Distributed Computing in IHEP

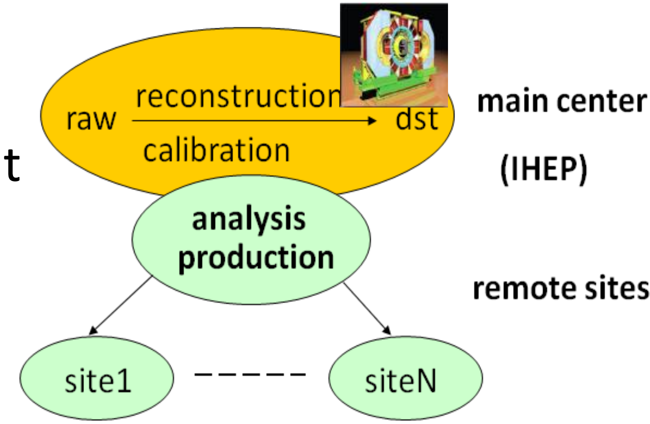
Xiaomei ZHANG
On behalf of BESIII distributed computing team
Institute of High Energy Physics

Fifth DIRAC User Workshop

Ferrara, May 2015

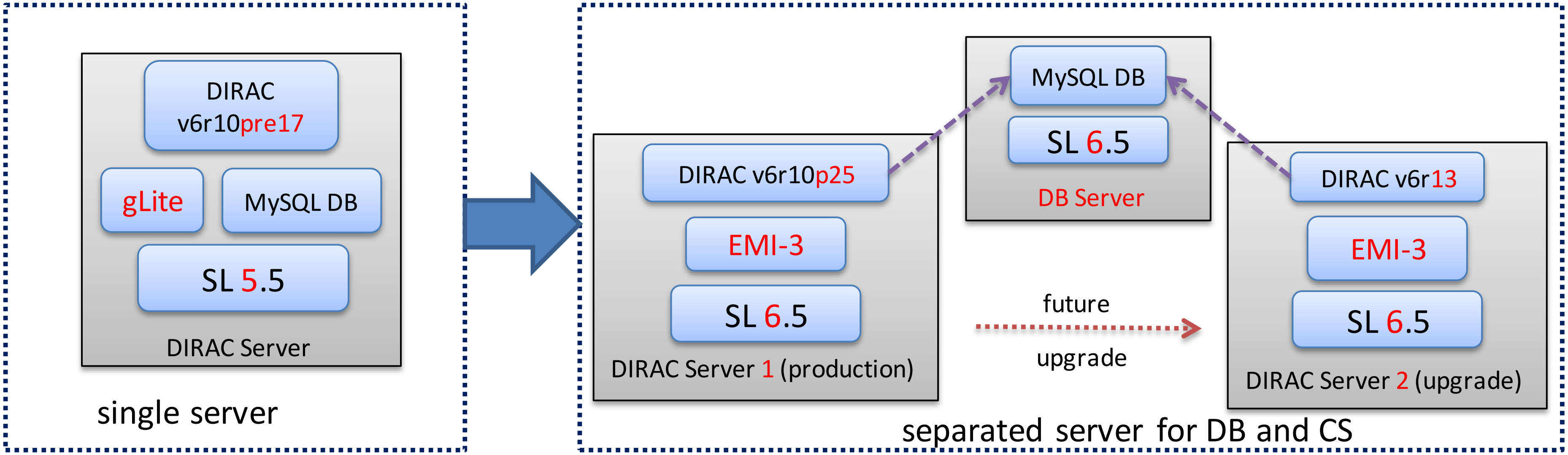
Introduction to BESIII

- The BESIII experiment studies electron-positron collisions in the tau-charm threshold region (1.0-2.3GeV) with the luminosity 1 x $10^{33}/cm^2/s$
- The BESIII distributed computing is set up based on DIRAC since 2012
  - With IHEP as central site for central storage and all the activities, other remote sites take care of MC production and analysis
  - The system integrated ~2000 CPU cores and ~400TB and is in good status
- Features and challenges
  - Lack of grid experiences among communities, the clusters are most common resources
  - SE is not affordable to all the sites so that Central SE plays a great role to share data and store output
  - Lack of manpower to maintain sites, monitoring is important to ensure stability of systems



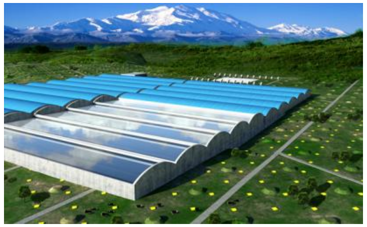
DIRAC set-up and upgrade

- Three set-up: production, test, development
- Production set-up
  - Two servers, separate DB from CS for better performance and easy upgrade
- Upgrade
  - From v6r10pre17 to v6r10p25 last June, including OS and grid middleware
  - Plan to upgrade to v6r13 soon this year
  - Upgrade is not so easy for us after the system is in production
    - Use a separated machine to do the upgrade, and quick transit with DNS exchange to avoid long downtime
    - If there are some migration tools provided for DB, it will make upgrade easier

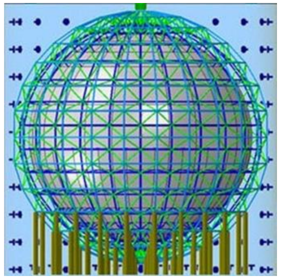


Multi-VO installations

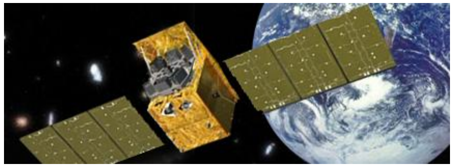
- Since 2014, single installation is extended to support multi-VO
- Motivation
  - With the experience of BESIII distributed computing, several new experiments in IHEP would like to try or use distributed computing in future
  - Would be difficult for them to afford man power so far to set up and maintain a new DIRAC system
  - Many sites joined more than one IHEP experiments. Multi-VO supports will let site management more convenient
  - Inspired by the idea "DIRAC as a Service", the existing BESDIRAC can be extended to support new VOs as soon as possible
- Current VOs supported besides BESIII
  - CEPC-SPPC
  - JUNO
- To be supported
  - LHAASO, HXMT



Large High Altitude Air Shower Observatory (LHAASO)



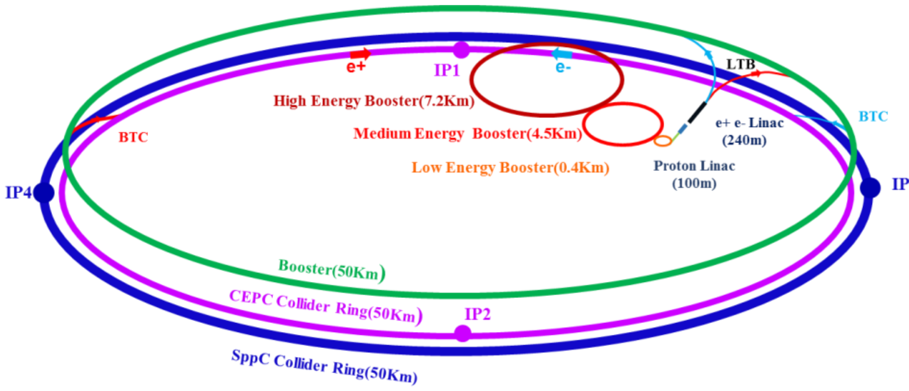
Jiangmen Underground Neutrino Observatory (JUNO)



Hard X-ray Modulation Telescope (HXMT)

Introduction to CEPC-SPPC

- The first phase CEPC (Circular Electron-Positon Collider) aims to be a Higgs factory to study Higgs properties at 250Gev in a 50km ring with the luminosity 2 x 10^{34}/cm^2/s
- The SPPC (Supper Proton-Proton Collider) is the second phase with 70 TeV in the same ring
- The data Volume is expected to be 100~2000 times higher than that of BESIII
- Data taking will be in 2025~2028
- In pre-study stage, they would like to use distributed computing in the detector design phase
  - The resource requirement is 1000 CPU cores, 2 PB per year
  - Nearly no local resources supports now, and collect resources from collaborations
- The current software framework they used is adopted from ILC
  - Whizard for generation
  - Mokka for simulation
  - Marlin for reconstruction and analysis
- The basic components have been set up
  - BESDIRAC is extended to support CEPC VO
  - VOMS and CVMFS
  - Job submission scripts
  - Central SE is extended to support CEPC VO



LTB : Linac to Booster

BTC : Booster to Collider Ring

What have done for multi-VO supports

- User and group management
  - Users are grouped
  - Groups are classified with VO properties
- Workload management
  - Generic pilot group has been defined for each VO
  - Site director has been created for each VO
  - The site director and Generic pilot for a VO take care of matching between jobs and resources for this VO
    - Site director find the matching resources to send pilots
    - Generic pilot pull the matching jobs
  - Resources are configured to support certain VOs
    - Cluster and grid can be controlled on VO level
    - Cloud only can be controlled on group level, matching "owner_group" between sites and jobs
- Dirac File Catalog
  - Root directories have been created for each VO
  - Permission control is done through group level

What to do for multi-VO supports

- VO information is missing in web pages
  - In “Site summary”, sites can’t be distinguished with different VOs
    - Users need to know the available resources for a certain VO
  - In “Accounting”, resource usage can’t be grouped with VO
    - Bills need to be provided for each VO
  - In “Job monitoring”, jobs for all the VO mix together and can’t be classified with VO
- Priority control on resources on VO and group level
  - Leave it to completely the sites on VO level
  - Try with JobShare property for multi-VO to do control on group level?

DIRAC extensions in use

- VMDIRAC
- WebAppDIRAC
- Web
- BESDIRAC, a BES extension to DIRAC
  - Hold BESIII-specific packages
    - Data managements tools, a wrapper of DIRAC commands for BESIII special case
      - BESIII dataset toolkits
      - Random trigger toolkits
    - Data transfer system, allow user requests for massive transfer
    - Monitoring system for sites, done by the JINR group
    - Task management

Task management

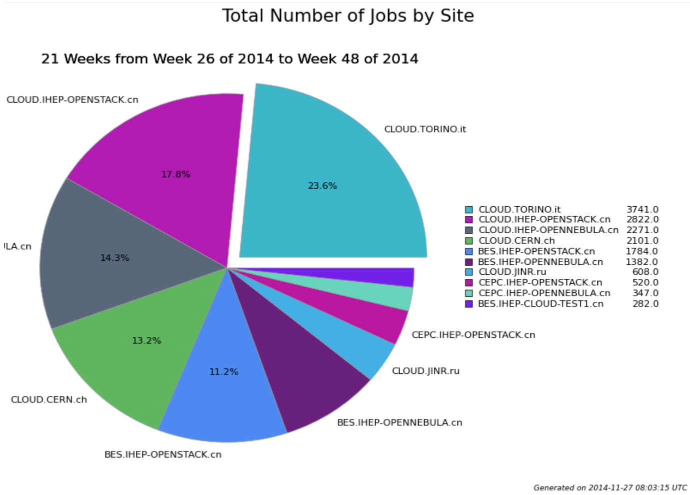
- Motivation
  - Users want to get task info in a quick way, not just individual jobs
  - Production managers want to have a review of task history
- Functions
  - Get the progress and info of the task
  - Reschedule/delete jobs by task
- Components
  - Database: TaskDB
    - Store the task information, its related jobs
  - Service: TaskManager
  - Agent: TaskAgent
    - Update the task status and keep it in the task history
- Available commands
  - besdirac-wms-task-list, besdirac-wms-task-show
  - besdirac-wms-task-reschedule,besdirac-wms-task-delete
- Future combinations with JobGroup?

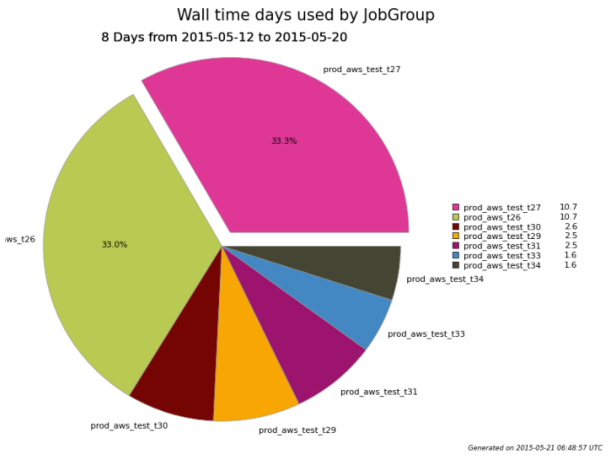
WebAppDIRAC

- We use both old and new web portal with different ports
- The complete transit from old web portal to new web portal still need time
  - Google API is not well supported in China so that new web portal has problems to open sometimes
  - Data transfer system extended by us need to rewrite from old web framework to new one

VMDIRAC

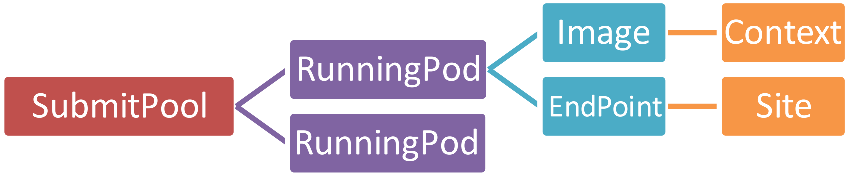
- Since July 2014, VMDIRAC started to be used in BESIII
- Cloud integration has been successfully done for BESIII distributed computing based on VMDIRAC
  - With the help of Ricardo Graciani, Victor Mendez, Victor fernandez
- Various Cloud type already added and used
  - OpenStack with nova
  - OpenNebula with rocci
  - AWS with EC2
- Tests have been done for private cloud, 16000 jobs with success rate can reach 99%
- Preliminary tests are doing for AWS cloud with different VM type
- So far we are satisfied with VMDIRAC

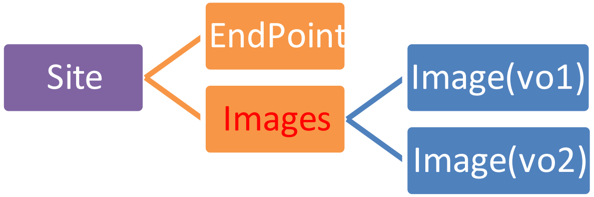




VMDIRAC

- Configuration can be optimized to make it easier
  - A little complicated for new users
  - In multi-VO, number of RunningPods will be large with many images to support
- Monitoring is not too convenient, and more functions need to be added to avoid frequent access to different clouds
  - The query of VMs need to have more filters with many cloud sites joined. Eg sitename, vo
  - Status of VM isn’t described completely
  - Clear up expired VM status in monitoring page, eg. one month
  - The delete control to VMs is added in current web portal
  - Establish connections between jobID and VMID





VMDIRAC

- Central information can be established to know cloud info easily
  - Centralize cloud information from different providers
    - eg. cpu, memory, instances, quotas, etc
  - Images management and query
- Some problems met
  - OwnerGroup can't add more than one group
  - Not all the VMs are not closed automatically when there are no jobs as expected
  - The start of VMs are out of control maybe because VMs status is not precisely got by VM monitor
  - Support for multi-core VMs is needed to enable multi jobs on one VMs
    - Start multi job agents to fill the cores
    - How about automatic start and shutdown of VMs with multi job agents?

DIRAC functionalities in use

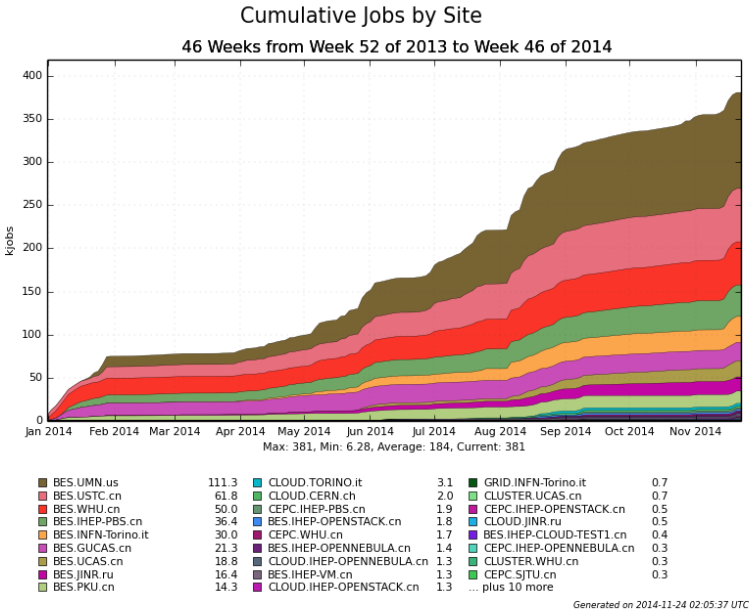
- Workload management
  - Quite interested in configurable pilots to do pre-check of sites to reduce failure rate
- Resources
  - Cluster (PBS, LSF, condor)
  - Cloud (openstack, opennebula)
  - Grid (creamce)
- Accounting
  - Very useful and important in multi-VO set-up
  - The information to be integrated into total accounting system in computing center for detail billing for each VOs
- DFC
  - BESIII file, metadata, dataset catalogue is built up based on DFC
  - Static dataset feature is added and query with dataset name is supported

DIRAC systems interested

- DIRAC Resource Status System
  - To implement site monitoring
- DIRAC Transformation System
  - To build production manager
- DIRAC Workflow
  - To take care of different job chains
- Server side job splitting
  - To reduce submission time for a large amount of jobs

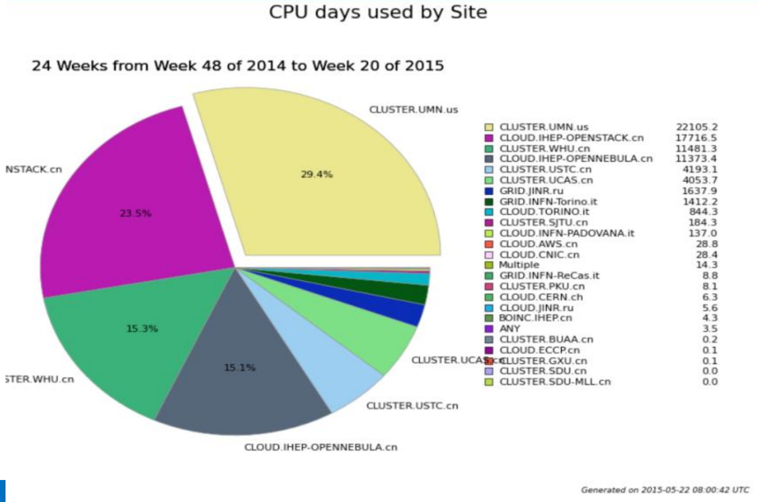
BESIII operations

- In 2014, 380K jobs have been completed and 70TB data have been transferred back to IHEP data center
- Cluster is the main resources, and the cloud resources grow fast


Cumulative Jobs by Site

46 Weeks from Week 52 of 2013 to Week 46 of 2014

Generated on 2014-11-24 02:05:37 UTC

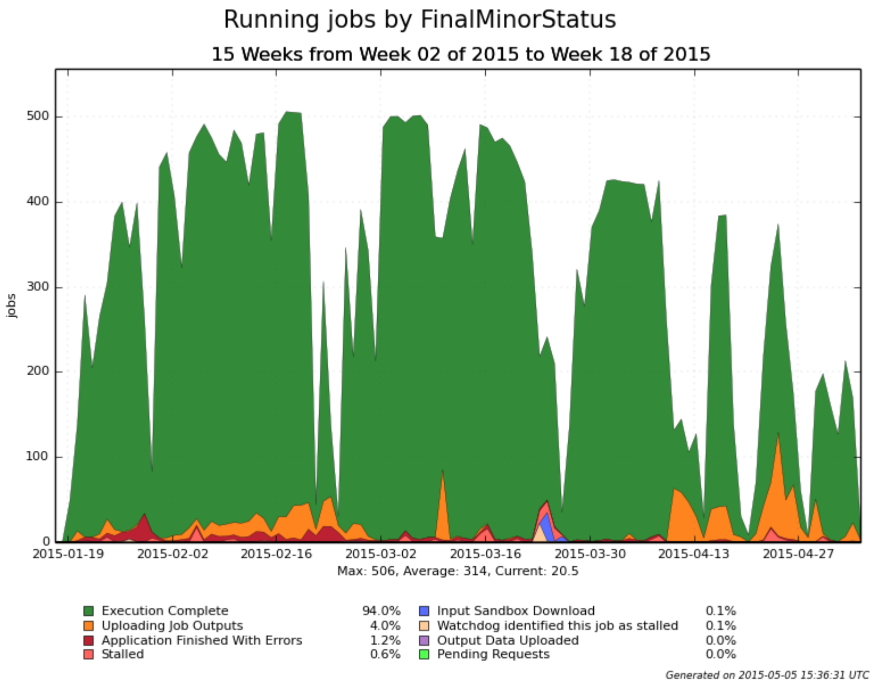

CPU days used by Site

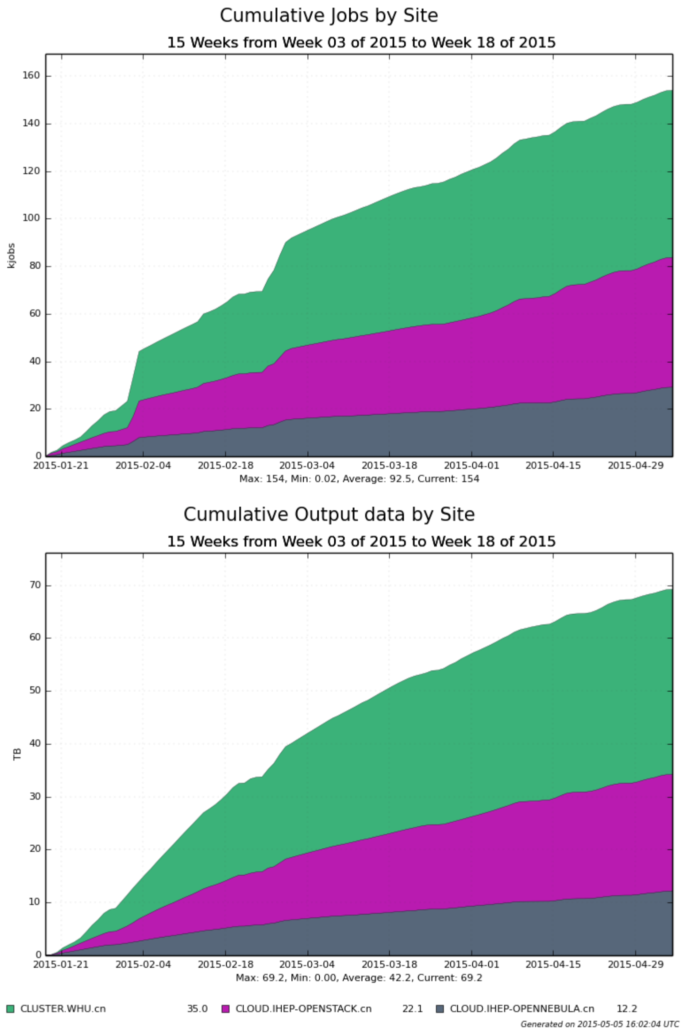
24 Weeks from Week 48 of 2014 to Week 20 of 2015

Generated on 2015-05-22 08:00:42 UTC

CEPC operations

- The system is in production status since Jan. 19th, 2015
- Three active sites, max 510 CPU cores.
- Job success rate: 94.0%
- ~ 150,000 jobs are done
- ~ 70 TB output data written to central storage StoRM+Lustre





Summary

- BESIII distributed computing is in good status
- Multi-VO supports have been set up for new experiments
- Cloud integration based on VMDIRAC is running well with private and commercial cloud
- But measures are still needed to make the system better

- THANK DIRAC TEAM FOR STRONG SUPPORTS AND USEFUL HELP!!!!