

Managing Data

*IHEP, Beijing,
13-15 November 2013*



Data management components

Storage Elements

File Catalogs

DIRAC conventions for user data

Data operation commands

Data bookkeeping with the File Catalog CLI

Replica Catalog

Metadata Catalog

Storage Elements

gLite/EGI Storage Elements

Standard SRM interface

Gridftp protocol

Need Globus libraries, limited number of platforms

Allow third party transfers between them

Managed by the site managers within EGI SLAs

DIRAC Storage Elements

DISET based components

DIPS (Dirac Secure Protocol)

Does not allow third party transfers

Replication through local cache

Third party transfers will be available in the future

▶ File Catalogs

LCG File Catalog (LFC)

Part of the EGI middleware

DIRAC File Catalog

- ▶ DASET based components
- ▶ Part of the DIRAC set of services
 - Community service
 - MySQL backend

Client tools: command line, CLI, Python API

Support of the User Metadata

For DIRAC users the use of any Storage Element or File Catalog is transparent

Community choice which components to use

Different SE types can be mixed together

Several File Catalogs can be used in parallel

- ▶ Complementary functionality
- ▶ Redundancy

Users see depending on the DIRAC Configuration

Logical Storage Elements

e.g. DIRAC-USER, IN2P3-disk

Logical File Catalog

Each file is identified by its Logical File Name (LFN)

Primary unique identifier

GUIDs are supported but their uniqueness is under the responsibility of user applications

This is different from LFC

Mostly for support of some applications, e.g. ROOT I/O

LFN construction

Starts always with the VO name

▶ /vo.france-asia.org/...

User data

/vo.france-asia.org/user/a/atsareg/...

PFN (Physical File Name) construction

Always contains LFN as its trailing part

- ▶ **dirac-dms-add-file**
 - ▶ Upload file to the grid SE (lcg-cr)
- ▶ **dirac-dms-get-file**
 - ▶ Download file to the grid SE (lcg-cp)
- ▶ **dirac-dms-replicate-lfn**
 - ▶ Make another replica of a file (lcg-rep)
- ▶ **dirac-dms-lfn-replicas**
 - ▶ List replicas of a given file (lcg-lr)
- ▶ **dirac-dms-user-lfns**
 - ▶ Get a list of all the user files
- ▶ Plus others ...
 - ▶ See tutorial materials

- ▶ Specialized shell with common commands collected together with a “file system” look-n-feel
 - ▶ Namespace browsing: cd, ls
 - ▶ Finding info: size, meta get
 - ▶ Data operations: add, get, replicate, rm
 - ▶ Metadata operations, meta (set,get,show), find

- ▶ File Catalog operations are generally synchronous
 - ▶ Quick, can wait for the prompt
- ▶ Physical data operations can take very long time
 - ▶ And even fail in the end
- ▶ For example, consider removing data:
 - ▶ Delete replicas on all the SEs
 - ▶ Delete files (lfns)
 - ▶ Delete directories (recursively)
- ▶ Long operations are performed asynchronously
 - ▶ Do not wait for completion
 - ▶ Make sure the operation is accomplished despite possible problems

Tutorial page

<https://github.com/DIRACGrid/DIRAC/wiki/DataManagement>

<https://github.com/DIRACGrid/DIRAC/wiki/FileCatalog>

With DIRAC command line tools

- ▶ Getting data files to the grid
- ▶ Downloading data files from the grid
- ▶ Replicating files
- ▶ Exploring the File Catalog console

Metadata can be associated with each directory as key:value pairs to describe its contents

Int, Float, String, DateTime value types

Some metadata variables can be declared indices

Those can be used for data selections

Subdirectories are inheriting the metadata of their parents

Data selection with metadata queries

Example:

```
find /dirac/user Meta1=Value1 Meta2>3 Meta2<5 Meta3=2,3,4
```

File metadata can also be defined

The functionality is similar to the AMGA gLite service

The internal structure is very different

Different scalability properties

BES Collaboration (IHEP, Beijing) performed an extensive comparison of DFC vs AMGA

Similar performance

▶ DFC is chosen for their Computing Model

Some features of DFC

Support for the data provenance information

Ancestor<->descendent relationships

Support for efficient storage usage reports

Real time

Necessary for the storage quota policies

Tutorial

<https://github.com/DIRACGrid/DIRAC/wiki/DataManagementAdvanced>

With File Catalog CLI:

- Upload several files in several directories
- Define directory metatags with values
- Define file metatags
- Find files by metadata