Multi-VO support in IHEP's distributed computing environment

Journal of Physics: Conference Series **664** (2015) 062068 doi:10.1088/1742-6596/664/6/062068

Multi-VO support in IHEP's distributed computing environment

T Yan[1], B Suo[1], X H Zhao[1], X M Zhang[1], Z T Ma[1], X F Yan[1], T Lin[1], Z Y Deng[1], W D Li[1], S Belov[2], I Pelevanyuk[2], A Zhemchugov[2] and H Cai[3]

[1] Institute of High Energy Physics, 19B Yuquan Road, Beijing 100049, R. P. China
[2] Joint Institute for Nuclear Research, Joliot-Curie 6, 141980 Dubna, Moscow region, Russia
[3] Department of Physics, Wuhan University, 299 Bayi Road, Wuhan 430072, P. R. China

E-mail: yant@ihep.ac.cn

Abstract. Inspired by the success of BESDIRAC, the distributed computing environment based on DIRAC for BESIII experiment, several other experiments operated by Institute of High Energy Physics (IHEP), such as Circular Electron Positron Collider (CEPC), Jiangmen Underground Neutrino Observatory (JUNO), Large High Altitude Air Shower Observatory (LHAASO) and Hard X-ray Modulation Telescope (HXMT) etc, are willing to use DIRAC to integrate the geographically distributed computing resources available by their collaborations. In order to minimize manpower and hardware cost, we extended the BESDIRAC platform to support multi-VO scenario, instead of setting up a self-contained distributed computing environment for each VO. This makes DIRAC as a service for the community of those experiments. To support multi-VO, the system architecture of BESDIRAC is adjusted for scalability. The VOMS and DIRAC servers are reconfigured to manage users and groups belong to several VOs. A lightweight storage resource manager StoRM is employed as the central SE to integrate local and grid data. A frontend system is designed for user's massive job splitting, submission and management, with plugins to support new VOs. A monitoring and accounting system is also considered to easy the system administration and VO related resources usage accounting.

1. Introduction

The Beijing Spectrometer III (BESIII) experiment located at the Institute of High Energy Physics (IHEP), China, is designed to study tau-charm physics [1, 2]. It started raw data taking in the year 2009. Since the data sample increases rapidly, the local computing resources in IHEP can't meet the requirements of all raw data processing and Monte-Carlo (MC) data production. As a supplement, the distributed computing environment has been setup and been in production status since 2012 [3]. It is built based on the middleware DIRAC (Distributed Infrastructure with Remote Agent Control), which is originally developed for LHCb experiment, but now a community grid solution [4, 5]. DIRAC provides a complete grid solution for both workload and data management system, and it is designed to minimize the effort of local sites and system maintainer. Thus DIRAC fits our situation that most of the sites are small and lack of experts in grid computing.

Recently, inspired by the success of BESDIRAC project, some other experiments operated by IHEP are willing to use DIRAC for their solution to distributed computing. These potential

DIRAC users are Circular Electron Positron Collider (CEPC) [6, 7], Jiangmen Underground Neutrino Observatory (JUNO) [8, 9], Large High Altitude Air Shower Observatory (LHAASO) [10], Hard X-ray Modulation Telescope (HXMT) [11]. Each experiment's community naturally forms a virtual organization (VO). They have various heterogeneous computing and storage resources located at geographically distributed universities or institutions worldwide, with larger or smaller size. They wish to integrate those resources and supply to their VOs users, especially at the early stage of the experiments.

In order to save manpower and hardware cost, we decided to extend the currently running BESDIRAC platform to support multi-VOs. That is to say, we are making DIRAC as a services for those user communities. The example of France-Grilles project shows that DIRAC is technically supporting such an extension [12].

This paper is organized as follows. Firstly we introduce the system of BESDIRAC in Sec. 2, then explain how we are motivated to the multi-VO solution in Sec. 3. In Sec. 4, the details on the techniques of multi-VO support are presented, including system architecture refinement, configuration on VOMS and DIRAC servers, employment of StoRM SE, as well as the frontend, monitoring and accounting systems. Finally it follows the conclusion.

2. Distributed computing environment for BESIII experiment

The distributed computing environment for BESIII experiment is built upon the middleware DIRAC, with BESDIRAC extensions for BES specific modules. It started running at the year 2012. About 3000 CPU cores and 400 TB storage are contributed by 10 sites from BESIII collaboration members, to serve MC simulation, reconstruction and analysis jobs, as well as data transfer between sites. As shown in Fig. 1, averagely there are about 50 thousand jobs are executed per month.

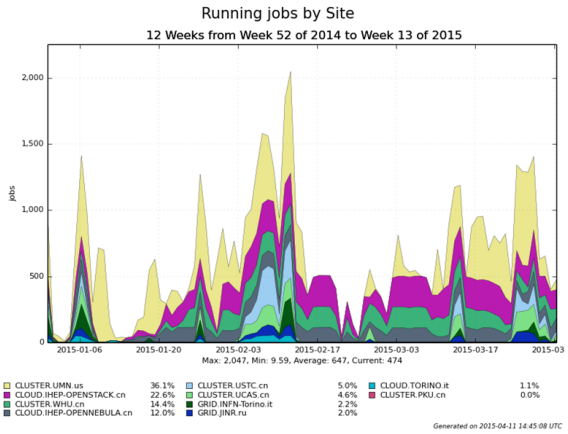

Figure 1. The running BESIII jobs by site in the last three months.

The system architecture of BESIII distributed computing environment is shown in Fig. 2. The frontend system for user job splitting and submission is ganga [13] with gangaBOSS extension [14], here BOSS is the BESIII offline software system. The central component is the middleware, DIRAC, it receives jobs from ganga and distributes them to remote sites. As shown in Fig. 1, we have three kinds of sites: 5 cluster sites, 2 grid sites and 3 cloud sites. Virtual machines in cloud sites are scheduled by a DIRAC extension called VMDIRAC [15]. The DIRAC server is hosted on a physical machine with 8 cores, 16 GB RAM, and 500GB disk. All the services are running in this single server with MySQL database in the same host. The VOMS server is running in a virtual machine, with only one VO (i.e. bes) configured. The public CVMFS server at CERN hosts a repository boss.cern.ch for our software deployment.

For the data management system, we use DIRAC File Catalog as the core catalogs [16]. A dCache [17] Storage Element (SE) is setup for central SE. It has a disk array of 126 TB capacity. The 20 TB random trigger data needed for reconstruction jobs is uploaded from IHEP's local Lustre [18] file system to this dCache SE and transferred to remote sites, by a high level dataset based data transfer system [19]. This transfer system is also used for DST data exchange among IHEP and remote sites. The output data of jobs are written directly back to the dCache central SE, then user download the output data from SE to Lustre by a tool based on rsync [20].

3. Motivations of multi-VO support

The above BESDIRAC platform is chosen to be extended to support multi-VOs, since the advantage of building DIRAC as a service is obvious in our case:

(i) DIRAC needs dedicated hardware and expert manpower to maintain, small VOs are not willing to afford that;

(ii) many universities in China joined several experiments hosted by IHEP, this means one site will belong to several VOs, a single DIRAC server will be easier to manage these resources;

(iii) the BESDIRAC setup is already in production status, thus it's much easier to extend it than build a new system for each VO;

(iv) once this multi-VO extension is done, it will be very convenient to support newly joined VOs in future.

Therefore, instead of setting up dedicated distributed computing environment for each experiment, operating a single setup for all the experiments is an effective way to make maximal use of shared resources and manpower while remaining flexible and open for new experiments to join.

4. Techniques of multi-VO support

4.1. System architecture refinements

As shown in Fig. 3, the system architecture of IHEP's distributed computing environment is refined to support multi-VOs.

The redundancy of DIRAC services is achieved by setting up three DIRAC servers, one is master and the other two are slaves. Each DIRAC server is hosted on a physical machine with 32 cores, 64 GB RAM and 1 TB disk. A geographically separated DIRAC mirror server located at Joint Institute for Nuclear Research is also in progress. The vital services such as Configuration Service, File Catalog Service are duplicated at several servers, the mirrors keeps synchronized with the master instance. Also, some services like Site Directors are distributed in several hosts for load balancing. The MySQL database is deployed in a standalone machine with 32 cores, 64 GB RAM and 600 GB 15k SAS Disk arranged in RAID-10. It will be easier to upgrade DIRAC or enlarge the system scale by adding new slaves when we have a separated DB server.

For the data management system, we setup a StoRM SE [21] with multi-VO support. It also acts as a frontend of the PB level local Lustre file system. The local data is exposed to remote sites through StoRM, uploading and downloading data between Lustre and SE is no longer needed. For software deployment, a CVMFS stratum 0 server is setup at IHEP, hosting the repositories of all the VOs supported. Repository of a newly coming VO can be added rapidly. We developed lightweight frontend system called JSUB, three VOs (bes, cepc, juno) have already been supported. A monitoring and accounting system is also added to increase the efficiency of system maintenance when more and more resources are joined.

4.2. Configurations on VOMS and DIRAC servers

The VOMS server is upgraded to the latest version and reconfigured to support multi-VOs, by using the `voms-configure` utility. Each VO has a unique port and it's own database.

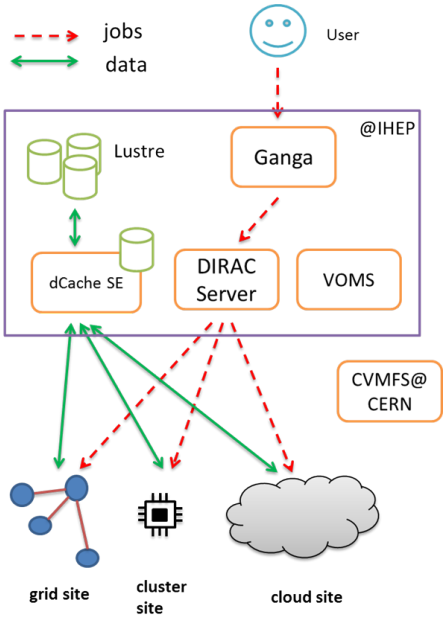

Figure 2. Architecture of BES-DIRAC.

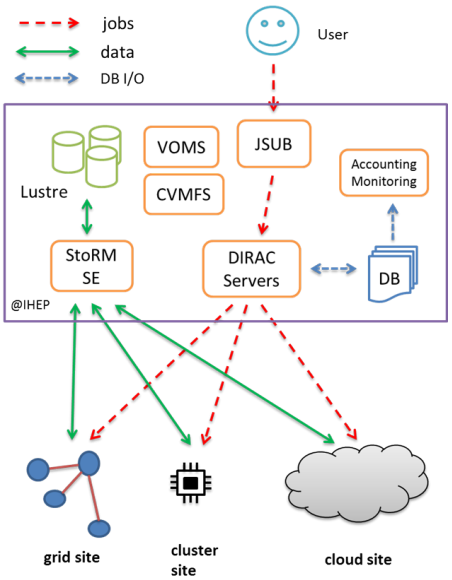

Figure 3. Refined architecture for multi-VO support.

In DIRAC server configuration system, each VO together with its related users and groups should be registered in `Registry` section. In `VOMS` and `VO` subsections, the Mapping, URLs and Servers information of each VO should be added. For each VO, a related user group, generic pilot group should be added and their VO properties should be specified.

In `Resources\Sites` subsection, one can specify which VOs are supported by a site or a queue in site. Multi-VO support in one site is also allowed. The cloud sites is somewhat different from cluster and grid sites. It's managed by VMDIRAC and explicit VO tag is not supported currently. A temporary solution is to specify the groups belong to the VO in `Requirements` of the `RunningPods`.

In workload management system, the site director agent is rearranged that one agent works for one VO, with its own configurations files. In these files, the VO property is given. The site director will send generic pilot job to the queue under the same VO as the job. At the worker node, the generic pilot can pull all jobs in the same VO.

4.3. StoRM SE

The StoRM is completely VOMS aware since it relies on user credential for what concern user authentication and authorization. It can use the VOMS extensions to define the access policy.

StoRM is configured with YAIM. A multi-VO setup is made by the following steps. In general, YAIM variable `VOS` lists all the supported VO names, and we can set the following variables on storage area for each VO:

- `STORM_STORAGEAREA_LIST`
- `STORM_{SA}_VONAME`
- `STORM_{SA}_ONLINE_SIZE`
- `STORM_{SA}_DEFAULT_ACL_LIST`

The users and groups for VOMS mapping as well as the vomses information should also be set.

4.4. Frontend system
Since ganga is too complicated for us, it takes a long time to add plugin for new VOs. Therefore, we developed a lightweight frontend JSUB based on DIRAC API. The user cases of BESIII, CEPC, JUNO experiments are taken into consideration. The design principle and goals of JSUB are

- lightweight, only necessary functionalities are implemented;
- easy to add plugin for new VO;
- support task-based job splitting, submission and management;
- support DIRAC and HTCondor backend.

A prototype of JSUB is accomplished and has been used by CEPC MC production group. We will continue the development of this frontend system to improve it's robustness and scalability.

4.5. Monitoring and accounting system
A monitoring system can help the maintainer on shift to find problems quickly and locate it exactly. This will greatly save manpower during the management of several VO's resources. Moreover, a VO-related accounting system is also necessary for the statistics of resource usage by different users and different VOs, especially in the case cloud sites. Such a monitoring and accounting system is under design recently and will be finished at the end of this year.

5. Conclustion
The distributed computing environment for BESIII experiment at IHEP has been extended to support multi-VO usage scenario. The middleware DIRAC has been made as a services for several experiments in IHEP. The experience of designing multi-VO support in IHEPs distributed computing environment may be interested and useful for other high energy physics center operating several experiments.

Acknowledgments
The authors would like to express their appreciation to Andrei Tsaregorodtsev and Ricardo G. Diaz for their help on DIRAC configurations and many discussions, Victor Mendez and Victor Fernandez for their help on VMDIRAC, and colleagues at IHEP computing center for their support. This work is funded in part by the National Natural Science Foundation of China (NSFC) under grant no. 11375221, Joint Funds of NSFC under grant no. U1232201 and U1232109, and Joint RFBR-NSFC project no.14-07-91152.

References
[1] BESIII Collaboration 2009 The construction of the BESIII experiment *Nuclear Instruments and Methods in Physics Research Section* A **598** 7–11
[2] Ablikim M *et al* 2010 Design and construction of the BESIII detector *Nuclear Instruments and Methods in Physics Research Section* A **614** 345–99
[3] Deng Z Y, Li W D, Lin L, Liu H M, Nicholson C, Sun Y Z, Zhang X M and Zhemchugov A 2012 Experience of BESIII data production with local cluster and distributed computing model *J. Phys: Conf. Series* **396** 032031
[4] Tsaregorodtsev A *et al* 2008 DIRAC: A Community Grid Solution *J. Phys: Conf. Series* **119** 062048
[5] Casajus A *et al* 2012 Status of the DIRAC project *J. Phys: Conf. Series* **396** 032107
[6] Gibney E 2014 China plans super collider *Nature* **511** 394–5
[7] The CEPC-SPPC study group 2015 *The CEPC-SPPC Preliminary Conceptual Design Report (pre-CDR)* http://cepc.ihep.ac.cn/preCDR/volume.html
[8] He M *et al* 2014 Jiangmen Underground Neutrino Observatory *Preprint arXiv:1412.4195*
[9] Li Y F 2014 Overview of the Jiangmen Underground Neutrino Observatory (JUNO) *Int. J. Mod. Phys. Conf. Ser.* **31** 1460300

[10] Cao Z 2010 A future project at tibet: the large high altitude air shower observatory (LHAASO) *Chin. Phys.* C **34** 249–52
[11] Lu Y, Zhang W Z, Qu J L, Song L M, Gao J, Zhang H M and Ou G 2010 Observation constraints of the hard X-ray modulation telescope HXMT *Science China Physics, Mechanics & Astronomy* **53** Suppl. 1, 31–5
[12] Tsaregorodtsev A 2014 DIRAC distributed computing services *J. Phys: Conf. Series* **513** 032096
[13] Moscicki J T 2009 Ganga: A tool for computational-task management and easy access to grid resources *Comp. Phys. Commun.* **180** 2303–16
[14] Antoniev I *et al* 2012 BESIII and SuperB: distributed job management with Ganga *J.Phys: Conf. Series* **396** 032120
[15] Munoz V M, Albor V F, Diaz R G, Ramo A C, Pena T F, Arevalo G M and Silva J J S 2012 The Integration of CloudStack and OCCI/OpenNebula with DIRAC *J.Phys: Conf. Series* **396** 032075
[16] Nicholson C, Lin L, Deng Z Y, Li W D, Zhang X M and Zheng Y H 2012 File and metadata management for BESIII distributed computing *J.Phys: Conf. Series* **396** 032078
[17] The dCache project. http://www.dcache.org
[18] The Lustre file system. http://lustre.org
[19] Lin T, Zhang X M, Li W D and Deng Z Y 2014 The High-level dataset-based data transfer system in BESDIRAC *J. Phys: Conf. Series* **513** 032059
[20] The rsync project http://rsync.samba.org
[21] The StoRM project http://italiangrid.github.io/storm/