DIRAC Distributed Computing Services

A Tsaregorodtsev
Centre de Physique des Particules de Marseille,
163 Avenue de Luminy Case 902 13288 Marseille, France

On behalf of the DIRAC Project[1]

Abstract. DIRAC Project provides a general-purpose framework for building distributed computing systems. It is used now in several HEP and astrophysics experiments as well as for user communities in other scientific domains. There is a large interest from smaller user communities to have a simple tool like DIRAC for accessing grid and other types of distributed computing resources. However, small experiments cannot afford to install and maintain dedicated services. Therefore, several grid infrastructure projects are providing DIRAC services for their respective user communities. These services are used for user tutorials as well as to help porting the applications to the grid for a practical day-to-day work. The services are giving access typically to several grid infrastructures as well as to standalone computing clusters accessible by the target user communities. In the paper we will present the experience of running DIRAC services provided by the France-Grilles NGI and other national grid infrastructure projects.

1. Introduction

DIRAC Project started in 2003 in order to develop a set of tools to manage large amounts of data for the LHCb experiment at CERN, Geneva. The main goal was to build a system for managing distributed computing resources available to the LHCb Collaboration [1]. The necessity for developing another software layer on top of the standard grid middleware was due to the general instability and lack of high-level features in the grid systems at that time. To cope with these shortcomings, DIRAC introduced a novel Workload Management System based on the *pilot job* concept, which allowed to considerably increase usage efficiency of the grid. Other systems developed for LHCb included Transformation System for automated data-driven workflow management which formed the basis of the LHCb data production activities; Data Management System for automated data replication with integrity checking needed to deal with the large volumes of the LHCb data; Accounting System for detailed reporting of all the LHCb computing activities, and several others. All the DIRAC systems were developed in a single framework with the same secure client/server protocols, authorization policy rules, Configuration System for service discovery, etc. Altogether this allowed building the whole LHCb Data Production System within the same software development environment (Figure 1). This

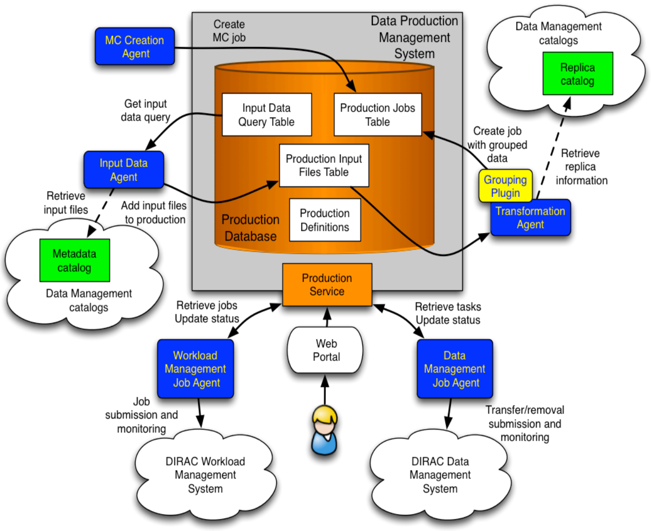
MC Creation Agent
Create MC job
Data Production Management System
Data Management catalogs
Replica catalog
Get input data query
Input Data Query Table
Production Jobs Table
Input Data Agent
Production Input Files Table
Create job with grouped data
Retrieve replica information
Grouping Plugin
Transformation Agent
Retrieve input files
Add input files to production
Production Database
Production Definitions
Metadata catalog
Data Management catalogs
Production Service
Retrieve jobs Update status
Retrieve tasks Update status
Web Portal
Workload Management Job Agent
Data Management Job Agent
Job submission and monitoring
Transfer/removal submission and monitoring
DIRAC Workload Management System
DIRAC Data Management System

Figure 1. LHCb Data Production System built in the DIRAC Framework

[1] *http://diracgird.org*

minimized the development and maintenance efforts making the software more robust and easily extendable [2].

The DIRAC based LHCb Data Production System has been successful in timely processing of the LHC data in the first years of its operations. On the other hand the base software tools were developed using a general purpose framework. Therefore, it was natural to offer these software and tools to other user communities than LHCb, in High Energy Physics (HEP) and other scientific domains. This allowed transferring the rich experience of exploiting complex computing system to other applications needing to manage large amounts of computing resources. An open-source DDIRAC Project was established to develop and disseminate the software for building general purpose distributed computing systems [3].

At present, the LHCb Collaboration still remains the largest community using the DIRAC software [4]. Daily LHCb activities show up to 50K simultaneously running jobs at more than hundred sites. The sites can be as different as usual grid sites, standalone clusters like the one provided by the Yandex Russian internet company or the LHCb on-line farm – a non-grid cluster specialized for the fast events filtering in the LHCb experiment Data Acquisition System. All these resources are integrated in one coherent system viewed by the users as a single large computing facility. The LHCb/DIRAC group of developers created a number of extensions to the core DIRAC software to deal with specific LHCb workflows and data. This is done by using a plug-in mechanism introduced into DIRAC for easy customization for the needs of a particular user community or application domain.

Other experiments also chose DIRAC as the basis for the data production systems. In the HEP domain, the first collaboration to adopt DIRAC was the Belle II [5] experiment at KEK, Japan. The computing model of the Belle II experiment included the requirement of having transparent access to various types of resources including the computing center at KEK, grid resources as well as the possibility to use cloud resources, and in particular the commercial clouds. Therefore, the choice of DIRAC after a thorough evaluation of the system was largely determined by its ability to aggregate heterogeneous resources in a simple way. It is important to mention that some special developments were done in the DIRAC Project to meet the requirements of the Belle II Collaboration. They have already started to test the prototype of the Data Production System in the new framework with two production runs performed in 2013 (Figure2) [6].

Among other HEP experiments using DIRAC we can mention the ILC/CLIC [7] Collaboration and the BES III experiment [8] in IHEP, China. Each of these experiments uses DIRAC in their specific way enriching the overall DIRAC user community with their experience, software contributions and general feedback, which allows developers to improve the quality of the software. Some new DIRAC systems were developments to the needs of these experiments. For example, the Belle Collaboration initiated the development of the Virtual Machine DIRAC scheduler (VMDIRAC) [9], which allows using various cloud computing resources. The DIRAC File Catalog service [10] was developed to meet the requirements of the ILC/CLIC Collaboration to have a robust bookkeeping of its data. Later the DIRAC File Catalog was adopted also by the BES III experiment computing system.

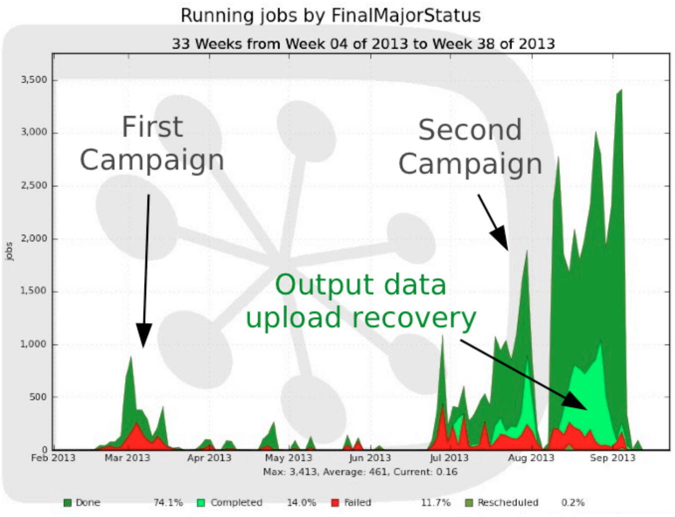

Figure 2. Jobs execution in the Belle II Data Production System

DIRAC is used now also by several experiments in the astrophysics domain (CTA, Fermi/LAT, Glast) [11] and few more started evaluation of DIRAC for their use cases (LSST, Auger).

The simplicity of DIRAC usage for the end-users made it a suitable tool for training programs introducing the concept of grid computing. A number of tutorials run by the DIRAC experts for users in different application domains showed a high interest in the system. However, it quickly became evident that small user groups are not able to install, configure and maintain such complex systems as DIRAC. In many cases, the level of general computing expertise of such groups is low to be able to run DIRAC services in a sustainable way. Therefore, it turned out to be an interesting opportunity for a large grid infrastructure project to provide DIRAC services targeting multiple, generally small user communities coming from various scientific domains. This provides them with an easy access to the grid and other computing resources without the need to understand and operate complex software and configurations. In the following we present the experience of multi-community DIRAC services provided by the France-Grilles National Grid Initiative (NGI) project.

2. DIRAC As A Service

In France by 2011 there were several DIRAC service installations used by different scientific or regional communities. There was also a DIRAC instance maintained by the France-Grilles NGI as part of its training and dissemination program. This allowed several teams of experts in different universities to gain experience with installation and operation of DIRAC services. However, the combined maintenance effort for multiple DIRAC service instances was quite high. Therefore, it was proposed to integrate independent DIRAC installations into a single national service to optimize operational costs. The responsibilities of different partners of the project were distributed as follows. The France-Grilles NGI (FG) ensures the overall coordination of the project. The IN2P3 Computing Centre (CC/IN2P3) hosts the service providing the necessary hardware and manpower. The service is operated by a distributed team of experts from several laboratories and universities participating to the project (Figure 3). After an initial preproduction period of 6 months during which all the necessary software and hardware were commissioned and tested and management procedures were defined, the France-Grilles DIRAC service (FG-DIRAC) was put into production in May 2012 [13].

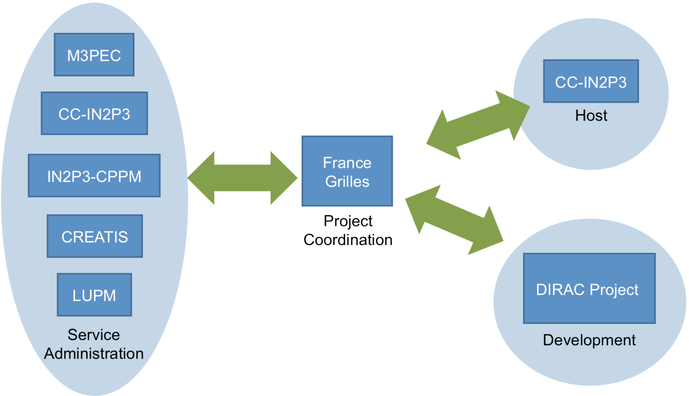

Figure 3. France-Grilles DIRAC National Service responsibilities

From the start, FG-DIRAC was conceived for the use by multiple user communities. After the first year in production, 15 different Virtual Organizations (VO) are supported by the service. The most important user communities are: *astro, biomed, esr, euasia, prod.vo.eu-eela.eu, vo.formation.idgrilles.fr, vo.france-asia.org, vo.france-grilles.fr*. Several VOs started to use the FG-DIRAC service in order to evaluate the properties of the DIRAC system but now they operate dedicated services. For example, VOs corresponding to large astrophysics experiments CTA and Glast, *vo.cta.in2p3.fr* and *glast.org* respectively, now continue with their own DIRAC instances.

Several other large experiments are now evaluating the system performance. Currently, there are more than 100 registered users in the FG-DIRAC service, some of them representing so called robots behaving themselves on behalf of large user groups.

The FG-DIRAC service hosted at CC/IN2P3 has the following hardware configuration:

- 6 virtual servers (on 3 physical hosts), each with 8 cores, 10 GB RAM, 20GB disk space
- 1 TB of disk space NFS mounted, shared by all the servers
- MySQL database server provided by CC/IN2P3 as an independent service

The servers are running the following services:

- *ccdirac01* – security sensitive services, Configuration Master service
- *ccdirac02* – Workload Management System services
- *ccdirac03* – Data Management System services
- *ccdirac04* – StorageElement, Accounting, Monitoring services
- *ccdirac05* – Web Portal (http://dirac.france-grilles.fr)
- *ccdirac06* – REST interface service

The overall FG-DIRAC service is complemented by several redundant supporting services to enhance the system reliability, e.g. Configuration System slave services, located in the partner universities.

During the first year of running, users of the FG-DIRAC service executed more than 5 millions jobs with the biomed community being the most active one (Figure 4). Almost half of the biomed grid activities during this year were carried out with the help of the FG-DIRAC service (Figure 5).

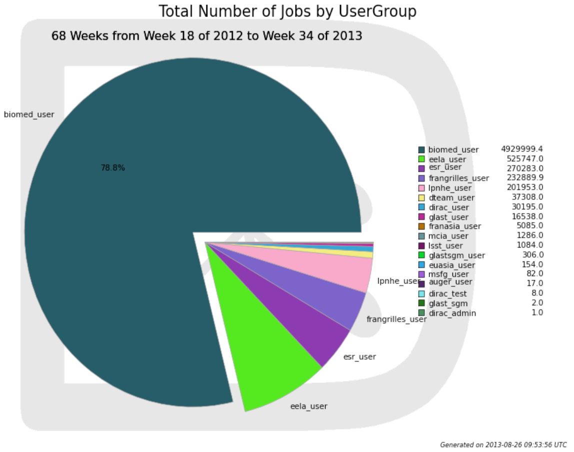

Figure 4. User jobs in the first year of the FG-DIRAC

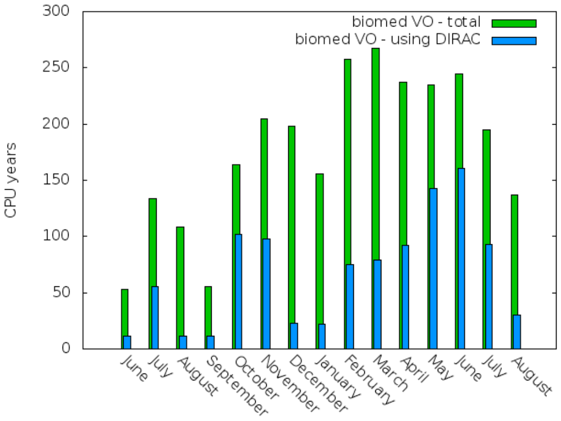

Figure 5. FG-DIRAC fraction of biomed community grid

The FG-DIRAC service saw up to 10K simultaneously running jobs on about 100 sites. If even this is not yet reaching the level of the LHC experiments, it is gradually approaching this scale.

Following the experience gained with the FG-DIRAC, similar services are now being deployed in several other countries. DIRAC services are operated by the IberGRID (Spain and Portugal), IGI (Italy), CNGrid (China). Prototype installations are available for GridPP (UK), VNGrid (Vietnam), and in some other grid infrastructures. It is important that all these projects are actively collaborating, sharing experience and getting help from the DIRAC user community. The feedback from multiple DIRAC service administrators helped to improve the system stability and enhance monitoring and management tools.

3. Resources and services available via FG-DIRAC

3.1. Resources

DIRAC was primarily developed in order to facilitate access to the grid computing resources. Therefore, access to grid sites was provided in the first place. Grid infrastructures based on gLite and ARC middleware are both supported. Grids based on other middleware types, e.g. UNICORE, GOS, can be easily incorporated if there will be interested users.

Sites that are not part of any grid infrastructure but still having computing resources that can be contributed to the interested user communities can be easily incorporated. For those sites it is enough to define a local user account available for login via *ssh* or *gsissh* mechanism and having the right to submit jobs to the local batch system (LRMS). The DIRAC Workload Management System (WMS) will use this user account through an *ssh tunnel* to deploy pilot jobs to the local LRMS worker nodes. Once deployed, pilots will integrate into the standard DIRAC WMS infrastructure [14]. Various LRMS types are supported: Torque/PBS, LSF, Condor, SGE, SLURM, OAR. The plug-in architecture of DIRAC makes it easy to add new LRMS types as needed. It is also possible to incorporate sites with just a set of worker hosts not arranged into any LRMS system. In this case each host will be accessed directly by the DIRAC WMS via its public IP address.

Since recently, a new type of computing resources, computing Clouds, are getting a lot of attention because of their flexibility and ease of management compared to traditional computing clusters. Although the cloud management tools are not yet stable and are under intensive development, DIRAC provides means to incorporate sites with various Cloud management systems: CloudStack, OpenStack, OpenNebula, Amazon EC2, and others [9]. The VMDIRAC project provides an intelligent virtual machine (VM) scheduler, which takes into account the status of the central Task Queue as well as the cost of the VM deployment. Once the VM is deployed it runs the same Job Agent as in all the pilot jobs on any other type of worker nodes and therefore it seamlessly incorporates into the DIRAC WMS making its access transparent to the users (Figure 6).

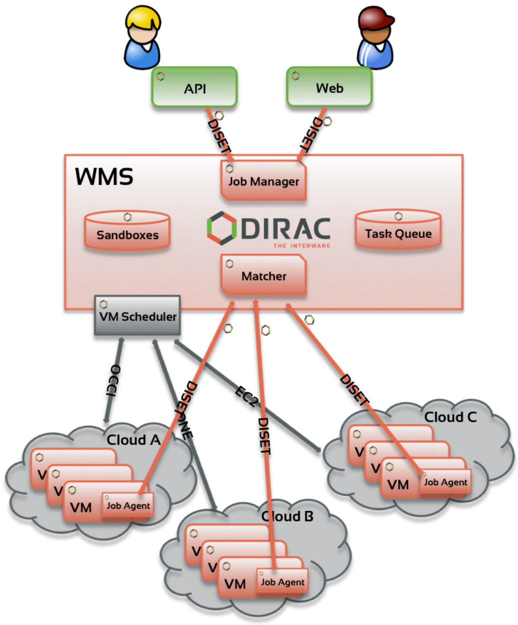

Figure 6. Integration of Cloud resources in the DIRAC Workload Management System

Another type of resources, which is becoming more and more popular, are volunteer grids based on the BOINC technology with virtualization of the volunteer nodes [15]. DIRAC provides access to volunteer grids making full use of the tools developed by the BOINC Project. Within the FG-DIRAC services there is a dedicated installation of the BOINC server. Volunteer clients have to install the standard BOINC client and the VirtualBox hypervisor software, configure the client to work with the FG-DIRAC BOINC server and allocate resources on the local computer. A special BOINC application is dispatched to the client and performs the following operations: downloads and caches the VM image; starts the VM in the VirtualBox hypervisor; customizes the environment of the VM. Once the VM started, it runs a standard pilot job, which is shipped in the input sandbox of the application. From this point on, the VM becomes a standard worker node of the DIRAC WMS.

Recently, the FG-DIRAC service got access also to the volunteer grid operated by the EDGI project [16]. This project provides a specialized service with the CREAM Computing Element (CE) interface. Submitting pilot jobs to this CE allows to fully incorporate the EDGI grid into the FG-DIRAC pool of resources.

The apparent simplicity of incorporation of new computing resources is due to the DIRAC Workload Management scheduling model where workload execution and the computing resource reservation are separated. The worker nodes in grids, clouds or clusters are first reserved by sending pilot jobs according to the requirements of user payloads in the central Task Queue. The Job Agents started in the pilot jobs are contacting the central Workload Management server to pick up user jobs most suitable for the given worker node properties and with a highest priority as defined by community policies. This late binding of payloads to resources is a key aspect of DIRAC flexibility and efficiency in the usage of different computing resources.

As for the storage resources, the FG-DIRAC service provides access to all standard grid Storage Elements (SE) with the SRM2 service interface, which are accessible by the FG-DIRAC users. In addition, it provides a DIRAC SE service with a relatively small capacity of 1TB for the user files.

3.2. Services

The FG-DIRAC Project provides all the basic DIRAC services sufficient to run most of the standard user tasks. The DIRAC WMS services allow to run, monitor and account user jobs with some advanced features enabled like bulk job submission.

The jobs have access to the user data stored in any of the grid SEs with the replica information provided by the one of the LCG File Catalogs (LFC), de facto standard grid file catalog service, or by the DIRAC File Catalog (DFC) service. The latter has similar functionalities as LFC but better performance and some advanced features with respect to the LFC service [10]. Dedicated DFC services can be defined for each user community if necessary.

The FG-DIRAC provides a Web Portal for users to submit and monitor their jobs as well as to retrieve their results. A special service provides a REST interface to selected DIRAC functionalities supporting OAUTH2 based authorization mechanism [18]. This service allows a language neutral access to DIRAC from environments like specialized application portals. Application portals are becoming more and more popular now, especially because they allow users to concentrate on their main work without the need to interact with the complex computing infrastructure directly.

As necessary more advanced services can be made available in the FG-DIRAC installation for a general-purpose usage. We can mention the following services that can be of interest for potential users:

- Support for the MPI jobs. This is done by means of a dedicated service in DIRAC to orchestrate a combined work of pilot jobs forming the MPI rings of workers [19]
- Transformation System for automated bulk submission of user jobs triggered by various events, e.g. by the new data registrations
- Data Replication service for automated data movement

In addition to the services provided by the DIRAC Project, the FG-DIRAC installation can also host the community specific services built in the DIRAC framework and provided as standard extensions.

4. Conclusions and outlook

The first year of operations of the FG-DIRAC Project in France demonstrated a high interest of various user communities in such services. They can reduce considerably the threshold for the access to grid and other distributed computing resources. This is well illustrated by the fact that other national DIRAC services started to appear providing similar functionalities to their users. The usage of the FG-DIRAC and other similar services is constantly increasing with many user communities starting to evaluate their functionalities.

The first experience with running a multi-VO DIRAC service showed quite a number of areas where further development is necessary. In particular, this concerns various tools to help managing the services, improve and simplify their installation and performance monitoring, quickly spot occasional problems. With more users involved, administrators of the FG-DIRAC service need specialized interfaces for managing the DIRAC Registry – the database of the service users, groups and VOs. A special attention should be given to the tools for management the computing resources available through the DIRAC service. This concerns the description of the resources with a clear attribution of the available computing and storage elements to different VOs. The resource status monitoring is also a very laborious activity, which should be automated as much as possible. This is in the focus of the newly developed Resource Status System (RSS) within the DIRAC Project. The RSS is now in use in the LHCb Collaboration and will be soon deployed also in the FG-DIRAC service as well.

5. Acknowledgments

The author would like to point out that the results presented in this paper is a joint work of a large number of developers of the DIRAC Project, managers of various community DIRAC installations as well as administrators of the multi-VO services.

References

[1] Tsaregorodtsev A et al 2008 DIRAC: a community grid solution *J. Phys.: Conf. Ser.* 119 062048

[2] Tsaregorodtsev A et al 2012 Status of the DIRAC Project *J. Phys.: Conf. Ser.* 396 032107

[3] DIRAC Project - http://diracgrid.org

[4] Stagni F and Charpentier Ph 2012 The LHCb DIRAC-based production and data management operations systems *J. Phys.: Conf. Ser.* 368 012010

[5] Graziani Dias R et al 2011 Belle-DIRAC Setup for Using Amazon Elastic Compute Cloud *J. of Grid Computing* **9**(1) 65-79

[6] Kuhr T, Hara T, Miyake H, Sevior M 2013 First Production with the Belle II Data Production System *Proceedings of the CHEP 2013 International Conference, Amsterdam*

[7] ILC Collaboration - http://www.linearcollider.org

[8[BES III Collaboration - http://bes.ihep.ac.cn/bes3

[9] Méndez Muñoz V et al 2012 The Integration of CloudStack and OCCI/OpenNebula with DIRAC *J. Phys.: Conf. Ser.* **396** 032075

[10] Poss S and Tsaregorodtsev A 2012 DIRAC File Replica and Metadata Catalog *J. Phys.: Conf. Ser.* 396 032108

[12] Arrabito L et al 2013 DIRAC framework evaluation for the Fermi-LAT and CTA experiment *Proceedings of the CHEP 2013 International Conference, Amsterdam*

[13] Arrabito L et al 2012 Instance nationale et multi-communauté de DIRAC pour France Grilles, Journées Scientifiques Mésocentres et France Grilles, oai:hal.archives-ouvertes.fr:hal-00766084

[14] Casajus Ramo A, Graciani Dias R, Paterson S, Tsaregorodtsev A 2010 DIRAC pilot framework and the DIRAC Workload Management System *J. Phys.: Conf. Ser.* **219** 062049

[16] BOINC Project - http://boinc.berkeley.edu

[17] EDGI Project - http://edgi-project.eu

[18] Casajus Ramo A et al 2012 DIRAC RESTful API *J. Phys.: Conf. Ser.* **396** 052019

[19] Tsaregorodtsev A and Hamar V 2012 MPI support in the DIRAC Pilot Job Workload Management System *J. Phys.: Conf. Ser.* **396** 032109