BESIII Distributed Computing with DIRAC

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

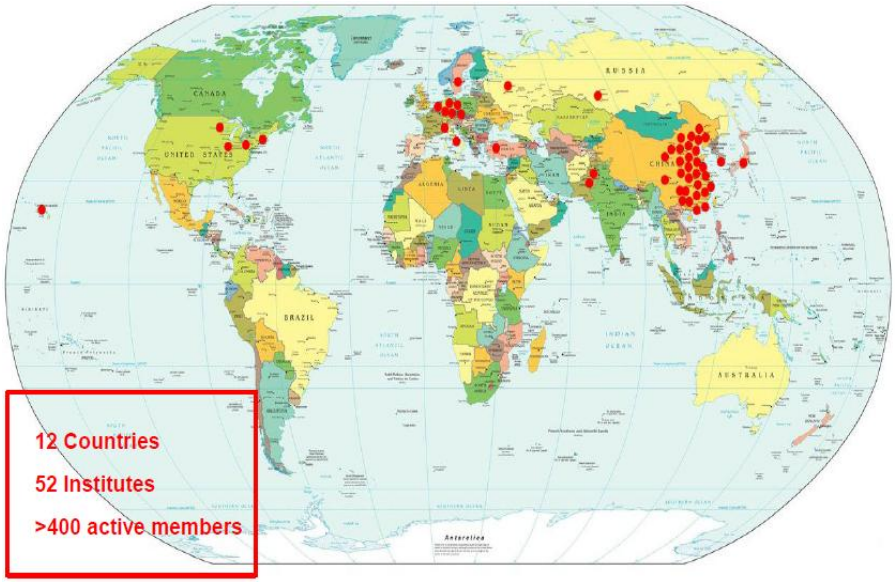
Fourth DIRAC User Workshop
CERN, May 2014

Part I

STATUS OF BESIII DISTRIBUTED COMPUTING

BESIII experiment

- Located in Beijing, study electron-positron collisions in the tau-charm threshold region. Accelerator: BEPCII Detector: BESIII
- Beam energy: 1.0-2.3 GeV
- Design luminosity: $1 \times 10^{33}/cm^2/s$ (100 times higher than BESII)
- About 12 countries, 52 institutes in the cooperation



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**

Working group

- IHEP (central services, management of sites)
  - Xiaomei Zhang, Tian Yan, Xianghu Zhao
- JINR (data management, site monitoring)
  - Alexey Zhemchugov, Sergy Belov, Igor Pelevanyuk
- SOOCHOW university (VMDIRAC and cloud)
  - Lingzhi Lin, Jing Wei

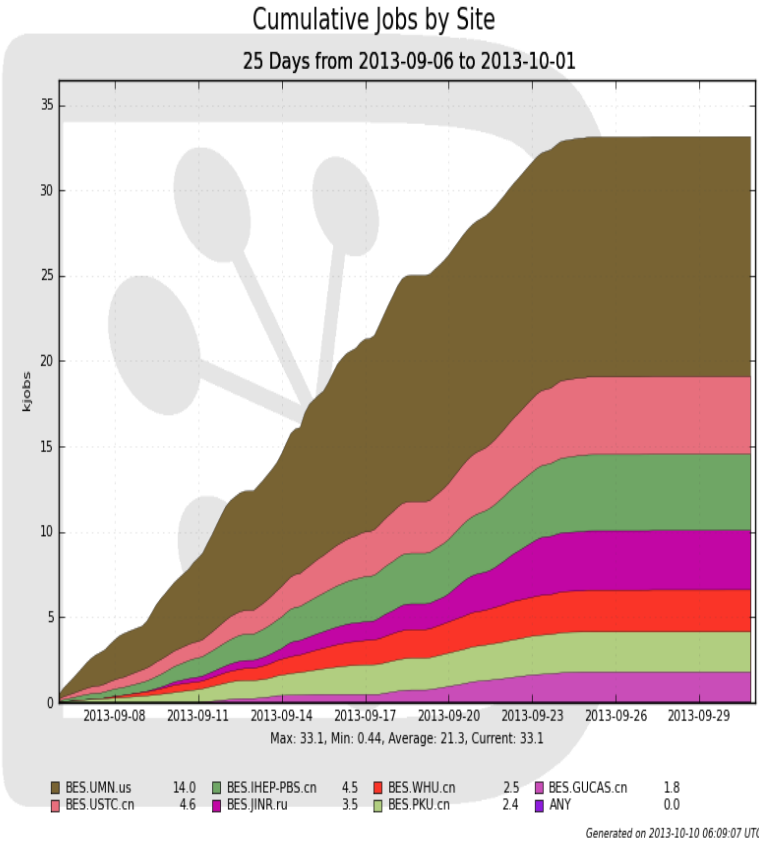
Resources joined

- 10 sites joined from the BESIII collaboration and 5 sites with SE
- About 2032 CPU cores and 246 TB are available

#	Site Name	CPU Cores	LRMS	SE
1	IHEP-PBS.cn	96	PBS	dCache 126 TB
2	GUCAS.cn	152	PBS	
3	USTC.cn	128 + 300~768	PBS + condor	dCache 24 TB
4	PKU.cn	88	PBS	
5	JINR.ru	40~128	gLite	dCache 7.3 TB
6	UMN.us	768	SGE+condor	BestMan 50TB
7	WHU.cn	30 + 100~300	PBS	StoRM 39 TB
8	INFN-Torino.it	~200	gLite	
9	SDU.cn	~102	PBS	
10	BUAA.cn	~256	PBS	
	Total	**2032~3086**		**246.3 TB**

Three production campaigns

- 2012.11
  - produced 200 million Psi(3770) bhabha events
  - more than 8000 jobs are submitted and run
  - Stop because the central SE has no enough space
- 2013.9
  - produced 800 million Jpsi inclusive events
  - more than 33000 jobs
  - success rate about 84.1%, main failure in data uploading and site failure
- 2013.12
  - produced 1,352.3 million Psi(3770) events
  - about 69904 Jobs
  - success rate about 90.4%, main failure in data uploading and site failure



Simulation+reconstruction challenges

- Plan to do the whole production in remote sites
- The whole MC production takes two steps: simulation-reconstruction
  - Reconstruction need to input random trigger data to mix with raw events as background
  - Random trigger data is stored as files classified by run number
  - One Job can cover multi runs(~10 runs), each job need to include about 10GB random trigger files
  - The size of random trigger needed by jobs is nearly the same as the inputted simulated data
- Challenges
  - It is difficult to download random trigger data with several runs in one job since the data to be downloaded is too big
  - Not all the sites owns their own SEs

Simulation+reconstruction tests

- Current methods taken
  - Random trigger is deployed to the site with SE
  - The site without SE download random trigger data directly from other Ses
  - Only single run in one job is allowed
- Advantage
  - Make simulation+reconstruction possible in remote sites
  - Downloading can use time of simulation step
- Disadvantage
  - CPU efficiency is not high since jobs are waiting for downloading the whole random trigger file
  - The range of runs taken care by each job can't be too much
- Cloud storage would be the better solution?
  - Open instead of download the whole file
  - Share random trigger among sites

Part II

BESDIRAC

DIRAC set-up

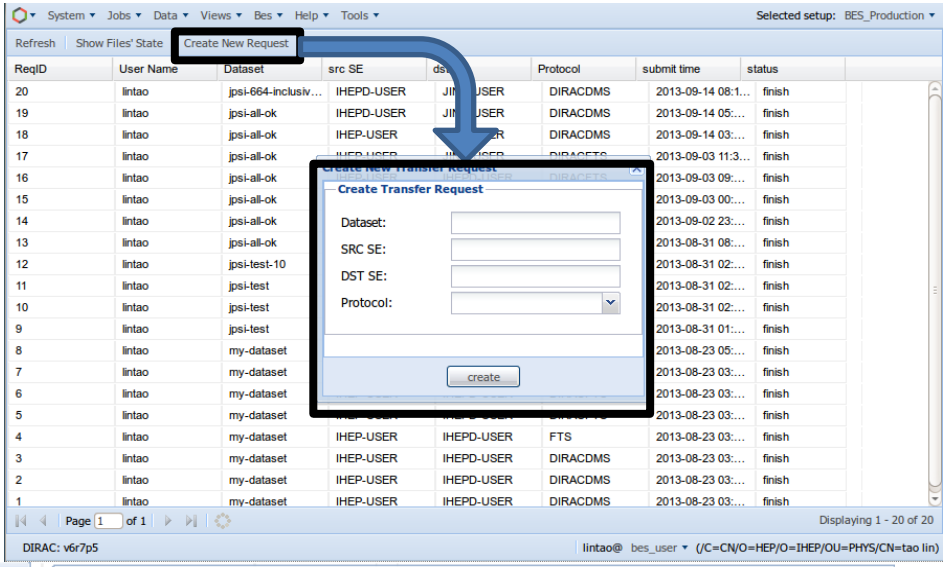
- One server for production
  - V6r10-pre17
  - DIRAC components used
    - Workload management
    - Dirac File Catalog
    - Accounting
- One separated server for test and development
  - Latest release
- Plan
  - Add one server for upgrade purposes
    - Different instances with one CS
  - Add one slave server for data transfer services

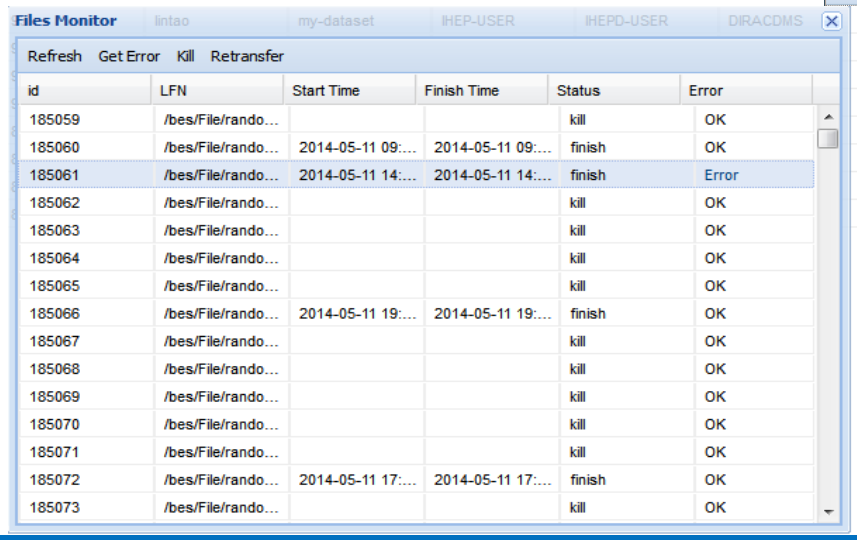
BESDIRAC

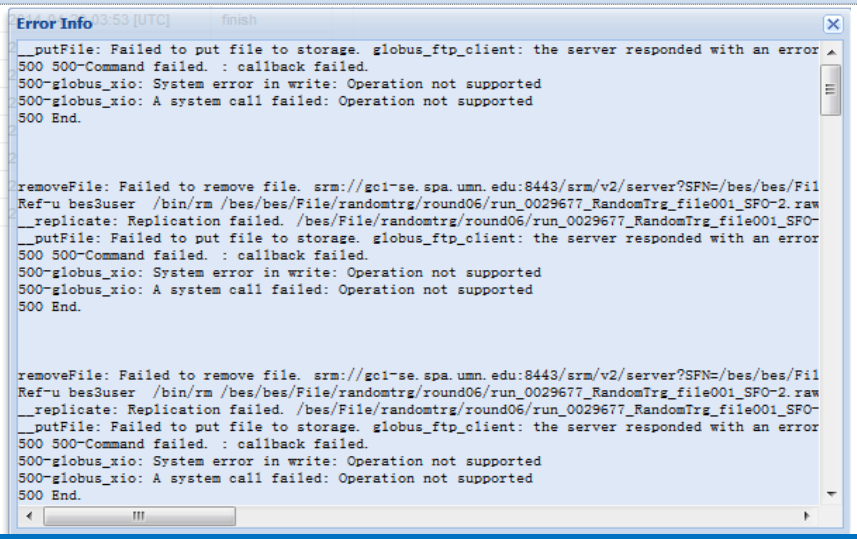
- An extension to DIRAC
  - V0r8
  - Hold BESIII-specific packages
    - Data managements tools
      - BESIII dataset toolkits
      - Random trigger toolkits
      - Special wrapper to DFC commands and APIs
    - Data transfer system
    - Monitoring system (to be included)

Data Transfer System (I)

- Goals
  - Transfer DST datasets to sites for analysis
  - Copy back MC production job outputs to IHEP central SE
- Usage
  - Users can submit and monitor transfer requests through web interface
  - Transfer service take care of bulk transfer automatically with dataset name

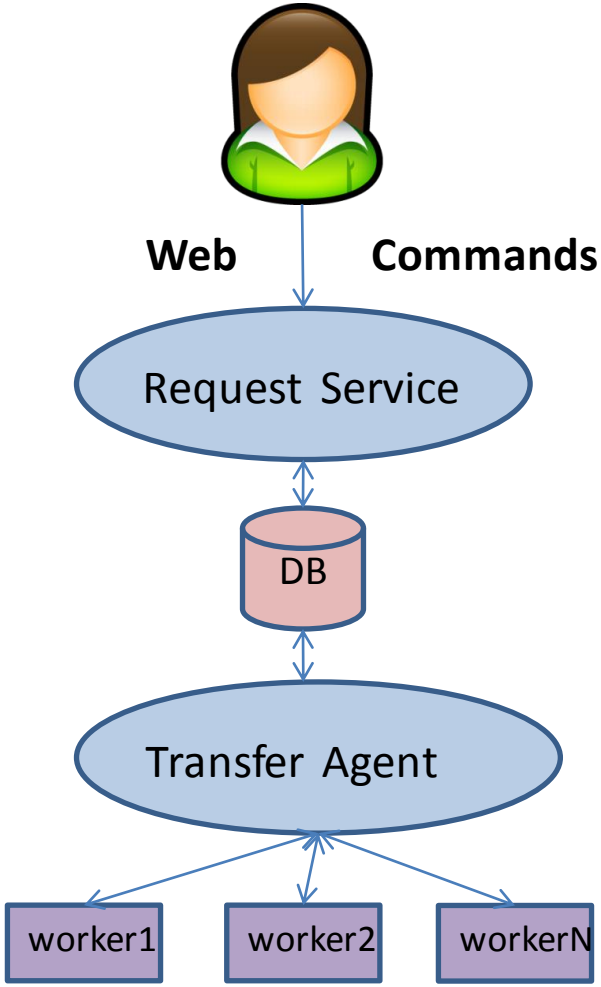






Data Transfer system (II)

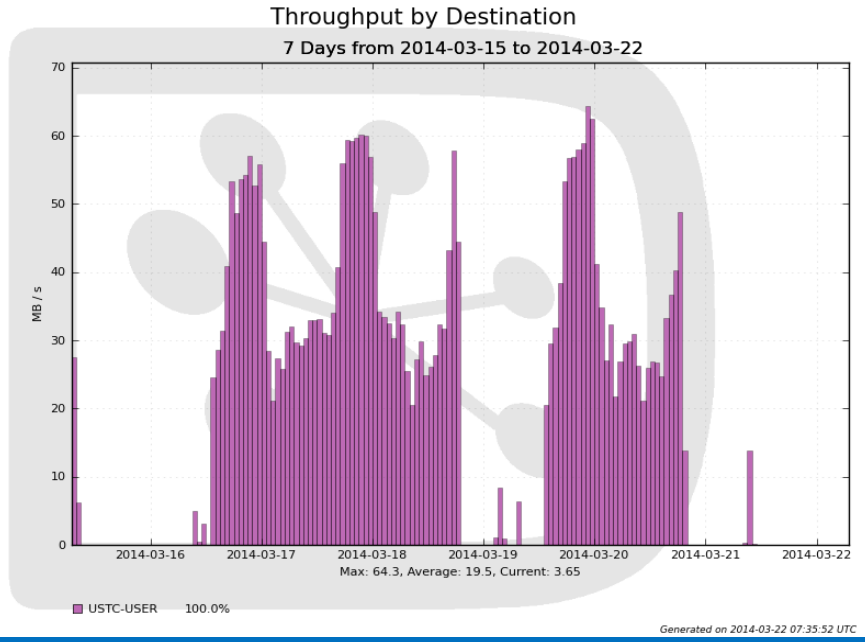
- Developed based on DIRAC framework
- Main components
  - Transfer Agent
    - scheduler to manage transfer workers
  - Transfer workers
    - Manage real transfers between SEs
  - Transfer Request Service
    - manage the transfer requests created by users
  - Transfer DB
    - communicate between the agent and service
  - Accounting
    - keep the transfer history



Data Transfer System (III)

- Transfer tests between remote SEs
  - IHEP SE, USTC SE , JINR SE, WHU SE, UMN SE
  - Average Speed is about 20MB/s
  - Success rate is above 99%
  - Most of failure comes from lost connection between workers and agent, need to be fixed

Batch	Destination SE	Files	Data Size	Average Speed	Success Rate
1	USTC-USER	12,468	4.04 TB	23.4 MB/s	99.37%
2	JINR-USER	12,468	4.04 TB	24.9 MB/s	99.76%
3	WHU-USER	12,468	4.04 TB	21.4 MB/s	99.76%
Total		**37,404**	**12.12 TB**		



Site monitoring

- Taken by the JINR group
- Similar to RSS? With BESIII special requirement?
- Igor Pelevanyuk would like to present the details

Part III

TO BE EXTENDED

Extensions

- WebAppDIRAC
- Cloud Storage
- VMDIRAC
- BONIC

WebAppDIRAC

- BESDIRAC still use old web portal
  - Data transfer extensions use old web framework
  - Site monitoring already use new one
- Need to develop new web applications
  - Query page of BESIII datasets
  - Task-based production accounting
- Plan
  - Migrate from old portal to new one before new developments
  - How to do the smooth migrations from old one to new one? Any guides?

Cloud storage

- Taken by Fabio Hernandez from IN2P3
- Goal
  - Determine if and how we can exploit cloud-based storage for research
  - Identify relevant use cases
- Cloud-based storage
  - Object storage system, standard access protocols (typically HTTP-based), accessible through wide area networks
  - Both commercial services and in-house deployments
  - Key points: immutable objects (no POSIX compliance), 2 level namespace (containers & objects)

Current status

- Developed an extension to ROOT for supporting transparent read of data stored in the cloud
  - OpenStack Swift and S3 (Amazon and Google)
  - No modifications to ROOT source code nor to experiment's code base
  - Supports all versions of ROOT since Oct. 2009
  - Source code and documentation:
    - https://github.com/airnandez/root-cloud

Current status (cont.)

- Developed a FUSE-based file system interface to cloud storage
  - Goal: to expose your files stored in the cloud as local files (Linux and MacOS X)
  - Usable both for batch jobs and for your own personal computer
  - Think of it as cloud storage to be the backend of your personal storage element
- Example use case: BES III distributed event reconstruction
  - Goal: run event reconstruction jobs in remote sites with slowish network links to IHEP computing center
  - Random trigger data (1GB-2GB), hosted at IHEP (Beijing), mounted in read-only mode in remote worker nodes
  - Unmodified reconstruction jobs transparently read chunks of those files
  - Network connectivity does not seem to be a limiting factor
  - Benefit: bulk download of whole file not needed, on-demand download only what the job actually reads, no local storage element in remote BES III sites
  - Lower the barrier for remote sites to contribute compute power to the DIRAC-managed BES III grid

What is the next?

- Vision
  - The DIRAC jobs should be able to interact with user's own cloud-based storage
    - For transparently storing to and retrieving data from my space
  - When the job finishes, it stores its output data in that storage and users immediately see the new files appear from my personal computer: double click to open
    - Even if the DIRAC client is not installed in my machine
- Would this be interesting for the DIRAC community?
  - Happy to hear from you: fabio@in2p3.fr

VMDIRAC

- Goals:
  - Sites has provided and would like to provide cloud-based resources
    - INFN-Torino, WHU, SOOCHOW
    - CAS cloud centers (in construction)
  - Triggered by some existing cloud projects
- Status:
  - VMDIRAC is installed and studied
  - Openstack cloud testbed has been set up

BOINC

- Status:
  - BONIC resources is connected and tested
  - Failure rate is high when accessing experiment software
- Problems:
  - It is slow to read experiment software in CVMFS first time when VMs start from User PCs
  - User PCs have not as good network connection with outside as worknodes

Summary

- BESIII distributed computing is already in production
- More efforts need to be done to make it robust and efficient such as monitoring
- Some challenges still exists like accessing random trigger data
- New technologies and resources are interested to make it more convenient and useful

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