# DIRAC 和 BESIII分布式计算

张晓梅

中国科学院高能物理研究所

2013年7月5日

# 内容

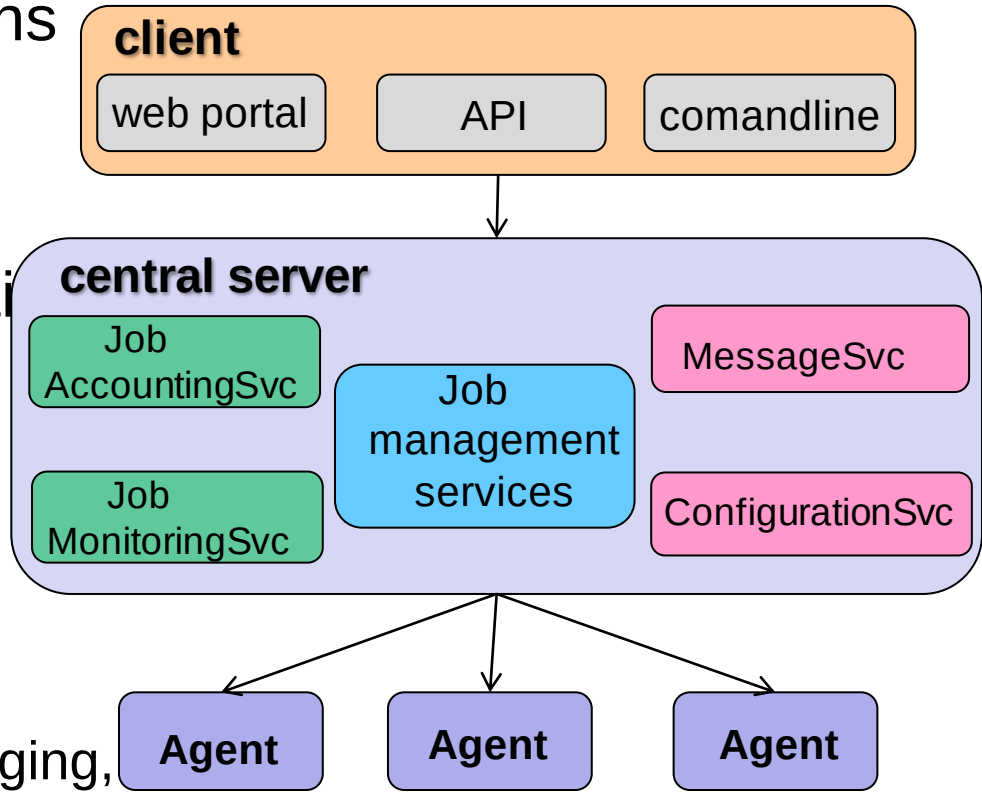
- DIRAC简介
- BESIII分布式计算的设计与组成
  - Computing Model and Evolution
  - Workload management
  - Data management
- BESIII分布式计算的进展
  - Site status
  - Recent tests
  - Ongoing activities and developments
  - Problems and Challenges
- Summary

# What is DIRAC

- Distributed Infrastructure with Remote Agent Control
  - a general purpose Open Source distributed computing framework
  - a complete solution for workload management and data management in distributed computing
- History
  - DIRAC was born as the LHCb distributed computing project
  - Since 2009 DIRAC became an independent project
  - Many projects are starting to use or evaluate
    - HEP: LHCb, ILC/CLIC, Belle II, BES III, ( SuperB )
    - Astrophysics: CTA, Glast, Fermi-LAT, LSST
    - Other communities: biomed, earth sciences, etc

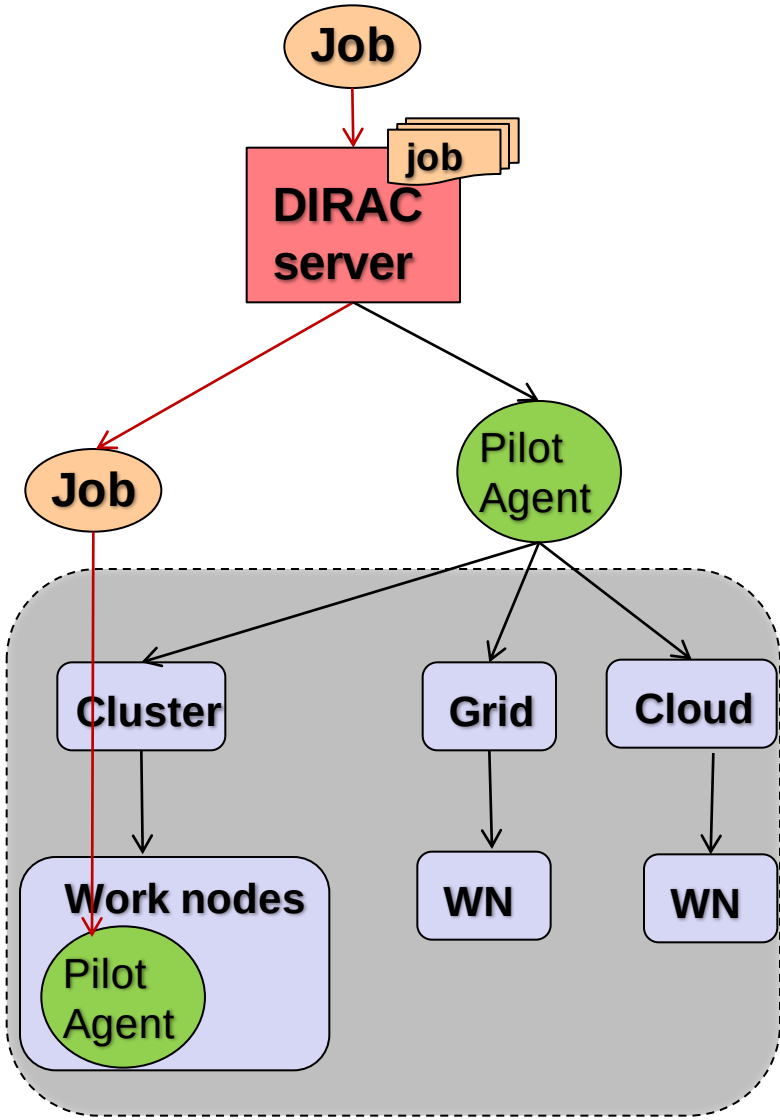
# DIRAC as a framework

- Provide modular architecture and well-defined components for developing and extensions
  - Services, agents, clients, databases
- Provide secure communicati framework between the distributed components
  - Client/server, DISET
- Provide basic services
  - Configuration, monitoring, logging, accounting, etc



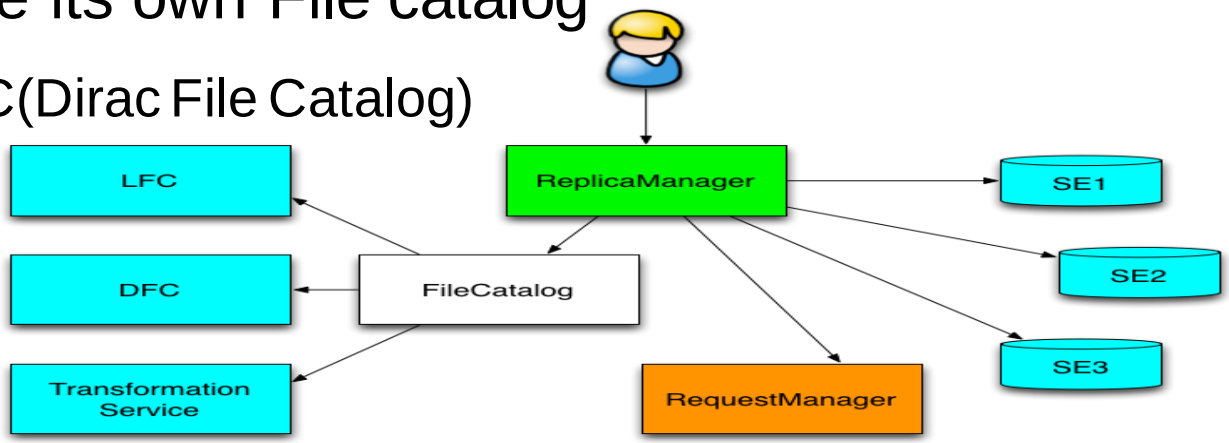
# DIRAC
# as a workload management system

- Use pull scheduling and pilot agent paradigm
  - fault tolerance and efficient
  - Commonly used in LCG
- Support the integration of various heterogeneous resources
  - Cluster, grid, cloud…..
- Provide priority management
  - Users and groups management
  - Assign job priorities for different groups and activities
- Provide basic APIs for developing job management frontend



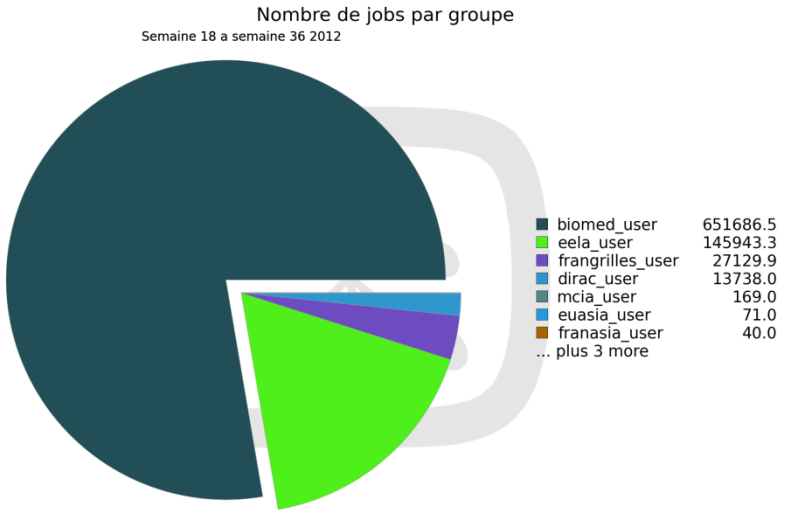
# DIRAC
# as a data management system

- Allow to integrate various Storage Elements and File Catalog
  - dCache, DPM, BestMan, LFC, DFC……
  - Transparent to users with unique commands and definitions
- Provide its own storage element interface
  - Dirac storage element based on DIP protocol
  - Not allow third party transfers with SRM-based SE
- Provide its own File catalog
  - DFC(Dirac File Catalog)



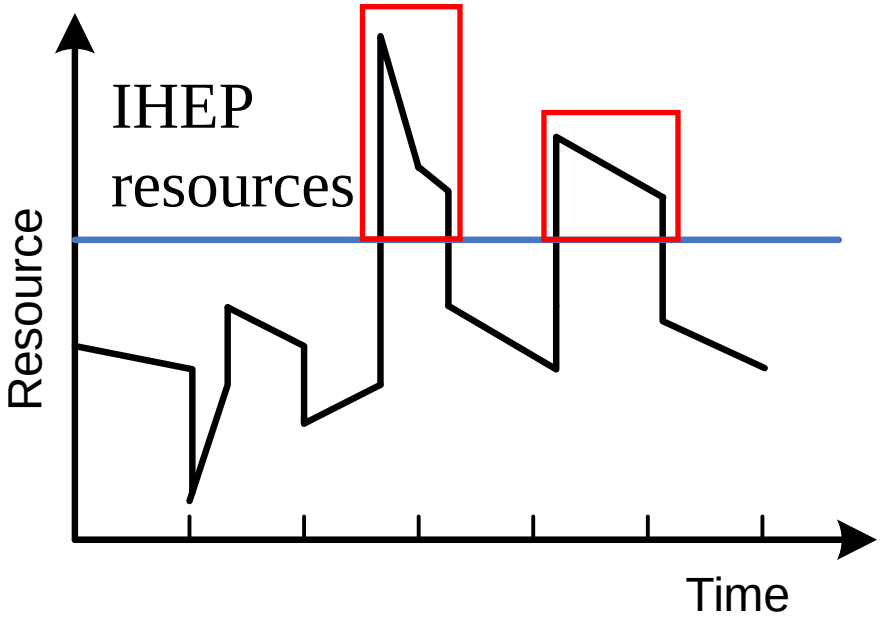
# DIRAC as a service

- One DIRAC installation can serve many VOs
  - Different policies for different VOs
  - Different resource priorities for different groups
- Use cases
  - Small user communities can not afford maintaining dedicated DIRAC services
  - One region can provide DIRAC services for their users
- Examples
  - France Grid DIRAC service hosted
    - 15VOs, 88users
    - In production since 2012 May



# Context of BESIII distributed computing

- Why?
  - Increasing data volume
    - 1 billion J/ψ and beyond
  - Peek needs in some periods
- Extra computing resources needed to supplement IHEP
  - **Resources from BESIII collaboration**
  - Cloud, volunteer computing…
  - Other available resources
- First focus on MC production; distributed analysis later

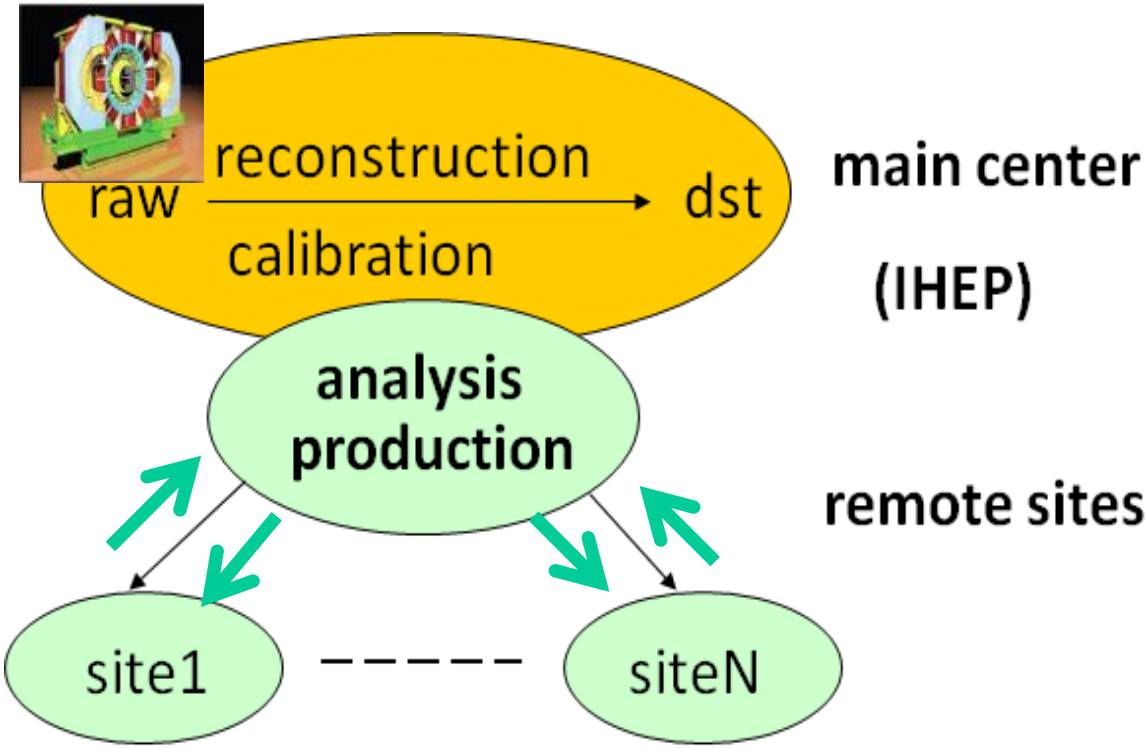


# Design philosophy

- Make it as simple as possible for sites to join
- Make it as convenient as possible for users to use
- Use existing and mature software and middleware wherever possible
  - easy to set up, maintain, extend and support
- Well fit the need of BESIII computing

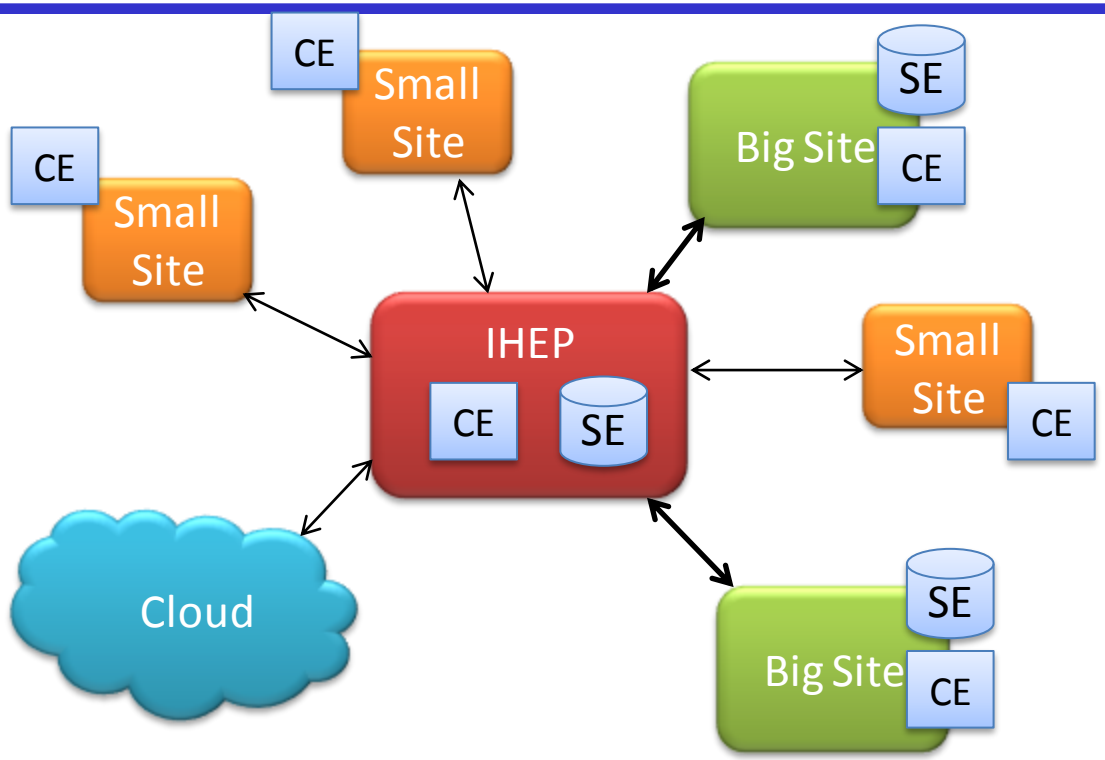
# Computing model

- IHEP as central site
  - Raw data processing, bulk reconstruction, analysis ….
  - Central storage for all the data
  - Central services
- Remote sites
  - MC production, analysis




- Basic Data flow
  - Simulation data produced in remote sites transferred back by transfer tools or directly written back to IHEP by jobs for permanent storage
  - Reconstructed data (DST) transferred to remote sites for particular analysis

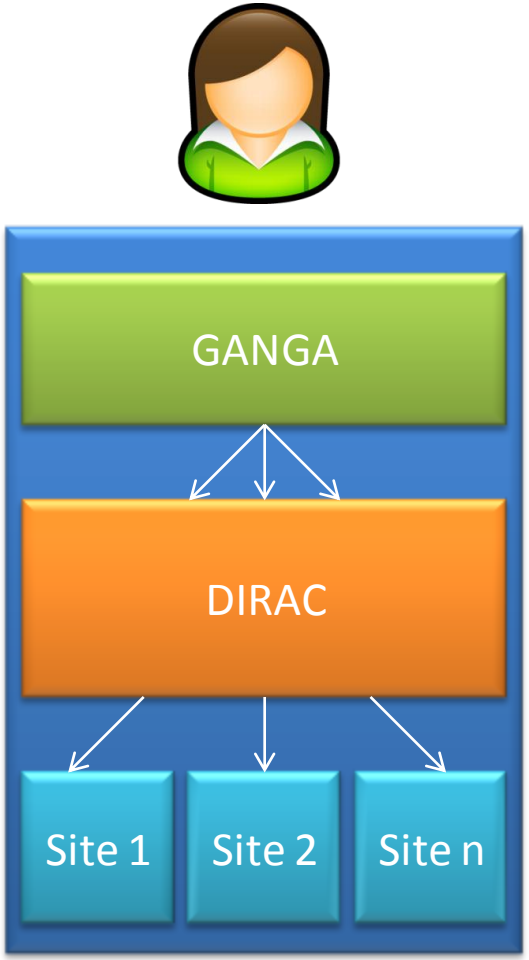
# Evolution of computing model




| Big | Small |
|---|---|
| Computing + Storage – can host random trigger files | Computing only – no need for large local SE |
| Simulation + reconstruction | Simulation only |
| DST files transferred / copied back to IHEP | RTRAW files written directly to IHEP in jobs |

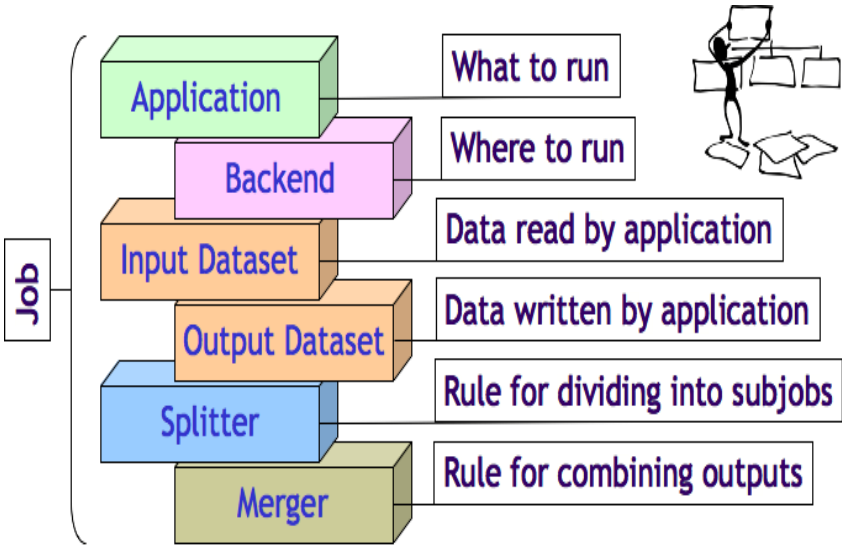
# Workload management

- Components:
  - GANGA for mass job submission , splitting and management
    - GangaBoss plug-in has been written to support BES software
    - Make mass job management convenient to end users
  - DIRAC for distributing computing jobs to various sites
    - DIRAC server is running at IHEP with clients at remote sites
  - CVMFS (CERN VM File System) for deploying BOSS on target sites
    - Commonly used in LCG and Cloud
    - Make the deployment of heavy software light-weight
    - Clients running at distributed sites can load BOSS version from server at IHEP



# GANGA and GangaBOSS

- GANGA is a user-friendly frontend that handles **mass** job definition and management
  - Jointly developed by ATLAS and LHCb experiments
  - with modular architecture
- GangaBOSS is the package developed based on basic modules
  - Auto handle complete life cycle of BOSS jobs in grid and cluster
  - Hide grid complexity to end users
  - Well work with DIRAC and local clusters to complete job management



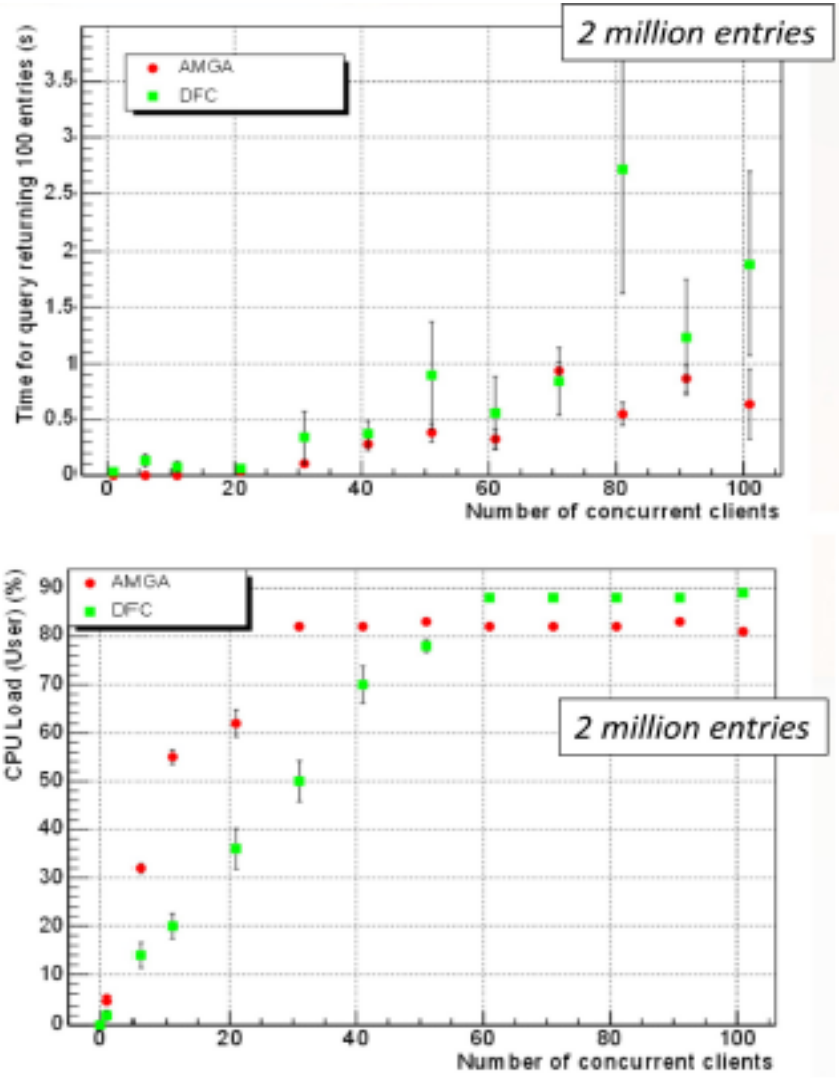

Build → Configure → Split → Submit → Monitor → Merge

# Data management

- Data storage, Data transfer and Data info management
- The purpose of Distributed data info management
  - Allow **production users** to readily register files and datasets
  - Allow **analysis users** to readily find and access files / datasets
  - Jobs running in distributed resources can find and access necessary files, then register output files as datasets
- Badger (BESIII Advanced Data Manager) being developed for BESIII file and metadata management
  - Replica Catalogue - map logical file names (LFNs) to physical file names (PFNs) at different sites
  - Metadata Catalogue – define metadata for datasets and files
  - Dataset Catalogue – define datasets and related operations

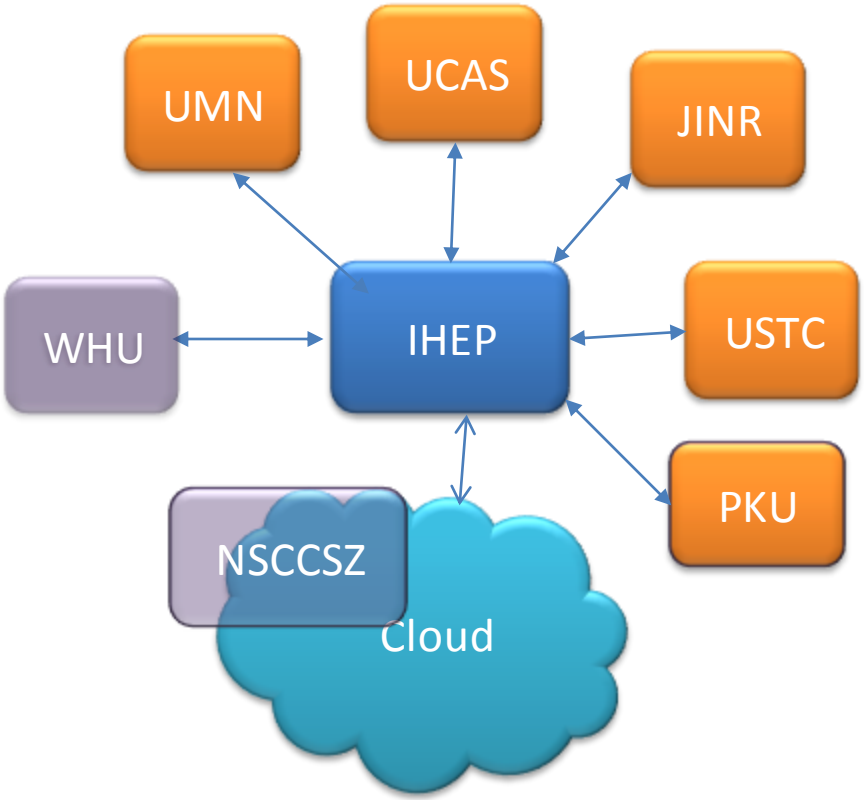
# Data Management

- Investigated two choices: AMGA (from gLite) and DIRAC File Catalog (DFC)
- DFC functionality best matches our needs
  - Performance of metadata query are similar with high number of clients
  - Can include everything in a single catalogue
  - DFC developers very responsive in helping develop extra functionality needed for BESIII, good working partnership has been established



# Site status

- Totally 8 sites are in active status now
- Two new sites based on virtualization technology are added successfully
  - NSCCSZ and WHU
  - A good experience for the sites who can't change basic OS
  - An interesting try of cloud resource
- Torino site in Italy is ready to join soon
  - The site has a good experience about grid and cloud



# Site status

| Site | Type | Job slots | Status |
|---|---|---|---|
| JINR | Grid | 2100 (shared with LHC) | Active for BOSS grid jobs |
| UCAS | PBS | 80 (+64) | Active for BOSS grid jobs |
| IHEP-PBS | PBS | 96+200 | Active for BOSS grid jobs |
| PKU | PBS | 168 | Active for BOSS grid jobs |
| USTC | PBS / Condor | 128 | Active for BOSS grid jobs |
| UMN | SGE | 400 | Active for BOSS grid jobs |
| WHU | PBS(virtual) | 100 | Active for BOSS grid jobs |
| NSCCSZ | PBS(virtual) | 10 | Active for BOSS grid jobs |
| Torino | Grid/Cloud | | In progress |

# IHEP site

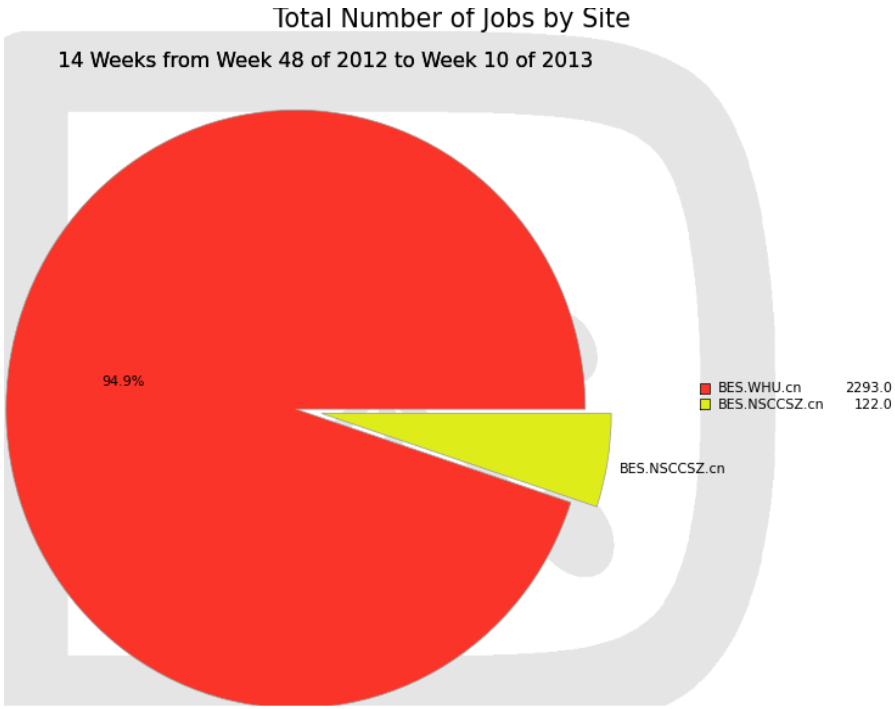
- Hold and maintain central main services
  - DIRAC, Ganga, CVMFS, squid, BESDIRAC,git
- Own IHEP-PBS and IHEP-LCG site
  - IHEP-PBS has 96 cores
  - IHEP-LCG just for testing
- Have dCache SE already running
  - with160TB storage space
  - Study of the integration of Lustre and dCache SE
    - Luster as a local interface and dCache as a grid interface
    - fasten the exchange of grid data and local data inside IHEP
  - See details in Xiaofei's slides

# Sites based on virtualization

- Two use cases
  - National Supercomputing Centre, ShenZhen
  - Wu Han university
- Sites are set up over VMs
  - Not completely "Cloud" site in some sense
  - Can be more flexible with cloud management middleware supported
- Why "virtual" sites?
  - Sites has difficulties to change basic OS
  - Sites are easy to be maintained and extended with virtual images
  - A start to explore how to use cloud resources which are based on virtualization

# Sites based on virtualization

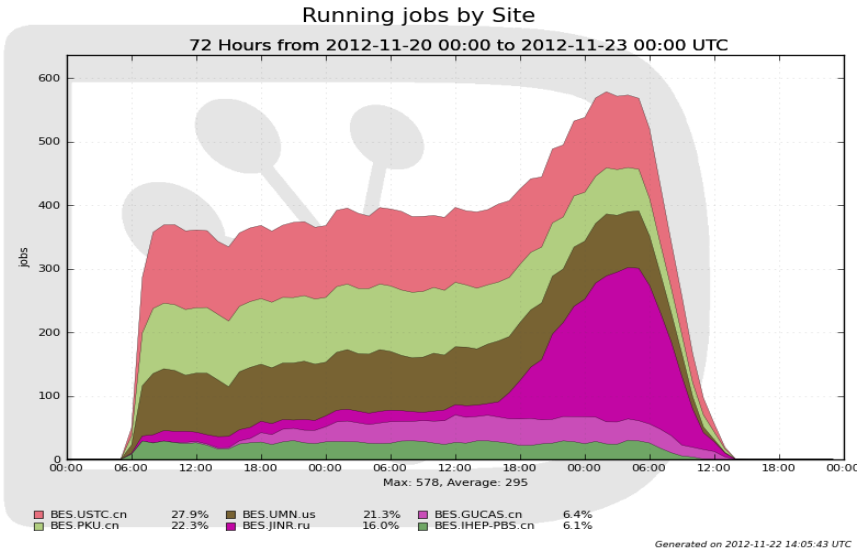
- Steps of setting up
  - Choose one of Hypervisors to build up virtual resources
    - WHU is using VMWare
    - NSCCSZ is using KVM
  - Set up Cluster over virtual resources
  - Register into DIRAC as a Cluster site
  - More reference in twiki




**http://docbes3.ihep.ac.cn/~offlinesoftware/index.php/WHU_experiences_on_setting_up_local_clusters_with_virtual_machines**

# First production

- 5 remote sites joined
- Production of 200 million ψ'' bhabha events
  - More than 8000 jobs are submitted and run
  - More than 1600 CPU days are used
- Data all transferred back to IHEP SE
  - About 237GB production data are transferred back to IHEP SE




| Name | Tier | GridType | Country | MaskStatus | Efficiency (%) | Status |
|---|---|---|---|---|---|---|
| China: 5 Sites | | | | | | |
| BES.GUCAS.cn | Tier-2 | BES | China | Active | 100.0 | Good |
| BES.IHEP-LCG.cn | Tier-2 | BES | China | Active | 0.0 | Idle |
| BES.PKU.cn | Tier-2 | BES | China | Active | 100.0 | Good |
| BES.IHEP-PBS.cn | Tier-2 | BES | China | Active | 100.0 | Good |
| BES.USTC.cn | Tier-2 | BES | China | Active | 100.0 | Good |
| Russia: 1 Site | | | | | | |
| BES.JINR.ru | Tier-2 | BES | Russia | Active | 99.1 | Good |
| United States: 1 Site | | | | | | |
| BES.UMN.us | Tier-2 | BES | United States | Active | 100.0 | Good |

# First Production

- Main failure reasons and experience
  - CVMFS service of Single node failed
  - IHEP SE is filled up
- Grid vs. local CPU time rough comparison
  - DIRAC reported CPU time is normalized according to processor type
  - Time reported includes DIRAC and BOSS setup on nodes

| | IHEP CC | Grid |
|---|---|---|
| No. of cores | 1000 | ~400 |
| Total (wall) time for 100 million events | 13.5 hours | ~ 57 hours (inc. data transfer) |
| Time / event | 0.0005 s | ~0.0020 s |
| CPU time / event | 0.43 s | 0.71 s |

# Ongoing activities and developments

- BESIII data transfers system
  - Seen in LinTao's slides
- Dataset in Badger
- Site monitoring
- DIRAC Accounting with NoSQL
  - Seen in ZhangGang's slides

# Dataset and BADGER

- BESIII File Catalog exists already, but ...
- ... at the moment we have already
  - > 400000 raw files
  - ~ 2000000 DST files
- Datasets (logical collections of files) are necessary to handle the BESIII data
- Dynamic datasets are provided by DIRAC
  - list of files in a dataset is generated with query each time

# Dataset and BADGER

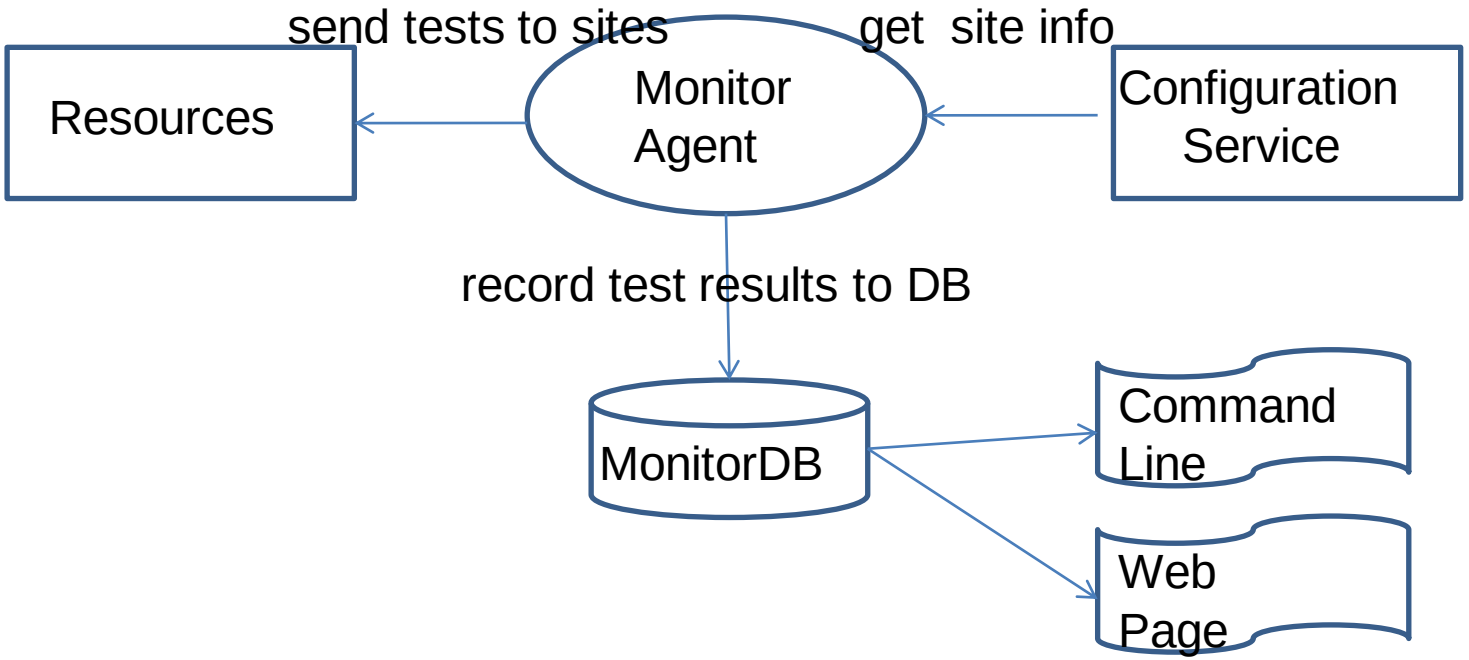
- Two features are missing in DFC
  - BESIII analysis use absolute normalization
    - Check table is considered to assure that the number of events and data files in a dataset is always the same
  - Operations related with Dataset name is required
- To manage datasets, we need API and a set of client programs
  - First priority: *createDataset, listDataset, checkDataset, removeDataset, copyDataset, downloadDataset*
  - Second priority: introduce replication (*createReplica, listReplica, removeReplica*)
- Need to provide web portal and clients for end users to query and manage dataset

# Monitoring

- Know real-time status of services and sites, track site problems
- Monitoring Types:
  - site
    - CVMFS, SE, CE, squid……
  - Job
    - Job number: pending, run, failed, completed
  - Storage monitoring
    - Space info from DFC and SE (physical, used, available)
  - Data transfer
    - Transfer rate , status (good, failed, active), transfer volume
  - Dataset Popularity monitoring
    - Make decisions for data replications and deletions

# Monitoring

- Start from site monitoring
  - Use SAM tests job to check the availability of sites
  - Design and developments will be based on DIRAC framework



# Problems and challenges

- Sites are easy to fail from time to time
  - Manpower is weak for maintenance of remote sites
  - Hope site monitoring can help improve it
- Many things seem to be difficult without SE in sites
  - CPU efficiency of jobs is low without SE in sites
    - Slow transfers happened during run-time of jobs
  - Simulation and Reconstruction chain is difficult to be done
  - Efficient transfers can’t be done

# Problems and challenges

- But difficult to let sites have SE
  - No experience to install and maintain high performance parallel file system
  - No experience for grid SE
  - Need funding for disks and machines
- Not enough manpower to join maintenance, upgrade and development of central services
  - Some experienced members left, or focus on other tasks
  - New members are just coming, need time to be familiar with

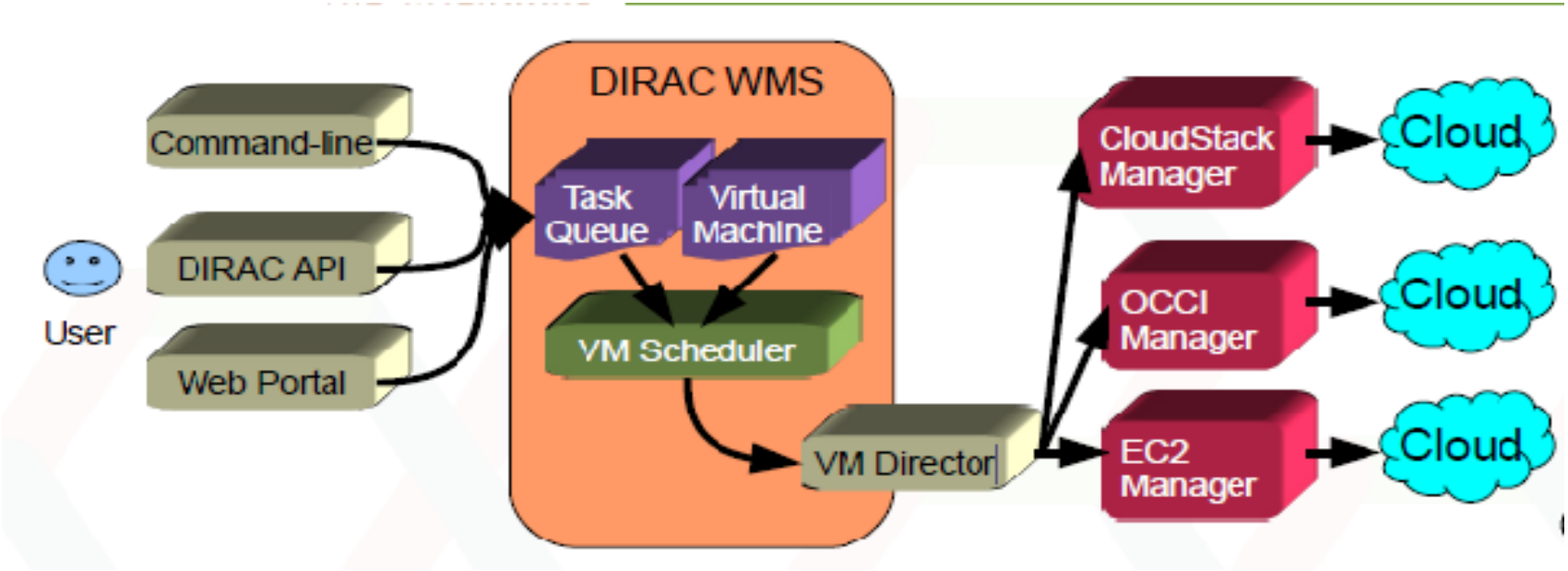
# Summary and Next plan

- First production has proved that the prototype system is workable
- Also gives us valuable experience
- Challenges still exist
- Man power is still a problem
- Next plan
  - Share DST files with remote sites for analysis
  - Second large scale production for real use
  - Explore new system and new technology to make systems more robust and more flexible

- Thank you!

# DIRAC and cloud

- VMDIRAC is developed to support cloud resources
  - A federated cloud model integration
- Cloud interface or type supported
  - OCCI, EC2, CloudStack, Openstack, Opennebula



# Torino site

- Own a private cloud managed by Opennebula
- Cloud both provide critical services and computing workforce
- Cloud resources are in full production mode since more than one year
  - Jobs efficiency is good comparing to no-cloud site