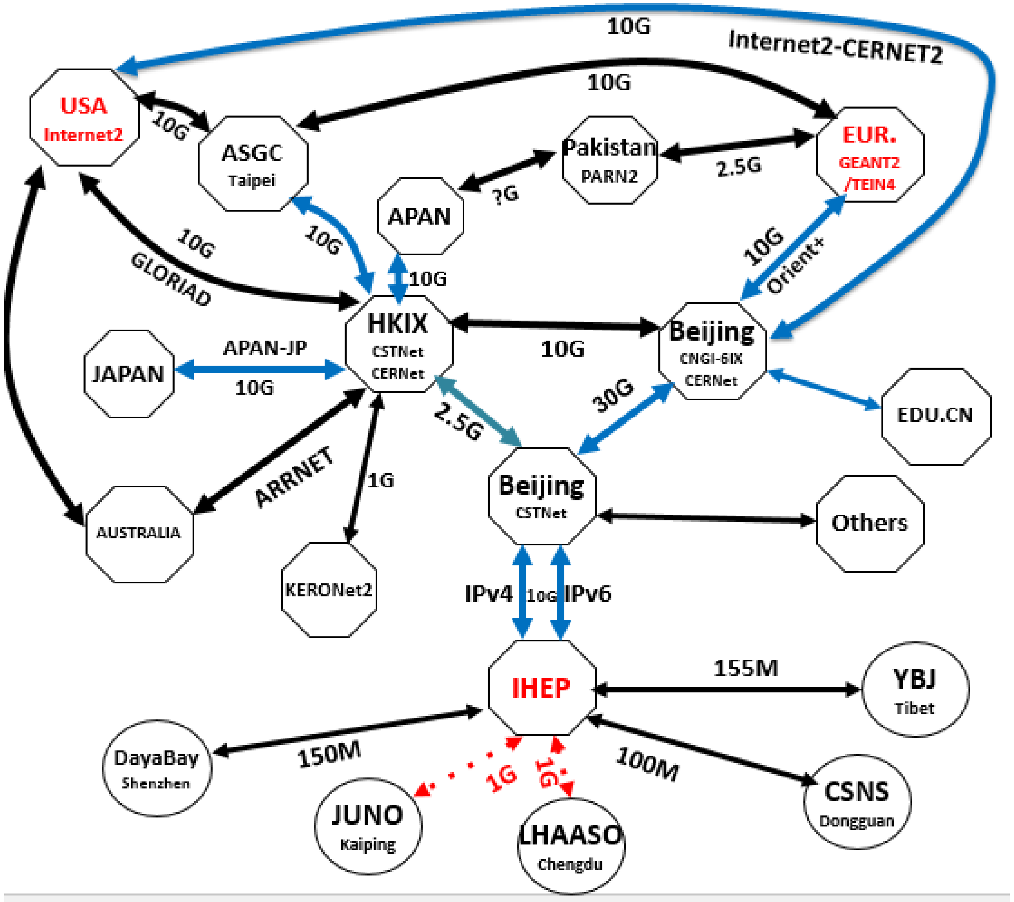




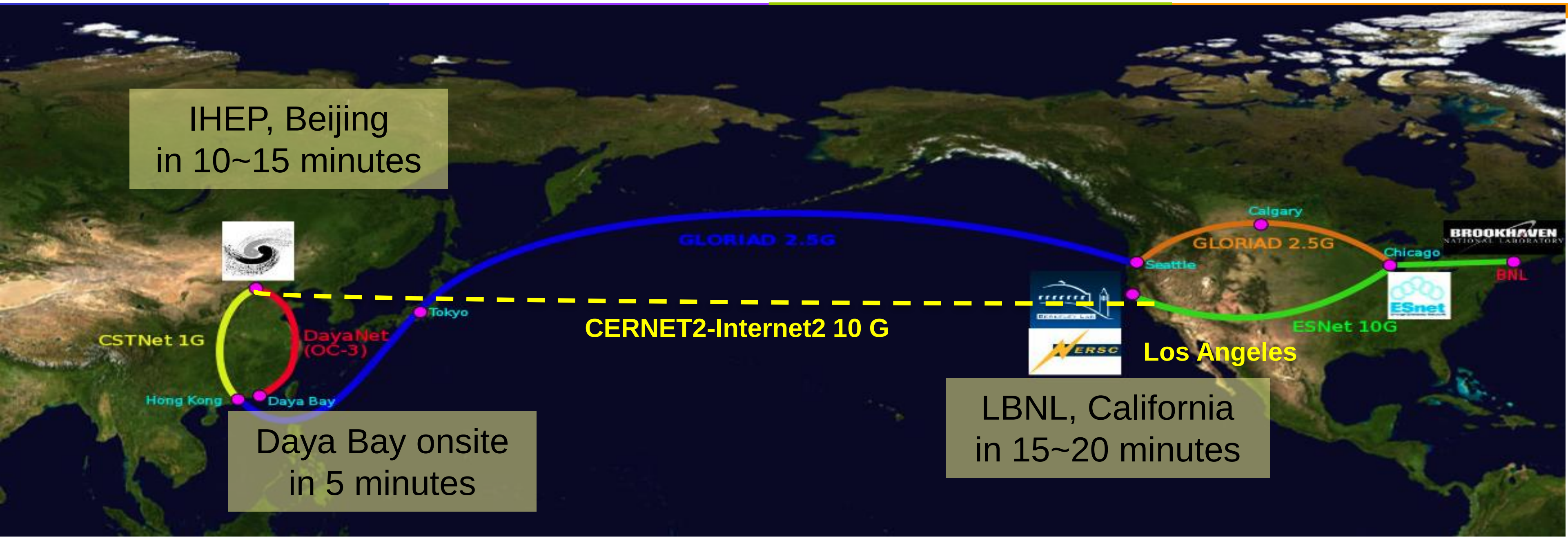
International and domestic links

- Dedicated Links for three other IHEP sites (two in the future)
  - Shenzhen (Dayabay)
  - Dongguan (CSNS)
  - Tibet (YBJ/ARGO)
  - Kaiping (JUNO)
  - Chengdu (LHAASO)

- Good Internet connections
  - IHEP-Europe: 10 Gbps
  - IHEP-USA: 10 Gbps
  - ~4 PB/year data exchange



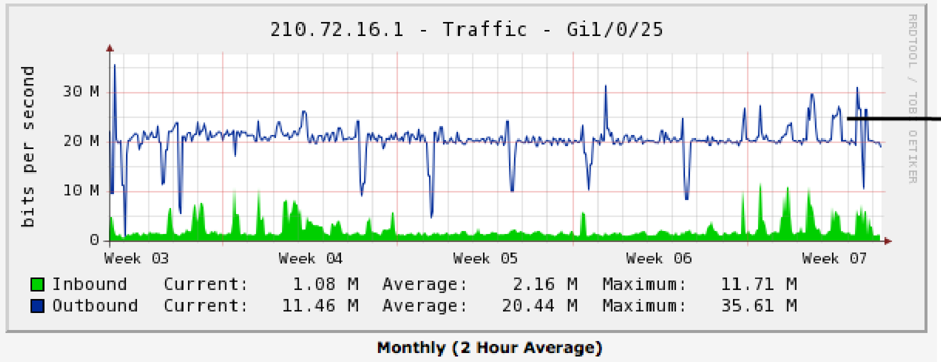
Data Transfer: DYB (1)



IHEP, Beijing
in 10~15 minutes

Daya Bay onsite
in 5 minutes

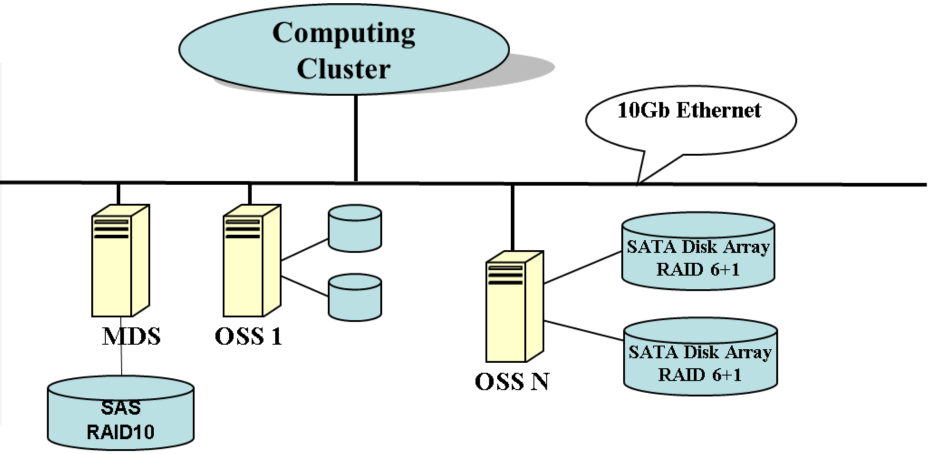
LBNL, California
in 15~20 minutes



```
🟩 Inbound   Current:   1.08 M  Average:   2.16 M  Maximum:  11.71 M
🟦 Outbound  Current:  11.46 M  Average:  20.44 M  Maximum:  35.61 M
```

Monthly (2 Hour Average)

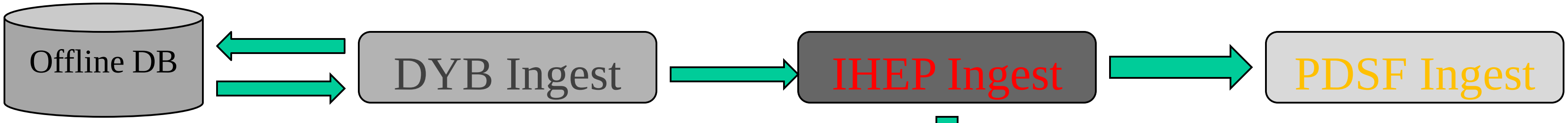
Daya Bay onsite network monitoring

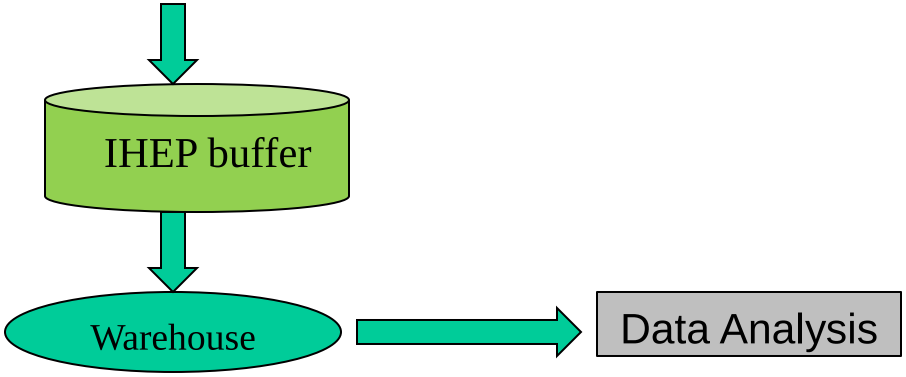


Infrastructure of data storage

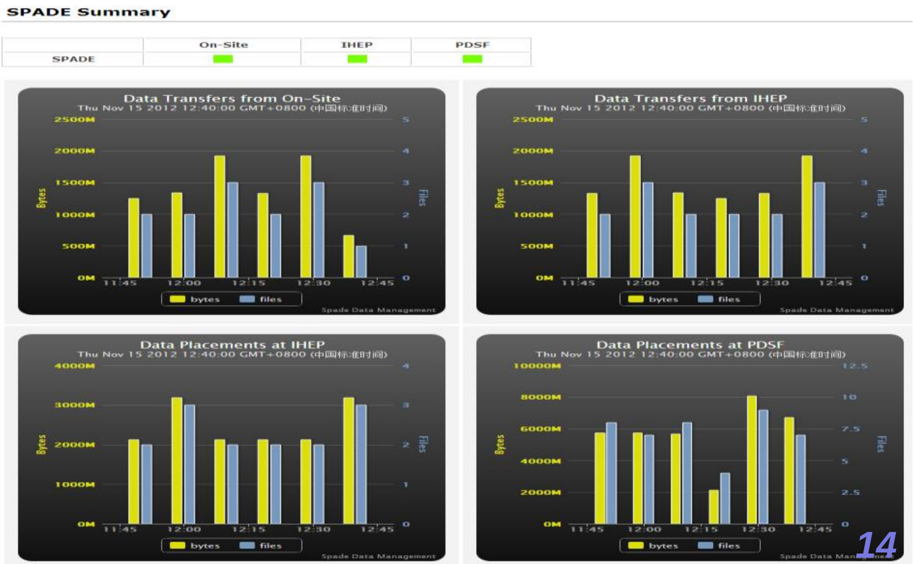
Data Transfer: DYB (2)

- 250 GB raw data per day





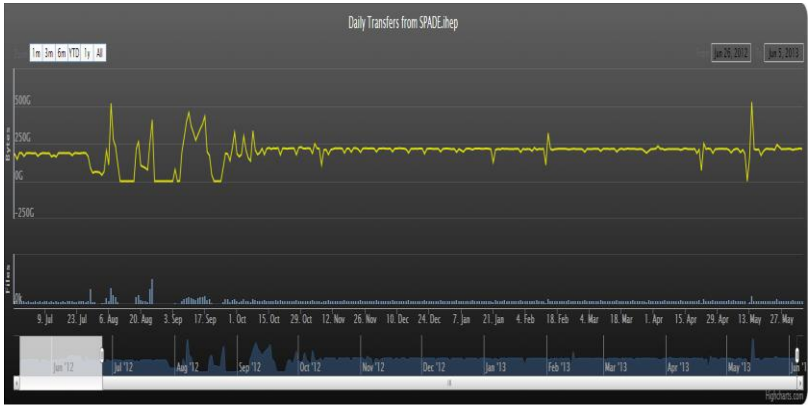
- The raw data are transferred from onsite back to IHEP and then relayed to the PDSF at LBNL.
- The data transfer system is called SPADE.



SPADE Summary

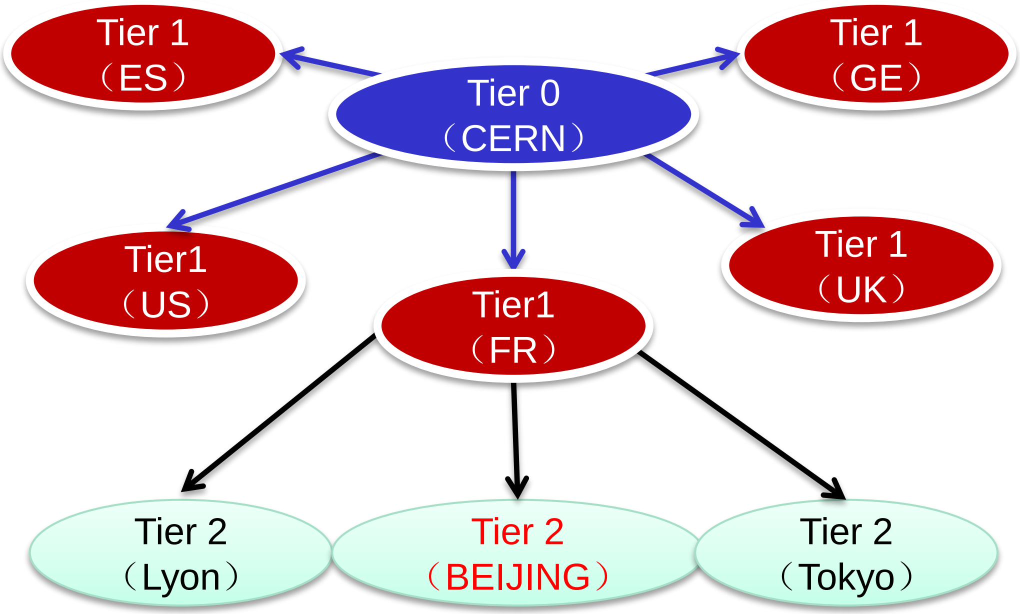
	On-Site	IHEP	PDSF
SPADE	🟩	🟩	🟩

Data Transfers from On-Site
Thu Nov 15 2012 12:40:00 GMT+0800 (中国标准时间)
Data Transfers from IHEP
Thu Nov 15 2012 12:40:00 GMT+0800 (中国标准时间)
Data Placements at IHEP
Thu Nov 15 2012 12:40:00 GMT+0800 (中国标准时间)
Data Placements at PDSF
Thu Nov 15 2012 12:40:00 GMT+0800 (中国标准时间)
Bytes
Files
2500M
2000M
1500M
1000M
500M
0M
4000M
3000M
10000M
8000M
6000M
11:45
12:00
12:15
12:30
12:45
bytes
files
Spade Data Management

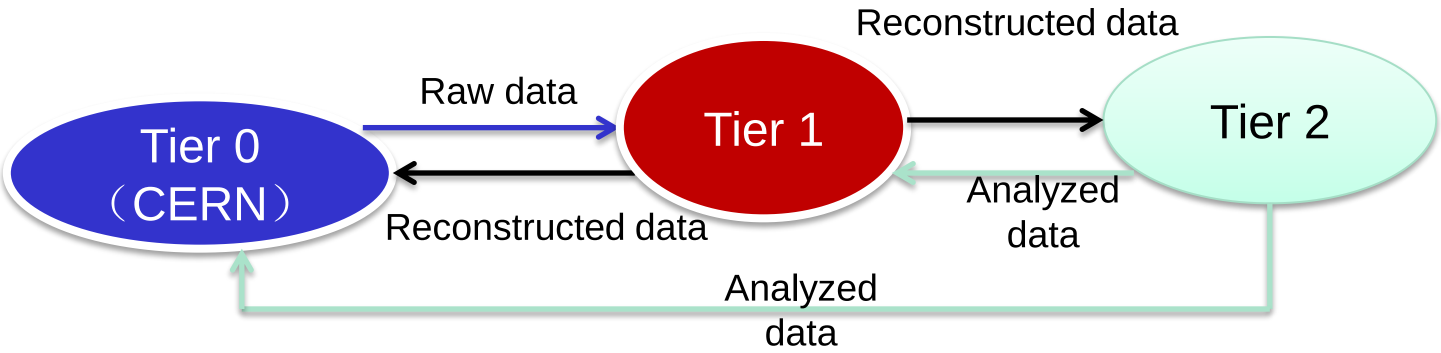


Grid computing for LHC

Raw data->Reconstructed data-> Analyzed data



- WLCG is consists of 1 Tier0, 11 Tier1s, over 100 Tier2s
- Raw data generated at Tier0, distributed to Tier1s
- Reconstructed data generated at Tier1s, transfer back to Tier0, and distributed to Tier2
- Analyzed data generated at Tier1/Tier2s, transfer back to Tier1s and Tier0

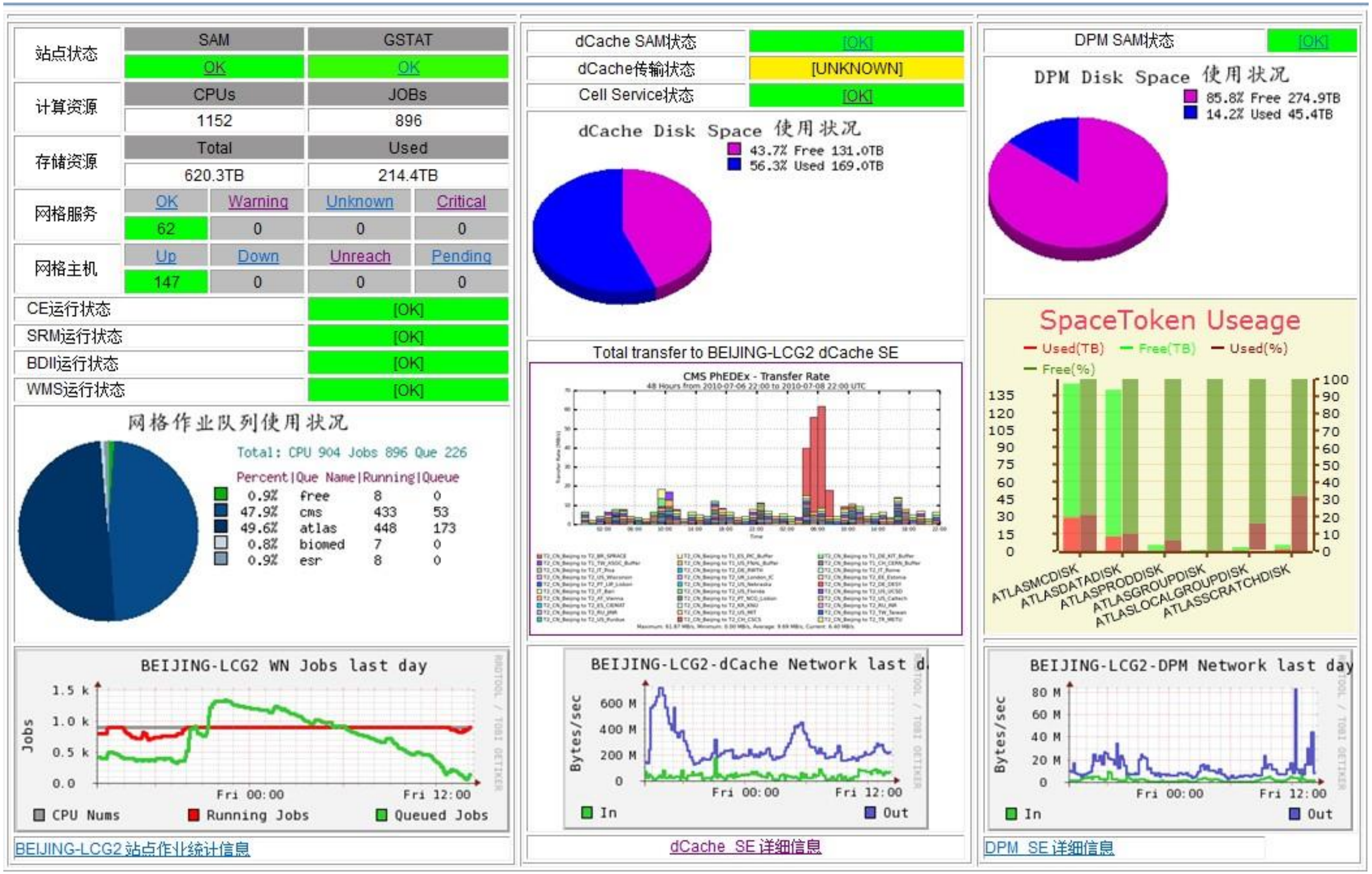


Beijing Tier-2 site (1)

- Tier 2 (BEIJING-LCG2) to support both CMS and Atlas
- ~1200 CPU resources shared between CMS and Atlas experiments
- 540TB for CMS dCache SE, 400 TB for Altas DPM SE
- In production since 2007, about 2M jobs every year

	CPU Hours (kSI2K-hours)	Jobs
2009	4.55 M	1.33M
2010	8.64 M	2.45 M
2011	11 M	4.79 M
2012	12 M	5.50 M
2013	7.7 M	1.87 M
2014	9.8 M	1.89 M
2015	7.0 M	2.15 M

Beijing Tier-2 site (2)



站点状态	SAM		GSTAT	
	OK		OK	
计算资源	CPUs		JOBs	
	1152		896	
存储资源	Total		Used	
	620.3TB		214.4TB	
网格服务	OK	Warning	Unknown	Critical
	62	0	0	0
网格主机	Up	Down	Unreach	Pending
	147	0	0	0

CE运行状态	[OK]
SRM运行状态	[OK]
BDII运行状态	[OK]
WMS运行状态	[OK]

网格作业队列使用状况

Total: CPU 904 Jobs 896 Que 226

	Percent	Que Name	Running	Queue
🟩	0.9%	free	8	0
■	47.9%	cms	433	53
■	49.6%	atlas	448	173
□	0.8%	biomed	7	0
■	0.9%	esr	8	0

BEIJING-LCG2 站点作业统计信息

dCache SAM状态	[OK]
dCache传输状态	[UNKNOWN]
Cell Service状态	[OK]

dCache Disk Space 使用状况

- 43.7% Free 131.0TB
- 56.3% Used 169.0TB

Total transfer to BEIJING-LCG2 dCache SE

CMS PhEDEx - Transfer Rate
48 Hours from 2010-07-06 22:00 to 2010-07-08 22:00 UTC
Transfer Rate (MB/s)
Time

dCache SE 详细信息

DPM SAM状态	[OK]

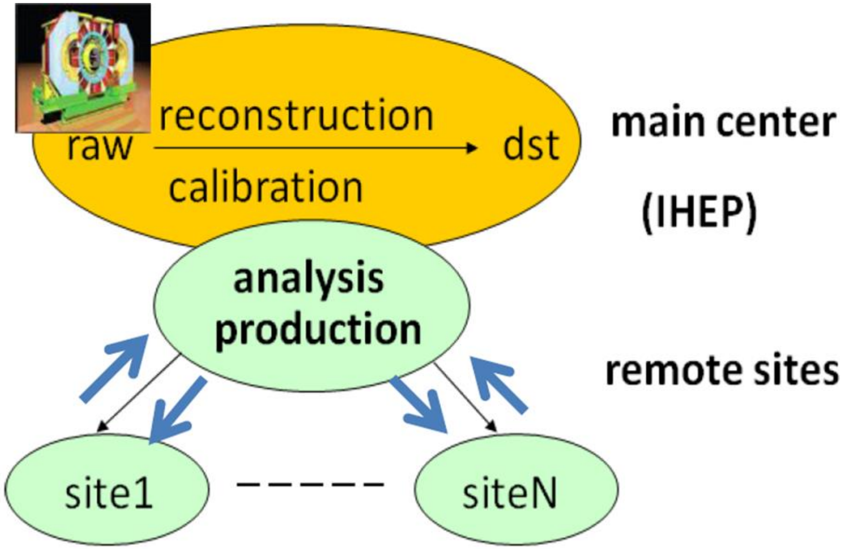
DPM Disk Space 使用状况

- 85.8% Free 274.9TB
- 14.2% Used 45.4TB

DPM SE 详细信息

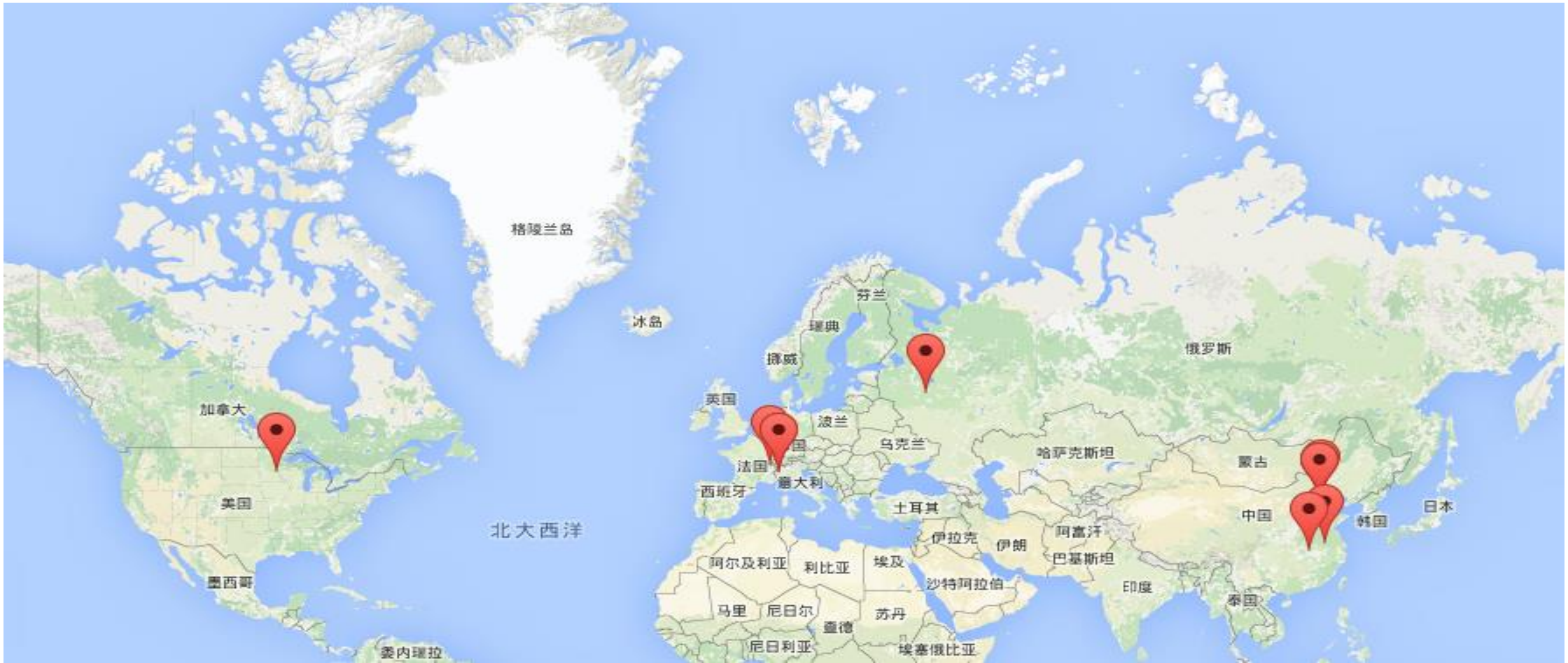
BESIII Grid Computing

- IHEP as central site
  - Raw data processing, bulk reconstruction, analysis etc
- Remote sites for peak needs
  - MC production, analysis
- Data flow
  - Central storage in IHEP
  - IHEP -> Sites, DST for analysis
  - Sites -> IHEP, MC data for backup



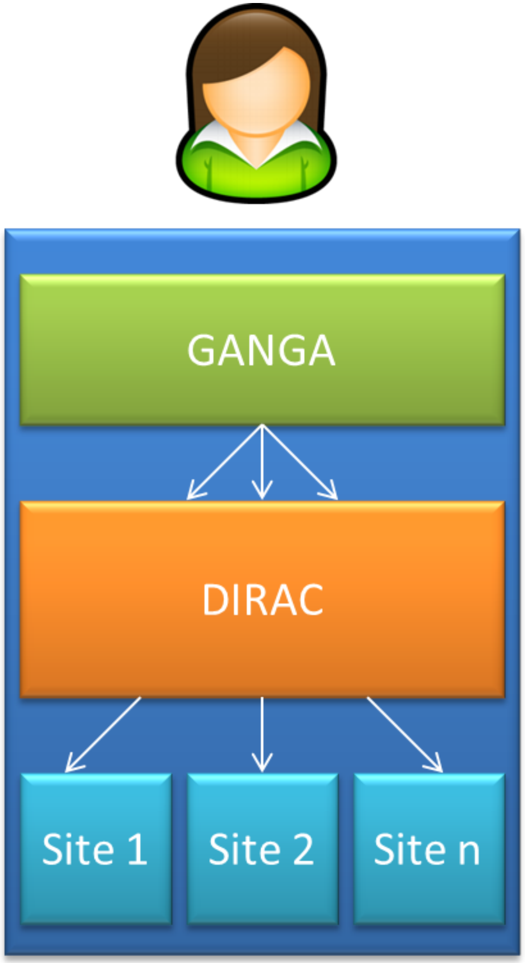
BESIII Grid resources

- About 14 sites from USA, Italy, Russia, China universities
- About 2000 cores CPU resources, 500 TB storage have been integrated
- 4 resource type resources are supported
  - Grid, Cluster, Cloud and Volunteer computing



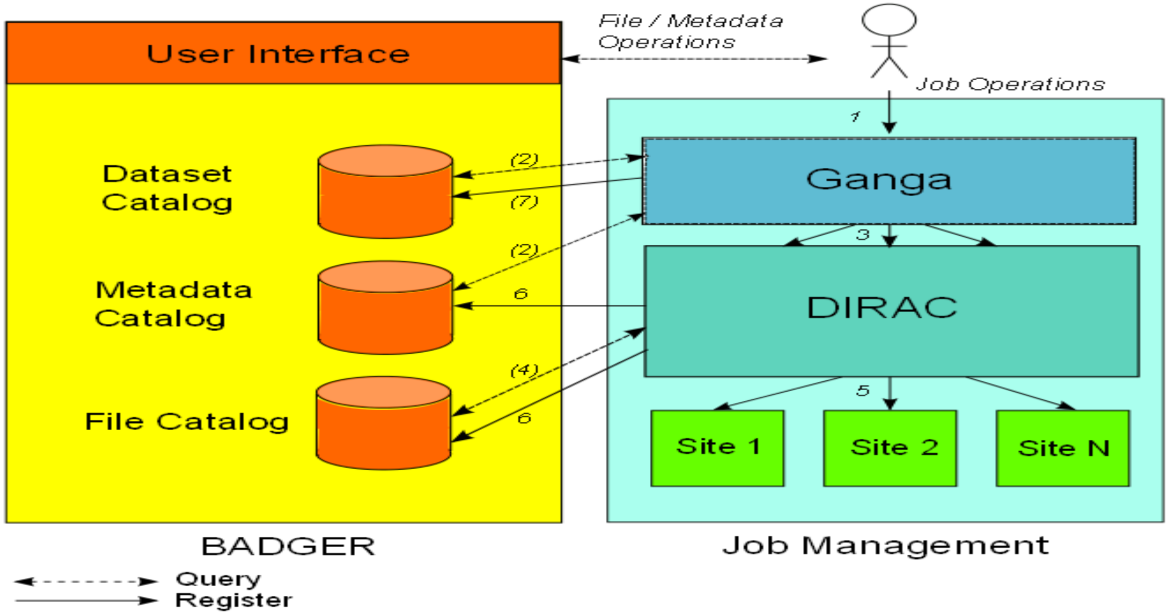
Workload management

- **Main Components**
  - DIRAC (**D**istributed **I**nfrastructure with **R**emote **A**gent **C**ontrol)
    - interware to cope with heterogeneous resources
  - GANGA and JSUB
    - Massive job submission user interface
  - CVMFS (CERN VM File System)
    - deploy experiment software to remote sites



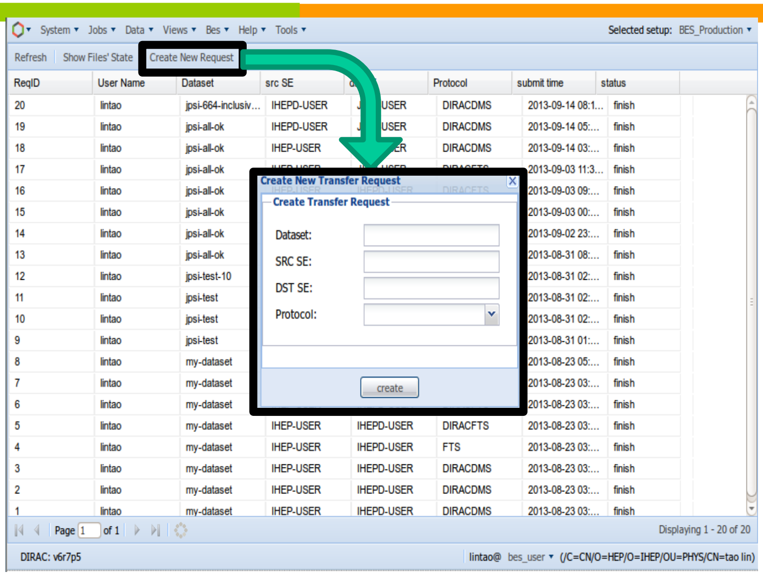
Data management

- Badger (BESIII Advanced Data Manager)
  - Based on DFC (Dirac File Catalogue)
  - Developed for BESIII file and metadata management
  - Replica Catalogue, Metadata Catalogue, Dataset Catalogue



Data transfer

- Data transfer system is designed and developed
  - Dataset supported
  - Massive transfer among sites
- Maximum speed can reach 1.9Gb/s at first production
  - close to IHEP outbound network bandwidth in 2014
- Each year, about 90TB data exchange



ReqID	User Name	Dataset	src SE	d	Protocol	submit time	status
20	lintao	jpsi-664-inclusiv...	IHEPD-USER	J USER	DIRACDMS	2013-09-14 08:1...	finish
19	lintao	jpsi-all-ok	IHEPD-USER	J USER	DIRACDMS	2013-09-14 05:...	finish
18	lintao	jpsi-all-ok	IHEP-USER	ER	DIRACDMS	2013-09-14 03:...	finish
17	lintao	jpsi-all-ok				2013-09-03 11:3...	finish
16	lintao	jpsi-all-ok				2013-09-03 09:...	finish
15	lintao	jpsi-all-ok				2013-09-03 00:...	finish
14	lintao	jpsi-all-ok				2013-09-02 23:...	finish
13	lintao	jpsi-all-ok				2013-08-31 08:...	finish
12	lintao	jpsi-test-10				2013-08-31 02:...	finish
11	lintao	jpsi-test				2013-08-31 02:...	finish
10	lintao	jpsi-test				2013-08-31 02:...	finish
9	lintao	jpsi-test				2013-08-31 01:...	finish
8	lintao	my-dataset				2013-08-23 05:...	finish
7	lintao	my-dataset				2013-08-23 03:...	finish
6	lintao	my-dataset				2013-08-23 03:...	finish
5	lintao	my-dataset	IHEP-USER	IHEPD-USER	DIRACFTS	2013-08-23 03:...	finish
4	lintao	my-dataset	IHEP-USER	IHEPD-USER	FTS	2013-08-23 03:...	finish
3	lintao	my-dataset	IHEP-USER	IHEPD-USER	DIRACDMS	2013-08-23 03:...	finish
2	lintao	my-dataset	IHEP-USER	IHEPD-USER	DIRACDMS	2013-08-23 03:...	finish
1	lintao	my-dataset	IHEP-USER	IHEPD-USER	DIRACDMS	2013-08-23 03:...	finish

Create New Transfer Request

Create Transfer Request

Dataset:

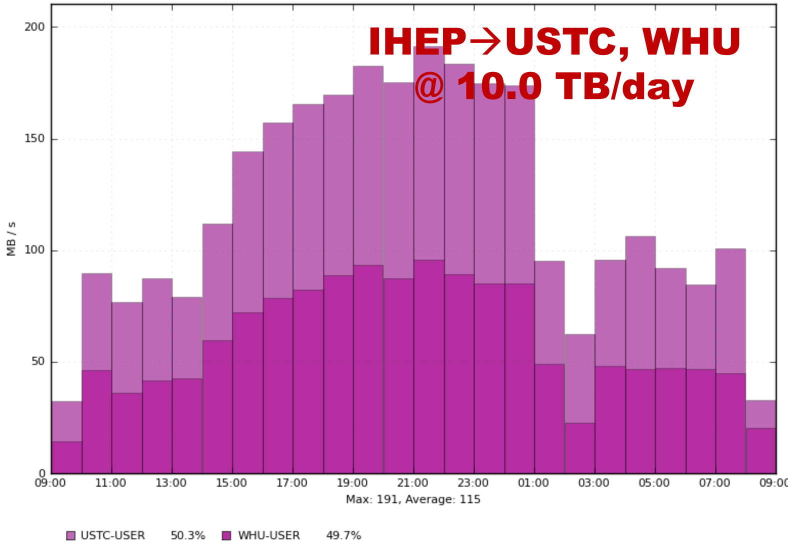
SRC SE:

DST SE:

Protocol:

Throughput by Destination

24 Hours from 2014-07-21 09:00 to 2014-07-22 09:00 UTC

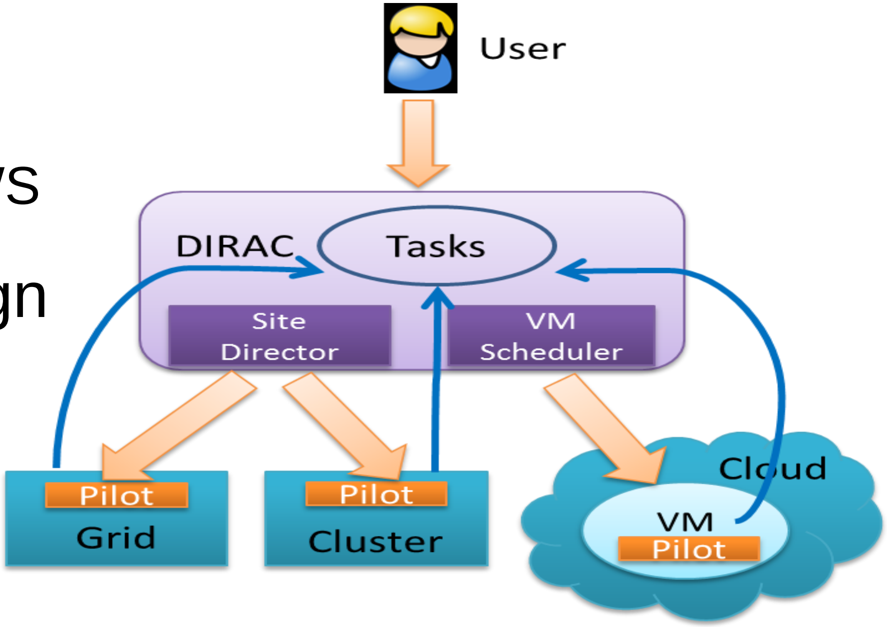


IHEP→USTC, WHU @ 10.0 TB/day

Generated on 2014-07-22 09:01:45 UTC

Cloud integration

- Elastic scheduling has been implemented for flexible resource allocation
  - Based on VMDIRAC1.0 with extra VM scheduler
- Cloud resources were in production since 2014, including
  - INFN, IHEP, JINR, CNIC
- Cloud types supported
  - OpenStack, OpenNebula, AWS
- VMDIRAC2.0 is under design
  - Easy configuration
  - Adopt new pilot tech

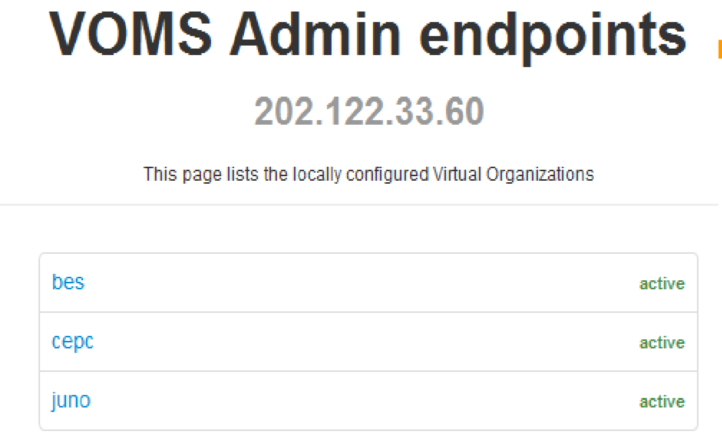


Integration of commercial clouds

- In June 2015, AWS cloud has been integrated
  - With the support of Amazon AWS China region
  - BOSS image created and upload to AWS
  - Connect with AWS API in VMDIRAC elastic scheduling
- Tests done and price evaluated
  - 400,000 BOSS rhopi events have been simulated with 100% success rate
  - c3.large is more suitable type than other CPU types
  - About 0.20 CNY for every 1000 events, mainly used by computing 92%
- Other domestic commercial clouds (eg. AliYun) are in the assessment process

Multi-VO supports (1)

- Motivation
  - More experiments express interests on using or evaluating distributed computing
  - Joint resources belongs to more than one experiments
  - Save manpower and simplify management of resources
- Multi-VO has been supported in one set-up
  - VOMS system to help classify different VO and groups
  - VO-based authentication and priority control to be added in DIRAC central scheduling system

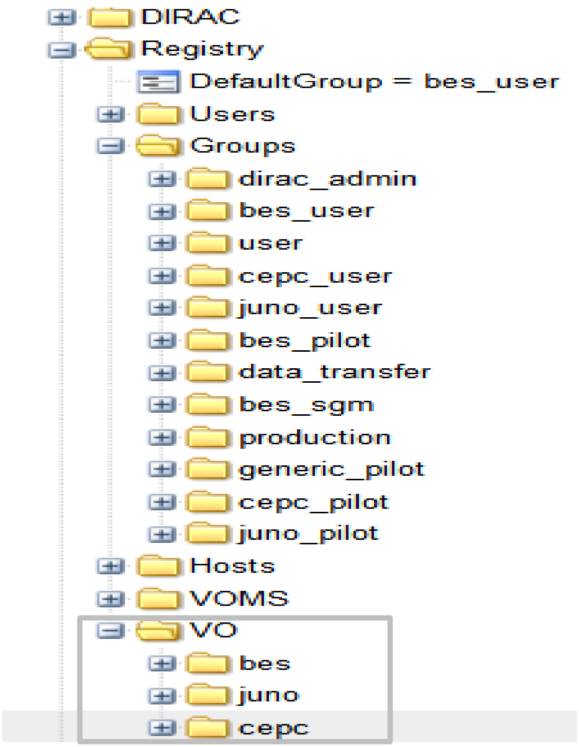


VOMS Admin endpoints

202.122.33.60

This page lists the locally configured Virtual Organizations

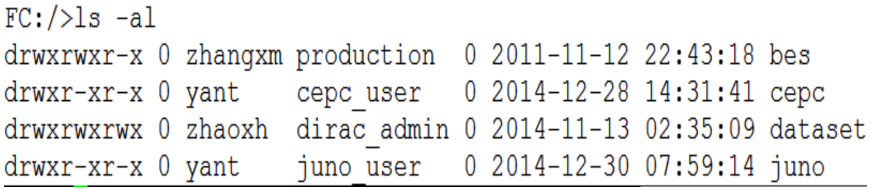
bes	active
cepc	active
juno	active



```
DIRAC
Registry
    DefaultGroup = bes_user
    Users
    Groups
        dirac_admin
        bes_user
        user
        cepc_user
        juno_user
        bes_pilot
        data_transfer
        bes_sgm
        production
        generic_pilot
        cepc_pilot
        juno_pilot
    Hosts
    VOMS
    VO
        bes
        juno
        cepc
```

Multi-VO supports (2)

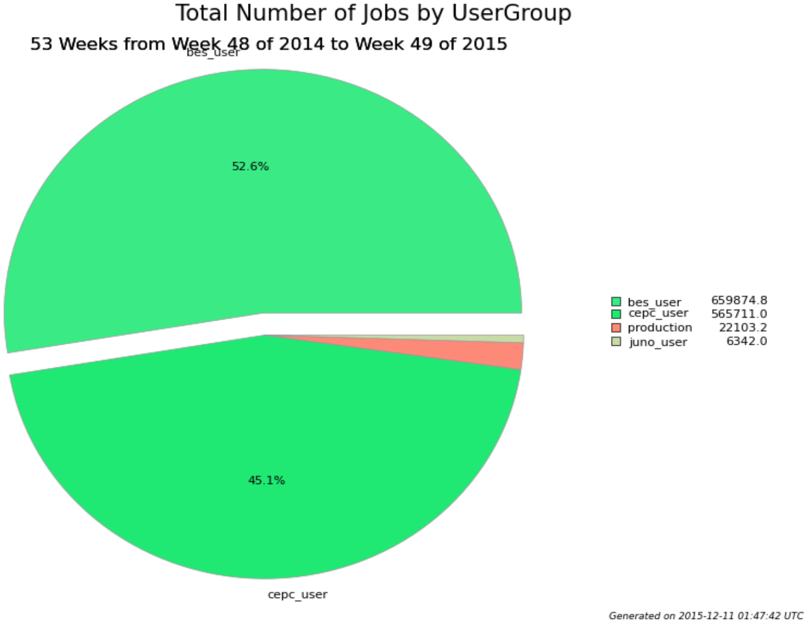
- Independent software publishing repositories defined in CVMFS
  - /cvmfs/boss.ihep.ac.cn, /cvmfs/cepc.ihep.ac.cn, /cvmfs/juno.ihep.ac.cn
- Badger and StoRM central storage have been extended to support multi-vo



```
FC:/>ls -al
drwxrwxr-x 0 zhangxm production  0 2011-11-12 22:43:18 bes
drwxr-xr-x 0 yant    cepc_user   0 2014-12-28 14:31:41 cepc
drwxrwxrwx 0 zhaoxh  dirac_admin 0 2014-11-13 02:35:09 dataset
drwxr-xr-x 0 yant    juno_user   0 2014-12-30 07:59:14 juno
```

❖ Current experiments supported

- BESIII, JUNO, CEPC

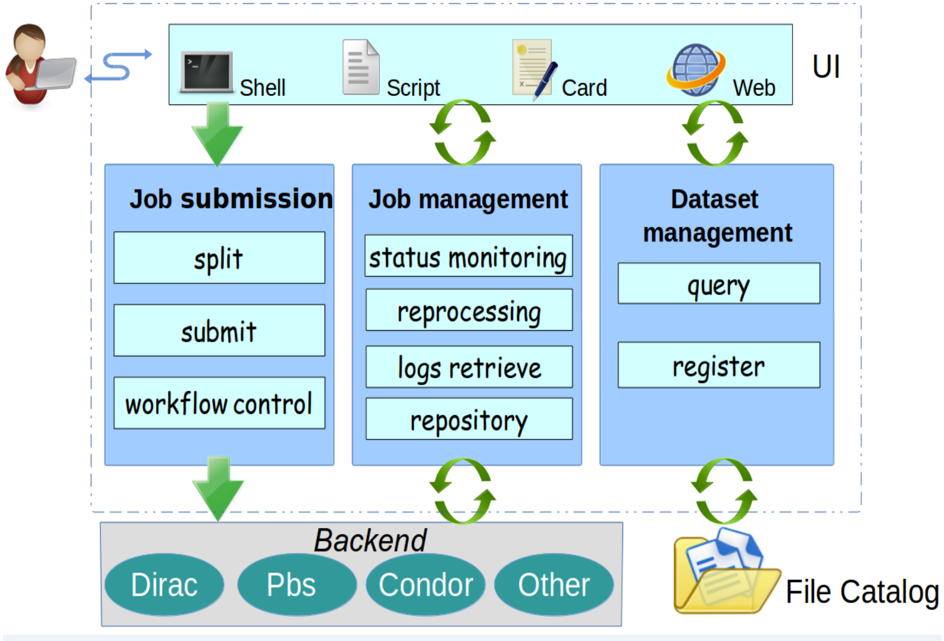


General task submission tool (JSUB)

- Aim to ease the procedure of experiments to use grid
- A general framework to take care of life cycle of tasks

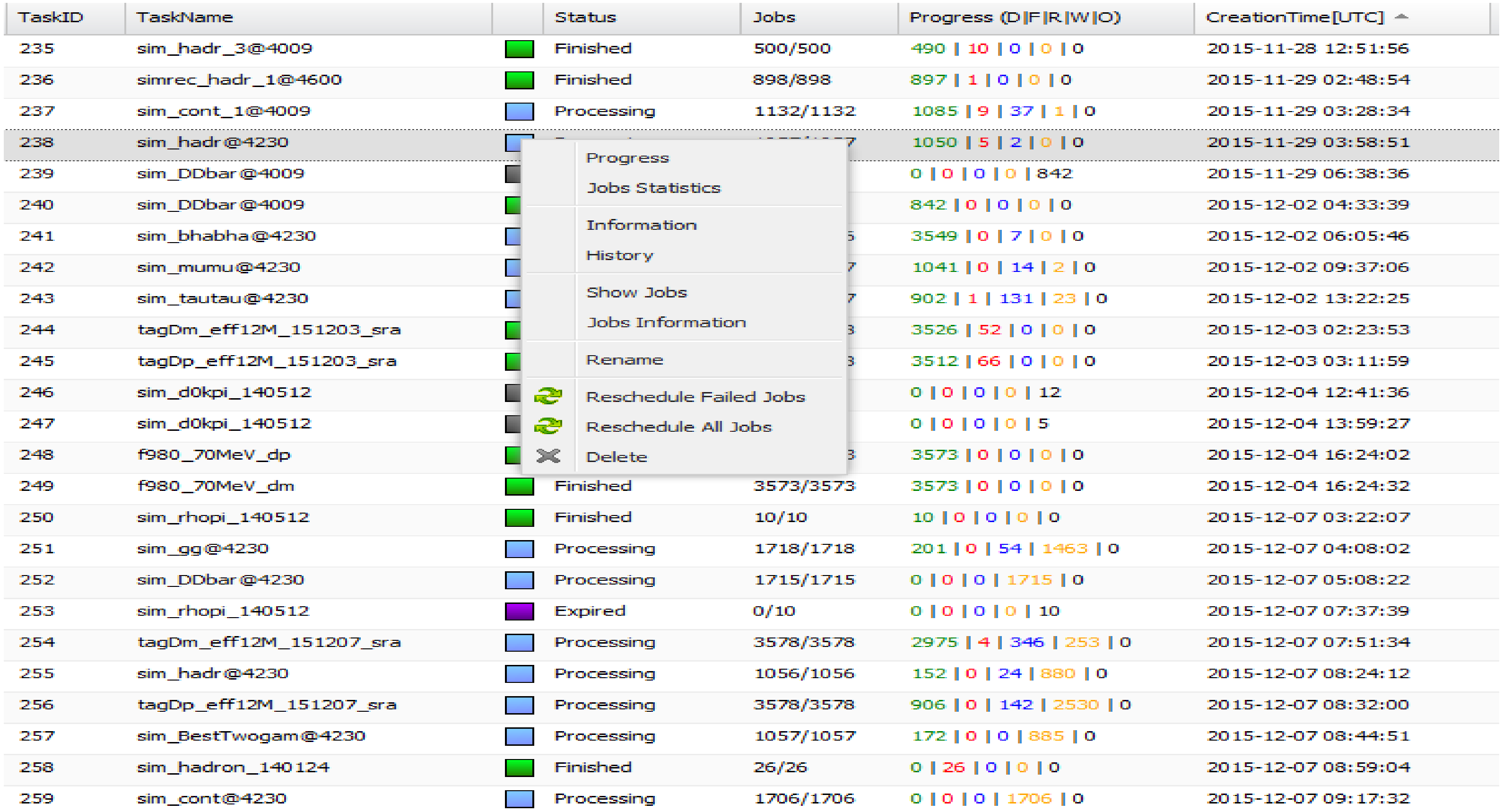
  split->submit->workflow control->status monitor->results retrieve -> reprocess

  - User interface
    - Use YAML, easy to parse with python, clear to users
  - Job submission
    - Support definition of experimental Job-splitting and workflow
  - Job management
  - Dataset management
    - Query Input dataset and register output dataset
  - Backend supports
    - DIRAC, PBS, Condor



General task submission tool (JSUB)

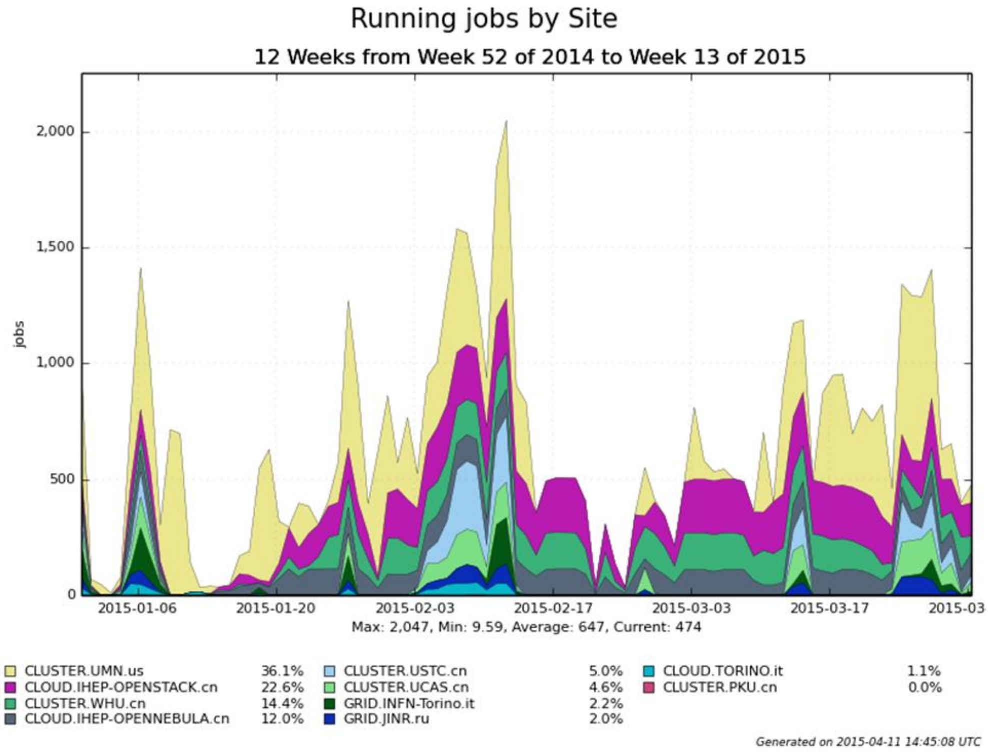
- Monitor and reprocess through web portal or commands
  - Task progress can be easily tracked, even to jobs and events
  - Reschedule and delete are provided

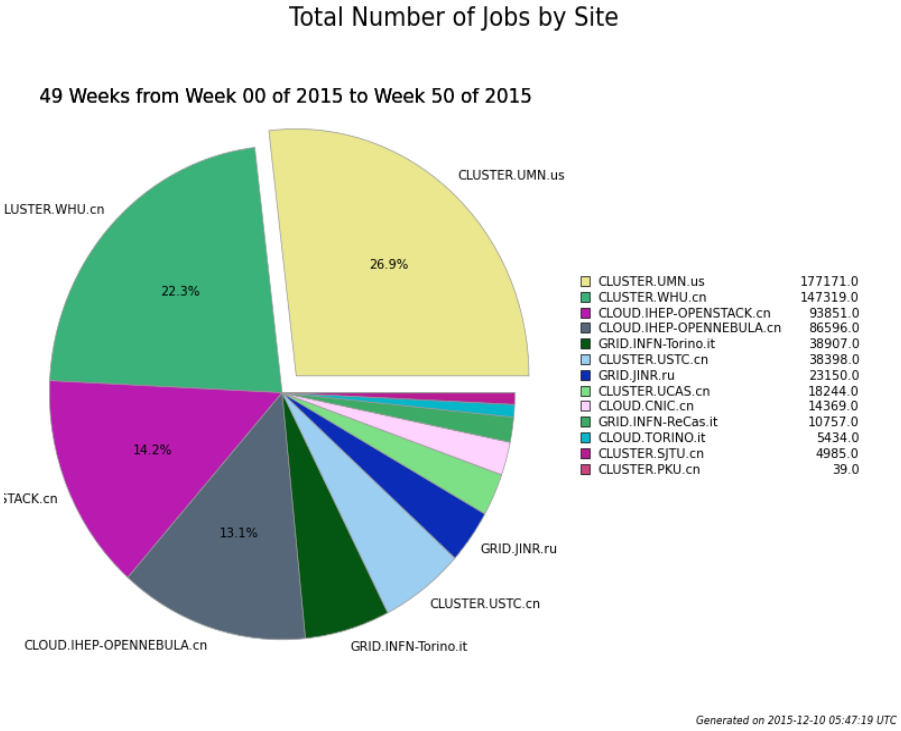


TaskID	TaskName		Status	Jobs	Progress (D|F|R|W|O)	CreationTime[UTC]
235	sim_hadr_3@4009	🟩	Finished	500/500	490 | 10 | 0 | 0 | 0	2015-11-28 12:51:56
236	simrec_hadr_1@4600	🟩	Finished	898/898	897 | 1 | 0 | 0 | 0	2015-11-29 02:48:54
237	sim_cont_1@4009	🟦	Processing	1132/1132	1085 | 9 | 37 | 1 | 0	2015-11-29 03:28:34
238	sim_hadr@4230	🟦	[illegible]	[illegible]	1050 | 5 | 2 | 0 | 0	2015-11-29 03:58:51
239	sim_DDbar@4009	⬛	[illegible]	[illegible]	0 | 0 | 0 | 0 | 842	2015-11-29 06:38:36
240	sim_DDbar@4009	🟩	[illegible]	[illegible]	842 | 0 | 0 | 0 | 0	2015-12-02 04:33:39
241	sim_bhabha@4230	🟦	[illegible]	[illegible]	3549 | 0 | 7 | 0 | 0	2015-12-02 06:05:46
242	sim_mumu@4230	🟦	[illegible]	[illegible]	1041 | 0 | 14 | 2 | 0	2015-12-02 09:37:06
243	sim_tautau@4230	🟦	[illegible]	[illegible]	902 | 1 | 131 | 23 | 0	2015-12-02 13:22:25
244	tagDm_eff12M_151203_sra	🟩	[illegible]	[illegible]	3526 | 52 | 0 | 0 | 0	2015-12-03 02:23:53
245	tagDp_eff12M_151203_sra	🟩	[illegible]	[illegible]	3512 | 66 | 0 | 0 | 0	2015-12-03 03:11:59
246	sim_d0kpi_140512	⬛	[illegible]	[illegible]	0 | 0 | 0 | 0 | 12	2015-12-04 12:41:36
247	sim_d0kpi_140512	⬛	[illegible]	[illegible]	0 | 0 | 0 | 0 | 5	2015-12-04 13:59:27
248	f980_70MeV_dp	🟩	[illegible]	[illegible]	3573 | 0 | 0 | 0 | 0	2015-12-04 16:24:02
249	f980_70MeV_dm	🟩	Finished	3573/3573	3573 | 0 | 0 | 0 | 0	2015-12-04 16:24:32
250	sim_rhopi_140512	🟩	Finished	10/10	10 | 0 | 0 | 0 | 0	2015-12-07 03:22:07
251	sim_gg@4230	🟦	Processing	1718/1718	201 | 0 | 54 | 1463 | 0	2015-12-07 04:08:02
252	sim_DDbar@4230	🟦	Processing	1715/1715	0 | 0 | 0 | 1715 | 0	2015-12-07 05:08:22
253	sim_rhopi_140512	🟪	Expired	0/10	0 | 0 | 0 | 0 | 10	2015-12-07 07:37:39
254	tagDm_eff12M_151207_sra	🟦	Processing	3578/3578	2975 | 4 | 346 | 253 | 0	2015-12-07 07:51:34
255	sim_hadr@4230	🟦	Processing	1056/1056	152 | 0 | 24 | 880 | 0	2015-12-07 08:24:12
256	tagDp_eff12M_151207_sra	🟦	Processing	3578/3578	906 | 0 | 142 | 2530 | 0	2015-12-07 08:32:00
257	sim_BestTwogam@4230	🟦	Processing	1057/1057	172 | 0 | 0 | 885 | 0	2015-12-07 08:44:51
258	sim_hadron_140124	🟩	Finished	26/26	0 | 26 | 0 | 0 | 0	2015-12-07 08:59:04
259	sim_cont@4230	🟦	Processing	1706/1706	0 | 0 | 0 | 1706 | 0	2015-12-07 09:17:32

Running Status

- The system is in production since the end of 2012
- Total Jobs are 665K in 2015, 340K in 2014
- Max running jobs can reach 2K (First season in 2015)



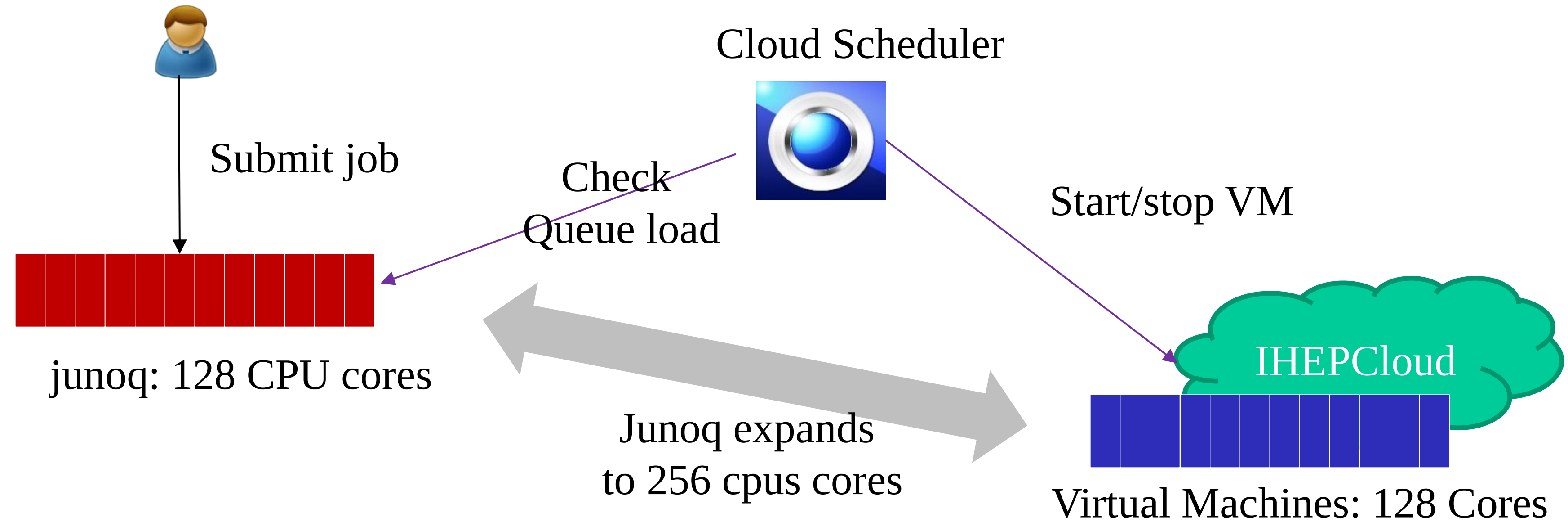


IHEPCloud: a Private IaaS platform

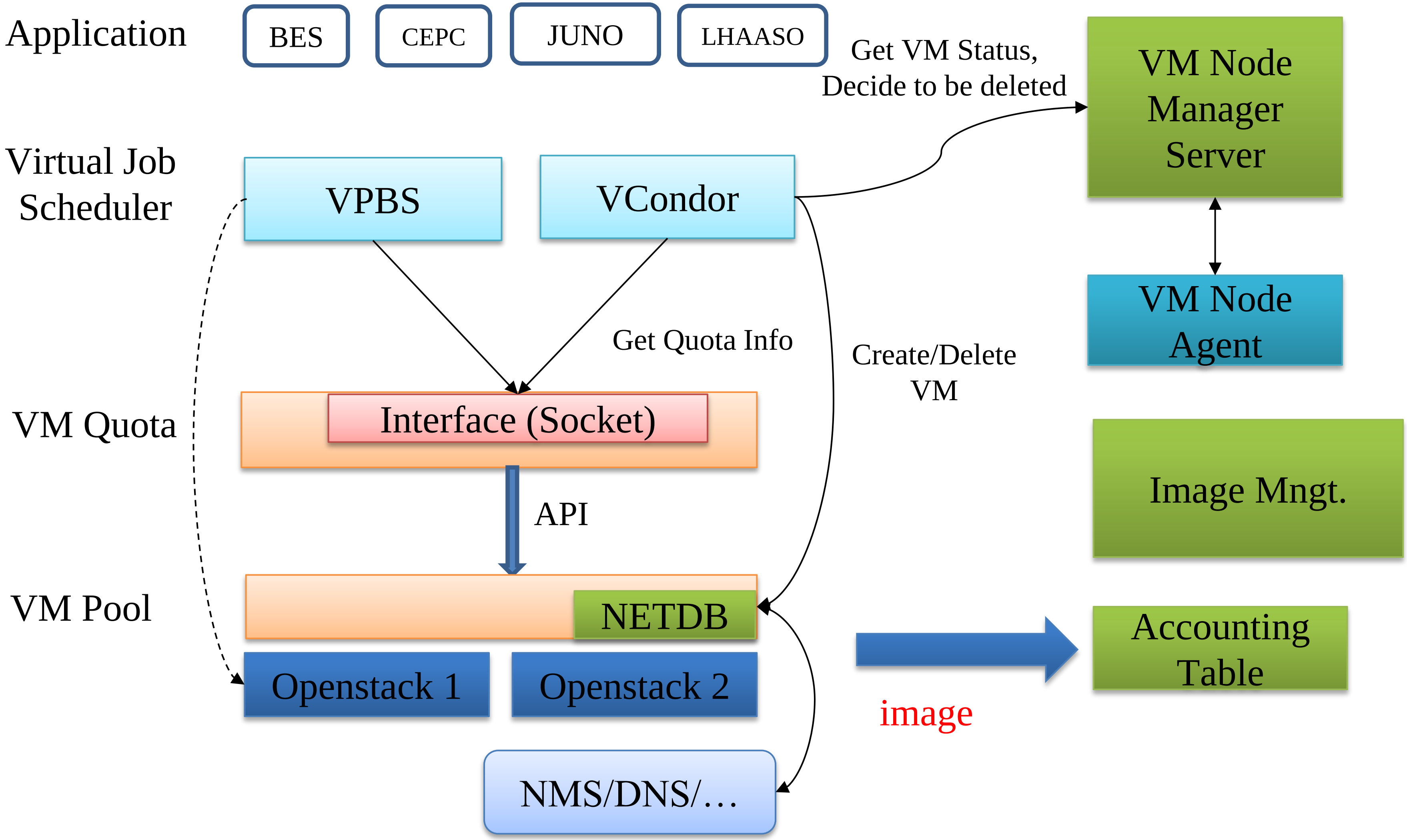
- Launched in May 2014
- Three use cases
  - User self-service virtual machine platform (IaaS)
    - User register and destroy VM on-demand
  - Virtual Computing Cluster
    - Combined with physical queue, jobs will be allocated to virtual queue automatically when physical one is busy.
  - Distributed computing system
    - Working as a cloud site: Dirac call cloud interface to start or stop virtual work nodes

Virtual computing cluster

- If a job queue is busy, new virtual machines will be created automatically to expand the queue
- Easy to be used for different experiments
- Provide dynamic virtual resource on demand
- Transparent to user, no change of user job submission

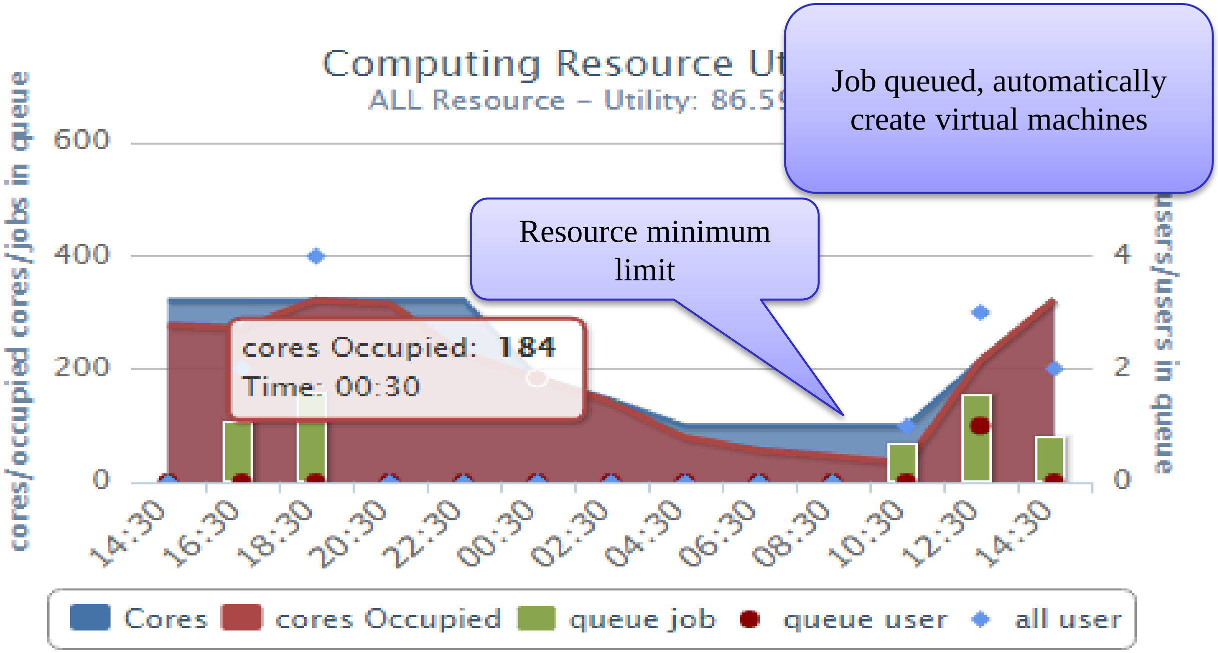


VM management



Dynamic scheduling

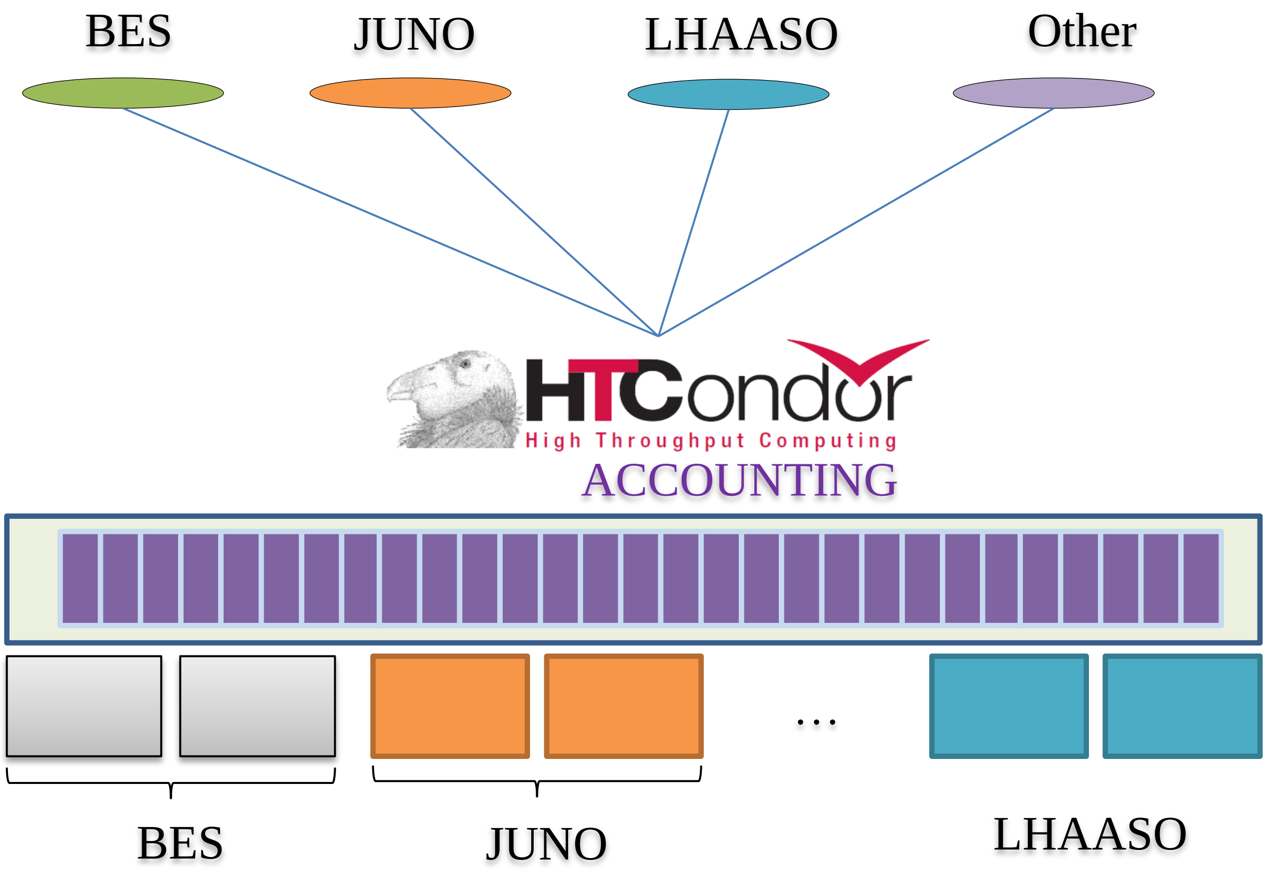
- Support multiple batch systems： PBS/Torque, HTCondor
- Dynamic VM provision： virtual machines are created and destroyed on demand
- Fair-share algorithm： guarantee resources are equally distributed among different experiments.



Job queued, automatically create virtual machines

Resource minimum limit

Future setup



Job Queues

Resource Scheduler

Virtual Machines

Physical Machines

High Performance Computing

- Needs from experiments and theoretical calculation
  - BESIII partial wave analysis
  - Geant4 detector simulation (CPU time and memory consuming)
  - Simulation and modelling for accelerator design
  - Lattice QCD calculation
- A HPC cluster at IHEP is being planned in 2017
  - NVIDIA Tesla GPUs
  - Xeon Phi coprocessors
  - Interconnected by the InfiniBand network
- A HPC prototype was set up and testing with the HybriLIT at JINR has been scheduled.

Summary

- Grid and cloud computing technologies were adopted to support various types of HEP experiments in China.
  - Dirac-based grid to integrate resources within an experiment
  - Cloud to promote sharing of resources among different experiments
- In collaboration with JINR, the BESIII Grid system has been developed and is running well in both M.C. data production and physics analysis.
- Hope we could continue to strengthen the collaboration with JINR on HEP computing.

Thank You！

谢谢