BESIII Distributed Computing with DIRAC

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

Third DIRAC User Workshop
Marseille, France, October 2012

Outline

- Introduction to BESIII experiment
- BESIII distributed computing and its status
  - Computing model
  - Job management
  - Data management
- Ongoing activities
  - DIRAC Accounting System with NoSQL
  - Production manager based on DIRAC
  - Data transfer with DIRAC data management system
  - Use Cloud and VM resources
- Summary

BESIII experiment

- Study electron-positron collisions in the tau-charm threshold region. Accelerator: BEPCII Detector: BESIII, Located in Beijing,
- Beam energy: 1.0-2.3 GeV
- Design luminosity: $1 \times 10^{33}/cm^2/s$ (100 times higher than BESII)
- Achieved luminosity: $0.7 \times 10^{33}/cm^2/s$

- Project timeline:
  - 2004 – Start of BEPC upgrade
  - 2006 - The detector installation
  - 2007 - BEPCII/BESIII commissioning
  - 2008 – Start of data taking
  - 2009 – Start of physics run data taking

BESIII collaboration

- China, Germany, Italy, Japan, JINR, Korea, Netherlands, Pakistan, Russia, Sweden, Turkey, USA

Data Volume

Type	Event size(KB)	2012(PB)	2020(PB)
Raw data (.raw)	~12	0.3	3.6
Reconstructed raw data(.dst)	~4	0.17	1.8
Simulated (.rtraw) and reconstructed data(.dst)	~6.5/23	0.13	1.0
Total		0.6	6.4

Not much comparing with LHC, but quite a lot for a single computing center, so we begin to consider distributed computing……

Challenges and Difficulties

- Most of sites have no grid experience, providing clusters in most cases
  - Cluster type are various: Condor, PBS, SGE, LSF……
- Many sites have no experience to set up and maintain parallel file systems or grid-enabled SE
  - Consider to allow some sites without SE
- Man power are very limited both in IHEP and remote sites
  - No full FTE, often a physics student as a part-time job
- Network between sites not good
  - 80Mb/s more or less
- Central storage system(Lustre and castor) is not allowed to open to outside, not grid-enabled
  - Put DPM or dCache between Lustre and WAN as a solution

Design principles

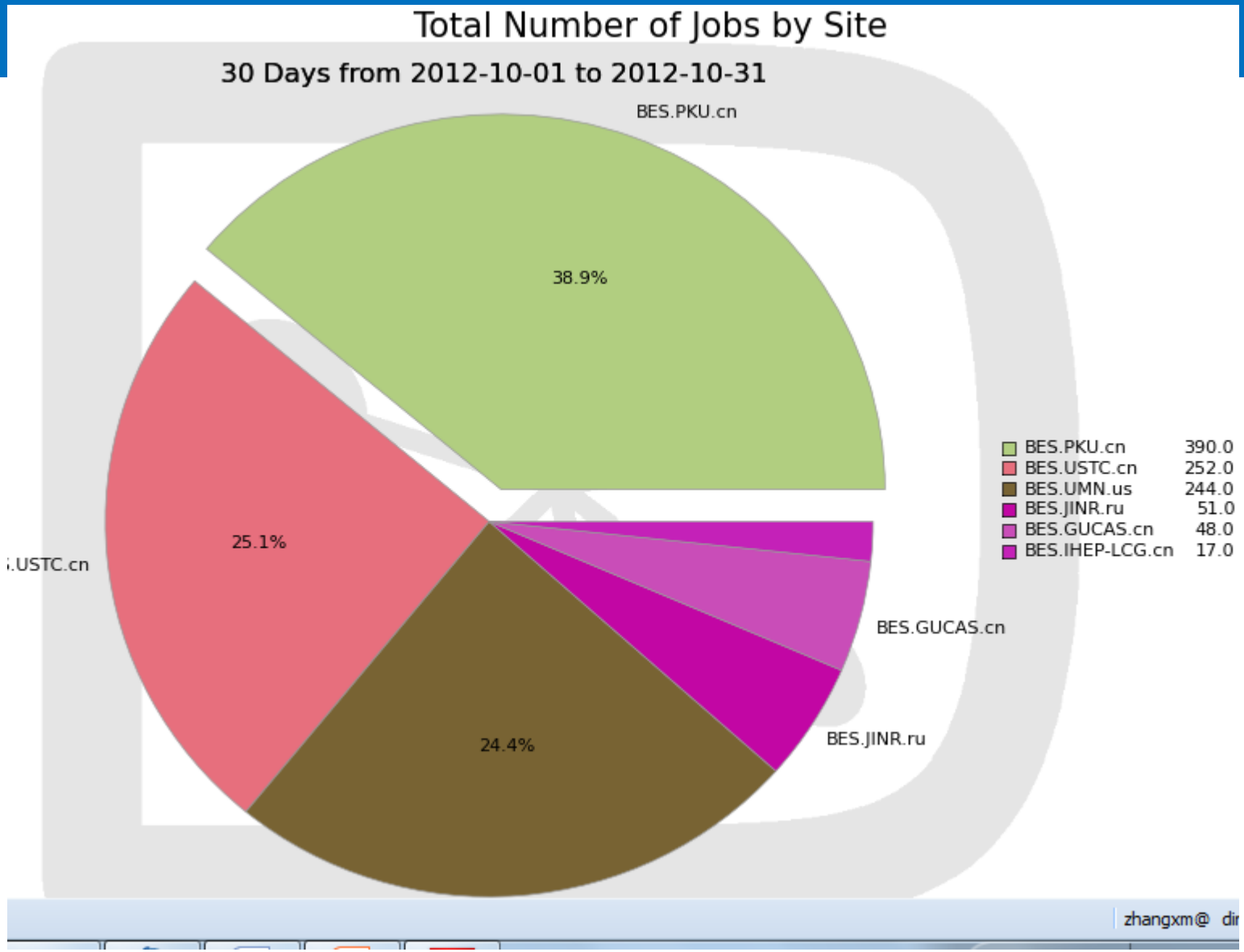
- Make it as simple as possible for sites to join and for users to use
- Use existing and mature software and middleware wherever possible, easy to set up, maintain and support
- Cope with low-bandwidth and unstable network
- Make everything as easy as possible to start with

BESIII Distributed Computing Model

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

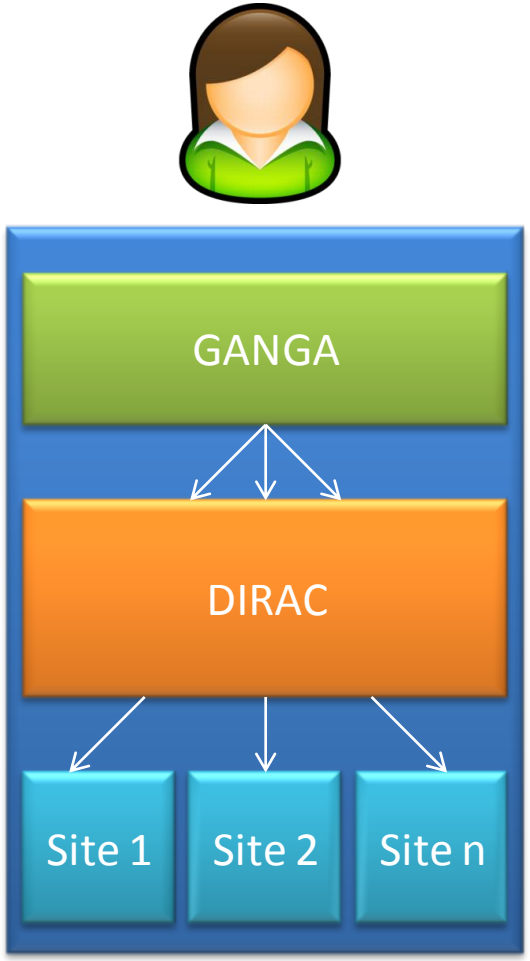
Site status

Site	Type	SE	Nationality	Status
JINR	GLite	dCache	Russia	Active
UCAS	PBS	BestMan	Beijing, China	Active
IHEP-PBS	PBS	DPM	Beijing, China	Active
PKU	PBS	No	Beijing, China	Active
USTC	PBS(Condor)	DPM	Anhui, China	Active
UMN	SGE	No(plan to have)	U.S.	Active
WHU	PBS	No	Wuhan, China	In progress
SDU	PBS	No	Shandong, China	In progress
NSCCSZ	LSF	No	Shenzhen, China	In progress


Total Number of Jobs by Site
30 Days from 2012-10-01 to 2012-10-31
BES.PKU.cn
38.9%
25.1%
.USTC.cn
BES.GUCAS.cn
BES.JINR.ru
24.4%
BES.PKU.cn 390.0
BES.USTC.cn 252.0
BES.UMN.us 244.0
BES.JINR.ru 51.0
BES.GUCAS.cn 48.0
BES.IHEP-LCG.cn 17.0
zhangxm@

Job Management

- **Decided to be based on DIRAC**
  - Comparing with GOS
  - More flexible, easy for sites, strong supports…
- **Components**
  - DIRAC for integrating site resources and distributing jobs
  - GANGA for user job submission interface
    - Consider to develop production manager system based on DIRAC
  - CVMFS (CERN VM File System) for deploying BOSS to remote sites



Status of BES DIRAC

- Use basic functionality, progress is slow
  - Basic workload management components
  - Dirac File Catalog
  - Accounting
- Try more functionality
  - Request management and data transfer
  - DFC web portal
  - Work flow and transformation
  - Resource status
  - VMDirac
- Consider to develop BES-specific extension
  - BESWeb
  - BES production manager

Data Management

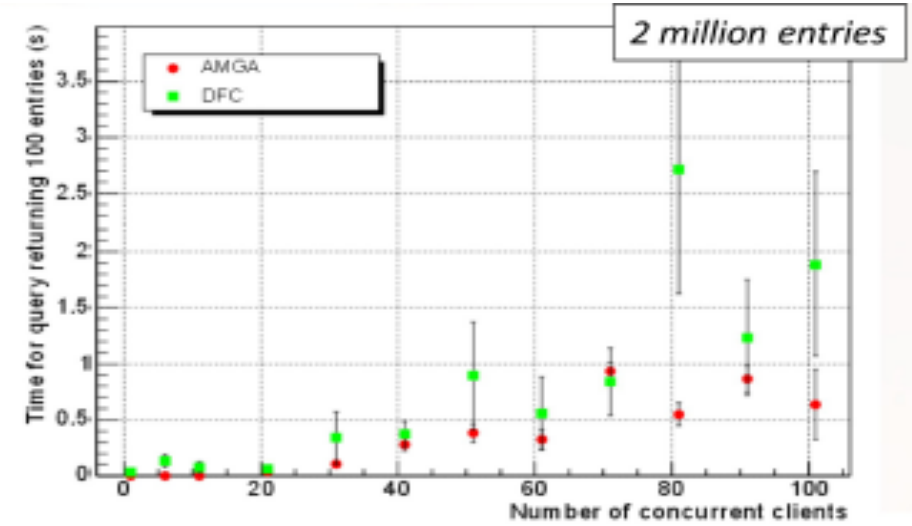
- Comparisons and tests have been done between AMGA and DFC
  - Both give acceptable performance
- Decide to use DFC for file / metadata / dataset catalogue
  - DFC meets more of BESIII requirements, more convenient to integrate with BESIII job management system
  - “Nice set of features, good performance, can keep everything in one catalogue”
  - “DIRAC team very helpful and responsive to our many questions, problems and feature requests!”
- BESIII grid data management tool, BADGER(BES-III Advanced Data manaGER) has been built up based on DFC , and been configured with BES schema

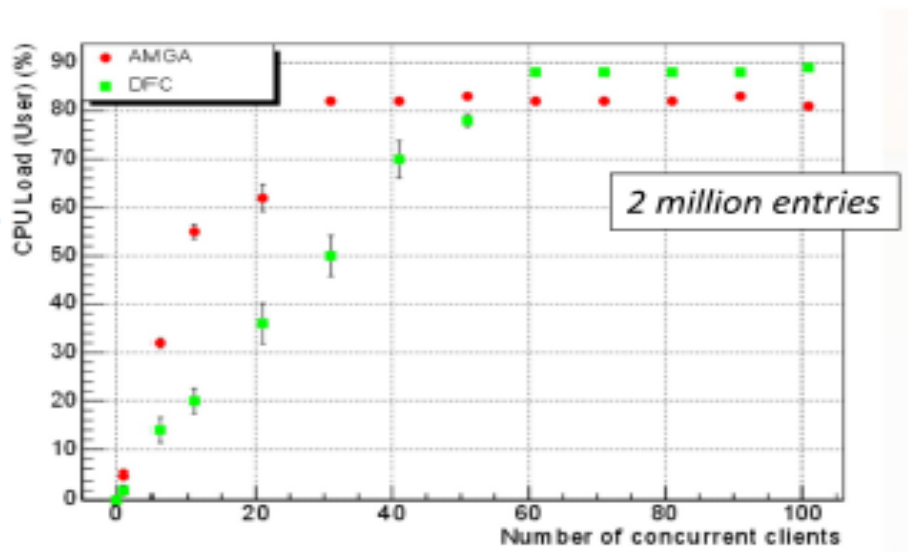
Data Management(Cont.)

- Need further supports and discussions in "dataset" issue
  - More dataset functionality (metaset), such as listing all the existing datasets
    - Dataset name is commonly used in old BESIII bookkeeping system
  - "Static dataset" issue, hot topic, not decided yet
    - Reason from physics group, most of BESIII analysis use absolute normalization including the exact event numbers involved to get branchings, cross-sections, etc

DFC vs. AMGA

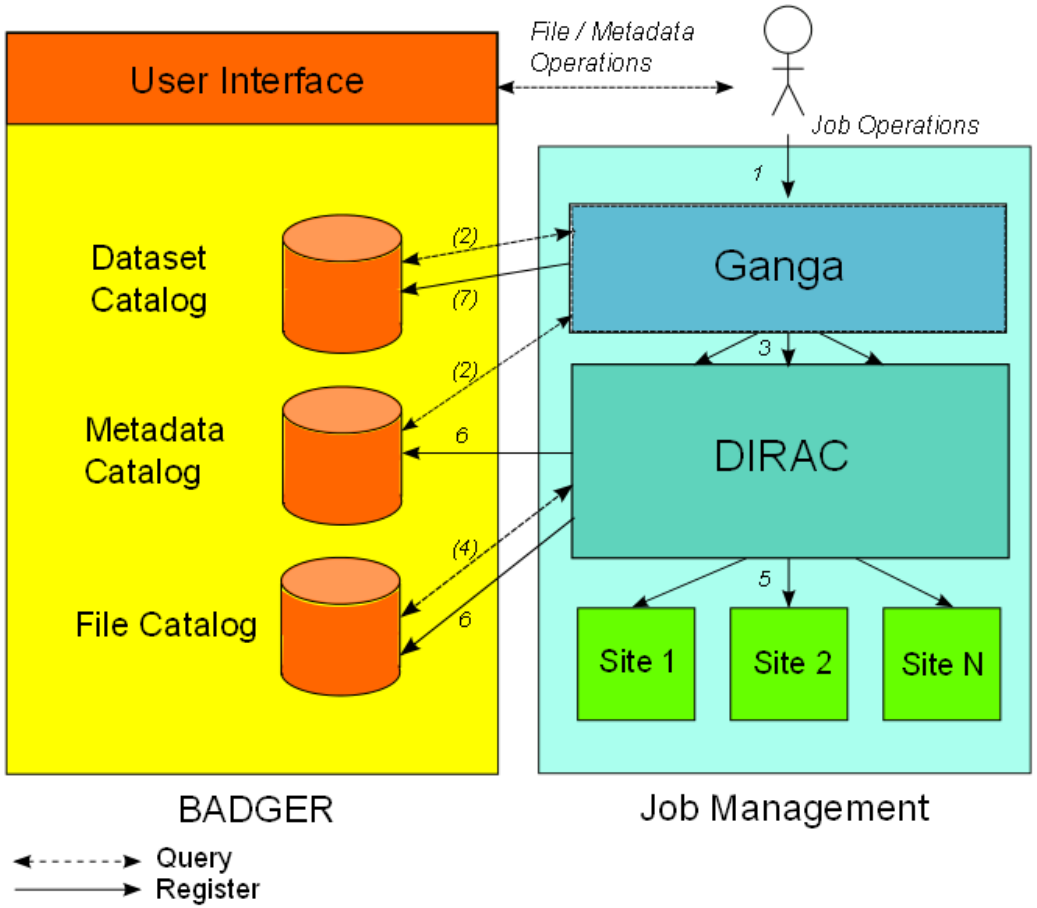
- **Conditions:**
  - BESIII Metadata schema
  - BESIII dataset
  - MySQL backend Optimized to support more than 100 processes
- **Results:**
  - With low number of clients, AMGA queries ~10x faster
  - With high number of clients, query times approx. equal
  - DFC CPU usage rises more slowly





BADGER and Job flow

- BADGER can work well with jobs
  - Data management and job management system are both based on DIRAC
- Jobs can automatically register output files as datasets in BADGER
- Jobs can easily find input files with metadata query or with dataset name from BADGER



Data transfer with DIRAC

- Have no data transfer system yet
  - Plan to use DIRAC transfer request management system with FTS
- Glite File Transfer Service (FTS2) has been established in Dubna , Russia
  - Some Channels (IHEP-JINR, JINR-IHEP, IHEP-USTC, USTC-IHEP) are established
  - Gridftp and SRM transfers are enabled
- Integrate with DIRAC request management (ongoing)
  - Direct FTS submission by DIRAC DMS commands(dirac-dms-fts-submit) is ok
  - Bulk FTS transfers with a list of LFN can be done through DIRAC Request Manager services
  - Transfer status can be seen from DIRAC request service portal

Data transfer (Cont.)

- To do and expect:
  - Transfer requests can be made using BESIII datasets defined in BADGER
  - Request and Transfer errors can be more clearly shown in monitoring page
  - Request monitoring web portal can be more useful and powerful
    - Clean up not useful parameters like JobID, requestType….
    - Show more detailed transfer info, such as dataset name to be transfered, transfer rate, completion status…….
  - Get Accounting system well configured for Data transfers

RequestId	JobID		Status	RequestType	Operation	OwnerDN	OwnerGroup	Error
11	0		Assigned	transfer	replicateAndRe...	/C=CN/O=HEP/OU=IHEP/CN=zhang xiaomei	bes_user	
4	0		Assigned	transfer	replicateAndRe...	/C=CN/O=HEP/O=IHEP/OU=PHYS/CN=tao lin	bes_user	

Production manager

- In preparation phase
  - Consider to use DIRAC workflow, transformation?
  - LHCb production system is studied, including work flow, transformation system,etc
  - Wait for detailed requirements from BESIII production group
  - User interface will be designed as parts of DIRAC web portal
- Good suggestions and useful help in face-to-face meeting
  - Good and simple examples from Stephane Poss about using DIRAC workflow
  - Try to start with it

DIRAC Accounting System with NoSQL

- **As a contribution to DIRAC**
- **Goal :** Explore a NoSQL data store to replace current Mysql for storing and exploiting DIRAC accounting records
- The student (Zhang Gang) is focused on that with the guide of Fabio and Ricardo
- **Status (from Zhang Gang):**
  - Get familiar with various NoSQL data stores, such as Hadoop Hbase, Cassandra, Riak, CouchDB, etc
  - Understand source code, data structure and the schema design of the DIRAC Accounting System
  - Choose Hadoop Hbase as a start point and do the comparisons with Mysql using LHCb data (ongoing)

DIRAC Accounting System with NoSQL(cont.)

- **Difficulties &Problems**
  - Still not quite sure which NoSQL database is most suitable in DIRAC case although starting with Hadoop Hbase
    - Suggestions? Easily set up with one server， Cassandra?
- **More details**
  - http://twiki.ihep.ac.cn/twiki/bin/view/Dirac/20120809
  - http://twiki.ihep.ac.cn/twiki/bin/view/Dirac/20120823
  - http://twiki.ihep.ac.cn/twiki/bin/view/Dirac/20120913

Use Cloud and VM resources

- Reasons
  - Some sites are not able to change original OS
    - Resources are shared by other applications
    - The original OS is not suitable for BESIII applications
  - Some collaborations may consider to buy commercial cloud resources as a contribution
  - Not avoid to buy commercial resources for peek requirements for some special periods
- Start with doing some practices
  - Add and configure VMDirac in BES DIRAC
  - Use OpenStack as an example
    - Testing environment is available in IHEP
    - Try to understand how to set up VM with CERNVM image and do the contextualization with OCCI interface provided by Openstack
  - Need more guides and help from Victors

A coming use case - NSCCSZ

- A good chance to learn how china cloud resources are provided
- NSCCSZ (National Super Computer Center at ShenZhen)
  - A national super computer center, located in ShenZhen
  - owns cluster resources and cloud resources
  - More than 10000 cpu cores in LSF cluster with MPI supports
  - More than 1000 blade servers for cloud
- For clusters, they provide Suse OS
  - BESIII software is developed and tested in SLC
  - Seem not possible to use clusters
- For cloud, details need to be known
  - Such as interface, API…..
  - Need to support CERNVM image, not available yet

Summary

- DIRAC has been used to build up BES Data management and Job management
- Basic functionality of BESIII distributed computing is in place
- Large scale tests close to real production are needed to study system stability and performance
- More efforts and developments are needed on data transfer system and production management system, on attracting more sites to join
- Try to add more DIRAC functionality to make the system more complete and powerful
- Happy to contribute to DIRAC more in the future

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