



# WORLDWIDE LHC COMPUTING GRID

## GLITE 3 USER GUIDE

### MANUALS SERIES

---

<b>Document identifier:</b>	<b>CERN-LCG-GDEIS-722398</b>
<b>EDMS id:</b>	<b>722398</b>
<b>Version:</b>	<b>1.1</b>
<b>Date:</b>	<b>January 17, 2007</b>
<b>Section:</b>	<b>Experiment Integration and Distributed Analysis</b>
<b>Document status:</b>	<b>PUBLIC</b>
<b>Author(s):</b>	Stephen Burke, Simone Campana, Antonio Delgado Peris, Flavia Donno, Patricia Méndez Lorenzo, Roberto Santinelli, Andrea Sciabà
<b>File:</b>	<b>gLite-3-UserGuide</b>

---

*Abstract: This guide is an introduction to the WLCG/EGEE Grid and to the gLite 3 middleware from a user's point of view.*

---

## Document Change Record

Issue	Item	Reason for Change
17/01/07	v1.1	Revised version
20/04/06	v1.0	First draft

## Files

Software Products	User files
PDF	<a href="https://edms.cern.ch/file/722398/1.1/gLite-3-UserGuide.pdf">https://edms.cern.ch/file/722398/1.1/gLite-3-UserGuide.pdf</a>
PS	<a href="https://edms.cern.ch/file/722398/1.1/gLite-3-UserGuide.ps">https://edms.cern.ch/file/722398/1.1/gLite-3-UserGuide.ps</a>
HTML	<a href="https://edms.cern.ch/file/722398/1.1/gLite-3-UserGuide.html">https://edms.cern.ch/file/722398/1.1/gLite-3-UserGuide.html</a>

## CONTENTS

<b>1</b>	<b>INTRODUCTION.....</b>	<b>10</b>
1.1	ACKNOWLEDGMENTS .....	10
1.2	OBJECTIVES OF THIS DOCUMENT .....	10
1.3	APPLICATION AREA.....	10
1.4	DOCUMENT EVOLUTION PROCEDURE .....	10
1.5	REFERENCE AND APPLICABLE DOCUMENTS .....	10
1.6	TERMINOLOGY .....	14
1.6.1	Glossary .....	15
<b>2</b>	<b>EXECUTIVE SUMMARY.....</b>	<b>18</b>
<b>3</b>	<b>OVERVIEW .....</b>	<b>19</b>
3.1	PRELIMINARY MATTERS.....	20
3.1.1	Code Development.....	20
3.1.2	Troubleshooting .....	20
3.1.3	User and VO utilities .....	20
3.2	THE WLCG/EGEE INFRASTRUCTURE.....	21
3.3	THE WLCG/EGEE ARCHITECTURE.....	21
3.3.1	Security.....	21
3.3.2	User Interface .....	22
3.3.3	Computing Element .....	23
3.3.4	Storage Element .....	23
3.3.5	Information Service .....	24

3.3.6	Data Management .....	28
3.3.7	Workload Management .....	30
3.4	JOB FLOW .....	30
3.4.1	Job Submission .....	30
3.4.2	Other Operations .....	32
<b>4</b>	<b>GRID SECURITY AND GETTING STARTED .....</b>	<b>33</b>
4.1	BASIC SECURITY CONCEPTS .....	33
4.1.1	Private and Public Keys .....	33
4.1.2	Encryption .....	33
4.1.3	Signing .....	33
4.1.4	Certificates .....	34
4.1.5	Certification Authorities .....	34
4.1.6	Proxies .....	34
4.1.7	VOMS Proxies .....	35
4.2	FIRST STEPS .....	35
4.3	OBTAINING A CERTIFICATE .....	36
4.3.1	X.509 Certificates .....	36
4.3.2	Requesting the Certificate .....	36
4.3.3	Getting the Certificate .....	37
4.3.4	Renewing the Certificate .....	38
4.3.5	Taking Care of Private Keys .....	38
4.4	REGISTERING WITH WLCG/EGEE .....	39
4.4.1	The Registration Service .....	39
4.4.2	Virtual Organisations .....	40

4.5	SETTING UP THE USER ACCOUNT.....	40
4.5.1	The User Interface .....	40
4.5.2	Checking a Certificate .....	41
4.6	PROXIES.....	44
4.6.1	Standard Proxies.....	44
4.6.2	VOMS Proxies .....	46
4.6.3	Proxy Renewal .....	48
<b>5</b>	<b>INFORMATION SERVICE.....</b>	<b>51</b>
5.1	THE MDS .....	51
5.1.1	lcg-infosites .....	51
5.1.2	lcg-info .....	54
5.1.3	The Local GRIS .....	56
5.1.4	Using the ldapsearch command to read the MDS.....	56
5.1.5	The Site GIIS/BDII .....	59
5.1.6	The top-level BDII .....	61
5.2	R-GMA .....	64
5.2.1	R-GMA concepts .....	65
5.2.2	The R-GMA Browser .....	65
5.2.3	The R-GMA CLI .....	66
5.2.4	R-GMA APIs.....	69
5.3	SERVICEDISCOVERY .....	69
5.3.1	Running a Service Discovery query .....	70
5.4	MONITORING .....	71
5.4.1	GridICE .....	71

<b>6</b>	<b>WORKLOAD MANAGEMENT .....</b>	<b>73</b>
6.1	INTRODUCTION .....	73
6.2	THE JOB DESCRIPTION LANGUAGE .....	73
6.3	THE COMMAND LINE INTERFACE .....	79
6.3.1	Single Job Submission.....	80
6.3.2	Job Operations.....	84
6.3.3	Advanced Sandbox Management.....	90
6.3.4	Real Time Output Retrieval .....	91
6.3.5	The BrokerInfo .....	94
6.4	ADVANCED JOB TYPES .....	95
6.4.1	Job Collections .....	95
6.4.2	Checkpointable Jobs .....	98
6.4.3	DAG jobs .....	98
6.4.4	Partitionable jobs .....	98
6.4.5	Parametric jobs .....	99
6.4.6	Interactive Jobs .....	99
6.4.7	MPI Jobs .....	100
6.5	COMMAND LINE INTERFACE CONFIGURATION .....	101
6.5.1	WMProxy Configuration.....	101
6.5.2	gLite Network Server Configuration.....	102
6.5.3	LCG-2 Network Server Configuration .....	104
<b>7</b>	<b>DATA MANAGEMENT.....</b>	<b>107</b>
7.1	INTRODUCTION .....	107

7.2	STORAGE ELEMENTS .....	107
7.2.1	Data Channel Protocols .....	107
7.2.2	Types of Storage Elements .....	107
7.2.3	The Storage Resource Manager interface.....	108
7.3	FILE NAMES IN GLITE 3 .....	108
7.4	FILE CATALOGUE IN GLITE 3 .....	109
7.4.1	LFC Commands .....	110
7.4.2	Access Control Lists .....	113
7.5	FILE AND REPLICA MANAGEMENT CLIENT TOOLS .....	115
7.5.1	LCG Data Management Client Tools .....	115
7.6	FILE TRANSFER SERVICE.....	122
7.6.1	Basic Concepts .....	122
7.6.2	Transfer job states.....	123
7.6.3	Individual file states .....	123
7.6.4	FTS Commands .....	124
7.7	LOW LEVEL DATA MANAGEMENT TOOLS.....	127
7.7.1	GSIFTP.....	128
7.7.2	CASTOR and RFIO .....	129
7.7.3	dCache and DCAP .....	130
7.8	JOB SERVICES AND DATA MANAGEMENT .....	130
<b>A</b>	<b>THE GRID MIDDLEWARE.....</b>	<b>133</b>
<b>B</b>	<b>ENVIRONMENT VARIABLES AND CONFIGURATION FILES .....</b>	<b>135</b>
<b>C</b>	<b>JOB STATUS DEFINITION.....</b>	<b>137</b>

<b>D</b>	<b>USER TOOLS .....</b>	<b>139</b>
D.1	INTRODUCTION .....	139
D.2	JOB MANAGEMENT FRAMEWORK .....	139
D.3	JOB MONITORING.....	139
D.4	JOB STATUS MONITORING .....	140
D.5	TIME LEFT UTILITY.....	140
<b>E</b>	<b>VO-WIDE UTILITIES.....</b>	<b>142</b>
E.1	INTRODUCTION .....	142
E.2	FREEDOM OF CHOICE FOR RESOURCES.....	142
E.3	SERVICE AVAILABILITY MONITORING.....	142
E.4	THE VO BOX .....	142
E.5	VO SOFTWARE INSTALLATION.....	143
<b>F</b>	<b>DATA MANAGEMENT AND FILE ACCESS THROUGH AN APPLICATION PRO- GRAMMING INTERFACE .....</b>	<b>144</b>
<b>G</b>	<b>THE GLUE SCHEMA.....</b>	<b>156</b>
G.1	BASIC CONCEPTS .....	156
G.2	MAPPINGS.....	157
G.3	INFORMATION PROVIDERS.....	157
G.4	GLUE ATTRIBUTES .....	157
G.4.1	Site information .....	158
G.4.2	Service information .....	158
G.4.3	Attributes for the Computing Element .....	159



---

G.4.4	Attributes for the Storage Element .....	163
G.4.5	Attributes for the CE-SE Binding .....	165

# 1. INTRODUCTION

## 1.1. ACKNOWLEDGMENTS

This work received support from the following institutions:

- Istituto Nazionale di Fisica Nucleare, Roma, Italy.
- Ministerio de Educación y Ciencia, Madrid, Spain.
- Particle Physics and Astronomy Research Council, UK.

## 1.2. OBJECTIVES OF THIS DOCUMENT

This document gives an overview of the gLite 3 middleware. It helps users to understand the building blocks of the Grid and the available interfaces to the Grid services in order to run jobs and manage data.

This document is neither an administration nor a developer guide.

## 1.3. APPLICATION AREA

This guide is addressed to WLCG/EGEE users and site administrators who would like to work with the gLite 3 middleware.

## 1.4. DOCUMENT EVOLUTION PROCEDURE

The guide reflects the current status of the gLite middleware, and will be modified as new gLite releases are produced. In some parts of the document, references to the foreseeable future of the gLite software are made.

## 1.5. REFERENCE AND APPLICABLE DOCUMENTS

# REFERENCES

- [1] Glossaries of Grid terms  
<http://www.gridpp.ac.uk/gas/>

- <http://egee-jra2.web.cern.ch/EGEE-JRA2/Glossary/Glossary.html>  
<http://grid-it.cnaf.infn.it/fileadmin/users/dictionary/dictionary.html>
- [2] EGEE – Enabling Grids for E-science  
<http://eu-egee.org/>
  - [3] gLite – Lightweight Middleware for Grid Computing  
<http://cern.ch/glite/>
  - [4] Worldwide LHC Computing Grid  
<http://cern.ch/LCG/>
  - [5] The DataGrid Project  
<http://www.edg.org/>
  - [6] DataTAG – Research & technological development for a Data TransAtlantic Grid  
<http://cern.ch/datatag/>
  - [7] The Globus Alliance  
<http://www.globus.org/>
  - [8] GriPhyN – Grid Physics Network  
<http://www.griphyn.org/>
  - [9] iVDgL – International Virtual Data Grid Laboratory  
<http://www.ivdgl.org/>
  - [10] Open Science Grid  
<http://www.opensciencegrid.org/>
  - [11] The Virtual Data Toolkit  
<http://vdt.cs.wisc.edu/>
  - [12] NorduGrid  
<http://www.nordugrid.org/>
  - [13] Ian Foster, Carl Kesselman, Steven Tuecke,  
The Anatomy of the Grid: Enabling Scalable Virtual Organizations  
<http://www.globus.org/alliance/publications/papers/anatomy.pdf>
  - [14] M. Dimou,  
LCG User Registration and VO Management  
<https://edms.cern.ch/document/428034/>
  - [15] EGEE CIC Operations Portal  
<http://cic.in2p3.fr/>
  - [16] Global Grid User Support  
<http://www.ggus.org/>

- [17] GOC Database 2.0  
<https://goc.grid-support.ac.uk/gridsite/gocdb2/>
- [18] GOC Monitoring links  
<http://goc.grid-support.ac.uk/gridsite/monitoring/>  
Google map  
<http://goc02.grid-support.ac.uk/googlemaps/sam.html>
- [19] Overview of the Grid Security Infrastructure  
<http://www-unix.globus.org/security/overview.html>
- [20] The Storage Resource Manager  
<http://sdm.lbl.gov/srm-wg/>
- [21] The GLUE schema  
<http://glueschema.forge.cnaf.infn.it/>
- [22] The GLUE LDAP schema  
[http://forge.cnaf.infn.it/plugins/scmsvn/viewcvs.php/v\\_1\\_2/mapping/ldap/schema/openldap-2-1/?root=glueschema](http://forge.cnaf.infn.it/plugins/scmsvn/viewcvs.php/v_1_2/mapping/ldap/schema/openldap-2-1/?root=glueschema)
- [23] MDS 2.2 Features in the Globus Toolkit 2.2 Release  
[http://www.globus.org/toolkit/mds/#mds\\_gt2](http://www.globus.org/toolkit/mds/#mds_gt2)
- [24] R-GMA: Relational Grid Monitoring Architecture  
<http://www.r-gma.org/index.html>
- [25] B. Tierney *et al.*,  
A Grid Monitoring Architecture,  
GGF, 2001 (revised 2002)  
<http://www-didc.lbl.gov/GGF-PERF/GMA-WG/papers/GWD-GP-16-2.pdf>
- [26] S. Campana, M. Litmaath, A. Sciabà,  
LCG-2 Middleware overview  
<https://edms.cern.ch/document/498079/>
- [27] F. Pacini,  
EGEE User's Guide – WMS Service  
<https://edms.cern.ch/document/572489/>
- [28] F. Pacini,  
EGEE User's Guide – WMPProxy service  
<https://edms.cern.ch/document/674643/>
- [29] WP1 Workload Management Software – Administrator and User Guide  
[http://www.infn.it/workload-grid/docs/DataGrid-01-TEN-0118-1\\_2.pdf](http://www.infn.it/workload-grid/docs/DataGrid-01-TEN-0118-1_2.pdf)

- [30] CESNET,  
EGEE User's Guide – Service Logging and Bookkeeping (L&B)  
<https://edms.cern.ch/document/571273/>
- [31] Using lxplus as an LCG-2 User Interface  
<http://grid-deployment.web.cern.ch/grid-deployment/documentation/UI-lxplus/>
- [32] GridICE: a monitoring service for the Grid  
<http://gridice.forge.cnaf.infn.it/>
- [33] Condor Classified Advertisements  
<http://www.cs.wisc.edu/condor/classad>
- [34] The Condor Project  
<http://www.cs.wisc.edu/condor/>
- [35] F. Pacini,  
Job Description Language HowTo  
[http://www.infn.it/workload-grid/docs/DataGrid-01-TEN-0102-0\\_2-Document.pdf](http://www.infn.it/workload-grid/docs/DataGrid-01-TEN-0102-0_2-Document.pdf)
- [36] F. Pacini,  
JDL Attributes  
[http://www.infn.it/workload-grid/docs/DataGrid-01-TEN-0142-0\\_2.pdf](http://www.infn.it/workload-grid/docs/DataGrid-01-TEN-0142-0_2.pdf)
- [37] F. Pacini,  
Job Description Language Attributes Specification for the gLite middleware (submission through Network Server)  
<https://edms.cern.ch/document/555796/1/>
- [38] F. Pacini,  
Job Description Language Attributes Specification for the gLite middleware (submission through WMPProxy service)  
<https://edms.cern.ch/document/590869/1/>
- [39] The EDG-Brokerinfo User Guide  
<http://www.infn.it/workload-grid/docs/edg-brokerinfo-user-guide-v2.2.pdf>
- [40] MPI Wiki page  
<http://grid.ie/mpi/wiki/>
- [41] GSIFTP Tools for the Data Grid  
<http://www.globus.org/toolkit/docs/2.4/datagrid/deliverables/gsiftp-tools.html>
- [42] RFIO: Remote File Input/Output  
<http://doc.in2p3.fr/doc/public/products/rfio/rfio.html>
- [43] CASTOR  
<http://cern.ch/castor/>

- [44] dCache  
<http://www.dCache.org/>
- [45] Scientific Linux  
<http://www.scientificlinux.org/>
- [46] User level tools documentation Wiki  
[http://goc.grid.sinica.edu.tw/gocwiki/User\\_tools](http://goc.grid.sinica.edu.tw/gocwiki/User_tools)
- [47] R. Santinelli, F. Donno,  
Experiment Software Installation  
<https://edms.cern.ch/document/498080/>  
[http://grid-deployment.web.cern.ch/grid-deployment/eis/docs/internal/chep04/SW\\_Installation.pdf](http://grid-deployment.web.cern.ch/grid-deployment/eis/docs/internal/chep04/SW_Installation.pdf)
- [48] GOC Wiki  
<http://goc.grid.sinica.edu.tw/gocwiki/FrontPage>
- [49] Freedom of Choice for Resources  
<https://lcg-fcr.cern.ch:8443/fcr/fcr.cgi>
- [50] Service Availability Monitoring  
[http://goc.grid.sinica.edu.tw/gocwiki/Service\\_Availability\\_Monitoring](http://goc.grid.sinica.edu.tw/gocwiki/Service_Availability_Monitoring)
- [51] Service Availability Monitoring Web Interface  
<https://lcg-sam.cern.ch:8443/sam/sam.py>
- [52] VO box How-to  
[http://goc.grid.sinica.edu.tw/gocwiki/VO-box\\_HowTo](http://goc.grid.sinica.edu.tw/gocwiki/VO-box_HowTo)
- [53] EGEE User's Guide – Service Discovery  
<http://hepunix.rl.ac.uk/egee/jra1-uk/sd/service-discovery.pdf>
- [54] Storage Classes in WLCG  
<http://glueschema.forge.cnaf.infn.it/uploads/Spec/V13/SE-Model-3.5.pdf>
- [55] LCG-2 User Guide  
<https://edms.cern.ch/document/454439/>

## 1.6. TERMINOLOGY

The Grid world has a lot of specialised jargon. Acronyms used in this document are explained below; for more acronyms and definitions see [1].

### 1.6.1. Glossary

<b><i>AFS:</i></b>	Andrew File System
<b><i>API:</i></b>	Application Programming Interface
<b><i>BDII:</i></b>	Berkeley Database Information Index
<b><i>CASTOR</i></b>	CERN Advanced STORage manager
<b><i>CE:</i></b>	Computing Element
<b><i>CERN:</i></b>	European Laboratory for Particle Physics
<b><i>ClassAd:</i></b>	Classified advertisement (Condor)
<b><i>CLI:</i></b>	Command Line Interface
<b><i>CNAF:</i></b>	INFN's National Center for Telematics and Informatics
<b><i>dcap:</i></b>	dCache Access Protocol
<b><i>DIT:</i></b>	Directory Information Tree (LDAP)
<b><i>DLI:</i></b>	Data Location Interface
<b><i>DN:</i></b>	Distinguished Name
<b><i>EDG:</i></b>	European DataGrid
<b><i>EDT:</i></b>	European DataTAG
<b><i>EGEE:</i></b>	Enabling Grids for E-science
<b><i>ESM:</i></b>	Experiment Software Manager
<b><i>FCR:</i></b>	Freedom of Choice for Resources
<b><i>FNAL:</i></b>	Fermi National Accelerator Laboratory
<b><i>FTS:</i></b>	File Transfer Service
<b><i>GFAL:</i></b>	Grid File Access Library
<b><i>GG:</i></b>	Grid Gate (aka gatekeeper)
<b><i>GGF:</i></b>	Global Grid Forum (now called OGF)
<b><i>GGUS:</i></b>	Global Grid User Support
<b><i>GIIS:</i></b>	Grid Index Information Server
<b><i>GLUE:</i></b>	Grid Laboratory for a Uniform Environment
<b><i>GMA:</i></b>	Grid Monitoring Architecture
<b><i>GOC:</i></b>	Grid Operations Centre
<b><i>GRAM:</i></b>	Grid Resource Allocation Manager
<b><i>GRIS:</i></b>	Grid Resource Information Service
<b><i>GSI:</i></b>	Grid Security Infrastructure
<b><i>gsidcap:</i></b>	GSI-enabled version of the dCache Access Protocol
<b><i>gsirfio:</i></b>	GSI-enabled version of the Remote File Input/Output protocol
<b><i>GUI:</i></b>	Graphical User Interface
<b><i>GUID:</i></b>	Grid Unique ID
<b><i>HSM:</i></b>	Hierarchical Storage Manager
<b><i>ID:</i></b>	Identifier
<b><i>INFN:</i></b>	Istituto Nazionale di Fisica Nucleare
<b><i>IS:</i></b>	Information Service
<b><i>JDL:</i></b>	Job Description Language
<b><i>kdcap:</i></b>	Kerberos-enabled version of the dCache Access Protocol

<b>LAN:</b>	Local Area Network
<b>LB:</b>	Logging and Bookkeeping Service
<b>LDAP:</b>	Lightweight Directory Access Protocol
<b>LFC:</b>	LCG File Catalogue
<b>LFN:</b>	Logical File Name
<b>LHC:</b>	Large Hadron Collider
<b>LCG:</b>	LHC Computing Grid
<b>LRC:</b>	Local Replica Catalogue
<b>LRMS:</b>	Local Resource Management System
<b>LSF:</b>	Load Sharing Facility
<b>MDS:</b>	Monitoring and Discovery Service
<b>MPI:</b>	Message Passing Interface
<b>MSS:</b>	Mass Storage System
<b>NS:</b>	Network Server
<b>OGF:</b>	Open Grid Forum (formerly called GGF)
<b>OS:</b>	Operating System
<b>PBS:</b>	Portable Batch System
<b>PFN:</b>	Physical File name
<b>PID:</b>	Process IDentifier
<b>POOL:</b>	Pool of Persistent Objects for LHC
<b>PPS:</b>	Pre-Production Service
<b>RAL:</b>	Rutherford Appleton Laboratory
<b>RB:</b>	Resource Broker
<b>RFIO:</b>	Remote File Input/Output
<b>R-GMA:</b>	Relational Grid Monitoring Archicecture
<b>RLI:</b>	Replica Location Index
<b>RLS:</b>	Replica Location Service
<b>RM:</b>	Replica Manager
<b>RMC:</b>	Replica Metadata Catalogue
<b>RMS:</b>	Replica Management System
<b>ROC:</b>	Regional Operations Centre
<b>ROS:</b>	Replica Optimization Service
<b>SAM:</b>	Service Availability Monitoring
<b>SASL:</b>	Simple Authorization & Security Layer (LDAP)
<b>SE:</b>	Storage Element
<b>SFN:</b>	Site File Name
<b>SMP:</b>	Symmetric Multi Processor
<b>SN:</b>	Subject Name
<b>SRM:</b>	Storage Resource Manager
<b>SURL:</b>	Storage URL
<b>TURL:</b>	Transport URL
<b>UI:</b>	User Interface
<b>URI:</b>	Uniform Resource Identifier
<b>URL:</b>	Uniform Resource Locator



**UUID:** Universal Unique ID  
**VDT:** Virtual Data Toolkit  
**VO:** Virtual Organization  
**WLCG:** Worldwide LHC Computing Grid  
**WMS:** Workload Management System  
**WN:** Worker Node  
**WPn:** Work Package #n

## 2. EXECUTIVE SUMMARY

This user guide is intended for users of the gLite 3 middleware. In these pages, the user will find an introduction to the services provided by the WLCG/EGEE Grid and a description of how to use them. Examples are given of the management of jobs and data, the retrieval of information about resources, and other functionality.

An introduction to the gLite 3 middleware is presented in Chapter 3. This chapter describes all the middleware components and provides most of the necessary terminology. It also presents the WLCG and the EGEE projects, which developed the gLite 3 middleware.

In Chapter 4, the preliminary procedures to follow before starting to use the Grid are described: how to get a certificate, join a Virtual Organisation and manage proxy certificates.

Details on how to get information about the status of Grid resources are given in Chapter 5, where the different information services and monitoring systems are discussed.

An overview of the Workload Management service is given in Chapter 6. This chapter explains the basic commands for job submission and management, as well as those for retrieving information on running and finished jobs.

Data Management services are described in Chapter 7. Not only the high-level interfaces are described, but also commands that can be useful in case of problems or for debugging purposes.

Finally, the appendices give information about the gLite 3 middleware components (Appendix A), the configuration files and environment variables for users (Appendix B), the possible states of a job during submission and execution (Appendix C), user tools for the Grid (Appendix D), VO-wide utilities (Appendix E), APIs for data management and file access (Appendix F), and the GLUE Schema used to describe Grid resources (Appendix G).

### 3. OVERVIEW

The *EGEE project* [2] has a main goal of providing researchers with access to a geographically distributed computing Grid infrastructure, available 24 hours a day. It focuses on maintaining and developing the *gLite* middleware [3] and on operating a large computing infrastructure for the benefit of a vast and diverse research community.

The *Worldwide LHC Computing Grid Project (WLCG)* [4] was created to prepare the computing infrastructure for the simulation, processing and analysis of the data of the *Large Hadron Collider (LHC)* experiments. The LHC, which is being constructed at the European Laboratory for Particle Physics (*CERN*), will be the world's largest and most powerful particle accelerator.

The WLCG and the EGEE projects share a large part of their infrastructure and operate it in conjunction. For this reason, we will refer to it as the *WLCG/EGEE infrastructure*.

The gLite 3 middleware comes from a number of Grid projects, like DataGrid [5], DataTag [6], Globus [7], GriPhyN [8], iVDGL [9], EGEE and LCG. This middleware is currently installed in sites participating in WLCG/EGEE.

In WLCG other Grid infrastructures exist, namely the *Open Science Grid (OSG)* [10], which uses the middleware distributed by VDT [11], and NorduGrid [12], which uses the ARC middleware. These are not covered by this guide.

The case of the LHC experiments illustrates well the motivation behind Grid technology. The LHC accelerator will start operation in 2007, and the experiments that will use it (*ALICE*, *ATLAS*, *CMS* and *LHCb*) will generate enormous amounts of data. The processing of this data will require huge computational and storage resources, and the associated human resources for operation and support. It was not considered feasible to concentrate all the resources at one site, and therefore it was agreed that the LCG computing service would be implemented as a geographically distributed *Computational Data Grid*. This means that the service will use computing and storage resources installed at a large number of computing sites in many different countries, interconnected by fast networks. The gLite middleware hides much of the complexity of this environment from the user, giving the impression that all of these resources are available in a coherent virtual computer centre.

The users of a Grid infrastructure are divided into *Virtual Organisations (VOs)* [13], abstract entities grouping users, institutions and resources in the same administrative domain [14].

The WLCG/EGEE VOs correspond to real organisations or projects, such as the four LHC experiments, the community of biomedical researchers, etc. An updated list of all the EGEE VOs can be found at the CIC portal [15].

## 3.1. PRELIMINARY MATTERS

### 3.1.1. Code Development

Many of the services offered by WLCG/EGEE can be accessed both by the user interfaces provided (CLIs or GUIs), or from applications by making use of various APIs. References to APIs used for particular services will be given later in the sections describing such services.

A totally different matter is the development of software that forms part of the gLite middleware itself. This falls outside the scope of this guide.

### 3.1.2. Troubleshooting

This document will also explain the meaning of the most common error messages and give some advice on how to avoid some common errors. This guide cannot, however, include all the possible failures a user may encounter while using gLite 3. These errors may be produced due to user mistakes, to misconfiguration of the Grid components, to hardware or network failures, or even to bugs in the gLite middleware.

Subsequent sections of this guide provide references to documents which go into greater detail about the gLite 3 components.

The *Global Grid User Support (GGUS)* [16] service provides centralised support for WLCG/EGEE users, by answering questions, tracking known problems, maintaining lists of frequently asked questions, providing links to documentation, etc. The GGUS portal is the key entry point for Grid users looking for help.

Finally, a user who thinks that there is a security risk in the Grid may directly contact the relevant site administrator if the situation is urgent, as this may be faster than going through GGUS. Information on the local site contacts can be obtained from the Information Service or from the GOC database [17], which is described in Chapter 4.

### 3.1.3. User and VO utilities

This guide mainly covers information useful for the average user. Thus, only core gLite 3 middleware is described. Nevertheless, there are several tools which are not part of the middleware, but may be very useful to users. Some of these tools are summarised in Appendix D.

Likewise, there are utilities that are only available to certain (authorised) users of the Grid. An example is the administration of the resources viewed by a VO or the installation of VO software on WLCG/EGEE nodes. Only authorised users can install software on the computing resources of

WLCG/EGEE: the installed software is also published in the Information Service, so that users can select sites where the software they need is installed. Information on such topics is given in Appendix E.

### 3.2. THE WLCG/EGEE INFRASTRUCTURE

WLCG/EGEE operates a production Grid distributed over more than 200 sites around the world, with more than 30,000 CPUs and 20 PB of data storage. The status of the Grid can be seen from the various monitoring pages linked from the Grid Operations Centre (GOC) monitoring page [18]; in particular look at the Google map for a quick overview. Sites can choose which VOs to support and at what level, so users will generally not have access to every site; later chapters describe how to find out which resources are available to a specific user.

Sites vary widely in the size of their computing and storage resources; for WLCG the largest sites are designated as Tier 1 and play a key role in storing and processing data. Sites are organised into geographical regions, co-ordinated by a Regional Operations Centre (ROC).

WLCG/EGEE also runs a smaller Pre-Production Service (PPS), a separate Grid where new versions of the middleware can be tested by both sites and users before being deployed on the main production Grid.

### 3.3. THE WLCG/EGEE ARCHITECTURE

This section provides a quick overview of the WLCG/EGEE architecture and services.

#### 3.3.1. Security

As explained earlier, the WLCG/EGEE user community is grouped into Virtual Organisations. Before WLCG/EGEE resources can be used, a user must read and agree to the WLCG/EGEE usage rules and any further rules for the VO he wishes to join, and register some personal data with a Registration Service.

Once the user registration is complete, he can access WLCG/EGEE. The *Grid Security Infrastructure (GSI)* in WLCG/EGEE enables secure authentication and communication over an open network [19]. GSI is based on public key encryption, X.509 certificates, and the Secure Sockets Layer (SSL) communication protocol, with extensions for single sign-on and delegation.

In order to authenticate himself to Grid resources, a user needs to have a digital X.509 certificate issued by a *Certification Authority (CA)* trusted by WLCG/EGEE; Grid resources are generally also issued with certificates to allow them to authenticate themselves to users and other services.

The user certificate, whose *private key* is protected by a password, is used to generate and sign a temporary certificate, called a *proxy certificate* (or simply a proxy), which is used for the actual authen-

tication to Grid services and does not need a password. As possession of a proxy certificate is a proof of identity, the file containing it must be readable only by the user, and a proxy has, by default, a short lifetime (typically 12 hours) to reduce security risks if it should be stolen.

The authorisation of a user on a specific Grid resource can be done in two different ways. The first is simpler, and relies on the *grid-mapfile* mechanism. The Grid resource has a local grid-mapfile which maps user certificates to local accounts. When a user's request for a service reaches a host, the *Subject Name* of the user (contained in the proxy) is checked against what is in the local grid-mapfile to find out to which local account (if any) the user certificate is mapped, and this account is then used to perform the requested operation [19]. The second way relies on the *Virtual Organisation Membership Service (VOMS)* and the *LCAS/LCMAPS* mechanism, which allow for a more detailed definition of user privileges, and will be explained in more detail later.

A user needs a valid proxy to submit jobs; those jobs carry their own copies of the proxy to be able to authenticate with Grid services as they run. For long-running jobs, the job proxy may expire before the job has finished, causing the job to fail. To avoid this, there is a proxy renewal mechanism to keep the job proxy valid for as long as needed. The *MyProxy server* is the component that provides this functionality.

### 3.3.2. User Interface

The access point to the WLCG/EGEE Grid is the *User Interface (UI)*. This can be any machine where users have a personal account and where their user certificate is installed. From a UI, a user can be authenticated and authorized to use the WLCG/EGEE resources, and can access the functionalities offered by the Information, Workload and Data management systems. It provides CLI tools to perform some basic Grid operations:

- list all the resources suitable to execute a given job;
- submit jobs for execution;
- cancel jobs;
- retrieve the output of finished jobs;
- show the status of submitted jobs;
- retrieve the logging and bookkeeping information of jobs;
- copy, replicate and delete files from the Grid;
- retrieve the status of different resources from the Information System.

In addition, the WLCG/EGEE APIs are also available on the UI to allow development of Grid-enabled applications.

### 3.3.3. Computing Element

A *Computing Element (CE)*, in Grid terminology, is some set of computing resources localized at a site (i.e. a cluster, a computing farm). A CE includes a *Grid Gate (GG)*<sup>1</sup>, which acts as a generic interface to the cluster; a *Local Resource Management System (LRMS)* (sometimes called *batch system*), and the cluster itself, a collection of *Worker Nodes (WNs)*, the nodes where the jobs are run.

There are two GG implementations in gLite 3: the *LCG CE*, developed by EDG and used in LCG-2<sup>2</sup>, and the *gLite CE*, developed by EGEE. Sites can choose what to install, and some of them provide both types. The GG is responsible for accepting jobs and dispatching them for execution on the WNs via the LRMS.

In gLite 3 the supported LRMS types are OpenPBS/PBSPRO, LSF, Maui/Torque, BQS and Condor, with work underway to support Sun GridEngine.

The WNs generally have the same commands and libraries installed as the UI, apart from the job management commands. VO-specific application software may be preinstalled at the sites in a dedicated area, typically on a shared file system accessible from all WNs.

It is worth stressing that, strictly speaking, a CE corresponds to a single *queue* in the LRMS, following this naming syntax:

```
CEId = <gg_hostname>:<port>/<gg_type>-<LRMS_type>-<batch_queue_name>
```

According to this definition, different queues defined in the same cluster are considered different CEs. This is currently used to define different queues for jobs of different lengths or other properties (e.g. RAM size), or for different VOs. Examples of CE names are:

```
ce101.cern.ch:2119/jobmanager-lcglsf-grid_alice
t2-ce-01.mi.infn.it:2119/jobmanager-lcgpbs-short
lcg02.sinp.msu.ru:2119/blah-pbs-atlas
cmsrv25.fnal.gov:2119/condor-condor-cms
gridgate.cs.tcd.ie:2119/jobmanager-lcgcondor-condor
mars-ce2.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-12hr
```

### 3.3.4. Storage Element

A *Storage Element (SE)* provides uniform access to data storage resources. The Storage Element may control simple disk servers, large disk arrays or tape-based *Mass Storage Systems (MSS)*. Most WLCG/EGEE sites provide at least one SE.

<sup>1</sup>For Globus-based CEs, it is called *Gatekeeper*.

<sup>2</sup>LCG-2 is the former middleware stack used by WLCG/EGEE.

Storage Elements can support different data access protocols and interfaces, described in detail in Section 7.2. Simply speaking, *GSIFTP* (a GSI-secure FTP) is the protocol for whole-file transfers, while local and remote file access is performed using *RFIO* or *gsidcap*.

Most storage resources are managed by a *Storage Resource Manager (SRM)* [20], a middleware service providing capabilities like transparent file migration from disk to tape, file pinning, space reservation, etc. However, different SEs may support different versions of the SRM protocol and the capabilities can vary.

There is a number of SRM implementations in use, with varying capabilities. The *Disk Pool Manager (DPM)* is used for fairly small SEs with disk-based storage only, while *CASTOR* is designed to manage large-scale MSS, with front-end disks and back-end tape storage. *dCache* is targeted at both MSS and large-scale disk array storage systems. Other SRM implementations are in development, and the SRM protocol specification itself is also evolving.

*Classic SEs*, which do not have an SRM interface, provide a simple disk-based storage model. They are in the process of being phased out.

The most common types of SEs currently present in WLCG/EGEE are summarized in the following table:

Type	Resources	File transfer	File I/O	SRM
Classic SE	Disk server	GSIFTP	insecure RFIO	No
DPM	Disk pool	GSIFTP	secure RFIO	Yes
dCache	Disk pool/MSS	GSIFTP	gsidcap	Yes
CASTOR	MSS	GSIFTP	insecure RFIO	Yes

### 3.3.5. Information Service

The *Information Service (IS)* provides information about the WLCG/EGEE Grid resources and their status. This information is essential for the operation of the whole Grid, as it is via the IS that resources are discovered. The published information is also used for monitoring and accounting purposes.

Much of the data published to the IS conforms to the *GLUE Schema* [21], which defines a common conceptual data model to be used for Grid resource monitoring and discovery. More details about the GLUE schema can be found in Appendix G.

Two IS systems are used in gLite 3: the *Globus Monitoring and Discovery Service (MDS)* [23], used for resource discovery and to publish the resource status, and the *Relational Grid Monitoring Architecture (R-GMA)* [24], used for accounting, monitoring and publication of user-level information.

#### MDS

The MDS implements the GLUE Schema using OpenLDAP, an open source implementation of the *Lightweight Directory Access Protocol (LDAP)*, a specialised database optimised for reading, browsing and searching information. Access to MDS data is insecure, both for reading (clients and users) and for writing (services publishing information), i.e. no Grid credentials are required.

The LDAP information model is based on *entries* (objects like a person, a computer, a server, etc.), each with one or more *attributes*. Each entry has a *Distinguished Name (DN)* that uniquely identifies it, and each attribute



has a type and one or more values.

A DN is formed from a sequence of attribute/value pairs, and based on their DNs entries can be arranged into a hierarchical tree-like structure, called a *Directory Information Tree (DIT)*.

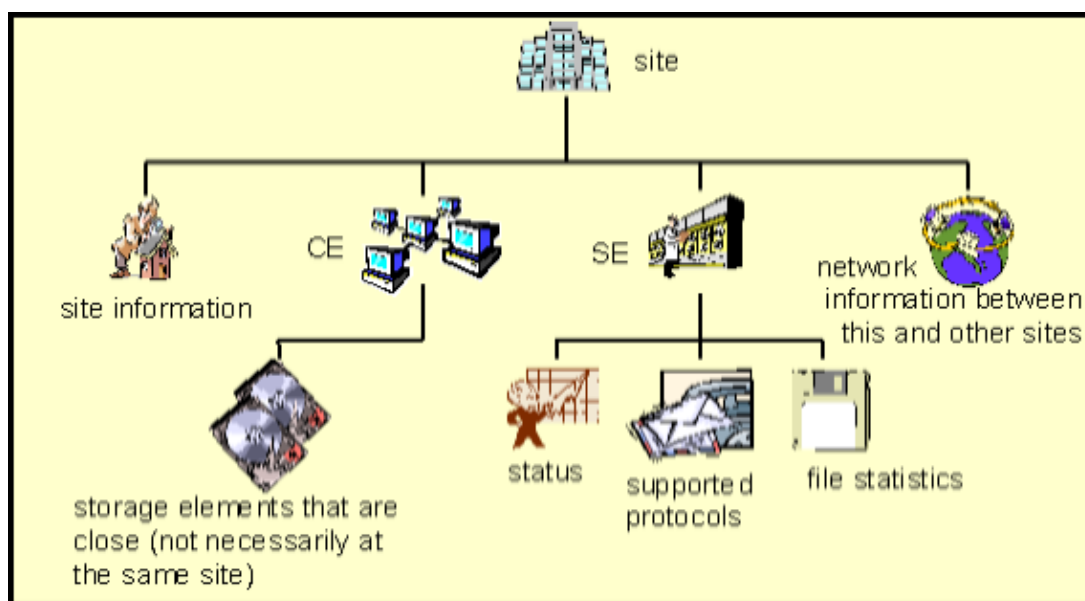


Figure 1: The Directory Information Tree (DIT)

Figure 1 schematically depicts the Directory Information Tree (DIT) of a site: the root entry identifies the site, and entries for site information, CEs and SEs and other services appear at lower levels. Appendix G describes the GLUE schema entries in more detail.

The *LDAP schema* describes the information that can be stored in each entry of the DIT and defines *object classes*, which are collections of mandatory and optional attribute names and value types. While a directory entry describes some object, an object class can be seen as a general description of an object, as opposed to the description of a particular instance.

Figure 2 shows the MDS architecture in WLCG/EGEE. Computing and storage resources at a site run a piece of software called an *Information Provider*, which generates the relevant information about the resource (both static, like the type of SE, and dynamic, like the used space in an SE). This information is published via an LDAP server called a *Grid Resource Information Server (GRIS)*, which normally runs on the resource itself.

At each site another LDAP server, called a *Site Grid Index Information Server (GIIS)*, collects the information from the local GRISes and republishes it. In WLCG/EGEE, the GIIS uses a *Berkeley Database Information Index (BDII)* to store data, which is more stable than the original Globus GIIS.

Finally, a BDII is also used as the top level of the hierarchy. BDIIs at this level are configured to read from a specific set of sites, which effectively defines a view of the overall Grid resources. These BDIIs query the GIISes at every site and act as a cache by storing information about the Grid status in their database. The BDIIs therefore contain all the available information about the Grid sites they look at. Nevertheless, it is always possible to get

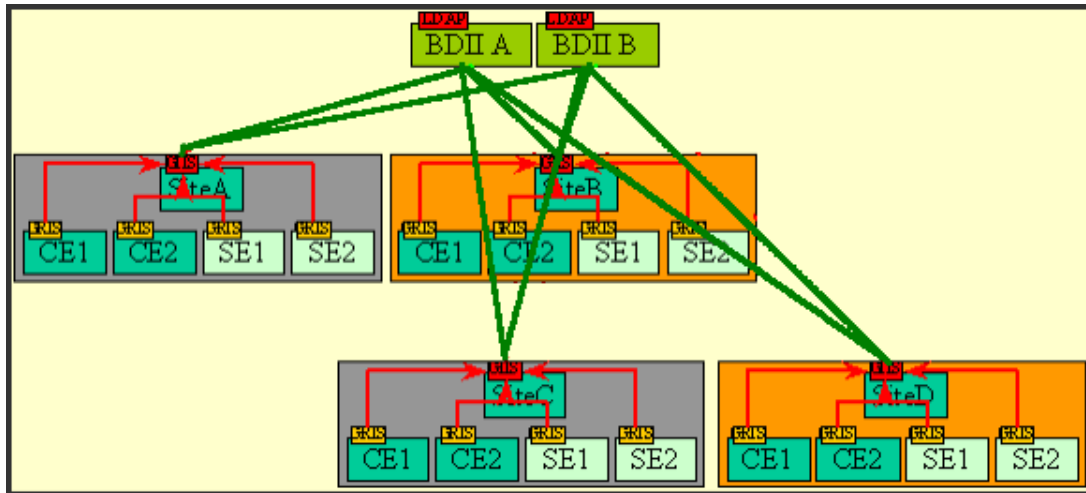


Figure 2: The MDS Information Service in WLCG/EGEE.

information about specific resources by directly contacting the GIISes or even the GRISes.

The top-level BDIs obtain information about the sites in the Grid from the Grid Operations Centre (GOC) database [17], where site managers can insert the contact address of their GIIS as well as other useful information about the site.

### R-GMA

R-GMA is an implementation of the *Grid Monitoring Architecture (GMA)* proposed by the *Global Grid Forum (GGF)* [25]. In R-GMA, information is in many ways presented as though it were in a global distributed relational database, although there are some differences (for example, a table may have multiple rows with the same primary key). This model is more powerful than the LDAP-based one, since relational databases support more advanced query operations. It is also much easier to modify the schema in R-GMA, making it more suitable for user information.

The architecture consists of three major components (Figure 3):

- The **Producers**, which provide the information, register themselves with the Registry and describe the type and structure of the information they provide.
- The **Consumers**, which request the information, can query the Registry to find out what type of information is available and locate Producers that provide such information. Once this information is known, the Consumer can contact the Producer directly to obtain the relevant data.
- The **Registry**, which mediates the communication between the Producers and the Consumers.

The Producers and Consumers are processes (servlets) running in a server machine at each site (sometimes known as a *MON box*). Users interact with these servlets using CLI tools or APIs on the WNs and UIs, and they in turn interact with the Registry, and with Consumers and Producers at other sites, on the user's behalf.

From the user's point of view the information and monitoring system appears like a large relational database and it can be queried as such. Hence, R-GMA uses a subset of SQL as a query language. The user publishes *tuples* (database rows) to a Producer with an SQL insert statement, and queries the Consumers using SQL select statements.

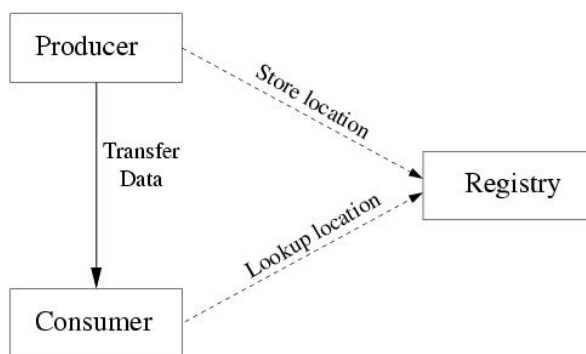


Figure 3: The R-GMA architecture.

R-GMA presents information as a single virtual database containing a set of virtual tables. The *Schema* contains the name and structure (column names, types and settings) of each virtual table in the system (Figure 4), and is normally co-located with the Registry. The Registry contains a list of Producers which publish information for each table. Producers may also register a *predicate*, i.e. a restriction on the tuples they produce, which enables more efficient selection of Producers which may be able to satisfy a query. A Consumer runs an SQL query on a table and the Registry selects the best Producers to answer the query through a process called *mediation*. The Consumer then contacts each Producer directly, combines the information and returns a set of tuples. The details of this process are hidden from the user, who just receives the tuples in response to a query.

An R-GMA system is defined by the Registry and the Schema: what information will be seen by a Consumer depends on what Producers are registered with the Registry. There is only one Registry and one Schema in the WLCG/EGEE production Grid (separate instances exist for the Pre-Production System).

There are two types of Producers: *Primary Producers*, which publish information coming from a user or an Information Provider, and *Secondary Producers*, which consume and republish information from Primary Producers and normally store it in a real database.

Producers can also be classified depending on the type of queries accepted:

- *Continuous* (formerly known as stream): information is sent directly to Consumers as it is produced;
- *Latest*: only the latest information (the tuple with the most recent timestamp for a given value of the primary key) is sent to the Consumer;
- *History*: all tuples within a configurable retention period are stored to allow subsequent retrieval by Consumers.

*Latest* queries correspond most directly to a standard query on a real database. Primary Producers are usually of type *Continuous*. Secondary Producers (which often use a real database to store the data) must be set up in

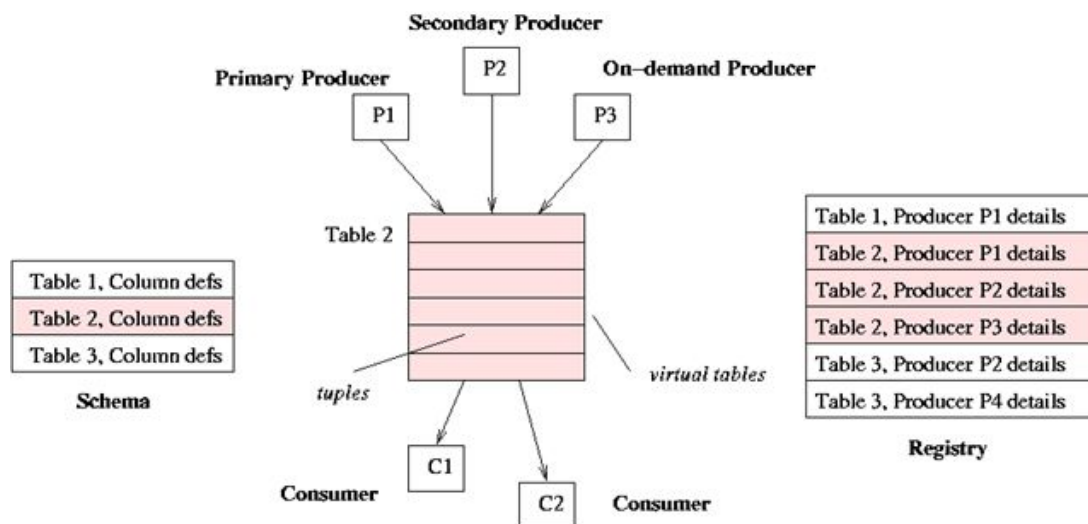


Figure 4: The virtual database of R-GMA

advance to archive information and be able to reply to *Latest* and/or *History* queries. Secondary Producers are also required for joins to be supported in the Consumer queries.

R-GMA is currently used for accounting and both system- and user-level monitoring. It also holds the same GLUE schema information as the MDS; this is not currently used to locate resources for job submission, though.

### 3.3.6. Data Management

The primary unit for Grid data management, as in traditional computing, is the *file*. In a Grid environment, files can have *replicas* at many different sites. Because all replicas must be consistent, Grid files cannot be modified after creation, only read and deleted. Ideally, users do not need to know where a file is located, as they use logical names for the files that the Data Management services use to locate and access them.

Files in the Grid can be referred to by different names: *Grid Unique Identifier (GUID)*, *Logical File Name (LFN)*, *Storage URL (SURL)* and *Transport URL (TURL)*. While the GUIDs and LFNs identify a file irrespective of its location, the SURLs and TURLs contain information about where a physical replica is located, and how it can be accessed.

A file can be unambiguously identified by its GUID; this is assigned the first time the file is registered in the Grid, and is based on the UUID standard to guarantee its uniqueness (UUIDs use a combination of a MAC address and a timestamp to ensure that all UUIDs are distinct). A GUID is of the form: `guid:<unique_string>` (e.g. `guid:93bd772a-b282-4332-a0c5-c79e99fc2e9c`).

In order to locate a file in the Grid, a user will normally use an LFN. LFNs are usually more intuitive, human-readable strings, since they are allocated by the user. Their form is: `lfn:<any_string>`, but the current WLCG/EGEE file catalogue uses strings which have the standard Unix hierarchical format, with elements separated by / characters. A Grid file can have many LFNs, in the same way that a file in a Unix file system can

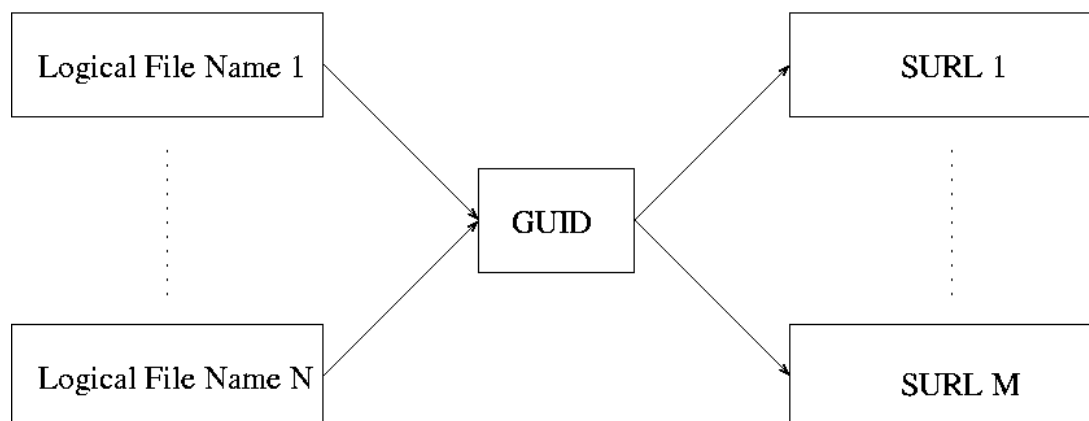


Figure 5: Different filenames in gLite 3.0.

have many links.

The SURL provides information about the physical location of a file replica. Currently, SURLs have the following formats:

```
sfn:<SE_hostname>/<path>
or
srm:<SE_hostname>/<path>
```

for files residing on a classic SE and on an SRM-enabled SE, respectively.

Finally, a TURL gives the necessary information to write or retrieve a physical replica, including hostname, path, protocol and port (as for any conventional URL), so that the application can open or copy it. The format is `<protocol>://<SE_hostname>:<port>/<path>`. There is no guarantee that the path, or even the hostname, in the SURL is the same as in the TURL for the same file. For a given file there may be as many TURLs as there are data access protocols supported by the SE - indeed there may be more, as SEs may hold multiple copies of a file for load-balancing. Figure 5 shows the relationship between the different names of a file.

The mappings between LFNs, GUIDs and SURLs are kept in a service called a *File Catalogue*, while the files themselves are stored in Storage Elements. Currently, the only file catalogue officially supported in WLCG/EGEE is the *LCG File Catalogue (LFC)*, although other catalogues exist.

The Data Management client tools are described in detail in Chapter 7. They allow a user to move data in and out of the Grid, replicate files between Storage Elements, interact with the File Catalogue and more. The high level data management tools shield the user from the complexities of Storage Element and catalogue implementations as well as transport and access protocols. Low level tools are also available, but should be needed only by expert users.

### 3.3.7. Workload Management

The purpose of the *Workload Management System (WMS)* is to accept user jobs, to assign them to the most appropriate Computing Element, to record their status and retrieve their output. The *Resource Broker (RB)* is the machine where the WMS services run [26] [27].

Jobs to be submitted are described using the *Job Description Language (JDL)*, which specifies, for example, which executable to run and its parameters, files to be moved to and from the Worker Node on which the job is run, input Grid files needed, and any requirements on the CE and the Worker Node.

The choice of CE to which the job is sent is made in a process called *match-making*, which first selects, among all available CEs, those which fulfill the requirements expressed by the user and which are close to specified input Grid files. It then chooses the CE with the highest *rank*, a quantity derived from the CE status information which expresses the “goodness” of a CE (typically a function of the numbers of running and queued jobs).

The RB locates the Grid input files specified in the job description using a service called the *Data Location Interface (DLI)*, which provides a generic interface to a file catalogue. In this way, the Resource Broker can talk to file catalogues other than LFC (provided that they have a DLI interface).

The most recent implementation of the WMS from EGEE allows not only the submission of single jobs, but also collections of jobs (possibly with dependencies between them) in a much more efficient way than the old LCG-2 WMS [29], and has many other new options.

Finally, the *Logging and Bookkeeping service (LB)* [30] tracks jobs managed by the WMS. It collects events from many WMS components and records the status and history of the job.

## 3.4. JOB FLOW

This section briefly describes what happens when a user submits a job to the WLCG/EGEE Grid to process some data, and explains how the different components interact.

### 3.4.1. Job Submission

Figure 6 illustrates the process that takes place when a job is submitted to the Grid. It refers to the LCG-2 WMS, but the gLite WMS is similar. The individual steps are as follows:

- a. After obtaining a digital certificate from a trusted Certification Authority, registering in a VO and obtaining an account on a User Interface, the user is ready to use the WLCG/EGEE Grid. He logs in to the UI and creates a proxy certificate to authenticate himself in subsequent secure interactions.
- b. The user submits a job from the UI to a Resource Broker. In the job description one or more files to be copied from the UI to the WN can be specified, and these are initially copied to the RB. This set of files is called the *Input Sandbox*. An event is logged in the LB and the status of the job is SUBMITTED.
- c. The WMS looks for the best available CE to execute the job. To do so, it interrogates the *Information Supermarket (ISM)*, an internal cache of information which in the current system is read from the BDII, to

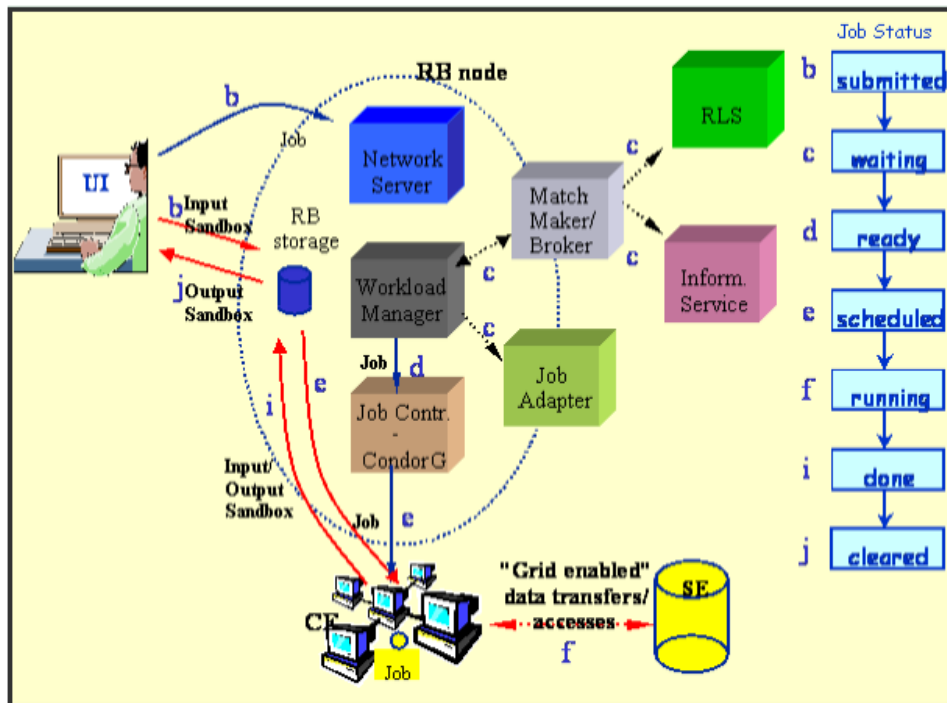


Figure 6: Job flow in the WLCG/EGEE Grid.

- determine the status of computational and storage resources, and the File Catalogue to find the location of any required input files. Another event is logged in the LB and the status of the job is WAITING.
- d. The RB prepares the job for submission, creating a wrapper script that will be passed, together with other parameters, to the selected CE. An event is logged in the LB and the status of the job is READY.
  - e. The CE receives the request and sends the job for execution to the local LRMS. An event is logged in the LB and the status of the job is SCHEDULED.
  - f. The LRMS handles the execution of jobs on the local Worker Nodes. The Input Sandbox files are copied from the RB to an available WN where the job is executed. An event is logged in the LB and the status of the job is RUNNING.
  - g. While the job runs, Grid files can be directly accessed from an SE using either the RFIO or gsidcap protocols, or after copying them to the local filesystem on the WN with the Data Management tools.
  - h. The job can produce new output files which can be uploaded to the Grid and made available for other Grid users to use. This can be achieved using the Data Management tools described later. Uploading a file to the Grid means copying it to a Storage Element and registering it in a file catalogue.
  - i. If the job ends without errors, the output (not large data files, but just small output files specified by the user in the so called *Output Sandbox*) is transferred back to the RB node. An event is logged in the LB and the status of the job is DONE.
  - j. At this point, the user can retrieve the output of his job to the UI. An event is logged in the LB and the status of the job is CLEARED.

- k. Queries for the job status can be addressed to the LB from the UI. Also, from the UI it is possible to query the BDII for the status of the resources.
- l. If the site to which the job is sent is unable to accept or run it, the job may be automatically resubmitted to another CE that satisfies the user requirements. After a maximum allowed number of resubmissions is reached, the job will be marked as aborted. Users can get information about the history of a job by querying the LB service.

### 3.4.2. Other Operations

While the Input and Output Sandboxes are a mechanism for transferring small data files needed to start a job or to check its results, large data files should be read and written from/to SEs and registered in a File Catalogue, and possibly replicated to other SEs. The LCG Data Management client tools are available for performing these tasks. In general, the user should not directly interact with the File Catalogue; instead, he should use the LCG tools.

The File Transfer Service (FTS) provides a managed way to move large numbers of files between SEs.

Users can interrogate the information system to retrieve static or dynamic information about the status of WLCG/EGEE resources and services. Although site GIISes/BDIIs, or even GRISes, can be directly queried, it is recommended to query only a central BDII (or R-GMA). Details and examples on how to interrogate GRIS, GIIS, BDII and R-GMA are given in Chapter 5.



## 4. GRID SECURITY AND GETTING STARTED

This section gives an overview of the security aspects of the WLCG/EGEE Grid, and describes the preliminary steps to gain access to the Grid.

### 4.1. BASIC SECURITY CONCEPTS

Grid security is a very complex area, but it is useful for users to understand at least the basics of how the system works. The following sections give a brief explanation of the most important concepts.

#### 4.1.1. Private and Public Keys

Grid security is based on the concept of *public key encryption*. Each user (or other entity like a server) has a *private key*, generated randomly. This is a number which can be used as a secret password to prove identity. The private key must therefore be kept totally secure; if someone can steal it they can impersonate the owner completely.

Each private key is mathematically related to another number called the *public key*. As the name suggests this can be known to everyone. Formally it's possible to calculate the private key from the public key, but in practice such a calculation is expected to take an unfeasibly long time (the time grows exponentially with the size of the keys). Conversely, calculating the public key from the private key is easy, hence these are sometimes referred to as *asymmetric keys*.

#### 4.1.2. Encryption

The keys are used with an *encryption* algorithm, i.e. a mathematical function which can be applied to any data to produce a coded version of the data. The algorithm has the property that data encrypted using the private key can be decrypted with the public key, and vice versa.

Among other things this can be used to prove identity. Imagine that Ada knows Ben's public key. Ada chooses a random piece of data, encrypts it with the public key and sends it to Ben. Ben decrypts it with the private key and sends it back to Ada. If it matches the number Ada first thought of it proves that Ben does indeed have the right private key.

#### 4.1.3. Signing

Private keys can also be used to *sign* a piece of data. This involves another mathematical function called a *hash function*. This is something which can be applied to data of any length, and produces a fixed-length number which is characteristic of the input data, like a digital fingerprint – in particular even a tiny change to the input would produce a completely different hash. It should also be difficult (i.e. take a very large amount of computer power) to find any data at all which would produce a given hash.

To sign a piece of data a hash is calculated from it, and the hash is then encrypted with the private key and the result attached to the data. Anyone else can then decrypt the hash with the public key, and compare it with one they calculate themselves. If the two hashes match they know two things: that the data was signed by someone who had the private key corresponding to that public key, and that the data has not been modified since it was signed.

#### 4.1.4. Certificates

To be useful, the public key has to be connected to some information about who the user (or server) is. This is stored in a specific format known as an *X.509 certificate* (X.509 being the name of the standard which specifies the format).

The most important thing in the certificate is the *Subject Name (SN)*, which is something which looks like:

```
/C=UK/O=eScience/OU=CLRC/L=RAL/CN=john smith
```

This is an example of a more general format called a *Distinguished Name (DN)*, which appears quite a lot in the Grid world. The idea is that a DN should uniquely identify the thing it names. The details of how to construct a DN have never been established as an international standard, but at least within the Grid it can be assumed that a DN is a unique name, and the SN in a certificate is the owner's name as far as the Grid is concerned.

A certificate also contains some other information, in particular an expiry date after which the certificate is no longer valid. User certificates are normally issued with an expiry date one year ahead, and have to be renewed before they expire. A renewed certificate will normally have new public and private keys, but will usually keep the same SN. In some circumstances, e.g. if the private key is stolen, a certificate may be *revoked*, i.e. added to a known list of certificates which should be considered invalid.

#### 4.1.5. Certification Authorities

Certificates are issued by a *Certification Authority (CA)*. There are many commercial CAs, e.g. Verisign and Thawte, but for Grid use there are special CAs run by academic organisations, generally serving users in a particular geographic region. The CA follows some defined procedures to make sure that it knows who users are and that they are entitled to have a certificate.

To allow people to verify the information in the certificate, the CA signs it with its own private key. Anyone who wants to check the validity of a certificate needs to know the public key of the CA, and the CA therefore has a certificate of its own. Potentially this could create an infinite regression, but this is prevented by the fact that CA certificates, known as *root certificates*, are self-signed, i.e. the CA signs its own certificate. These root certificates are then distributed in some secure way, which in the Grid is typically as Linux RPMs from a trusted repository. (The root certificates of many commercial CAs are often pre-installed in web browsers.)

#### 4.1.6. Proxies

To interact directly with a remote service a certificate can be used to prove identity. However, in the Grid world it is often necessary for a remote service to act on a user's behalf, e.g. a job running on a remote site needs to be able

to talk to other servers to transfer files, and it therefore needs to prove that it is entitled to use the user's identity (this is known as *delegation*). On the other hand, since the private key is so vital it should not be sent to remote machines which might be insecure.

The solution is the use of something called a *proxy*. Strictly speaking a proxy is also a certificate, but usually the unqualified term "certificate" is reserved for something issued by a CA. To make a proxy a new public/private key pair is created, a new certificate is built containing the public key with an SN like:

```
/C=UK/O=eScience/OU=CLRC/L=RAL/CN=john smith/CN=proxy
```

and it is signed with the long-term private key. Proxies normally have a rather short lifetime, typically 12 hours. Note that proxy creation is a purely local process, there is no contact with any remote service.

When a job is submitted, the proxy certificate, the private key for the proxy and the normal certificate (but **not** the long-term private key) are sent with it. When the job wants to prove its delegated identity to another service it sends it the proxy certificate and the standard certificate, but (usually) not the proxy private key. It can then use the chain of certificates to prove that it is entitled to use the delegated SN. In some circumstances a job may even create a new proxy itself, so the chain can potentially be longer.

In security terms a proxy is a compromise. Since the private key is sent with it anyone who steals it can impersonate the owner, so proxies need to be treated carefully. Also there is no mechanism for revoking proxies, so in general even if someone knows that one has been stolen there is little they can do to stop it being used. On the other hand, proxies usually have a lifetime of only a few hours so the potential damage is fairly limited.

#### 4.1.7. VOMS Proxies

A system called *VOMS (VO Management Service)* is used in WLCG/EGEE to manage information about the roles and privileges of users within a VO. This information is presented to services via an extension to the proxy. At the time the proxy is created one or more VOMS servers are contacted, and they return a mini certificate known as an *Attribute Certificate (AC)* which is signed by the VO and contains information about group membership and any associated roles within the VO.

To create a VOMS proxy the ACs are embedded in a standard proxy, and the whole thing is signed with the private key of the parent certificate. Services can then decode the VOMS information and use it as required, e.g. a user may only be allowed to do something if he has a particular role from a specific VO. One consequence of this method is that VOMS attributes can only be used with a proxy, they cannot be attached to a CA-issued certificate.

One other thing to be aware of is that each AC has its own lifetime. This is typically 12 hours as for the proxy, but it is possible for ACs to expire at different times to each other and to the proxy as a whole.

## 4.2. FIRST STEPS

Before using the WLCG/EGEE Grid, the user must do the following:

- a. Obtain an X.509 certificate from a (CA) recognized by WLCG/EGEE;

- b. Get registered with WLCG/EGEE by joining one or more Virtual Organisations;
- c. Obtain an account on a machine which has the WLCG/EGEE User Interface software installed, and copy the certificate to it;
- d. Create a proxy on the UI.

Steps a. to c. need to be executed only once to have access to the Grid, although the certificate will usually need to be renewed once a year, and VO membership may also need to be re-confirmed periodically.

Step d. needs to be executed each day the first time a request to the Grid is submitted, as it generates a proxy valid for a limited period of time (usually 12 hours). After the proxy expires a new proxy must be created before Grid services can be used again.

The following sections provide details on these prerequisites.

### **4.3. OBTAINING A CERTIFICATE**

#### **4.3.1. X.509 Certificates**

The first requirement the user must fulfill is to be in possession of a valid X.509 certificate issued by a recognized Certification Authority (CA). The role of a CA is to guarantee that a user is who he claims to be and is entitled to own his certificate. It is up to the user to discover which CA he should contact. In general CAs are organised geographically and by research institute. Each CA has its own procedure to release certificates.

The following URL maintains an updated list of recognised CAs, as well as detailed information on how to request certificates from a particular CA:

<http://lcg.web.cern.ch/LCG/users/registration/certificate.html>

For many purposes it may be useful to install the root certificates of Grid CAs in a web browser and/or email client, as this will enable the validation of Grid certificates used in web servers and to sign email. (The way to do this is specific to each piece of software and hence cannot be covered here.) In particular users should usually install the root certificate for their own CA. The root certificates can be obtained from this URL:

<https://www.tacar.org/certs.html>

#### **4.3.2. Requesting the Certificate**

In order to obtain a certificate, a user must create a request to a CA. The request is normally generated using either a web-based interface or console commands. Details of which type of request a particular CA accepts can be found on each CA's website.

For a web-based certificate request, a form must usually be filled in with information such as the name of the user, home institute, etc. After submission, a pair of private and public keys are generated, together with a request

for the certificate containing the public key and the user data. The request is then sent to the CA, while the private key stays in the browser, hence the same browser must be used to retrieve the certificate once it is issued.

**Note:** The user must usually install the CA root certificate in his browser first. This is because the CA has to sign the user certificate using its private key, and the user's browser must be able to validate the signature.

For some CAs the certificate requests are generated using a command line interface. The following discussion describes a common scenario for command-line certificate application using a hypothetical `grid-cert-request` command. Again, details of the exact command and the requirements of each CA will vary and can be found on the CA's website.

The `grid-cert-request` command would create, for example, the following 3 files:

<code>userkey.pem</code>	contains the private key associated with the certificate ( <b>This should be set with permissions so that only the owner can read it</b> , i.e. <code>chmod 400 userkey.pem</code> );
<code>userreq.pem</code>	contains the request for the user certificate (essentially the public key);
<code>usercert.pem</code>	a placeholder, to be replaced by the actual certificate when received from the CA (this can be readable by anyone).

Then the `userreq.pem` file has to be sent (usually by e-mail) to the desired CA.

#### 4.3.3. Getting the Certificate

After a request is generated and sent to a CA, the CA will have to confirm that the user asking for a certificate is who he claims he is. This usually involves a physical meeting, or sometimes a phone call, with a **Registration Authority (RA)**, somebody delegated by the CA to verify the legitimacy of a request, and approve it if so. The RA is usually someone at the user's home institute, and will generally need some kind of ID card to prove the user's identity.

After approval, the certificate is generated and delivered to the user. This can be done via e-mail, or by giving instructions to the user to download it from a web page. If the certificate was directly installed in the user's browser then it must be exported (saved) to disk for Grid use. Details of how to do this will depend on the browser, and are usually described on the CA web site.

The received certificate will usually be in one of two formats: **PEM** (extension `.pem`) or **PKCS12** (extension `.p12` or `.pfx`). The latter is the most common for certificates exported from a browser, but the PEM format is currently needed on a WLCG/EGEE UI. The certificates can be converted from one format to the other using the `openssl` command.

If the certificate is in PKCS12 format, then it can be converted to PEM using:

```
$ openssl pkcs12 -nocerts -in my_cert.p12 -out userkey.pem
$ openssl pkcs12 -clcerts -nokeys -in my_cert.p12 -out usercert.pem
```

where:

<code>my_cert.p12</code>	is the input PKCS12 format file;
<code>userkey.pem</code>	is the output private key file;
<code>usercert.pem</code>	is the output PEM certificate file.

The first command creates only the private key (due to the `-nocerts` option), and the second one creates the user certificate (`-clcerts -nokeys` option).

The `grid-change-pass-phrase -file <private_key_file>` command changes the pass phrase that protects the private key. This command will work even if the original key is not password protected. It is important to know that if the user loses the pass phrase, the certificate will become unusable and a new certificate will have to be requested.

Once in PEM format, the two files, `userkey.pem` and `usercert.pem`, should be copied to a User Interface. This is described later.

#### 4.3.4. Renewing the Certificate

CAs issue certificates with a limited duration (usually one year); this implies the need to renew them periodically. The renewal procedure usually requires that the certificate holder sends a request for renewal signed with the old certificate and/or that the request is confirmed by a phone call; the details depend on the policy of the CA. The certificate usually needs to be renewed **before** the old certificate expires; CAs may send an email to remind users that renewal is necessary, but users should try to be aware of the renewal date, and avoid times when they may be away for extended periods.

Renewed certificates have the same SN as the old ones; failing to renew the certificate usually implies the loss of the SN and the necessity to request a completely new certificate with a different SN, which is effectively a new Grid identity.

#### 4.3.5. Taking Care of Private Keys

A private key is the essence of a Grid identity. Anyone who steals it can impersonate the owner, and if it is lost it is no longer possible to do anything in the Grid, so taking care of it is vital. Certificates are issued personally to individuals, and must **never** be shared with other users. To use the Grid a user must agree to an Acceptable Use Policy, which among other things requires him to keep his private key secure.

Proxies also contain private keys, and although these are less valuable, as the lifetime of a proxy is short, it is still important to look after them.

On a UNIX UI the certificate and private key are stored in two files. Typically they are in a directory called `$HOME/.globus` and are named `usercert.pem` and `userkey.pem`, although these can be changed. The certificate is public and can be world-readable, although there is usually no need for it, but the key must only be readable by the owner (Grid commands will check this and refuse to work if the permissions are wrong). Ideally the key should be stored on a disk local to the UI rather than e.g. an NFS-mounted disk, although this is not always possible. If a certificate has been exported from a browser there may also be a PKCS12-format file (`.p12` or `.pfx`) which also contains the private key, and hence this must also be protected.

If a private key is stored under AFS, e.g. on LXPLUS at CERN, be aware that access is controlled by the AFS ACLs rather than the normal file permissions, so users must ensure that the key is not in a publically-readable area.

Web browsers also store private keys internally, and these also need to be protected. The details vary depending on the browser, but password protection should be used if available – this may not be the default (it is not with Internet Explorer). The most secure mode is one in which every use of the private key needs the password to be entered, but this can cause problems as some web sites ask for the certificate many times. If the key is not password-protected it is especially important to take care that no-one else can get access to a browser session.

It is important not to lose a private key, as this implies loss of all access to the Grid, and registration will have to be started again from scratch. This generally implies having several copies in different places – this is often useful anyway, e.g. to use the certificate from a web browser and several UI machines. However, all copies must be stored securely.

A private key stored on a UI must be encrypted, meaning that a passphrase must be typed whenever it is used. A key must **never** be stored without a passphrase. The passphrase should follow similar rules to any computer passwords, but in general should if anything be longer and harder to guess as it gives access to a much larger set of resources than a typical computer system. Usually it is only necessary to type the passphrase once or twice a day to create a proxy, so having a long passphrase is not a major overhead. Users should be aware of the usual risks, like people watching them type or transmitting the passphrase over an insecure link.

A proxy, which includes its own private key, can be stored anywhere, but is typically under `/tmp` (note that this is usually a local area, so when using systems like LXPLUS with many front-end machines, sessions in different windows may not all see the same `/tmp`). The file name is usually `x509up_u1234` where 1234 is the uid, but again this can vary. In any event, like the certificate key a proxy must only be readable by the owner. However, there is no passphrase protection.

## 4.4. REGISTERING WITH WLCG/EGEE

### 4.4.1. The Registration Service

Before a user can use the WLCG/EGEE infrastructure, registration of some personal data and acceptance of some usage rules are necessary. In the process, the user must also choose a *Virtual Organisation (VO)*. The VO must ensure that all its members have provided the necessary information, which will be stored in a database maintained by the VO, and have accepted the usage rules. The procedure through which this is accomplished may vary from VO to VO: pointers to all the VOs in WLCG/EGEE can be found at the Grid operations web site:

<http://cic.gridops.org/>

Note that some VOs are local and are not registered with WLCG/EGEE as a whole; in this case users should consult local documentation for information about registration procedures.

As an example of a registration service, the *LCG Registrar* serves the VOs of the LHC experiments. For detailed information please visit the following URL:

<http://lcg-registrar.cern.ch/>

The registration procedure normally requires the use of a web browser with the user certificate loaded, to enable the request to be properly authenticated. Browsers normally use the PKCS12 certificate format: if the certificate was issued to a user in the PEM format it has to be converted to PKCS12. The following command can be used to perform that conversion:

```
openssl pkcs12 -export -inkey userkey.pem -in usercert.pem \  
-out my_cert.p12 -name "My certificate"
```

where:

<code>userkey.pem</code>	is the path to the private key file;
<code>usercert.pem</code>	is the path to the PEM certificate file;
<code>my_cert.p12</code>	is the path for the output PKCS12-format file to be created;
<code>"My certificate"</code>	is an optional name which can be used to select this certificate in the browser after the user has uploaded it if the user has more than one certificate available.

Once in PKCS12 format the certificate can be loaded into the browser. Instructions about how to do this for some popular browsers are available at:

<http://lcg.web.cern.ch/LCG/users/registration/load-cert.html>

#### 4.4.2. Virtual Organisations

A VO is an entity which typically corresponds to a particular organisation or group of people in the real world. Membership of a VO grants specific privileges to a user. For example, a user belonging to the *atlas* VO will be able to read ATLAS files or to exploit resources reserved for the ATLAS collaboration.

At present, VO names are generally short strings like *cms* or *biomed*. However, it is likely that future VOs will have names in the style of DNS names, e.g. *newvo.cern.ch*, to ensure that different VOs will always have distinct names.

Becoming a member of a VO usually requires membership of the corresponding experiment; in any case a user must comply with the rules of the VO to gain membership. A user may be expelled from a VO if he fails to comply with these rules.

It is possible to belong to more than one VO, although this is an unusual case. However, using a single certificate with more than one VO requires the recognition of VOMS proxies which is not yet the case for all WLCG/EGEE middleware and services, hence it is currently necessary to have a separate certificate for each VO.

### 4.5. SETTING UP THE USER ACCOUNT

#### 4.5.1. The User Interface

Apart from registering with WLCG/EGEE, a user must also have an account on a WLCG/EGEE User Interface in order to access the Grid. To obtain such an account, a local system administrator must be contacted, either at the



user's own site or at a central site like CERN.

As an example, the CERN LXPLUS service can be used as a UI as described in [31]. This use could be extended to other (non LXPLUS) machines mounting AFS.

It is also possible for a user to install the UI software on his own machine, but this is outside the scope of this document.

Once the account has been created, the user certificate must be installed. The usual procedure is to create a directory named `.globus` under the user home directory and put the user certificate and key files there, naming them `usercert.pem` and `userkey.pem` respectively, with permissions 444 for the former, and 400 for the latter. A directory listing should give a result similar to this:

```

ls -l $HOME/.globus
total 13
-r--r--r--  1 doe      xy           4541 Aug 23  2006 usercert.pem
-r-----  1 doe      xy           963 Aug 23  2006 userkey.pem
  
```

#### 4.5.2. Checking a Certificate

To verify that a certificate is not corrupted and print information about it, the command `grid-cert-info` can be used from the UI. The `openssl` command can also be used to verify the validity of a certificate with respect to the certificate of the certification authority that issued it. The command `grid-proxy-init` can be used to check if there is a mismatch between the private key and the certificate.

##### *Example 4.5.2.1 (Retrieving information on a user certificate)*

With the certificate properly installed in the `$HOME/.globus` directory of the user's UI account, issue the command:

```
$ grid-cert-info
```

If the certificate is properly formed, the output will be something like:

```

Certificate:
  Data:
    Version: 3 (0x2)
    Serial Number: 5 (0x5)
    Signature Algorithm: md5WithRSAEncryption
    Issuer: C=CH, O=CERN, OU=cern.ch, CN=CERN CA
    Validity
      Not Before: Sep 11 11:37:57 2002 GMT
      Not After : Nov 30 12:00:00 2003 GMT
    Subject: O=Grid, O=CERN, OU=cern.ch, CN=John Doe
  
```

```

Subject Public Key Info:
  Public Key Algorithm: rsaEncryption
  RSA Public Key: (1024 bit)
    Modulus (1024 bit):
      00:ab:8d:77:0f:56:d1:00:09:b1:c7:95:3e:ee:5d:
      c0:af:8d:db:68:ed:5a:c0:17:ea:ef:b8:2f:e7:60:
      2d:a3:55:e4:87:38:95:b3:4b:36:99:77:06:5d:b5:
      4e:8a:ff:cd:da:e7:34:cd:7a:dd:2a:f2:39:5f:4a:
      0a:7f:f4:44:b6:a3:ef:2c:09:ed:bd:65:56:70:e2:
      a7:0b:c2:88:a3:6d:ba:b3:ce:42:3e:a2:2d:25:08:
      92:b9:5b:b2:df:55:f4:c3:f5:10:af:62:7d:82:f4:
      0c:63:0b:d6:bb:16:42:9b:46:9d:e2:fa:56:c4:f9:
      56:c8:0b:2d:98:f6:c8:0c:db
    Exponent: 65537 (0x10001)
X509v3 extensions:
  Netscape Base Url:
    http://home.cern.ch/globus/ca
  Netscape Cert Type:
    SSL Client, S/MIME, Object Signing
  Netscape Comment:
    For DataGrid use only
  Netscape Revocation Url:
    http://home.cern.ch/globus/ca/bc870044.r0
  Netscape CA Policy Url:
    http://home.cern.ch/globus/ca/CPS.pdf
Signature Algorithm: md5WithRSAEncryption
  30:a9:d7:82:ad:65:15:bc:36:52:12:66:33:95:b8:77:6f:a6:
  52:87:51:03:15:6a:2b:78:7e:f2:13:a8:66:b4:7f:ea:f6:31:
  aa:2e:6f:90:31:9a:e0:02:ab:a8:93:0e:0a:9d:db:3a:89:ff:
  d3:e6:be:41:2e:c8:bf:73:a3:ee:48:35:90:1f:be:9a:3a:b5:
  45:9d:58:f2:45:52:ed:69:59:84:66:0a:8f:22:26:79:c4:ad:
  ad:72:69:7f:57:dd:dd:de:84:ff:8b:75:25:ba:82:f1:6c:62:
  d9:d8:49:33:7b:a9:fb:9c:1e:67:d9:3c:51:53:fb:83:9b:21:
  c6:c5

```

The `grid-cert-info` command takes many options. Use the `-help` option for a full list. For example, the `-subject` option returns the Subject Name:

```

$ grid-cert-info -subject
/O=Grid/O=CERN/OU=cern.ch/CN=John Doe

```

or to check the certificate expiration date:

```

$ grid-cert-info -enddate
Oct 15 05:37:09 2006 GMT

```

or to know which CA issued the certificate:

```
$ grid-cert-info -issuer  
/C=CH/O=CERN/OU=GRID/CN=CERN CA
```

**Example 4.5.2.2 (Verifying a user certificate)**

To verify a user certificate, issue the following command from the UI:

```
$ openssl verify -CApath /etc/grid-security/certificates ~/.globus/usercert.pem
```

and if the certificate is valid and properly signed, the output will be:

```
/home/does/.globus/usercert.pem: OK
```

If the certificate of the CA that issued the user certificate is not found in `-CApath`, an error message like this will appear:

```
usercert.pem: /O=Grid/O=CERN/OU=cern.ch/CN=John Doe  
error 20 at 0 depth lookup:unable to get local issuer certificate
```

If the environment variable `X509_CERT_DIR` is defined, use its value in place of `/etc/grid-security/certificates`.

**Example 4.5.2.3 (Verifying the consistency between private key and certificate)**

If for some reason the user is using a certificate (`usercert.pem`) which does not correspond to the private key (`userkey.pem`), strange errors may occur. To test if this is the case, run the command:

```
grid-proxy-init -verify
```

In case of mismatch, the output will be:

```
Your identity: /C=CH/O=CERN/OU=GRID/CN=John Doe  
Enter GRID pass phrase for this identity:  
Creating proxy ..... Done
```

```
ERROR: Couldn't verify the authenticity of the user's credential to  
generate a proxy from.  
Use -debug for further information.
```

## 4.6. PROXIES

### 4.6.1. Standard Proxies

At this point, the user is able to generate a proxy using the command `grid-proxy-init`, which prompts for the user passphrase, as in the next example.

**Note:** with the introduction of VOMS, the `grid-proxy-init` command can be replaced by `voms-proxy-init`, as this is fully backward-compatible.

#### *Example 4.6.1.1 (Creating a proxy)*

To create a proxy, issue the command:

```
$ grid-proxy-init
```

If the command is successful, the output will be like

```
Your identity: /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
Enter GRID pass phrase for this identity:
Creating proxy ..... Done
Your proxy is valid until: Tue Jun 24 23:48:44 2003
```

and the proxy will be written in `/tmp/x509up_u<uid>`, where `<uid>` is the Unix UID of the user, unless the environment variable `X509_USER_PROXY` is defined, in which case its value is taken as the proxy file path.

If the user gives a wrong pass phrase, the output will be

```
ERROR: Couldn't read user key. This is likely caused by
either giving the wrong pass phrase or bad file permissions
key file location: /home/does/.globus/userkey.pem
Use -debug for further information.
```

If the proxy file cannot be created, the output will be

```
ERROR: The proxy credential could not be written to the output file.
Use -debug for further information.
```

If the user certificate files are missing, or the permissions of `userkey.pem` are not correct, the output is:

```
ERROR: Couldn't find valid credentials to generate a proxy.
Use -debug for further information.
```

By default, the proxy has a lifetime of 12 hours. To specify a different lifetime, the `-valid H:M` option can be used (the proxy is valid for `H` hours and `M` minutes –default is 12:00). When a proxy has expired, it becomes useless and a new one has to be created with `grid-proxy-init`. However, longer lifetimes imply bigger security risks, and the Grid Acceptable Use Policy generally limits proxy lifetimes to 24 hours — some services may reject proxies with lifetimes which are too long.

Use the option `-help` for a full listing of options.

It is also possible to print information about an existing proxy, or to destroy it before its expiration, as in the following examples.

#### ***Example 4.6.1.2 (Printing information on a proxy)***

To print information about a proxy, for example the Subject Name or the time left before expiration, give the command:

```
$ grid-proxy-info
```

The output, if a valid proxy exists, will be similar to

```
subject : /O=Grid/O=CERN/OU=cern.ch/CN=John Doe/CN=proxy
issuer  : /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
type    : full
strength : 512 bits
path    : /tmp/x509up_u7026
timeleft : 11:59:56
```

If a proxy does not exist, the output is:

```
ERROR: Couldn't find a valid proxy.
Use -debug for further information.
```

#### ***Example 4.6.1.3 (Destroying a proxy)***

To destroy an existing proxy before its expiration, it is enough to do:

```
$ grid-proxy-destroy
```

If no proxy exists, the result will be:

```
ERROR: Proxy file doesn't exist or has bad permissions
Use -debug for further information.
```

#### 4.6.2. VOMS Proxies

The *Virtual Organisation Membership Service (VOMS)* is a system which allows a proxy to have *extensions* containing information about the VO, the groups the user belongs to in the VO, and any roles the user is entitled to have.

In VOMS terminology, a *group* is a subset of the VO containing members who share some responsibilities or privileges in the project. Groups are organised hierarchically like a directory tree, starting from a VO-wide root group. A user can be a member of any number of groups, and a VOMS proxy contains the list of all groups the user belongs to, but when the VOMS proxy is created the user can choose one of these groups as the “primary” group.

A *role* is an attribute which typically allows a user to acquire special privileges to perform specific tasks. In principle, groups are associated to privileges that the user always has, while roles are associated to privileges that a user needs to have only from time to time. Note that roles are attached to groups, i.e. roles in different groups with the same role name are distinct.

The groups and roles are defined by each VO; they may be assigned to a user at the initial registration, or added subsequently.

To map groups and roles to specific privileges, what counts is the group/role combination, which is sometimes referred to as an FQAN (short form for Fully Qualified Attribute Name). The format is:

```
FQAN = <group name>[/Role=<role name>]
```

for example, /cms/HeavyIons/Role=production.

##### **Example 4.6.2.1** (Creating a VOMS proxy)

The `voms-proxy-init` command generates a Grid proxy, contacts one or more VOMS servers, retrieves the user attributes and includes them in the proxy. If used without arguments, it works exactly as `grid-proxy-init`.

To create a basic VOMS proxy, without requiring any special role or primary group, use:

```
$ voms-proxy-init -voms <vo>
```

where `<vo>` is the VO name. The output is similar to:

```
Your identity: /C=CH/O=CERN/OU=GRID/CN=John Doe
Enter GRID pass phrase:
Creating temporary proxy ..... Done
Contacting lcg-voms.cern.ch:15002 [/C=CH/O=CERN/OU=GRID/CN=host/lcg-voms.cern.ch]
"cms" Done
Creating proxy ..... Done
Your proxy is valid until Thu Mar 30 06:17:27 2006
```

Note that there are two steps: a standard Grid proxy is created first and used to authenticate to the VOMS server, and the full VOMS proxy is then created using information returned by it. If a valid proxy already exists the `-noregen` option can be used to avoid the first step, including typing the passphrase.

One clear advantage of VOMS proxies over standard proxies is that the middleware can find out to which VO the user belongs from the proxy, while using a normal proxy the VO has to be explicitly specified by other means.

To create a proxy with a given role (e.g. `production`) and primary group (e.g. `/cms/HeavyIons`), the syntax is:

```
$ voms-proxy-init -voms <alias>:<group name>[Role=<role name>]
```

where `alias` specifies the server to be contacted (see below), and usually is the name of the VO. For example:

```
$ voms-proxy-init -voms cms:/cms/HeavyIons/Role=production
```

`voms-proxy-init` uses a configuration file, whose path can be specified in several ways; if the path is a directory, the files inside it are concatenated and taken as the actual configuration file. A user-level configuration file, which must be owned by the user, is looked for in these locations:

- the argument of the `-userconf` option;
- the file `$HOME/.glite/vomses`.

If it is not found, a system-wide configuration file, which must be owned by root, is looked for in these locations:

- the argument of the `-confile` option;
- the file `$GLITE_LOCATION/etc/vomses`;
- the file `/opt/glite/etc/vomses`.

The configuration file must contain lines with the following syntax:

```
alias host port subject vo
```

where the items are respectively an alias (usually the name of the VO), the host name of the VOMS server, the port number to contact for a given VO, the DN of the server host certificate, and the name of the VO. For example:

```
"dteam" "lcg-voms.cern.ch" "15004"
"/C=CH/O=CERN/OU=GRID/CN=host/lcg-voms.cern.ch" "dteam"
```

**Example 4.6.2.2**    *(Printing information on a VOMS proxy)*

The `voms-proxy-info` command is used to print information about an existing VOMS proxy. Two useful options are `-all`, which prints everything, and `-fqan`, which prints the groups and roles in FQAN format. For example:

```

$ voms-proxy-info -all
subject   : /C=CH/O=CERN/OU=GRID/CN=John Doe/CN=proxy
issuer    : /C=CH/O=CERN/OU=GRID/CN=John Doe
identity  : /C=CH/O=CERN/OU=GRID/CN=John Doe
type      : proxy
strength  : 512 bits
path      : /tmp/x509up_u10585
timeleft  : 11:59:58
=== VO cms extension information ===
VO        : cms
subject   : /C=CH/O=CERN/OU=GRID/CN=John Doe
issuer    : /C=CH/O=CERN/OU=GRID/CN=host/lcg-voms.cern.ch
attribute : /cms/Role=NULL/Capability=NULL
timeleft  : 11:59:58
  
```

Note that there are separate times to expiry for the proxy as a whole and the VOMS extension, which can potentially be different.

### 4.6.3. Proxy Renewal

Proxies created as described in the previous section pose a problem: if a job does not finish before the expiration time of the proxy used to submit it, is aborted. This can easily happen, for example, if the job takes a very long time to execute, or if it stays in a queue for a long time. The easiest solution to the problem would be to use very long-lived proxies, but at the expense of an increased security risk. Moreover, the duration of a VOMS proxy is limited by the VOMS server and cannot be made arbitrarily long.

To overcome this limitation, a proxy credential repository system is used, which allows the user to create and store a long-term proxy in a dedicated server (a *MyProxy server*). The WMS will then be able to use this long-term proxy to periodically renew the proxy for a submitted job before it expires and until the job ends (or the long-term proxy expires).

To see if a WLCG/EGEE site has a MyProxy Server, the GOC database [17] may be consulted; MyProxy servers have a node type of `PROX`. A UI may have a default server defined in the `MYPROXY_SERVER` environment variable.

As the renewal process starts 30 minutes before the old proxy expires, it is necessary to generate an initial proxy long enough, or the renewal may be triggered too late, after the job has failed with the following error:

```

Status Reason: Got a job held event, reason: Globus error 131:
the user proxy expired (job is still running)
  
```

The minimum recommended time for the initial proxy is 30 minutes, and in most circumstances it should be



substantially longer. Job submission is forbidden for proxies with a remaining lifetime less than 20 minutes: an error message like the following will be produced:

```

**** Error: UI_PROXY_DURATION ****
Proxy certificate will expire within less then 00:20 hours.
  
```

Management of the proxy renewal functionality is available via the `myproxy` commands. The user must either specify the host name of a MyProxy server, or define it as the value of the `MYPROXY_SERVER` environment variable.

For the WMS to know which MyProxy server to use in the proxy renewal process, the name of the server must be included in an attribute of the job's JDL file (see Chapter 6). If the user does not add it manually, the name of the default MyProxy server is added automatically when the job is submitted. This default is defined in a VO-specific configuration file.

**Note:** the machine where the WMS runs must be trusted by the MyProxy server for renewal to be allowed.

**Example 4.6.3.1** *(Creating a long-term proxy and storing it in a MyProxy Server)*

To create and store a long-term proxy, the user must do, for example:

```
$ myproxy-init -s <myproxy_server> -d -n
```

where `-s <myproxy_server>` specifies the hostname of the machine where a MyProxy Server runs, the `-d` option instructs the server to associate the user DN to the proxy, and the `-n` option avoids the use of a passphrase to access the long-term proxy, so that the WMS can perform the renewal automatically.

The output will be similar to:

```

Your identity: /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
Enter GRID pass phrase for this identity:
Creating proxy ..... Done
Your proxy is valid until: Thu Jul 17 18:57:04 2003
A proxy valid for 168 hours (7.0 days) for user /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
now exists on myproxy.cern.ch.
  
```

By default, the long-term proxy lasts for one week and the proxies created from it last 12 hours. These lifetimes can be changed using the `-c` and the `-t` option respectively, but cannot be longer than the lifetime of the user certificate.

If the `-s <myproxy_server>` option is missing, the command will try to use the `MYPROXY_SERVER` environment variable to determine the MyProxy Server.

**Note:** If the hostname of the MyProxy Server is wrong, or the service is unavailable, the output will be similar to:

```
Your identity: /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
Enter GRID pass phrase for this identity:
Creating proxy ..... Done
Your proxy is valid until: Wed Sep 17 12:10:22 2003
Unable to connect to adc0014.cern.ch:7512
```

where only the last line reveals that an error occurred.

### **Example 4.6.3.2** (Retrieving information about a long-term proxy)

To get information about a long-term proxy stored in a Proxy Server, the following command may be used:

```
$ myproxy-info -s <myproxy_server> -d
```

where the `-s` and `-d` options have the same meaning as in the previous example. The output is similar to:

```
username: /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
owner: /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
timeleft: 167:59:48 (7.0 days)
```

Note that there must be a valid proxy on the UI, created with `grid-proxy-init` or `voms-proxy-init`, to successfully interact with the long-term proxy on the MyProxy server.

### **Example 4.6.3.3** (Deleting a long-term proxy)

Deleting a stored long-term proxy is achieved by doing:

```
$ myproxy-destroy -s <myproxy_server> -d
```

and the output is:

```
Default MyProxy credential for user /O=Grid/O=CERN/OU=cern.ch/CN=John Doe
was successfully removed.
```

Again, a valid proxy must exist on the UI.

## 5. INFORMATION SERVICE

The architecture of the gLite 3 Information Services, both MDS and R-GMA, was described in Chapter 3. In this chapter, we have a closer look at the structure of the information published by those services, and we examine some tools that can be used to get information from them.

Most middleware components (for Data and Workload Management) currently rely on information from MDS. However, R-GMA is also in use and many applications, especially for accounting and monitoring purposes, depend on it.

Some information about tools used for Grid monitoring in gLite 3 is also provided here.

### 5.1. THE MDS

In the following sections examples are given on how to interrogate the MDS Information Service in gLite 3. In particular, the different servers from which the information can be obtained are discussed. These are the local GRISes, the site GIISes/BDIIs and the global (or top-level) BDIIs. Of these, the top-level BDII is usually the one queried, since it contains all the interesting information for a VO in a single place.

Before the procedure to directly query the MDS is described, two higher level tools, `lcg-infosites` and `lcg-info`, are presented. These tools should be enough for most common user needs and will usually avoid the necessity of raw LDAP queries (although these are very useful for more complex or subtle requirements).

As explained in Chapter 3, the data in the MDS in WLCG/EGEE conforms to the LDAP implementation of the GLUE Schema, although for historical reasons some extra attributes are also currently published and may be queried and used by clients of the IS. The current implementation relates to version 1.2 of the GLUE schema, but version 1.3, which adds some new information in a backward-compatible way, will be deployed during early 2007. For a list of the defined object classes and their attributes, as well as for a reference on the Directory Information Tree used to publish those attributes, please check Appendix G.

As usual, the tools to query the IS shown in this section are command-line based. There exist, however, graphical tools that can be used to browse LDAP servers. As an example, the program `gq` is open source and can be found in some Linux distributions by default. Some comments on this tool are given in Section 5.1.6.

#### 5.1.1. `lcg-infosites`

The `lcg-infosites` command can be used to obtain VO-specific information on existing Grid resources. The syntax is the following:

```
lcg-infosites --vo <vo> <option> -v <verbosity> -f <site> --is <bdii>
```

This is the definition of the command options and arguments:

- `--vo <vo>`: the name of the VO to which the information to print is related (mandatory);

- <option>: specifies what information has to be printed. It can take the following values:
  - ce: the number of CPUs, running jobs, waiting jobs and CE names (global, no VO-specific information);
    - v 1: only the CE names;
    - v 2: the cluster names, the amount of RAM, the operating system name and version and the processor model;
  - se: the names of the SEs supporting the VO, the type of storage system and the used and available space;
    - v 1: only the SE names;
  - all: the information given by ce and se; together;
  - closeSE: the names of the CEs supporting the VO and their close SEs;
  - tag: the software tags published by each CE supporting the VO;
  - lfc: the hostname of the LFC catalogues available to the VO;
  - lfcLocal: the hostname of the local LFC catalogues available to the VO;
  - rb: the hostname and port of the RBs available to the VO;
  - dli: the Data Location Index servers available to the VO;
  - dliLocal: the local Data Location Index servers available to the VO;
  - vobox: the VO boxes available to the VO;
  - fts: the endpoints of the FTS servers available to the VO;
  - sitenames: the names of all WLCG/EGEE sites;
- --is <bdii>: the BDII to query. If not specified, the BDII defined in the environment variable LCG\_GFAL\_INFOSYS will be queried.
- -f <site>: restricts the information printed to the specified site (it applies only to options rb, dli, vobox and fts).

**Example 5.1.1.1 (Obtaining information about computing resources)**

The way to get information relating to the computing resources for the *alice* VO is:

```
$ lcg-infosites --vo alice ce
```

A typical output is as follows:

#CPU	Free	Total	Jobs	Running	Waiting	ComputingElement
15	4	0		0	0	ce002.ipp.acad.bg:2119/jobmanager-lcgpbs-alice
15	4	0		0	0	ce001.ipp.acad.bg:2119/blah-pbs-alice
80	8	0		0	0	ce02.grid.acad.bg:2119/jobmanager-pbs-alice
10	10	0		0	0	ce.hpc.iit.bme.hu:2119/blah-pbs-alice
96	94	0		0	0	grid109.kfki.hu:2119/jobmanager-lcgpbs-alice
3409	6	493		493	0	ce101.cern.ch:2119/jobmanager-lcglsf-grid_alice

```
3409      6      493      493      0      ce102.cern.ch:2119/jobmanager-lcglsf-grid_alice
3409      6      493      493      0      ce105.cern.ch:2119/jobmanager-lcglsf-grid_alice
[...]
```

**Example 5.1.1.2 (Obtaining information about storage resources)**

To get the status of the storage resources:

```
$ lcg-infosites --vo atlas se
```

Avail Space(Kb)	Used Space(Kb)	Type	SEs
39657488	106362948	n.a	se.phy.bg.ac.yu
31400000	18580000	n.a	se1.egee.man.poznan.pl
569586792	47148288	n.a	clrauvergridse01.in2p3.fr
1200000000	410000000	n.a	koala.unimelb.edu.au
22903032	42994124	n.a	se-lcg.sdg.ac.cn
457865076	663121389	n.a	atlasse01.ihep.ac.cn
29593756	80561288	n.a	se001.grid.bas.bg
931135488	41943040	n.a	se001.ipp.acad.bg
[...]			

**Example 5.1.1.3 (Listing the close Storage Elements)**

The option `closeSE` will give an output as follows:

```
$ lcg-infosites --vo dteam closeSE

Name of the CE: g02.phy.bg.ac.yu:2119/blah-pbs-dteam
se.phy.bg.ac.yu

Name of the CE: ce.phy.bg.ac.yu:2119/jobmanager-pbs-dteam
se.phy.bg.ac.yu

Name of the CE: fangorn.man.poznan.pl:2119/jobmanager-lcgpbs-dteam
se1.egee.man.poznan.pl
se1.egee.man.poznan.pl

Name of the CE: obsauvergridce01.univ-bpclermont.fr:2119/jobmanager-lcgpbs-dteam
clrauvergridse01.in2p3.fr
[...]
```

**Example 5.1.1.4** (Listing local LFC servers)

In order to retrieve the hostnames of the local LFC servers for a certain VO, use the command as follows:

```
$ lcg-infosites --vo atlas lfcLocal
lxb2038.cern.ch
pps-lfc.cnaf.infn.it
cclcglfcli03.in2p3.fr
[...]
```

**5.1.2. lcg-info**

The `lcg-info` command can be used to list either CEs or SEs satisfying a given set of conditions on their attributes, and to print, for each of them, the values of a given set of attributes. The information is taken from the BDII specified by the `LCG_GFAL_INFOSYS` environment variable or in the command line.

The general format of the command for listing CE or SE information is:

```
$ lcg-info [--list-ce | --list-se] [--query <query>] [--attrs <attrs>]
```

where either `--list-ce` or `--list-se` must be used to indicate if CEs or SEs should be listed. The `--query` option introduces a filter (conditions to be fulfilled) to the elements of the list, and the `--attrs` option may be used to specify which attributes to print. If `--list-ce` is specified then only CE attributes are considered (others are just ignored), and the reverse is true for `--list-se`.

The attributes supported (which may be included with `--attrs` or within the `--query` expression) are only a subset of the attributes present in the GLUE schema, those that are most relevant for a user.

The `--vo` option can be used to restrict the query to CEs and SEs which support the given VO; it is mandatory when querying for attributes which are inherently related to a VO, like `AvailableSpace` and `UsedSpace`.

Apart from the listing options, the `--help` option can be specified (alone) to obtain a detailed description of the command, and the `--list-attrs` option can be used to get a list of the supported attributes.

**Example 5.1.2.1** (Get the list of supported attributes)

To have a list of the supported attributes, use:

```
$ lcg-info --list-attrs
```

The output is similar to:

```
Attribute name  Glue object class  Glue attribute name
```

```
EstRespTime      GlueCE          GlueCEStateEstimatedResponseTime
WorstRespTime    GlueCE          GlueCEStateWorstResponseTime
TotalJobs        GlueCE          GlueCEStateTotalJobs
TotalCPUs        GlueCE          GlueCEInfoTotalCPUs
[...]
```

For each attribute, the simplified attribute name used by `lcg-info`, the corresponding object class and the attribute name in the GLUE schema are given.

**Example 5.1.2.2** *(List all the CEs satisfying given conditions and print the desired attributes)*

Suppose one wants to know how many jobs are running and how many free CPUs there are on CEs that have an Athlon CPU and have Scientific Linux:

```
$ lcg-info --list-ce --query 'Processor=*athlon*,OS=*Scientific*' \
--attrs 'RunningJobs,FreeCPUs'
```

The output could be:

```
- CE: alice003.nipne.ro:2119/jobmanager-lcgpbs-alice
  - RunningJobs      0
  - FreeCPUs         2

- CE: alice003.nipne.ro:2119/jobmanager-lcgpbs-dteam
  - RunningJobs      0
  - FreeCPUs         2
[...]
```

It must be stressed that `lcg-info` only supports a logical AND of logical expressions, separated by commas, and the only allowed operator is `=`. In equality comparisons of strings the `*` wildcard matches any number of characters.

Another useful query is one to know which CEs have installed a particular version of an experiment's software. That would be something like:

```
$ lcg-info --vo cms --list-ce --attrs Tag --query 'Tag=*ORCA_8_7_1*'
```

Note that this lists all tags for all VOs for the matching CEs.

**Example 5.1.2.3** *(List the close CEs for all the SEs)*

Similarly, suppose that you want to know which CEs are close to each SE:

```
$ lcg-info --list-se --vo cms --attrs CloseCE
```

the output will be like:

```

- SE: SE.pakgrid.org.pk
  - CloseCE          CE.pakgrid.org.pk:2119/jobmanager-lcgpbs-ops
                    CE.pakgrid.org.pk:2119/jobmanager-lcgpbs-cms
                    CE.pakgrid.org.pk:2119/jobmanager-lcgpbs-dteam

- SE: aliserv1.ct.infn.it
  - CloseCE          _UNDEF_

- SE: arxiloxos2.inp.demokritos.gr
  - CloseCE          arxiloxos1.inp.demokritos.gr:2119/jobmanager-lcgpbs-dteam
                    arxiloxos1.inp.demokritos.gr:2119/jobmanager-lcgpbs-cms
                    arxiloxos1.inp.demokritos.gr:2119/jobmanager-lcgpbs-ops

[...]
```

A value `_UNDEF_` means that the attribute is not defined for that SE or CE.

The `--bdii` option can be used to specify a particular BDII (e.g. `--bdii exp-bdii.cern.ch:2170`), and the `--sed` option can be used to output the results of the query in a format easy to parse in a script, in which values for different attributes are separated by `%` and values of list attributes are separated by `&`.

### 5.1.3. The Local GRIS

The first level of MDS information publication is the GRIS, which provides specific information for a particular service. The GRIS normally runs on the same node as the CE, SE or other service for which it publishes, although it may be on a different node. There is usually no need to query a GRIS directly except for detailed debugging, and in some cases site firewalls may prevent access from external sites.

In order to interrogate the GRIS on a specific node, the hostname and the TCP port where the GRIS run must be specified. The port is normally either 2135 or 2170. The following command can be used:

```
$ ldapsearch -x -h <hostname> -p 2135 -b "mds-vo-name=local, o=grid"
```

where the `-x` option indicates that simple authentication (instead of LDAP's SASL) should be used; the `-h` and `-p` options precede the hostname and port respectively; and the `-b` option is used to specify the initial entry for the search in the LDAP tree. If the port is 2170, the `-b` option should be `mds-vo-name=resource, o=grid` (for port 2170).

### 5.1.4. Using the `ldapsearch` command to read the MDS

For the LDAP implementation of the GLUE schema, the root of the DIT is always `o=grid`. At the GRIS level the next entry is (for historical reasons) either `mds-vo-name=local` or `mds-vo-name=resource`, but at the site level



this is replaced with `mds-vo-name=<sitename>`, and a top-level BDII has site entries under `mds-vo-name=<sitename>`, `mds-vo-name=local,o=grid`. The GLUE entries themselves are at lower levels and always have the same DN structure. For details, please refer to Appendix G.

The same effect as the command above can be obtained with:

```
$ ldapsearch -x -H <ldap_uri> -b "mds-vo-name=local, o=grid"
```

where the hostname and port are included in the `-H <ldap_uri>` option, avoiding the use of `-h` and `-p`.

#### **Example 5.1.4.1 (Interrogating the GRIS on a Computing Element)**

The command used to interrogate the GRIS located on host `lxb2006.cern.ch` is:

```
$ ldapsearch -x -h lxb2006.cern.ch -p 2135 -b "mds-vo-name=local, o=grid"
```

or:

```
$ ldapsearch -x -H ldap://lxb2006.cern.ch:2135 -b "mds-vo-name=local, o=grid"
```

In order to restrict the search, a filter of the form `attribute operator value` can be used. The operator is one of those defined in the following table (note that `<` and `>` are not included):

Operator	Description
=	Entries whose attribute is equal to the value
>=	Entries whose attribute is greater than or equal to the value
<=	Entries whose attribute is less than or equal to the value
=*	Entries that have any value set for that attribute
~=	Entries whose attribute value approximately matches the specified value

Furthermore, complex search filters can be formed by using boolean operators to combine constraints. The boolean operators that can be used are “AND” (`&`), “OR” (`|`) and “NOT” (`!`). The syntax for such expressions is the following:

```
( "&" or "|" or "!" (filter1) [(filter2) ...] )
```

Example of search filters are:

```
(& (Name=Smith) (Age>=32))
(! (GlueHostMainMemoryRAMSize<=1000))
```

It is possible to construct complex queries, but the syntax is not very intuitive so some experimentation may be needed. Be aware that filters may need to be escaped to prevent special characters being interpreted by the shell.

In LDAP, a special attribute `objectClass` is defined for each directory entry. It indicates which object classes are defined for that entry in the LDAP schema. This makes it possible to filter entries that contain a certain object class. The filter for this case would be:

```
'objectclass=<name>'
```

Apart from filtering the search, a list of attribute names can be specified, in order to limit the values returned. As shown in the next example, only the value of the specified attributes will be returned. Alternatively, `grep` or other Unix tools can be used to postprocess the output.

A description of all objectclasses and their attributes usable with the `ldapsearch` command can be found in Appendix G.

**Example 5.1.4.2** (Getting information about the site name from the GRIS on a CE)

```
$ ldapsearch -x -h lcgbdi02.gridpp.rl.ac.uk -p 2170 -b o=grid \
  '(&(objectclass=GlueSite)(GlueSiteName=ral*))' GlueSiteWeb \
  GlueSiteLatitude GlueSiteLongitude
version: 2

#
# filter: (&(objectclass=GlueSite)(GlueSiteName=ral*))
# requesting: GlueSiteWeb GlueSiteLatitude GlueSiteLongitude
#
# RAL-LCG2, RAL-LCG2, local, grid
dn: GlueSiteUniqueID=RAL-LCG2,mds-vo-name=RAL-LCG2,mds-vo-name=local,o=grid
GlueSiteLatitude: 51.57
GlueSiteLongitude: -1.32
GlueSiteWeb: http://www.gridpp.ac.uk/tier1a/

# search result
search: 2
result: 0 Success

# numResponses: 2
# numEntries: 1
```

By adding the `-LLL` option, it is possible to avoid the comments and the version information in the reply:

```
$ ldapsearch -LLL -x -h lcgbdi02.gridpp.rl.ac.uk -p 2170 -b o=grid \
  '(&(objectclass=GlueSite)(GlueSiteName=ral*))' GlueSiteWeb \
  GlueSiteLatitude GlueSiteLongitude
dn: GlueSiteUniqueID=RAL-LCG2,mds-vo-name=RAL-LCG2,mds-vo-name=local,o=grid
```

GlueSiteLatitude: 51.57  
GlueSiteLongitude: -1.32  
GlueSiteWeb: <http://www.gridpp.ac.uk/tier1a/>

### 5.1.5. The Site GIIS/BDII

At each site, a site GIIS or BDII collects information about all resources present at a site (i.e. data from all GRISes at the site). Site BDIIs are preferred to site GIISes and are the default in gLite 3 releases. In this section we explain how to query a site GIIS/BDII.

Often the site GIIS/BDII runs on a Computing Element, although it may be on a separate node. The port used to interrogate a site BDII is usually 2170. The DIT base name is based on the site name. However, it is sufficient to use a base of `o=grid` in ldap queries.

For a list of all sites and all resources present, refer to the GOC database [17]. The complete contact string for the site BDII is published in the GOC page for the site. For example, the following URL:

<https://goc.grid-support.ac.uk/gridsite/gocdb2/index.php?siteSelect=95>

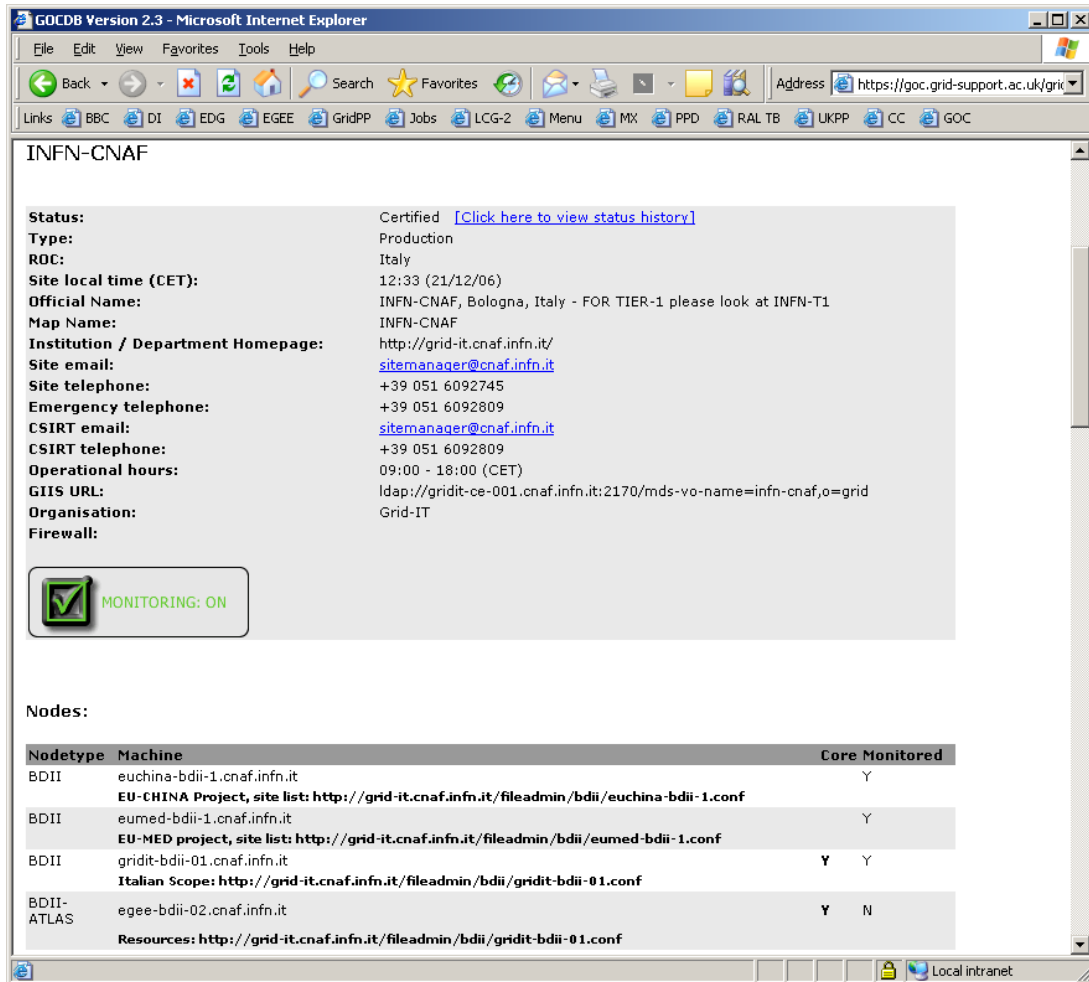
refers to the site information for the CNAF site (at Bologna in Italy), as shown in Figure 7. The GIIS (BDII) URL is:

`ldap://gridit-ce-001.cnaf.infn.it:2170/mds-vo-name=infn-cnaf,o=grid`

In order to interrogate it, use the `ldapsearch` command as follows.

#### **Example 5.1.5.1** (Interrogating a site BDII)

```
$ ldapsearch -x -h gridit-ce-001.cnaf.infn.it -p 2170 \  
-b mds-vo-name=infn-cnaf,o=grid  
version: 2  
  
#  
# filter: (objectclass=*)  
# requesting: ALL  
#  
  
# INFN-CNAF, grid  
dn: mds-vo-name=INFN-CNAF,o=grid  
objectClass: GlueTop  
  
# gridit-ce-001.cnaf.infn.it:2119/jobmanager-lcgpbs-lcg, INFN-CNAF, grid  
dn: GlueCEUniqueID=gridit-ce-001.cnaf.infn.it:2119/jobmanager-lcgpbs-lcg,mds-v  
o-name=INFN-CNAF,o=grid  
objectClass: GlueCETop
```



**INFN-CNAF**

**Status:** Certified [\[Click here to view status history\]](#)

**Type:** Production

**ROC:** Italy

**Site local time (CET):** 12:33 (21/12/06)

**Official Name:** INFN-CNAF, Bologna, Italy - FOR TIER-1 please look at INFN-T1

**Map Name:** INFN-CNAF

**Institution / Department Homepage:** <http://grid-it.cnaf.infn.it/>

**Site email:** [sitemanager@cnaf.infn.it](mailto:sitemanager@cnaf.infn.it)

**Site telephone:** +39 051 6092745

**Emergency telephone:** +39 051 6092809

**CSIRT email:** [sitemanager@cnaf.infn.it](mailto:sitemanager@cnaf.infn.it)


**CSIRT telephone:** +39 051 6092809

**Operational hours:** 09:00 - 18:00 (CET)

**GIIS URL:** <ldap://gridit-ce-001.cnaf.infn.it:2170/mds-vo-name=infn-cnaf,o=grid>

**Organisation:** Grid-IT

**Firewall:**

 **MONITORING: ON**

**Nodes:**

Nodetype	Machine	Core Monitored
BDII	euchina-bdii-1.cnaf.infn.it <b>EU-CHINA Project, site list:</b> <a href="http://grid-it.cnaf.infn.it/fileadmin/bdii/euchina-bdii-1.conf">http://grid-it.cnaf.infn.it/fileadmin/bdii/euchina-bdii-1.conf</a>	Y
BDII	eumed-bdii-1.cnaf.infn.it <b>EU-MED project, site list:</b> <a href="http://grid-it.cnaf.infn.it/fileadmin/bdii/eumed-bdii-1.conf">http://grid-it.cnaf.infn.it/fileadmin/bdii/eumed-bdii-1.conf</a>	Y
BDII	gridit-bdii-01.cnaf.infn.it <b>Italian Scope:</b> <a href="http://grid-it.cnaf.infn.it/fileadmin/bdii/gridit-bdii-01.conf">http://grid-it.cnaf.infn.it/fileadmin/bdii/gridit-bdii-01.conf</a>	Y Y
BDII-ATLAS	egee-bdii-02.cnaf.infn.it <b>Resources:</b> <a href="http://grid-it.cnaf.infn.it/fileadmin/bdii/gridit-bdii-01.conf">http://grid-it.cnaf.infn.it/fileadmin/bdii/gridit-bdii-01.conf</a>	Y N

Figure 7: The GOCDB information page for the INFN-CNAF site

```

objectClass: GlueCE
objectClass: GlueSchemaVersion
objectClass: GlueCEAccessControlBase
objectClass: GlueCEInfo
objectClass: GlueCEPolicy
objectClass: GlueCEState
objectClass: GlueInformationService
objectClass: GlueKey
GlueCEHostingCluster: gridit-ce-001.cnaf.infn.it
GlueCEName: lcg
GlueCEUniqueID: gridit-ce-001.cnaf.infn.it:2119/jobmanager-lcgpbs-lcg
GlueCEInfoGatekeeperPort: 2119
GlueCEInfoHostName: gridit-ce-001.cnaf.infn.it
GlueCEInfoLRMSType: pbs

```

```
GlueCEInfoLRMSVersion: torque_1.0.1p5
GlueCEInfoTotalCPUs: 10
GlueCEInfoJobManager: lcgpbs
GlueCEInfoContactString: gridit-ce-001.cnaf.infn.it:2119/jobmanager-lcgpbs-lcg
GlueCEInfoApplicationDir: /opt/exp_soft
GlueCEInfoDataDir: unset
GlueCEInfoDefaultSE: grid007g.cnaf.infn.it
GlueCEStateEstimatedResponseTime: 0
GlueCEStateFreeCPUs: 8
GlueCEStateRunningJobs: 0
GlueCEStateStatus: Draining
GlueCEStateTotalJobs: 0
GlueCEStateWaitingJobs: 0
GlueCEStateWorstResponseTime: 0
GlueCEStateFreeJobSlots: 0
GlueCEPolicyMaxCPUTime: 2880
GlueCEPolicyMaxRunningJobs: 0
GlueCEPolicyMaxTotalJobs: 0
GlueCEPolicyMaxWallClockTime: 4320
GlueCEPolicyPriority: 1
GlueCEPolicyAssignedJobSlots: 0
GlueCEAccessControlBaseRule: VO:atlas
GlueCEAccessControlBaseRule: VO:alice
GlueCEAccessControlBaseRule: VO:lhcb
GlueCEAccessControlBaseRule: VO:cms
GlueForeignKey: GlueClusterUniqueID=gridit-ce-001.cnaf.infn.it
GlueInformationServiceURL: ldap://gridit-ce-001.cnaf.infn.it:2135/mds-vo-name=
  local,o=grid
GlueSchemaVersionMajor: 1
GlueSchemaVersionMinor: 2

[...]
```

### 5.1.6. The top-level BDII

A top-level BDII collects all information coming from site GIISes/BDIIs and stores them in a cache. The top-level BDII can be configured to collect published information from resources in all sites in a Grid (usually derived from the GOC DB), or just from a subset of them. The site list is normally filtered to include only sites which are currently operational, and VOs can also apply their own filters to exclude sites which are currently failing certain critical tests, so the sites visible in a BDII may fluctuate.

In order to find the location of a top-level BDII at a site (if any), consult the GOCDB page for the site. The BDII will be listed with the rest of the nodes of the site (refer to Figure 7, node type BDII), and the entry may also include comments about the purpose and content of the BDII. One general-purpose top-level BDII is `lcg-bdii.cern.ch`.

A BDII can be interrogated using the base name `mds-vo-name=local,o=grid` (although it suffices to use `o=grid`) and port 2170. The sub-tree corresponding to a particular site appears under an entry with a DN like:

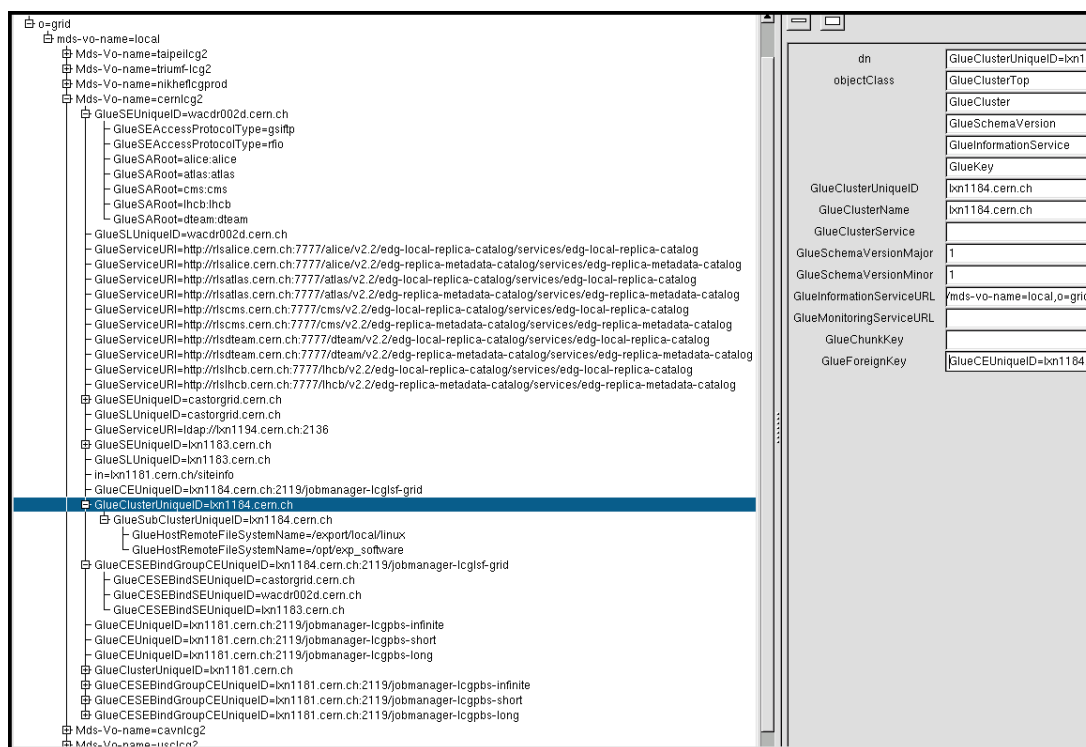


Figure 8: The LDAP directory of a gLite 3 BDII

Mds-Vo-name=<sitename>, mds-vo-name=local, o=grid

In Figure 8, a view of the DIT of a BDII in gLite 3 is shown. In the figure, only the sub-tree that corresponds to the CERN site is expanded. The DN for every entry in the DIT is shown. Entries for storage and computing resources, as well as for the bindings between CEs and SEs and for various services, can be seen in the figure.

Each entry can contain attributes from different object classes. This can be seen in the entry with DN `GlueClusterUniqueID=lxn1184.cern.ch, Mds-Vo-name=cernlcg2, mds-vo-name=local, o=grid`, which is highlighted in the figure. This entry contains several attributes from the object classes `GlueClusterTop`, `GlueCluster`, `GlueSchemaVersion`, `GlueInformationService` and `GlueKey`. However, one of the object classes is the primary one for the object, in this case `GlueCluster`, and an attribute from it is used to form the DN. Since every object in the tree must have a unique DN the attribute used must be unique at least within its branch of the tree.

In the right-hand side of the window, the DN of the selected entry and the names and values (in the cases where they exist) of the attributes for this entry are shown. Notice how the special `objectclass` attribute gives information about all the object classes that are applied to this entry.

As can be seen, a graphical tool can be quite useful to examine the structure (and indeed the details) of an MDS directory. In addition, the schema (object classes and attributes) can be also examined.

### **Example 5.1.6.1 (Interrogating a BDII)**

In this example, a query is sent to a BDII in order to retrieve two attributes from the `GlueCESEBind` object class for all sites:

```
$ ldapsearch -x -LLL -H ldap://lxn1187.cern.ch:2170 -b "o=grid" \
'objectclass=GlueCESEBind' GlueCESEBindCEUniqueID GlueCESEBindSEUniqueID

dn: GlueCESEBindSEUniqueID=castor.grid.sinica.edu.tw,GlueCESEBindGroupCEUnique
ID=tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-atlas,mds-vo-name=resource
,mds-vo-name=Taiwan-PPS,mds-vo-name=local,o=grid
GlueCESEBindSEUniqueID: castor.grid.sinica.edu.tw
GlueCESEBindCEUniqueID: tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-atlas

dn: GlueCESEBindSEUniqueID=castor.grid.sinica.edu.tw,GlueCESEBindGroupCEUnique
ID=tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-dteam,mds-vo-name=resource
,mds-vo-name=Taiwan-PPS,mds-vo-name=local,o=grid
GlueCESEBindSEUniqueID: castor.grid.sinica.edu.tw
GlueCESEBindCEUniqueID: tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-dteam

dn: GlueCESEBindSEUniqueID=dpm01.grid.sinica.edu.tw,GlueCESEBindGroupCEUniqueI
D=tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-biomed,mds-vo-name=resource
,mds-vo-name=Taiwan-PPS,mds-vo-name=local,o=grid
GlueCESEBindSEUniqueID: dpm01.grid.sinica.edu.tw
GlueCESEBindCEUniqueID: tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-biomed

dn: GlueCESEBindSEUniqueID=castor.grid.sinica.edu.tw,GlueCESEBindGroupCEUnique
ID=tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-biomed,mds-vo-name=resourc
e,mds-vo-name=Taiwan-PPS,mds-vo-name=local,o=grid
GlueCESEBindSEUniqueID: castor.grid.sinica.edu.tw
GlueCESEBindCEUniqueID: tb009.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-biomed

[...]
```

### **Example 5.1.6.2 (Listing all the CEs which publish a given tag)**

The attribute `GlueHostApplicationSoftwareRunTimeEnvironment` can be used to publish experiment-specific information (*tags*) for a CE, for example to indicate that a given set of experiment software is installed. To list all the CEs which publish a given tag, a query to a BDII can be performed. In this example, the information is retrieved for all subclusters:

```
$ ldapsearch -h lxn1187.cern.ch -p 2170 -b "o=grid" -x 'objectclass=GlueSubCluster' \
GlueChunkKey GlueHostApplicationSoftwareRunTimeEnvironment
```

### **Example 5.1.6.3 (Listing all the SEs which support a given VO)**

A Storage Element *supports* a VO if users of that VO are allowed to store files on that SE. It is possible to find out which SEs support a VO with a query to the BDII. For example, to have the list of all SEs supporting the *alice* VO, together with the storage space available in each of them, a query similar to this can be used:

```
$ ldapsearch -LLL -h lxn1187.cern.ch -p 2170 -b \
"mds-vo-name=local,o=grid" -x "GlueSAAccessControlBaseRule=alice" \
GlueChunkKey GlueSAStateAvailableSpace GlueSAStateUsedSpace
```

where the `GlueSAAccessControlBaseRule` attribute contains the name of the supported VO. The obtained result will be something like the following:

```
dn: GlueSALocalID=alice,GlueSEUniqueID=gw38.hep.ph.ic.ac.uk,mds-vo-name=UKI-LT
  2-IC-HEP-PPS,mds-vo-name=local,o=grid
GlueSAStateAvailableSpace: 275474688
GlueSAStateUsedSpace: 35469432
GlueChunkKey: GlueSEUniqueID=gw38.hep.ph.ic.ac.uk

dn: GlueSALocalID=alice,GlueSEUniqueID=grid08.ph.gla.ac.uk,mds-vo-name=UKI-ScottGrid-Gla-PPS,mds-vo-name=local,o=grid
GlueSAStateAvailableSpace: 3840000000
GlueSAStateUsedSpace: 1360000000
GlueChunkKey: GlueSEUniqueID=grid08.ph.gla.ac.uk

dn: GlueSALocalID=alice,GlueSEUniqueID=grid13.csl.ee.upatras.gr,mds-vo-name=Pr
  eGR-02-UPATRAS,mds-vo-name=local,o=grid
GlueSAStateAvailableSpace: 186770000
GlueSAStateUsedSpace: 10090000
GlueChunkKey: GlueSEUniqueID=grid13.csl.ee.upatras.gr
[...]
```

## 5.2. R-GMA

As explained in section 3.3.5, R-GMA is an alternative information system to MDS. The standard GLUE information is published in R-GMA, together with various monitoring data, and the system is also available for users to publish their own data. The system can be used via a command-line interface or APIs for C, C++, Python and Java, and for queries there is also a web browser interface. Several applications already use R-GMA, especially for accounting and monitoring purposes.

This section gives a brief overview of R-GMA, but for more information see [24].



### 5.2.1. R-GMA concepts

From a user point of view, R-GMA is very similar to a standard relational database. Data are organised in relational tables, and inserted and queried with SQL-style `INSERT` and `SELECT` statements (the allowed syntax is a subset of SQL, but reasonably complete for most purposes). However, there are some differences to bear in mind. The most basic is that a standard relational database can only have one row (tuple) with a given primary key value, but R-GMA usually has more than one. Related to this is the fact that R-GMA supports three different query types. Each tuple has a timestamp, and for a given primary key value you can query the most recent tuple (*Latest query*), a history of all tuples within some defined retention period (*History query*), or ask for tuples to be streamed to you as they are published (*Continuous query*). Continuous queries can also return a limited amount of historical (“old”) data.

There are also some differences depending on how and where the data are stored. Each site has an R-GMA server which deals with all R-GMA interaction from clients at that site. The servers store data published from local clients (known as *primary producers*), and may also collect data from other sites and re-publish it (*secondary producers*). Generally speaking, primary producers answer Continuous queries and secondary producers answer Latest and History queries; the latter query types are only supported if someone has created a secondary producer for the table(s) concerned (this is normally the case for standard tables, e.g. GLUE). The data may be stored either in memory or in a real database, and some queries, notably joins, are only possible if all the required data can be found in a single real database. Such producers are known as *archivers*.

The local R-GMA servers store all the data and deal with all the client interactions, so in this sense R-GMA is a distributed system. However, there is also a central server known as the *Registry*, which holds the *schema* (the definitions of all the tables), and has lists of all consumers and producers to allow them to find each other. At present the Registry is a unique service in the Grid.

Users are free to create and use their own tables. However, at present there is only a single namespace for tables, so users should try to choose distinctive table names, e.g. prefixed with the VO or application name. There is a standard table called `userTable` which can be used for simple tests.

R-GMA is a secure service to the extent that you need a valid proxy to use it (or a valid certificate in your web browser). However, there is currently no authorisation control, so anyone can read and publish to any table. This is expected to be added in future releases.

### 5.2.2. The R-GMA Browser

The *R-GMA browser* is usually installed on each R-GMA server. It allows the user to easily navigate the schema (to see what tables are available and how they are defined), see all available producers for a table and query the (selected) producers. All this can be achieved using a web interface.

Figure 9 shows this R-GMA browser web interface. It is accessible via the following URL:

```
https://lcgmon01.gridpp.rl.ac.uk:8443/R-GMA/index.html
```

You can replace the hostname with the name of your local server to get a better response. In the left-hand bar you have a list of predefined tables to query; selecting one of them will give a drop-down list of matching items you can select, or you can just hit the Query button to see everything.

Alternatively, selecting the “Table Sets” link gives a complete list of all tables in the schema. Clicking on a table name gives a page where you can query the table definition, enter an SQL query, select the query type, and see and select from a list of all producers for that table. If you simply hit the Query button you get a Latest query for all data in the table, which corresponds to the intuitive idea of the current content of the table.

The best way to get some understanding of R-GMA is to play with the query interface for one of the standard tables, e.g. GlueSite, as the interface is reasonably intuitive. The browser is read-only so you can’t do any damage.

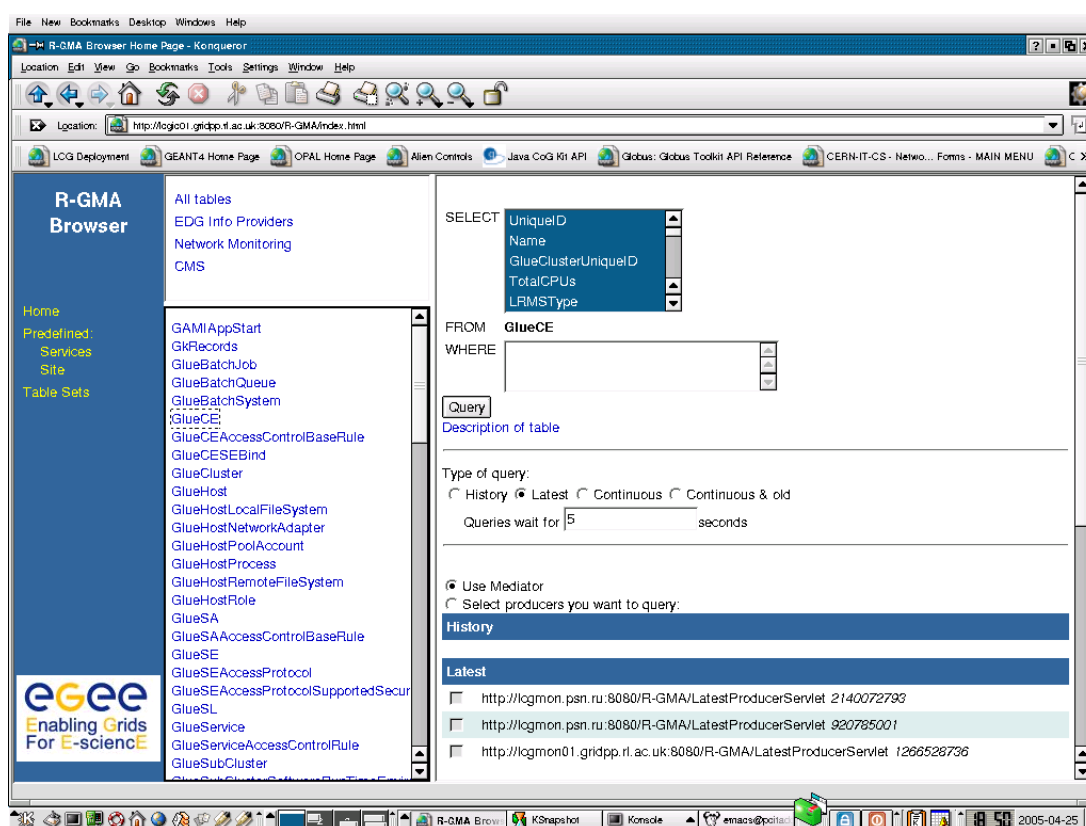


Figure 9: The R-GMA Web Interface

### 5.2.3. The R-GMA CLI

An R-GMA CLI is available on every UI and WN. This interface allows the user to perform queries and also to publish new information. It includes a consumer and can initiate both primary and secondary producers, although it does not provide all the detailed options available in the APIs.

The user can interact with the CLI directly from the command line by using the `-c` option:

```
rgma -c ``select Web from GlueSite where UniqueId='lcmgmon01.gridpp.rl.ac.uk'``
```

```
+-----+
| Web |
+-----+
| http://www.gridpp.ac.uk/tier1a/ |
+-----+
1 rows
```

If you simply type `rgma` an interactive shell is started:

```
Welcome to the R-GMA virtual database for Virtual Organisations.
=====
```

Your local R-GMA server is:

```
https://lcgmon01.gridpp.rl.ac.uk:8443/R-GMA
```

You are connected to the following R-GMA Registry services:

```
https://lcgic01.gridpp.rl.ac.uk:8443/R-GMA/RegistryServlet
```

You are connected to the following R-GMA Schema service:

```
https://lcgic01.gridpp.rl.ac.uk:8443/R-GMA/SchemaServlet
```

Type `''help''` for a list of commands.

```
rgma> select Web from GlueSite where UniqueId='lcmmon01.gridpp.rl.ac.uk'
```

```
+-----+
| Web |
+-----+
| http://www.gridpp.ac.uk/tier1a/ |
+-----+
1 rows
rgma>
```

As shown, the CLI reports the location of the registry, which holds pointers to all the R-GMA producers for all sites and VOs. Queries will collect information from the appropriate producers wherever they are located.

The syntax of all the commands available in the R-GMA interface can be obtained using the `help` command to get a list of the supported commands, and typing `help <command>` to get information on a particular command. A list of the most important commands is as follows:

Command	Description
help [<command>]	Display information (general or about a specific command)
exit / quit / CTRL-D	Exit the R-GMA command line shell
show [tables   producers of <table>]	Show the tables in the schema, or the current producers for a given table
describe <table>	Show the column names and types for the specified table
select	Query R-GMA (SQL syntax)
set query latest   continuous   history	Set the type of subsequent queries
insert	Insert a tuple into a primary producer (SQL syntax)
secondaryproducer <table>	Declare a table to be consumed and republished by a secondary producer
set [secondary]producer latest   continuous   history	Set the supported query type for the primary or secondary producer
set [timeout   maxage] <timeout> [<units>]	Set the timeout for queries or the maximum age of tuples to return

A simple example of how to query the R-GMA virtual database follows.

**Example 5.2.3.1 (Querying the R-GMA Information System)**

Inside the interface you can easily perform any query using SQL syntax:

```
rgma> set query continuous
Set query type to continuous
rgma> set timeout 120 seconds
Set timeout to 120 seconds
rgma> select UniqueID, TotalCPUs from GlueCE

+-----+-----+
| UniqueID                               | TotalCPUs |
+-----+-----+
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-atlas | 498      |
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-dteam | 498      |
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-lhcb  | 498      |
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-babar | 498      |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-lhcb    | 68       |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-cms     | 68       |
+-----+-----+
```

```

| grid001.fi.infn.it:2119/jobmanager-lcgpbs-atlas | 68 |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-lhcb | 68 |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-cms | 68 |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-atlas | 68 |
| grid012.ct.infn.it:2119/jobmanager-lcglsf-alice | 174 |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-lhcb | 68 |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-cms | 68 |
| grid001.fi.infn.it:2119/jobmanager-lcgpbs-atlas | 68 |
| grid012.ct.infn.it:2119/jobmanager-lcglsf-infinite | 174 |
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-atlas | 498 |
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-dteam | 498 |
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-lhcb | 498 |
| hepgrid2.ph.liv.ac.uk:2119/jobmanager-lcgpbs-babar | 498 |
+-----+-----+-----+
19 rows

```

In this example, we first set the type of query to continuous. That is, new tuples are received as they are published, and the query will not terminate unless the user aborts or a maximum time for the query is reached. This timeout is then defined as 120 seconds. Finally, we query for the ID and the number of CPUs of all CEs publishing information into R-GMA in the two minutes following the query.

#### 5.2.4. R-GMA APIs

There exist R-GMA APIs in Java, C, C++ and Python. They include methods for creating consumers, as well as primary and secondary producers; setting the types of queries and of producers, retention periods and time outs; retrieving tuples, and inserting data. The APIs are beyond the scope of this introduction, but detailed documentation exists for all APIs, including example code [24].

### 5.3. SERVICE DISCOVERY

The gLite 3 Service Discovery (SD) API makes it possible to access service details published to the Information Systems. The main purpose is to answer questions like: *I am at CERN, in the ops VO. Where is a MyProxy server?* It therefore represents a simplified view of the Grid Information System to locate resources/services and query their properties. The SD interface supports several information systems, currently MDS, R-GMA, Globus MDS4, and a local `service.xml` file. The client access mode to the underlying infrastructure that holds the information is set by environment variables.

- **R-GMA:** the R-GMA client software must be installed. R-GMA will look in `GLITE_LOCATION` for its configuration files. For connection to a secure server, either `X509_USER_PROXY` or `TRUSTFILE` must also be defined;
- **BDII:** the environment variable `LCG_GFAL_INFOSYS` must contain the hostname and port number of the BDII service to query;

- **FILE:** the default configuration file is in `$GLITE_LOCATION/etc/services.xml` and can be overridden by `$HOME/.glite/etc/services.xml`.

The gLite 3 Service Discovery API provides interfaces for the Java and C/C++ programming languages, and a command line interface (`glite-sd-query`). The gLite 3 Service Discovery User Guide [53] offers a comprehensive documentation of the SD APIs.

### 5.3.1. Running a Service Discovery query

In order to use Service Discovery, the user has to set the `GLITE_SD_PLUGIN` variable to specify the Information System(s) to be queried. To use all of the R-GMA, BDII and XML file-based information systems set:

```
GLITE_SD_PLUGIN="rgma,bdii,file"
```

The API will then try each in turn until one of them returns something.

The `glite-sd-query` command allows the listing of basic information about known services. To list detailed information about all services at a site, use the `-s` option instead:

```
glite-sd-query -x -s "cern.ch"
```

If a user wants to know more details about a specific service of type *myproxy* located at CERN, the same command is used with the following options:

```
> glite-sd-query -t myproxy -s CERN-PROD
```

```
Name: myproxy-fts.cern.ch:7512  
Type: MyProxy  
Endpoint: myproxy-fts.cern.ch:7512  
Version: 1.1.0
```

```
Name: prod-px.cern.ch:7512  
Type: MyProxy  
Endpoint: prod-px.cern.ch:7512  
Version: 1.1.0
```

**Note:** as an alternative to the `-s` option, the environment variable `GLITE_SD_SITE` can be used to restrict the search to a given site. The value of the variable can be either the name of the site in the GOC DB or the DNS domain of the site.

**Note:** the type of service is specified as part of the GLUE schema, and may have a more complex format: for example, the FTS service has a type of `org.glite.FileTransfer`.

## 5.4. MONITORING

The ability to monitor resource related parameters is currently considered a necessary functionality in any network. In such a heterogeneous and complex system as the Grid, this necessity becomes fundamental. A monitoring system implies the existence of a central repository of operational information (in WLCG/EGEE, the GOCDB). The monitoring system should be able to collect data from the resources in the system, in order to analyze the usage, behavior and performance of the Grid, detect and notify fault conditions, contract violations and user-defined events.

The GOC web page contains a whole section concerning monitoring information for WLCG/EGEE. Apart from R-GMA, several different monitoring tools are in use, including general-purpose monitoring tools and Grid specific systems like GridICE [32].

Also important are the web pages publishing the results of functional tests applied periodically to the all the sites registered within WLCG/EGEE. The results of these tests show if a site is responding correctly to standard Grid operations; otherwise, an investigation on the cause of the unexpected results is undertaken. Some VOs may even decide to automatically exclude from their BDII the sites that are not passing the functional tests successfully, so that they do not appear in the IS and are not considered for possible use by their applications.

**Note:** please do not report problems occurring with a site if this site is marked as having failures in the standard test reports. If that is the case, the site will already have been notified of the problems by the grid operations staff. Also, the site details in the GOCDB will show if the site is currently in scheduled downtime. The results of some sets of functional sites can be checked in the following URLs:

<http://goc.grid.sinica.edu.tw/gstat/>

<https://lcg-sam.cern.ch:8443/sam/sam.py>

or for the PPS:

<https://lcg-sam.cern.ch:8443/sam-pps/sam.py>

In the following section, as an example of a monitoring system, the GridICE service is described.

### 5.4.1. GridICE

The GridICE monitoring service is structured in a five layer architecture. The resource information is obtained from the gLite 3 Information Service, namely MDS. The information model for the retrieved data is an extended GLUE Schema, where some new objects and attributes have been added to the original model. Please refer to the documentation presented in [32] for details on this.

GridICE not only periodically retrieves the last information published in MDS, but also collects historical monitoring data in a persistent storage. This allows the observation of the evolution in time of the published data. In addition, GridICE will provide performance analysis, usage level and general reports and statistics, as well as the possibility to configure event detection and notification actions, although these two functionalities are still at an early development stage.

**Note:** All the information retrievable using GridICE (including the extensions of the GLUE schema) is also

obtainable through R-GMA, by defining the proper archivers. This represents an alternative way to obtain the information.

The GridICE web page that shows the monitoring information for WLCG/EGEE is accessible at the following URL (also linked from the GOC web site):

<http://gridice2.cnaf.infn.it:50080/gridice/site/site.php>

In the initial page (site view) a summary of the current status of the computing and storage resources per site is presented. This includes the load of the site network, the number of jobs being run or waiting to be run, and the amount of total and available storage space at the site. If a particular site is selected, then several pieces of information regarding each one of the services present on each of the nodes of the site are shown. The nodes are classified as Resource Brokers, CE access nodes or SE access nodes.

There are also other types of views: Geo, Gris and VO views. The Geo view presents a geographical representation of the Grid. The Gris view shows current and historical information about the status (on or off) of every node. Finally, the VO view holds the same information as the site view, but here nodes are classified on a per VO basis. The user can specify a VO name, and get the data about all the nodes that support it.

Finally, the job monitoring section of GridICE provides figures about the number of jobs for each VO that are running or are queued at each Grid site.



## 6. WORKLOAD MANAGEMENT

### 6.1. INTRODUCTION

The *Workload Management System (WMS)* is the gLite 3 component that allows users to submit *jobs*, and performs all tasks required to execute them, without exposing the user to the complexity of the Grid. It is the responsibility of the user to describe his jobs and their requirements, and to retrieve the output when the jobs are finished.

In the WLCG/EGEE Grid, two different workload management systems are deployed: the legacy LCG-2 system, developed in the EDG project, and the new system from the EGEE project, which is an evolution of the former and therefore has more functionalities.

In the following sections, we will describe the basic concepts of the language used to describe a job, the basic command line interface to submit and manage simple jobs, a description of more advanced job types, details on how to configure the command line interface, and some user tools related to job management.

### 6.2. THE JOB DESCRIPTION LANGUAGE

The *Job Description Language (JDL)* is a high-level language based on the *Classified Advertisement (ClassAd) language* [33], used to describe jobs and aggregates of jobs with arbitrary dependency relations. The JDL is used in WLCG/EGEE to specify the desired job characteristics and constraints, which are taken into account by the WMS to select the best resource to execute the job.

The fundamentals of the JDL are given in this section. A complete description of the JDL syntax is out of the scope of this guide, and can be found in [35]. The description of the JDL attributes for the LCG-2 WMS is in [36], and for the gLite WMS is in [36][38].

A job description is a file (called *JDL file*) consisting of lines having the format:

*attribute = expression;*

Expressions can span several lines, but only the last one must be terminated by a semicolon. Literal strings are enclosed in double quotes. If a string itself contains double quotes, they must be escaped with a backslash (e.g.: `Arguments = "\"hello\" 10"`). The character “'” cannot be used in the JDL.

Comments must be preceded by a sharp character (#) or a double slash (//) at the beginning of each line. Multi-line comments must be enclosed between “/\*” and “\*/” .

**Attention!** The JDL is sensitive to blank characters and tabs. No blank characters or tabs should follow the semicolon at the end of a line.

**Example 6.2.1** (Define a simple job)

To define a job which runs the `hostname` command on the WN, write a JDL like this:

```
Executable = "/bin/hostname";
StdOutput = "std.out";
StdError = "std.err";
```

The `Executable` attribute specifies the command to be run by the job. If the command is already present on the WN, it must be expressed as a absolute path; if it has to be copied from the UI, only the file name must be specified, and the path of the command on the UI should be given in the `InputSandbox` attribute. For example:

```
Executable = "test.sh";
InputSandbox = {"/home/does/test.sh"};
StdOutput = "std.out";
StdError = "std.err";
```

The `Arguments` attribute can contain a string value, which is taken as argument list for the executable:

```
Arguments = "fileA 10";
```

In the `Executable` and in the `Arguments` attributes it may be necessary to use special characters, such as `&`, `\`, `|`, `>`, `<`. If these characters should be escaped in the shell (for example, if they are part of a file name), they should be preceded by triple `\` in the JDL, or specified inside quoted strings.

The attributes `StdOutput` and `StdError` define the name of the files containing the standard output and standard error of the executable, once the job output is retrieved.

For the standard input, an input file can be similarly specified:

```
StdInput = "std.in";
```

If files have to be copied from the UI to the execution node, they must be listed in the `InputSandbox` attribute:

```
InputSandbox = {"test.sh", "fileA", "fileB", ...};
```

Only the file specified as `Executable` will have automatically the execution flag: if other files in the input sandbox have such flag on the UI, they will lose it when copied to the WN.

Finally, the files to be transferred back to the UI after the job is finished can be specified using the `OutputSandbox` attribute:

```
OutputSandbox = {"std.out", "std.err"};
```

Wildcards are allowed only in the `InputSandbox` attribute. The list of files in the Input Sandbox is relative to the current directory in the UI. Absolute paths cannot be specified in the `OutputSandbox` attribute. The `InputSandbox` cannot contain two files with the same name, even if they have a different absolute path, as when transferred they would overwrite each other.

The shell environment of the job can be modified using the `Environment` attribute. For example:

```
Environment = {"CMS_PATH=$HOME/cms", "CMS_DB=$CMS_PATH/cmdb"};
```

The `VirtualOrganisation` attribute can be used to explicitly specify the VO of the user:

```
VirtualOrganisation = "cms";
```

but is superseded by the VO contained in the user proxy, if a VOMS proxy is used. For normal proxies, the VO can either be specified in the JDL, in the UI configuration files or as argument to the job submission command (see section 6.3.1).

**Note:** a common error is to write `VirtualOrganization`. It will not work.

To summarise, a typical JDL for a simple Grid job would look like:

```
Executable = "test.sh";  
Arguments = "fileA fileB";  
StdOutput = "std.out";  
StdError = "std.err";  
InputSandbox = {"test.sh", "fileA", "fileB"};  
OutputSandbox = {"std.out", "std.err"};
```

where `test.sh` could be:

```
#!/bin/sh  
echo "First file:"  
cat $1  
echo "Second file:"  
cat $2
```

In section 6.3.1 it is explained how to submit such job.

### Example 6.2.2 (Specifying requirements on the CE)

The `Requirements` attribute can be used to express constraints on the resources where the job should run. Its value is a Boolean expression that must evaluate to `true` for a job to run on that specific CE. For that purpose all the GLUE attributes of the IS can be used, by prepending the `other.` string to the attribute name. For a list of GLUE attributes, see Appendix G.

**Note:** Only one `Requirements` attribute can be specified (if there are more than one, only the last one is considered). If several conditions must be applied to the job, then they all must be combined in a single `Requirements` attribute.

For example, let us suppose that the user wants to run on a CE using PBS as batch system, and whose WNs have at least two CPUs. He will write then in the job description file:

```
Requirements = other.GlueCEInfoLRMSType == "PBS" && other.GlueCEInfoTotalCPUs > 1;
```

The WMS can be also asked to send a job to a particular queue in a CE with the following expression:

```
Requirements = other.GlueCEUniqueID == "lxshare0286.cern.ch:2119/jobmanager-pbs-short";
```

or to any queue in a CE:

```
Requirements = other.GlueCEInfoHostName == "lxshare0286.cern.ch";
```

**Note:** as explained in 6.5, normally the condition that a CE is in production state is automatically added to the `Requirements` attribute. Thus, CEs that do not correctly publish this will not match. This condition is, nevertheless, configurable.

If the job duration is significant, it is strongly advised to put a requirement on the maximum CPU time, or the wallclock time (expressed in minutes), needed for the job to complete. For example, to express the fact that the job needs at least eight CPU hours and twelve wallclock hours:

```
Requirements = other.GlueCEPolicyMaxCPUTime > 480 &&
                other.GlueCEPolicyMaxWallClockTime > 720;
```

**Note:** if a job exceeds the time limits of the queue where it is running, it will be killed by the batch system. Currently, the WMS does not correctly report to the user that the job failed due to exceeded time limits, and it cannot distinguish this case from an abrupt death of the job due to other causes.

**Note:** the CPU time needed by a job is inversely proportional to the “speed” of the CPU, which is expressed by the `GlueHostBenchmarkSI00` attribute. To take into account the differences in speed of the CPUs in different CEs, the CPU time should be rescaled to the speed. If, for example, the job needs 720 minutes on a CPU with a speed of 1000, the correct requirement should be

```
Requirements = other.GlueCEPolicyMaxCPUTime >
                (720 * 1000 / other.GlueHostBenchmarkSI00);
```

If the job must run on a CE where a particular experiment software is installed and this information is published by the CE, something like the following must be written:

```
Requirements = Member("CMSIM-133", other.GlueHostApplicationSoftwareRunTimeEnvironment);
```

**Note:** the `Member` operator is used to test if its first argument (a scalar value) is a member of its second argument (a list). In fact, the `GlueHostApplicationSoftwareRunTimeEnvironment` attribute is a list of strings and is used to publish any VO-specific information relative to the CE (typically, information on the VO software available on that CE).

### *Example 6.2.3 (Specifying requirements using wildcards)*

It is also possible to use regular expressions when expressing a requirement. Let us suppose for example that the user wants all his jobs to run on any CE in the domain `cern.ch`. This can be achieved putting in the JDL file the following expression:

```
Requirements = RegExp("cern.ch", other.GlueCEUniqueID);
```

The opposite can be required by using:

```
Requirements = (!RegExp("cern.ch", other.GlueCEUniqueID));
```

### *Example 6.2.4 (Specifying requirements on a close SE)*

In order to specify requirements on the SE “close” to the CE where the job should run, the WMS uses a special match-making mechanism, called *gang-matching*[37]. For example, to ensure that the job runs on a CE with at least 200 MB of free disk space on a close SE, the following JDL expression can be used:

```
Requirements = anyMatch(other.storage.CloseSEs, target.GlueSASStateAvailableSpace > 204800);
```

**Attention!** At the time of writing, the *gang-matching* must **not** be used to submit jobs to the gLite WMS: due to a bug, this would block the WMS for several hours. However, it is still possible to require, for example, to run on a CE close to a given SE by using an expression like:

```
Member("castorsrm.pic.es", other.GlueCESEBindGroupSEUniqueID);
```

It is not possible, though, to write down requirements on SE properties other than their name. As a last note, the above requirement will not work with the LCG-2 WMS.

**Example 6.2.5** (A complex requirement used in gLite 3)

The following example has been actually used by the *alice* VO in order to find a CE that has some software packages installed (VO-*alice-AliEn* and VO-*alice-ALICE-v4-01-Rev-01*), and that allows the job to run for up to one day (i.e., so that the job is not aborted before it has time to finish).

```
Requirements = other.GlueHostNetworkAdapterOutboundIP==true &&
Member("VO-alice-AliEn", other.GlueHostApplicationSoftwareRunTimeEnvironment) &&
Member("VO-alice-ALICE-v4-01", other.GlueHostApplicationSoftwareRunTimeEnvironment) &&
(other.GlueCEPolicyMaxWallClockTime > 1440 );
```

**Example 6.2.6** (Using the automatic resubmission)

It is possible to have the WMS automatically resubmitting jobs which, for some reason, are aborted by the Grid. Two kinds of resubmission are available for the gLite 3 WMS: the **deep resubmission** and the **shallow resubmission** (only the former is available in the LCG-2 WMS). The resubmission is deep when the job fails after it has started running on the WN, and shallow otherwise.

The user can limit the number of times the WMS should resubmit a job by using the JDL attributes `RetryCount` and `ShallowRetryCount` for the deep and shallow resubmission respectively. For example, to disable the deep resubmission and limit the attempts of shallow resubmission to 3:

```
RetryCount = 0;
ShallowRetryCount = 3;
```

It is advisable to disable the deep resubmission, as it may happen that a job fails after it has already done something (for example, creating a Grid file), or the WMS thinks that a still running job has failed; depending on the job, the resubmission of an identical job might generate inconsistencies. On the other hand, the shallow resubmission is extremely useful to improve the chances of a job being correctly executed, and it is strongly recommended to use it.

The values of the `MaxRetryCount` and `MaxShallowRetryCount` parameters in the WMS configuration file represent both the default and the maximum limits for the number of resubmissions.

**Example 6.2.7** (Using the automatic proxy renewal)

The proxy renewal feature of the WMS is automatically enabled, as long as the user has stored a long-term proxy in the default MyProxy server (usually defined in the `MYPROXY_SERVER` environment variable for the MyProxy client commands, and in the UI configuration for the WMS commands). However it is possible to indicate to the WMS a different MyProxy server in the JDL file:

```
MyProxyServer = "myproxy.cern.ch";
```

The proxy renewal can be disabled altogether by adding to the JDL:

```
MyProxyServer = "";
```

### Example 6.2.8 (Defining the “goodness” of a CE)

The choice of the CE where to execute the job, among all the ones satisfying the requirements, is based on the *rank* of the CE, a quantity expressed as a floating-point number. The CE with the highest rank is the one selected.

By default, the rank is equal to `-other.GlueCEStateEstimatedResponseTime`, where the estimated response time is an estimation of the time interval between the job submission and the beginning of the job execution. However, the user can redefine the rank with the `Rank` attribute as a function of the CE attributes. For example:

```
Rank = other.GlueCEStateFreeCPUs;
```

which will rank best the CE with the most free CPUs. The next one is a more complex expression:

```
Rank = ( other.GlueCEStateWaitingJobs == 0 ? other.GlueCEStateFreeCPUs :
  -other.GlueCEStateWaitingJobs );
```

In this case, the selected CE will be the one with the least waiting jobs, or, if there are no waiting jobs, the one with the most free CPUs.

## 6.3. THE COMMAND LINE INTERFACE

In this section, all commands available for the user to manage jobs are described. For completeness, both the gLite CLI [28][27] and the LCG-2 CLI [29] are described.

The gLite WMS implements two different services to manage jobs: the *Network Server* and the *WMProxy*. The Network Server is the same service used in the LCG-2 WMS, and offers basically the same functionalities in both systems. The WMProxy, on the contrary, implements several new functionalities, among which:

- submission of job collections;
- faster authentication;
- faster match-making;
- faster response time for users;
- higher job throughput.

The three job management systems (the LCG-2 WMS, the gLite WMS via Network Server and the gLite WMS via WMProxy) offer each one a different set of commands with very similar functionalities and syntax. The following table summarizes these commands with their most commonly used options, and the use of these commands will be described in the following sections.

Function	LCG-2 WMS	gLite WMS via NS	gLite WMS via WMProxy
Submit a job	edg-job-submit [-o joblist] jdlfile	glite-job-submit [-o joblist] jdlfile	glite-wms-job-submit [-d delegID] [-a] [-o joblist] jdlfile
See job status	edg-job-status [-v verbosity] [-i joblist] jobIDs	glite-job-status [-v verbosity] [-i joblist] jobIDs	glite-wms-job-status [-v verbosity] [-i joblist] jobIDs
See job logging information	edg-job-get-logging-info [-v verbosity] [-i joblist] jobIDs	glite-job-logging-info [-v verbosity] [-i joblist] jobIDs	glite-wms-job-logging-info [-v verbosity] [-i joblist] jobIDs
Retrieve job output	edg-job-get-output [-dir outdir] [-i joblist] jobIDs	glite-job-output [-dir outdir] [-i joblist] jobIDs	glite-wms-job-output [-dir outdir] [-i joblist] jobIDs
Cancel a job	edg-job-cancel [-i joblist] jobID	glite-job-cancel [-i joblist] jobID	glite-wms-job-cancel [-i joblist] jobID
List available resources	edg-job-list-match jdlfile	glite-job-list-match jdlfile	glite-wms-job-list-match [-d delegID] [-a] jdlfile
Delegate proxy			glite-wms-job-delegate-proxy -d delegID

**Attention!** The recommended method to manage jobs is through the gLite WMS via WMProxy, because it gives the best performance and allows to use the most advanced functionalities.

### 6.3.1. Single Job Submission

To submit a job to the WLCG/EGEE Grid, the user must have a valid proxy certificate in the User Interface machine (as described in Chapter 4) and use one of the following commands:

```
glite-wms-job-submit -a jdlfile          (gLite WMS via WMProxy)
glite-job-submit jdlfile                 (gLite WMS via NS)
edg-job-submit jdlfile                   (LCG-2 WMS)
```

where `jdlfile` is a file containing the job description, usually with extension `.jdl`. The `-a` option for the WM-Proxy command is necessary to automatically delegate a user proxy to the WMProxy server (see later).

#### *Example 6.3.1.1 (Submitting a simple job)*

Create a file `test.jdl` with this content:

```
Executable = "/bin/hostname";
StdOutput = "std.out";
StdError = "std.err";
OutputSandbox = {"std.out", "std.err"};
```

It describes a simple job that will execute `/bin/hostname`. Standard output and standard error are redirected to the files `std.out` and `std.err` respectively; the `OutputSandbox` attribute ensures that they are transferred back to the User Interface after the job is finished.

Now, submit the job via WMProxy by doing:



```
$ glite-wms-job-submit -a test.jdl
```

If the submission is successful, the output is similar to:

```
Connecting to the service https://rb102.cern.ch:7443/glite_wms_wmproxy_server
```

```
===== glite-wms-job-submit Success =====
```

```
The job has been successfully submitted to the WMPProxy  
Your job identifier is:
```

```
https://rb102.cern.ch:9000/vZKKk3gdBla6RySximq_vQ
```

```
=====
```

In case of failure, an error message will be displayed and an exit status different from zero will be returned.

The command returns to the user the job identifier (*jobID*), which uniquely defines the job and can be used to perform further operations on the job, like interrogating the system about its status, or canceling it. The format of the jobID is:

```
https://<LB_hostname>[:<port>]/<unique_string>
```

where <unique\_string> is guaranteed to be unique and <LB\_hostname> is the host name of the Logging and Bookkeeping (LB) server for the job, which usually sits on the WMS used to submit the job.

**Note:** the jobID does **not** identify a web page.

**Note:** to submit jobs via WMPProxy, it is required to have a VOMS proxy, as with a standard proxy the submission will fail with an error like:

```
Error - Operation failed  
Unable to delegate the credential to the endpoint:  
https://rb102.cern.ch:7443/glite_wms_wmproxy_server  
User not authorized:  
unable to check credential permission (/opt/glite/etc/glite_wms_wmproxy.gacl)  
(credential entry not found)  
credential type: person  
input dn: /C=CH/O=CERN/OU=GRID/CN=John Doe
```

If the command returns the following error:

```
Error - WMPProxy Server Error  
LCMAPS failed to map user credential
```

Method: getFreeQuota  
Error code: 1208

it means that there are authentication problems between the UI and the WMPProxy server (you may not be authorized to use that WMPProxy server).

If the job is submitted via Network Server (both to an LCG-2 or a gLite WMS), an authentication problem will produce instead the following message:

```

**** Warning: API_NATIVE_ERROR ****
Error while calling the "NSClient::multi" native api
AuthenticationException: Failed to establish security context...

**** Error: UI_NO_NS_CONTACT ****
Unable to contact any Network Server
  
```

Many options are available to the job submission commands.

If the user proxy does not have VOMS extensions (allowed only for submission via Network Server), the user can specify his VO with the `--vo <vo_name>` option; otherwise the default VO is taken from the standard configuration files (see 6.5). If a VO name is not specified anywhere, at submission time this error will be returned:

```

**** Error: UI_NO_VOMS ****
Unable to determine a valid user's VO
  
```

The `-o <file_path>` option allows users to specify a file to which the jobID of the submitted job will be appended. This file can be given to other job management commands to perform operations on more than one job with a single command, and it is a convenient way to keep trace of one's jobs.

The `-r <CEId>` option is used to directly send a job to a particular CE. If used, the match making will not be carried out (see Section 6.3.5). The drawback is that the BrokerInfo file, which provides information about the evolution of the job, will not be created, and therefore the use of this option is discouraged.

A CE is identified by `<CEId>`, which is a string with the following format:

```

<CE_hostname>:<port>/jobmanager-<service>-<queue>           (for a LCG CE)
<CE_hostname>:<port>/blah-<service>-<queue>                 (for a gLite CE)
  
```

where `<CE_hostname>` and `<port>` are the host name of the machine and the port where the Grid Gate is running (the Globus Gatekeeper for the LCG CE and CondorC+BLAH for the gLite CE), `<queue>` is the name of one of the corresponding LRMS queue, and `<service>` is the LRMS type, such as `lsf`, `pbs`, `condor`. Examples of CEId are:

```

adc0015.cern.ch:2119/jobmanager-lcgpbs-infinite             (LCG CE)
prep-ce-01.pd.infn.it:2119/blah-lsf-atlas                  (gLite CE)
  
```

**Note:** The LCG-2 WMS is able to submit jobs only to LCG Computing Elements. On the other hand, the gLite WMS is capable to submit jobs to both the LCG and gLite CEs.



where <delegID> is a string chosen by the user. Subsequent invocations of `glite-wms-job-submit` and `glite-wms-job-list-match` can bypass the delegation of a new proxy if the same <delegID> is given to the `-d` option. For example, to delegate a proxy:

```
$ glite-wms-job-delegate-proxy -d mydelegID
```

```
Connecting to the service https://rb102.cern.ch:7443/glite_wms_wmproxy_server
```

```
===== glite-wms-job-delegate-proxy Success =====
```

```
Your proxy has been successfully delegated to the WMPProxy:  
https://rb102.cern.ch:7443/glite_wms_wmproxy_server
```

```
with the delegation identifier: mydelegID
```

```
=====
```

Alternatively, we can have the system to generate a random <delegID> by doing instead:

```
$ glite-wms-job-delegate-proxy -a
```

```
Connecting to the service https://rb102.cern.ch:7443/glite_wms_wmproxy_server
```

```
===== glite-wms-job-delegate-proxy Success =====
```

```
Your proxy has been successfully delegated to the WMPProxy:  
https://rb102.cern.ch:7443/glite_wms_wmproxy_server
```

```
with the delegation identifier: 2cBscH0taSqCYcH8fNYncw
```

```
=====
```

Then, to submit a job:

```
$ glite-wms-job-submit -d mydelegID test.jdl
```

**Note:** due to a current bug, if many WMPProxy servers are indicated in the UI configuration, `glite-wms-job-delegate-proxy` will delegate a proxy to one of them chosen at random, not to all of them. This limits the usability of the explicit delegation.

### 6.3.2. Job Operations

After a job is submitted, it is possible to see its current status, to retrieve a complete log of the job history, to recover its output when it is finished, and if needed to cancel it if it has not yet finished. The following examples

explain how.

**Example 6.3.2.1 (Retrieving the status of a job)**

Given a submitted job whose job identifier is <jobID>, the command is:

<code>glite-wms-job-status &lt;jobID&gt;</code>	(gLite WMS via WMPProxy)
<code>glite-job-status &lt;jobID&gt;</code>	(gLite WMS via NS)
<code>edg-job-status &lt;jobID&gt;</code>	(LCG-2 WMS)

and an example of a possible output from the gLite LB is:

```
$ glite-wms-job-status https://rb102.cern.ch:9000/fNdD4FW_Xxkt2s2aZJeoeg
```

```
*****
BOOKKEEPING INFORMATION:

Status info for the Job : https://rb102.cern.ch:9000/fNdD4FW_Xxkt2s2aZJeoeg
Current Status:      Done (Success)
Exit code:           0
Status Reason:       Job terminated successfully
Destination:         cel.inrne.bas.bg:2119/jobmanager-lcgpbs-cms
Submitted:           Mon Dec  4 15:05:43 2006 CET
*****
```

which contains the time when the job was submitted, the current status of the job, and the reason for being in that status (which may be especially helpful for the ABORTED status). The possible states in which a job can be found were introduced in Section 3.4.1, and are summarized in Appendix C. Finally, the `Destination` field contains the CEId of the CE where the job has been submitted and the job exit code, if the job is finished.

The verbosity level controls the amount of information provided. The value of the `-v` option ranges from 0 to 3 (the default is configurable in the UI). See [29] for detailed information on each of the fields returned.

The commands to get the job status can have several jobIDs as arguments, i.e.:

```
glite-wms-job-status <jobID1> ... <jobIDN>
```

or, more conveniently, the `-i <file_path>` option can be used to specify a file with a list of jobIDs (possibly created by the `-o` option of a job submission command). In this case, the command asks the user interactively the status of which job(s) should be printed. The `--noinput` option suppresses the interactivity and all the jobs are considered.

If the `--all` option is used instead, the status of all the jobs owned by the user submitting the command is retrieved. As the number of jobs owned by a single user may be large, there are some options that limit that job selection. The `--from/--to [MM:DD:]hh:mm[:[CC]YY]` options make the command query LB for jobs that were

submitted after/before the specified date and time. The `--status <status>` option makes the command retrieve only the jobs that are in the specified status, and the `--exclude <status>` option makes it retrieve jobs that are not in the specified status. This two last options are mutually exclusive, although they can be used with `--from` and `--to`.

In the following examples, the first command retrieves all jobs of the user that are in the status DONE or RUNNING, and the second retrieves all jobs that were submitted before the 17:35 of the current day, and that were not in the CLEARED status.

```
$ glite-wms-job-status --all -s DONE -s RUNNING
$ glite-wms-job-status --all -e CLEARED --to 17:35
```

**Note:** for the `--all` option to work, it is necessary that an index by owner is created in the LB server; otherwise, the command will fail, since it will not be possible for the LB server to identify the user's jobs. Such index can only be created by the LB server administrator, as explained in [29].

Finally, with the option `-o <file_path>` the command output can be written to a file.

### Example 6.3.2.2 (Canceling a job)

A job can be canceled before it ends using the commands

```
glite-wms-job-cancel <jobID>           (gLite WMS via WMPProxy)
glite-job-cancel <jobID>               (gLite WMS via NS)
edg-job-cancel <jobID>                 (LCG-2 WMS)
```

A job must be canceled using the command corresponding to the WMS flavour used to submit the job.

These commands require as arguments one or more JobIDs. For example:

```
$ glite-wms-job-cancel https://rb102.cern.ch:9000/P1c60RFsrIZ9mnBALa7yZA
glite-wms-job-cancel https://rb102.cern.ch:9000/P1c60RFsrIZ9mnBALa7yZA

Are you sure you want to remove specified job(s) [y/n]y : y

Connecting to the service https://128.142.160.93:7443/glite_wms_wmproxy_server

===== glite-wms-job-cancel Success =====

The cancellation request has been successfully submitted for the following job(s):

- https://rb102.cern.ch:9000/P1c60RFsrIZ9mnBALa7yZA

=====
```

If the cancellation is successful, the job will terminate in status CANCELLED.

**Example 6.3.2.3 (Retrieving the output of a job)**

If the job has successfully finished (it has reached the DONE status), its output can be copied to the UI with the commands

<code>glite-wms-job-output &lt;jobID&gt;</code>	(gLite WMS via WMProxy)
<code>glite-job-output &lt;jobID&gt;</code>	(gLite WMS via NS)
<code>edg-job-get-output &lt;jobID&gt;</code>	(LCG-2 WMS)

The job output must be retrieved using the command corresponding to the WMS flavour used to submit the job.

For example:

```
$ glite-wms-job-output https://rb102.cern.ch:9000/yabp72aERhofLA6W2-LrJw
```

```
Connecting to the service https://128.142.160.93:7443/glite_wms_wmproxy_server
```

```
=====
```

```
JOB GET OUTPUT OUTCOME
```

```
Output sandbox files for the job:
https://rb102.cern.ch:9000/yabp72aERhofLA6W2-LrJw
have been successfully retrieved and stored in the directory:
/tmp/dae_yabp72aERhofLA6W2-LrJw
```

```
=====
```

The default location for storing the outputs (normally /tmp) is defined in the UI configuration, but it is possible to specify in which directory to save the output using the `--dir <path_name>` option.

**Note:** the output of a job will be removed from the WMS machine after a certain period of time. How long this period is may vary depending on the administrator of the WMS, but the currently suggested time is 10 days, so users should try always to retrieve their jobs within one week after job completion (to have a safe margin).

**Example 6.3.2.4 (Retrieving logging information about submitted jobs)**

A complete history of a job is permanently stored in the Logging & Bookkeeping service and can be retrieved using the commands:

<code>glite-wms-job-logging-info &lt;jobID&gt;</code>	(gLite WMS via WMProxy)
<code>glite-job-logging-info &lt;jobID&gt;</code>	(gLite WMS via NS)
<code>edg-job-get-logging-info &lt;jobID&gt;</code>	(LCG-2 WMS)

This functionality is especially useful in the analysis of job failures, although the information provided is sometimes difficult to interpret.

The argument of this command is a list of one or more job identifiers. The `-i` and `-o` options work as in the previous commands.

The following is the typical sequence of events for a successful job:

```
$ glite-wms-job-logging-info https://rb102.cern.ch:9000/hk0VSbNhe59j0Buo24G_qw
```

```
*****
```

```
LOGGING INFORMATION:
```

```
Printing info for the Job : https://rb102.cern.ch:9000/hk0VSbNhe59j0Buo24G_qw
```

```

    ---
Event: RegJob
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:03 2006 CET
    ---
Event: RegJob
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:04 2006 CET
    ---
Event: UserTag
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:04 2006 CET
    ---
Event: UserTag
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:04 2006 CET
    ---
Event: UserTag
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:04 2006 CET
    ---
Event: Accepted
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:08 2006 CET
    ---
Event: EnQueued
- result           = START
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:08 2006 CET
    ---
Event: EnQueued
- result           = OK
- source           = NetworkServer
- timestamp        = Thu Dec 14 14:35:08 2006 CET
    ---
Event: DeQueued
- source           = WorkloadManager

```



```

- timestamp          = Thu Dec 14 14:35:09 2006 CET
  ---
Event: Match
- dest_id           = fangorn.man.poznan.pl:2119/jobmanager-lcgpbs-cms
- source            = WorkloadManager
- timestamp         = Thu Dec 14 14:35:18 2006 CET
  ---
Event: EnQueued
- result            = START
- source            = WorkloadManager
- timestamp         = Thu Dec 14 14:35:18 2006 CET
  ---
Event: EnQueued
- result            = OK
- source            = WorkloadManager
- timestamp         = Thu Dec 14 14:35:18 2006 CET
  ---
Event: DeQueued
- source            = JobController
- timestamp         = Thu Dec 14 14:35:18 2006 CET
  ---
Event: Transfer
- destination       = LogMonitor
- result            = START
- source            = JobController
- timestamp         = Thu Dec 14 14:35:18 2006 CET
  ---
Event: Transfer
- destination       = LogMonitor
- result            = OK
- source            = JobController
- timestamp         = Thu Dec 14 14:35:19 2006 CET
  ---
Event: Accepted
- source            = LogMonitor
- timestamp         = Thu Dec 14 14:35:29 2006 CET
  ---
Event: Transfer
- destination       = LRMS
- result            = OK
- source            = LogMonitor
- timestamp         = Thu Dec 14 14:35:50 2006 CET
  ---
Event: Running
- source            = LogMonitor
- timestamp         = Thu Dec 14 14:38:09 2006 CET
  ---
Event: ReallyRunning
- source            = LogMonitor

```

```
- timestamp          = Thu Dec 14 14:41:26 2006 CET
  ---
Event: Done
- source            = LogMonitor
- timestamp          = Thu Dec 14 14:41:26 2006 CET
```

\*\*\*\*\*

**Note:** in order to make easier to debug problems with the WMS, when asking for help for problems related to job submission and management, it is highly advisable to send the output of

```
glite-wms-job-logging-info -v 3 <jobID>
```

that is, using the highest level of verbosity.

### 6.3.3. Advanced Sandbox Management

A new feature introduced by the gLite WMS is the possibility to indicate input sandbox files stored not on the UI, but on a GridFTP server, and, similarly, to specify that output files should be transferred to a GridFTP server when the job finishes. This has several advantages:

- the input files do not have to be on the host from which the job is submitted;
- the output files are immediately available when the job ends, without having to issue a command to retrieve them;
- the sandbox files do not have to go through the WMS host, which otherwise can easily become a bottleneck.

The following examples show how to use this feature.

#### *Example 6.3.3.1 (Using input files on a GridFTP server)*

If the job input files are stored on a GridFTP server, it is possible to specify those files as GridFTP URI in the `InputSandbox` attribute:

```
InputSandbox = {"gsiftp://lxb0707.cern.ch/cms/does/data/fileA",
               "fileB"};
```

where `fileA` is located on the GridFTP server and `fileB` in the current directory on the UI.

It is also possible to specify a base GridFTP URI with the attribute `InputSandboxBaseURI`: in this case, files expressed as simple file names or as relative paths will be looked for under that base URI. Local files can still be defined using the `file://<path>` URI format. For example:

```

InputSandbox = {"fileA", "data/fileB", "file:///home/does/fileC"};
InputSandboxBaseURI = "gsiftp://lxb0707.cern.ch/cms/does";
  
```

is equivalent to

```

InputSandbox = {"gsiftp://lxb0707.cern.ch/cms/does/fileA",
                "gsiftp://lxb0707.cern.ch/cms/does/data/fileB",
                "/home/does/fileC"};
  
```

### **Example 6.3.3.2** (Storing output files in a GridFTP server)

In order to store the output sandbox files to a GridFTP server, the `OutputSandboxDestURI` attribute must be used together with the usual `OutputSandbox` attribute. The latter is used to list the output files created by the job in the WN to be transferred, and the former is used to express where the output files are to be transferred. For example:

```

OutputSandbox = {"fileA", "data/fileB", "fileC"};
OutputSandboxDestURI = {"gsiftp://lxb0707.cern.ch/cms/does/fileA",
                        "gsiftp://lxb0707.cern.ch/cms/does/fileB",
                        "fileC"};
  
```

where the first two files have to be copied to a GridFTP server, while the third file will be copied back to the WMS with the usual mechanism. Clearly, `glite-wms-job-output` will retrieve only the third file.

Another possibility is to use the `OutputSandboxBaseDestURI` attribute to specify a base URI on a GridFTP server where the files listed in `OutputSandbox` will be copied. For example:

```

OutputSandbox = {"fileA", "fileB"};
OutputSandboxBaseDestURI = "gsiftp://lxb0707.cern.ch/cms/does/";
  
```

will copy both files under the specified GridFTP URI.

**Note:** the directory on the GridFTP where the files have to be copied must already exist.

### **6.3.4. Real Time Output Retrieval**

It is possible to see the files produced by a job while it is still running by using the *Job Perusal* functionality, only available via WMPProxy. For the LCG-2 WMS a set of commands providing the same functionality can be used, although limited to the files containing the standard output and the standard error.

### **Example 6.3.4.1** (Inspecting the job output in real time with WMPProxy)

The user can enable the job perusal by setting the attribute `PerusalFileEnable` to true in the job JDL. This makes the WN to upload, at regular time intervals (defined by the `PerusalTimeInterval` attribute and expressed in seconds), a copy of the output files specified using the `glite-wms-job-perusal` command to the WMS machine (by default), or to a GridFTP server specified by the attribute `PerusalFilesDestURI`.

The following example shows how to use the job perusal. The JDL file should look like this:

```
Executable = "job.sh";
StdOutput = "stdout.log";
StdError = "stderr.log";
InputSandbox = {"job.sh"};
OutputSandbox = {"stdout.log", "stderr.log", "testfile.txt"};
PerusalFileEnable = true;
PerusalTimeInterval = 30;
RetryCount = 0;
```

After the job has been submitted with `glite-wms-job-submit`, the user can choose which output files should be inspected:

```
$ glite-wms-job-perusal --set -f stdout.log -f stderr.log -f testfile.txt \
https://rb102.cern.ch:9000/B02xR3EQg9ZHHoRc-1nJkQ

Connecting to the service https://128.142.160.93:7443/glite_wms_wmproxy_server

Connecting to the service https://128.142.160.93:7443/glite_wms_wmproxy_server

===== glite-wms-job-perusal Success =====

Files perusal has been successfully enabled for the job:
https://rb102.cern.ch:9000/B02xR3EQg9ZHHoRc-1nJkQ

=====
```

and, when the job starts, the user can see one output file:

```
$ glite-wms-job-perusal --get -f testfile.txt \
https://rb102.cern.ch:9000/B02xR3EQg9ZHHoRc-1nJkQ

Connecting to the service https://137.138.45.79:7443/glite_wms_wmproxy_server

Connecting to the service https://137.138.45.79:7443/glite_wms_wmproxy_server

===== glite-wms-job-perusal Success =====
```

The retrieved files have been successfully stored in:  
 /tmp/dae\_OoDVmWCAnhX\_HiSPvASGsg

```

=====
-----
file 1/1: testfile.txt-20061220115405_1-20061220115405_1
-----
  
```

```

This is a test file
Data : Wed Dec 20 11:53:37 CET 2006
Host : c01-017-103
  
```

Subsequent invocations of `glite-wms-job-perusal --get` will retrieve only the part of the file that was written after the previous invocation. To have the complete file, the `--all` option can be used. Only one file can be retrieved at a time. Finally, the job perusal can be disabled for all jobs using the `--unset` option.

**Example 6.3.4.2 (Inspecting the job output in real time with the LCG-2 WMS)**

A separate tool can be used to access the standard output and the standard error of a running job while it is still running.

The tool consists of three commands:

- `grid-stdout-mon`: this command runs on the WN at the same time that the user job (and must be invoked at the beginning of the job script). Every 10, 20, 30 minutes and then every hour it copies the partial standard output and standard error files to a SE. This SE can be specified in the JDL by setting the environment variable `LCG_STDOUT_MON_SE`; if it is not defined by the user, it defaults to the value of `DPM_HOST` in the WN. The upload mechanism is enabled by setting the environment variable `LCG_STDOUT_MON_FLAG` to `ON_DEMAND` in the JDL file. The actual upload of the output and error files begins when the user executes the command `grid-stdout-mon-on`;
- `grid-stdout-mon-on`: this command must be run from the UI to create a special file and a directory for each job to be monitored on the SE where the outputs are to be uploaded. It has to be executed after the job submission to allow the job outputs to be stored in the SE. If the upload has been enabled in the JDL, `grid-stdout-mon` will upload the job outputs only if the above special file exists.
- `grid-stdout-mon-get`: this command can be run from the UI to retrieve the job output from a specified SE. It has to be explicitly re-run to refresh the outputs while the jobs are running.

More information on this tool can be found in the User level tools Wiki[46].

### 6.3.5. The BrokerInfo

The *BrokerInfo file* is a mechanism to access, at job execution time, certain information concerning the job, for example the name of the CE, the files specified in the `InputData` attribute, the SEs where they can be found, etc.

The `BrokerInfo` file is created in the job working directory (that is, the current directory on the WN for the executable) and is named `.BrokerInfo`. Its syntax is based on Condor ClassAds and the information contained is not easy to read; however, it is possible to get it by means of a CLI, whose description follows.

**Note:** remember that using the `-r <CEId>` option of the job submission commands prevents the creation of the `BrokerInfo` file.

The commands to print the information in the `BrokerInfo` file are

```

glite-brokerinfo           (gLite WMS)
edg-brokerinfo             (LCG-2 WMS)
  
```

A detailed description of these commands and of the corresponding API can be found in [39].

The `glite-brokerinfo` command has the following syntax:

```
glite-brokerinfo [-v] [-f filename] function [parameter] [parameter] ...
```

where `function` is one of the following:

- `getCE`: returns the name of the CE where the job is running;
- `getDataAccessProtocol`: returns the protocol list specified in the `DataAccessProtocol` JDL attribute;
- `getInputData`: returns the file list specified in the `InputData` JDL attribute;
- `getSEs`: returns the list of the SEs with contain a copy of at least one file among those specified in `InputData`;
- `getCloseSEs`: returns a list of the SEs close to the CE where the job is running;
- `getSEMountPoint <SE>`: returns the access point for the specified close SE;
- `getSEFreeSpace <SE>`: returns the free space on the SE;
- `getLFN2SFN <LFN>`: returns the SURL of the specified LFN, listed in the `InputData` attribute;
- `getSEProtocols <SE>`: returns the list of the protocols available to transfer data in the specified SE;
- `getSEPort <SE> <Protocol>`: returns the port number used by the SE for the specified data transfer protocol;
- `getVirtualOrganisation`: returns the name of the VO specified in the JDL.

The `-v` option produced a more verbose output, and the `-f <filename>` option tells the command to parse the `BrokerInfo` file specified by `<filename>`. If the `-f` option is not used, the command tries to parse the file `$GLITE_WL_RB_BROKERINFO` or the file `./BrokerInfo`.

The `edg-brokerinfo` has exactly the same syntax, but the environment variable to specify the location of the `BrokerInfo` file is `EDG_WL_RB_BROKERINFO`.

## 6.4. ADVANCED JOB TYPES

This section describes how to use more advanced job types.

### 6.4.1. Job Collections

One of the most useful functionalities of WMPProxy is the ability to submit job *collections*, defined as sets of independent of jobs. This greatly speeds up the job submission time, compared to the submission of individual jobs, and together with the proxy delegation mechanisms, it saves a lot of processing time by reusing the same authentication for all the jobs in the collection.

#### *Example 6.4.1.1 (Submitting a job collection)*

The simplest way to submit a collection is to put the JDL files of all the jobs in the collection in a single directory, and use the `--collection <dirname>`, where `<dirname>` is the name of that directory. For example, suppose that there are two JDL files in the `jdl` directory

```
$ ls -l jdl/
job1.jdl  job2.jdl
```

We can submit both jobs at the same time by doing:

```
$ glite-wms-job-submit -a --collection jdl
```

```
Connecting to the service https://rb102.cern.ch:7443/glite_wms_wmproxy_server
```

```
===== glite-wms-job-submit Success =====
```

```
The job has been successfully submitted to the WMPProxy
Your job identifier is:
```

```
https://rb102.cern.ch:9000/n1JoZ8WbyJBrW3-pTU3f4A
```

```
=====
```

The jobID returned refers to the collection itself. To know the status of the collection and of all the jobs belonging to it, it is enough to use `glite-wms-job-status` as for any other kind of job:

```
$ glite-wms-job-status https://rb102.cern.ch:9000/n1JoZ8WbyJBrW3-pTU3f4A
```

\*\*\*\*\*

BOOKKEEPING INFORMATION:

```
Status info for the Job : https://rb102.cern.ch:9000/n1JoZ8WbyJBrW3-pTU3f4A
Current Status:      Done (Exit Code !=0)
Exit code:          1
Status Reason:      Warning: job exit code != 0
Destination:        dagman
Submitted:          Thu Dec 14 18:26:42 2006 CET
```

\*\*\*\*\*

- Nodes information:

```
Status info for the Job : https://rb102.cern.ch:9000/1SugblV08Ge30nIW07FAYw
Node Name:          job1_jd1
Current Status:     Done (Success)
Exit code:          0
Status Reason:      Job terminated successfully
Destination:        arxiloxos1.inp.demokritos.gr:2119/jobmanager-lcgpbs-cms
Submitted:          Thu Dec 14 18:26:42 2006 CET
```

\*\*\*\*\*

```
Status info for the Job : https://rb102.cern.ch:9000/_3SvxrOLsg5L5QVZMmzTrg
Node Name:          job2_jd1
Current Status:     Aborted
Status Reason:      hit job shallow retry count (0)
Destination:        ce-lcg.sdg.ac.cn:2119/jobmanager-lcgpbs-cms
Submitted:          Thu Dec 14 18:26:42 2006 CET
```

\*\*\*\*\*

In this example, one job succeeded and one failed, which explains why the status of the collection itself reports and exit code different from zero.

**Note:** executing `glite-wms-job-status` for the collection is the only way to know the jobIDs of the jobs in the collection.

The behaviour of the other job management commands is as follows:

- `glite-wms-job-output <collID>` retrieves the output of all the jobs in the collection `<collID>` which finished correctly;
- `glite-wms-job-cancel <collID>` cancels all the jobs in the collection;
- `glite-wms-job-logging-info <collID>` returns the logging information for the collection, but not for the jobs which compose it.

Once the jobIDs of the single jobs are known, the job management commands can be used with them exactly as for any other job.



### Example 6.4.1.2 (Advanced collections)

A more flexible way to define a job collection is illustrated in the following JDL file. Its structure includes a global set of attributes, which are inherited by all the sub-jobs, and a set of attributes for each sub-job, which supersede the global ones.

```
[
  Type = "Collection";
  VirtualOrganisation = "cms";
  MyProxyServer = "myproxy.cern.ch";
  InputSandbox = {"myjob.exe", "fileA"};
  OutputSandboxBaseDestURI = "gsiftp://lxb0707.cern.ch/data/does";
  DefaultNodeShallowRetryCount = 5;
  Requirements = Member("VO-cms-CMSSW_1_2_0",
    other.GlueHostApplicationSoftwareRunTimeEnvironment");
  Nodes = [
    [
      Executable = "myjob.exe";
      InputSandbox = {root.InputSandbox,
        "fileB"};
      OutputSandbox = {"myoutput1.txt"};
      Requirements = other.GlueCEPolicyMaxWallClockTime > 1440;
    ],
    [
      NodeName = "mysubjob";
      Executable = "myjob.exe";
      OutputSandbox = {"myoutput2.txt"};
      ShallowRetryCount = 3;
    ],
    [
      File = "/home/does/test.jdl";
    ]
  ]
]
```

The interpretation of this JDL file is as follows:

- it describes a collection (`Type = "Collection";`);
- the jobs belong to the *cms* VO;
- the MyProxy server to use for proxy renewal is `myproxy.cern.ch`;
- all the jobs in the collection have by default the executable `myjob.exe` and the file `fileA` in their sandbox (*shared input sandbox*);
- all the output files must be copied to a GridFTP server;
- the default maximum number of shallow resubmissions is 5;
- all the jobs must run on CEs with a given version of a software (`CMSSW_1_2_0`);

- the input sandbox of the first job (or node) has all the default files (`root.InputSandbox`), plus an additional file, `fileB`, while the second job has only the default files;
- the first job must run on a CE allowing at least one day of wallclock time;
- the second job has a limit of 3 for the number of shallow resubmissions;
- the third job is described by another JDL file, `/home/does/test.jdl`;
- the three jobs have names `node0`, `mysubjob` and `node2`.

The biggest advantage of this way to build a collection is the possibility to specify a shared input sandbox when the jobs have one or more in the input sandbox which are the same for each job.

A full description of the JDL syntax for collections is available at [38].

#### 6.4.2. Checkpointable Jobs

Job submitted to the gLite and the LCG-2 WMS can be *checkpointable*: in other words, the WMS can save intermediate states of the job in the LB, so that, if the job ends prematurely, it can be resubmitted starting the execution of the job from the last saved state rather than from the beginning.

Checkpointable jobs are specified by setting the JDL `JobType` attribute to `Checkpointable`. The user can specify the number (or list) of steps in which the job should be decomposed, and the step from where to start. This can be done by setting respectively the JDL attributes `JobSteps` and `CurrentStep`. The `CurrentStep` attribute is set automatically to 0 if it has not been defined in the JDL.

To resubmit an incomplete job from an intermediate state, the `--chkpt <filepath>` option of the job submission command must be used, where `<filepath>` is a JDL file containing a checkpoint state generated from a previous execution of the job.

The checkpoint state must be retrieved from the LB server by means of the `glite-wms-job-get-chkpt` command.

More information on checkpointable jobs can be found at [38] and [36].

#### 6.4.3. DAG jobs

The gLite WMS provides an implementation of for *direct acyclic graphs (DAG)*, which are sets of jobs linked by relative dependencies. A job A is said to depend on job B if A is not allowed to run before the job B is successfully completed. A complete description of the JDL syntax for DAGs will not be given here, but it is available elsewhere [38].

#### 6.4.4. Partitionable jobs

If a job can be decomposed into a set of independent sub-jobs, plus an initial *pre-job* and a final *aggregator job*, it can be described as a *partitionable job* and submitted as such to WMProxy.

Internally, a partitionable job is interpreted as a DAG, with the independent sub-jobs depending on the pre-job, and the aggregator job depending on all the sub-jobs. It is implicitly assumed that a partitionable job is also checkpointable.

A complete description of the JDL syntax for partitionable jobs will not be given here, but it is available elsewhere [38].

#### 6.4.5. Parametric jobs

A *parametric job* is a job collection where the jobs are identical but for the value of a running parameter. It is described by a single JDL, where attribute values may contain the current value of the running parameter. An example of a JDL for a parametric job follows:

```
[
JobType = "Parametric";
Executable = "myjob.exe";
StdInput = "input_PARAM.txt";
StdOutput = "output_PARAM.txt";
StdError = "error_PARAM.txt";
Parameters = 100;
ParameterStart = 1;
ParameterStep = 1;
InputSandbox = {"myjob.exe", "input_PARAM.txt";
OutputSandbox = {"output_PARAM.txt", "error_PARAM.txt"};
]
```

The attribute `Parameters` can be either a number, or a list of items (typically strings, but not enclosed within double quotes): in the first case, the value represents the maximum value of the running parameter `_PARAM_`; in the second case, it is the list of the values the parameter must take.

The attribute `ParameterStart` is the initial number of the running parameter, and the attribute `ParameterStep` is the increment of the running parameter between consecutive jobs. Both attributes can be set only if `Parameters` is a number.

#### 6.4.6. Interactive Jobs

Both the gLite and the LCG-2 WMS support the possibility to send *interactive jobs*, that is jobs that open a real time connection with a remote host (usually the UI from which the job was submitted) and the job standard streams (stdin, stdout, stderr) are redirected from/to the remote host. Simply said, the user can communicate in real time with these jobs.

When an interactive job is submitted, the `glite-job-submit` command forks a *Grid console shadow*, or *listener process*, which listens on a port for the job standard streams. A graphical window is opened, where the job streams are forwarded. The port on which the shadow process listens is assigned by the operating system, unless it is explicitly specified using the `ListenerPort` attribute in the JDL.

As the command in this case opens an X window, the user should make sure the `DISPLAY` environment variable is correctly set, an X server is running on the local machine and, if he is connected to the UI node from a remote machine (e.g. with ssh), secure X11 tunneling is enabled. If this is not possible, the user can specify the `--nogui` option, which makes the command provide a simple standard non-graphical interaction with the running job.

#### **Example 6.4.6.1** (A simple interactive job)

The following `interactive.jdl` file contains the description of a very simple interactive job.

The `OutputSandbox` is not necessary, since the output will be sent to the interactive window.

```
[
JobType = "Interactive" ;
Executable = "interactive.sh" ;
InputSandbox = {"interactive.sh"} ;
]
```

The executable specified in this JDL is the `interactive.sh` script, which follows:

```
#!/bin/sh
echo "Welcome!"
echo "Please tell me your name: "
read name
echo "That is all, $name."
echo "Bye bye."
exit 0
```

The `interactive.sh` script prints a message and then asks for an input. After the user has entered a name, this is shown back just to check that the input was received correctly (see Figure 10).

If an interactive job is submitted using the `--nolisten` option, the job standard streams coming from the WN are connected to named pipes on the UI, and their names are returned to the user together with the process ID of the listener. This allows the user to interact with the job using his own tools. It is important to note that when this option is specified, the UI has no more control over the listener process that has hence to be killed by the user when the job is finished.

If, for some reason, the listener process for an interactive job dies, it can be restarted using the command `glite-job-attach`.

More information on interactive jobs is available elsewhere [27] [29].

### **6.4.7. MPI Jobs**

The *Message Passing Interface (MPI)* is a commonly used standard library for parallel programming. The gLite WMS natively supports the submission of MPI jobs, which are jobs composed of a number of processes running

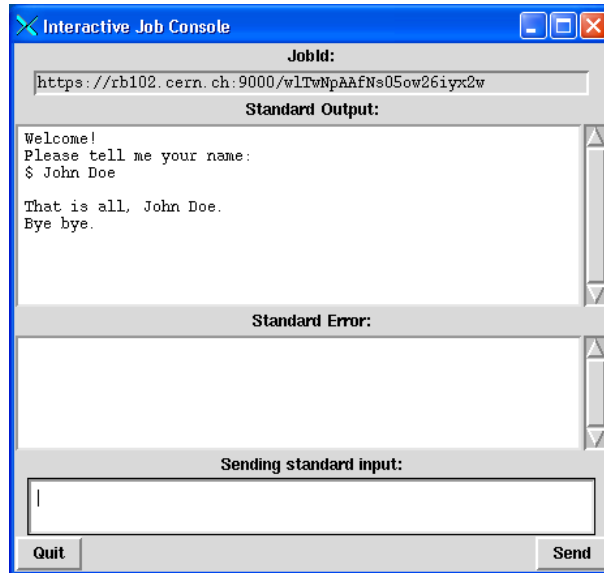


Figure 10: X window for an interactive job

on different WNs in the same CE. However, this support is still experimental and this functionality will not be described in this User Guide. The most complete source of information regarding MPI jobs on the Grid is currently the MPI Wiki page [40].

## 6.5. COMMAND LINE INTERFACE CONFIGURATION

The command line interface of the WMS can be configured using appropriate configuration files. In this section it is explained how to use and customize these configuration files.

### 6.5.1. WMPProxy Configuration

The WMPProxy commands (`glite-wms-*`) look for configuration files in these locations, in order of precedence:

- a. the file specified by the `--config` option;
- b. the file pointed by the `$GLITE_WMS_CLIENT_CONFIG` environment variable;
- c. the file `$HOME/.glite/<vo>/glite_wms.conf`, where `<vo>` is the user's VO name in lowercase;
- d. the file `$GLITE_LOCATION/etc/<vo>/glite_wms.conf`;
- e. the file `$GLITE_LOCATION/etc/glite_wms.conf`.

The settings in files with higher precedence supersede settings in files with lower precedence.

A typical configuration file looks as follows:

```
[  
VirtualOrganisation = "cms";  
Requirements = other.GlueCEStateStatus == "Production";  
Rank = - other.GlueCEStateEstimatedResponseTime;  
MyProxyServer = "myproxy.cern.ch";  
WMProxyEndpoints = {  
    "https://rb102.cern.ch:7443/glite_wms_wmproxy_server"};  
OutputStorage = "/tmp";  
ErrorStorage = "/tmp";  
ListenerStorage = "/tmp";  
AllowZippedISB = true;  
PerusalFileEnable = false;  
RetryCount = 0;  
ShallowRetryCount = 3;  
]
```

In this example file, the following properties are configured:

- the default VO;
- the default requirements;
- the default rank;
- the default MyProxy server;
- the endpoints of the WMProxy servers to be used to submit jobs;
- the path where to store the job output files;
- the path where to write log files;
- the path where to create listener input/output pipes for interactive jobs;
- whether the input sandbox should be zipped by default before being uploaded to the WMProxy server;
- whether the job file perusal support should be enabled by default;
- the default maximum number of deep resubmissions;
- the default maximum number of shallow resubmissions.

### 6.5.2. gLite Network Server Configuration

The gLite WMS commands for commands submitted via Network Server (*glite-\**) have two separate configuration files, a generic configuration and a VO-specific configuration.

The generic configuration file is looked for in these locations, in order of precedence:

- a. the file specified by the `--config` option;

- b. the file pointed by the `$GLITE_WMSUI_CONFIG_VAR` environment variable;
- c. the file `$GLITE_LOCATION/etc/glite_wmsui_cmd_var.conf`.

The settings in files with higher precedence supersede settings in files with lower precedence.

A typical generic configuration file looks as follows:

```
[
Requirements = other.GlueCEStateStatus == "Production";
Rank = - other.GlueCEStateEstimatedResponseTime;
RetryCount = 0;
OutputStorage="/tmp";
ErrorStorage= "/tmp";
ListenerStorage = "/tmp";
LoggingTimeout = 30;
LoggingSyncTimeout = 30;
NSLoggerLevel = 0;
DefaultStatusLevel = 1;
DefaultLogInfoLevel = 1;
DefaultVo = "dteam";
]
```

This configures these properties (excluding those whose meaning is the same as for the WMPProxy configuration):

- the timeout for the asynchronous logging function called by the UI when logging events to the LB server;
- the timeout for the synchronous logging function called by the UI when logging events to the LB server;
- the default quantity of information logged by the NS client;
- the default verbosity level for `glite-job-status`;
- the default verbosity level for `glite-job-logging-info`.

The VO-specific configuration is looked for in these locations, in order of precedence:

- a. the file specified by the `--config-vo` option;
- b. the file pointed by the `$GLITE_WMSUI_CONFIG_VO` environment variable;
- c. the file `$GLITE_LOCATION/etc/<vo>/glite_wmsui.conf`, where `<vo>` is the user's VO name in lowercase.

The settings in files with higher precedence supersede settings in files with lower precedence.

A typical generic configuration file looks as follows:

```
[
```

```
VirtualOrganisation = "cms";  
NSAddresses = {"rb102.cern.ch"};  
LBAddresses = {"rb102.cern.ch"};  
MyProxyServer = "myproxy.cern.ch";  
]
```

The configured settings (excluding the known ones) are:

- the list of WMS instances to be used;
- the list of the LB servers to be used for each WMS instance (a list of lists).

### 6.5.3. LCG-2 Network Server Configuration

The LCG-2 WMS commands (`edg-*`) have two separate configuration files, a generic configuration and a VO-specific configuration.

The generic configuration file is looked for in these locations, in order of precedence:

- a. the file specified by the `-config` option;
- b. the file pointed by the `$EDG_WL_UI_CONFIG_VAR` environment variable;
- c. the file `$EDG_LOCATION/etc/edg_wl_ui_cmd_var.conf`.

The settings in files with higher precedence supersede settings in files with lower precedence.

The VO-specific configuration is looked for in these locations, in order of precedence:

- a. the file specified by the `-config-vo` option;
- b. the file pointed by the `$EDG_WL_UI_CONFIG_VO` environment variable;
- c. the file `$EDG_LOCATION/etc/<vo>/edg_wl_ui.conf`, where `<vo>` is the user's VO name in lowercase.

The settings in files with higher precedence supersede settings in files with lower precedence.

Examples of configuration files are not provided, as they are almost identical to those for the gLite WMS via Network Server.

#### *Example 6.5.3.1 (Using several WMS)*

It is possible to configure the WMS CLI to use several WMS instances for improved redundancy. A job submission command will pick a WMS at random from a list and, if it is unavailable, it will pick another until it succeeds submitting a job or exhausts the list of WMS.

A list of WMS can be specified as follows:



- **WMProxy:** in a configuration file, define a list as value for `WMProxyEndpoints`:

```

WMProxyEndpoints = {
    "https://rb102.cern.ch:7443/glite_wms_wmproxy_server",
    "https://rb109.cern.ch:7443/glite_wms_wmproxy_server",
    "https://rb110.cern.ch:7443/glite_wms_wmproxy_server"
};

```

- **gLite WMS via Network Server:** in the VO-specific configuration file, define a list as value for `NSAddresses` and a corresponding list of lists for `LBAddresses`, where each element is the list of LB servers to be used for each NS:

```

NSAddresses = {
    "rb102.cern.ch",
    "rb109.cern.ch",
    "rb110.cern.ch"
};
LBAddresses = {
    {"rb102.cern.ch"},
    {"rb109.cern.ch"},
    {"rb110.cern.ch"}
};

```

- **LCG-2 WMS:** as in the previous case.

### **Example 6.5.3.2** (Using a separate LB server with WMProxy)

Normally, an LB server is installed on the same host as the WMS, and this configuration is appropriate for a light use of the WMS. However, when the WMS is very busy with managing a large number of jobs, the WMS services and the LB server might experience a performance degradation. In this case, it is advisable to configure the CLI in order to use an LB server on a separate machine. This is done using in the JDL the `LBAddress` attribute, whose value is a string representing the address (`<host>[:<port>]`) of the LB server.

If the user is submitting a job collection, the `LBAddress` attribute must be put in the common part of the collection JDL, because it must be the same for all the jobs in the collection. For example:

```

[
  Type = "Collection";
  LBAddress = "lxb7026.cern.ch:9000";
  ...
  Nodes = [
    ...
  ]
]

```

### **Example 6.5.3.3** (Using a separate LB server with the gLite WMS via Network Server)

Configuring the CLI to use a separate LB is done differently for the Network Server commands. In this case, more LB servers can be defined for each WMS, and one of them is picked randomly at each job submission. Moreover, this cannot be done in the job JDL, but only in the VO-specific configuration file. This is an example of a configuration to use a separate LB:

```
NSAddresses = {  
    "rb102.cern.ch",  
    "rb109.cern.ch",  
};  
LBAddresses = {  
    {"rb102.cern.ch", "lxb7026.cern.ch"},  
    {"lxb7026.cern.ch"},  
};
```

With this configuration, jobs submitted to `rb102.cern.ch` will use either `rb102.cern.ch` or `lxb7026.cern.ch` as LB server, while jobs submitted to `rb109.cern.ch` will use only `lxb7026.cern.ch` as LB server.

## 7. DATA MANAGEMENT

### 7.1. INTRODUCTION

This chapter describes Data Management clients and services available in gLite 3. An overview of the available Data Management APIs is also given in Appendix F.

### 7.2. STORAGE ELEMENTS

The *Storage Element* is the service which allows a user or an application to store data for future retrieval. All data in a SE must be considered *read-only* and therefore can not be changed unless physically removed and replaced. Different VOs might enforce different policies for space quota management. Contact your VO Data Management Administrator for details.

#### 7.2.1. Data Channel Protocols

The data transfer and access protocols supported in gLite 3 are summarized in the next table:

Protocol	Type	GSI secure	Description	Optional
GSIFTP	File Transfer	Yes	FTP-like	No
gsidcap	File I/O	Yes	Remote file access	Yes
insecure RFIO	File I/O	No	Remote file access	Yes
secure RFIO (gsirfio)	File I/O	Yes	Remote file access	Yes

The *GSIFTP*[41]<sup>3</sup> protocol offers the functionalities of FTP, but with support for GSI. It is responsible for secure, fast and efficient file transfers to/from Storage Elements. It provides third party control of data transfer as well as parallel stream data transfer. Every WLCG/EGEE SE runs at least one GridFTP server. For direct remote access of files stored in the SEs, the protocols currently supported by gLite 3 are the *Remote File Input/Output protocol (RFIO)* [42] and the *GSI dCache Access Protocol (gsidcap)*. RFIO was developed to access tape archiving systems, such as CASTOR (CERN Advanced STORAge manager)[43] and it comes in a secure and an insecure version. More information about RFIO can be found in Appendix F. The gsidcap protocol is the GSI enabled version of the dCache[44] native access protocol, *dcap*. The *file* protocol was used in the past for local file access to network filesystems. Currently this option is not supported anymore and the file protocol is only used to specify a file on the local machine (i.e. in a UI or a WN), but not stored in a Grid SE.

#### 7.2.2. Types of Storage Elements

In WLCG/EGEE, different types of Storage Elements are available:

<sup>3</sup>In the literature, the terms *GridFTP* and *GSIFTP* are sometimes used interchangeably. Strictly speaking, GSIFTP is a subset of GridFTP.

- **Classic SE:** it consists of a GridFTP server and an insecure RFIO daemon in front of a physical single disk or disk array. Very soon, the Classic SE will not be supported anymore;
- **CASTOR:** it consists in a disk buffer frontend to a tape mass storage system. A virtual filesystem (namespace) shields the user from the complexities of the disk and tape underlying setup. File migration between disk and tape is managed by a process called “stager”. The native storage protocol, the insecure RFIO, allows access of files in the SE. Since the protocol is not GSI-enabled, only RFIO access from a location in the same LAN of the SE is allowed. With the proper modifications, the CASTOR disk buffer can be used also as disk-only storage system;
- **dCache:** it consists of a server and one or more pool nodes. The server represents the single point of access to the SE and presents files in the pool disks under a single virtual filesystem tree. Nodes can be dynamically added to the pool. The native gsidcap protocol allows POSIX-like data access. dCache is widely employed as disk buffer frontend to many mass storage systems, like HPSS and Enstore, as well as a disk-only storage system.
- **LCG Disk pool manager:** is a lightweight disk pool manager, suitable for relatively small sites (max 10 TB of total space). Disks can be added dynamically to the pool at any time. Like in dCache and CASTOR, a virtual filesystem hides the complexity of the disk pool architecture. The secure RFIO protocol allows file access from the WAN.

### 7.2.3. The Storage Resource Manager interface

The Storage Resource Manager (SRM) has been designed to be the single interface (through the corresponding SRM protocol) for the management of disk and tape storage resources. Any type of Storage Element in WLCG/EGEE offers an SRM interface except for the Classic SE, which is being phased out. SRM hides the complexity of the resources setup behind it and allows the user to request files, keep them on a disk buffer for a specified lifetime (SRM 2.2 only), reserve space for new entries, and so on. SRM offers also a third party transfer protocol between different endpoints, not supported however by all SE implementations. It is important to notice that the SRM protocol is a storage management protocol and not a file access one.

## 7.3. FILE NAMES IN GLITE 3

As an extension of what was introduced in Chapter 3, the different types of file names that can be used within the gLite 3 file catalogue are summarized below:

- the **Grid Unique Identifier (GUID)**, which identifies a file uniquely, is of the form:

```
guid:<36_bytes_unique_string>
guid:38ed3f60-c402-11d7-a6b0-f53ee5a37e1d
```

- the **Logical File Name (LFN)** or User Alias, which can be used to refer to a file in the place of the GUID (and which should be the normal way for a user to refer to a file), has this format:

```
lfn:<any_string>
lfn:importantResults/Test1240.dat
```

In the case of the LCG File Catalogue (see Section 7.4), the LFNs are organized in a hierarchical directory-like structure, and they will have the following format:

```
lfn:/grid/<MyVO>/<MyDirs>/<MyFile>
```

- the **Storage URL (SURL)**, also known as **Physical File Name (PFN)**, which identifies a replica in a SE, is of the general form:

```
<sfm|srm>://<SE_hostname>/<some_string>
```

where the prefix is `sfm` for files located in SEs without a SRM interface and `srm` for SRM-managed SEs.

In the case of the `sfm` prefix, the string after the host name is the path to the location of the file and can be decomposed in the SE's access-point (path to the storage area of the SE), the relative path to the VO of the file's owner and the relative path to the file.

```
sfm://<SE_hostname><SE_Accesspoint><VO_path><filename>
sfm://tbed0101.cern.ch/data/dteam/doe/file1
```

In the case of SRM-managed SEs, one cannot assume that the SURL will have any particular format, other than the `srm` prefix and the host name. In general, SRM-managed SEs can use virtual file systems and the name a file receives may have nothing to do with its physical location (which may also vary with time). An example of this kind of SURL follows:

```
srm://srm.cern.ch/castor/cern.ch/grid/dteam/doe/file1
```

- the **Transport URL (TURL)**, which is a valid URI with the necessary information to access a file in a SE, has the following form:

```
<protocol>://<some_string>
gsiftp://tbed0101.cern.ch/data/dteam/doe/file1
```

where `<protocol>` must be a valid protocol (supported by the SE) to access the contents of the file (GSIFTP, RFIO, gsidcap), and the string after the double slash may have any format that can be understood by the SE serving the file. While SURLs are in principle invariable (they are entries in the file catalogue, see Section 7.4), TURLs are obtained dynamically from the SURL through the Information System or the SRM interface (for SRM managed SEs). The TURL therefore can change with time and should be considered only valid for a relatively small period of time after it has been obtained.

## 7.4. FILE CATALOGUE IN GLITE 3

Users and applications need to locate files (or replicas) on the Grid. The **File Catalogue** is the service which maintains mappings between LFN(s), GUID and SURL(s). The **LCG File Catalogue (LFC)** is the File Catalogue adopted by gLite 3.

The catalogue publishes its endpoint (service URL) in the Information Service so that it can be discovered by Data Management tools and other services (the WMS for example). LFC could either be used as a Local File Catalogue, holding only replicas stored at a given group of site, or a Global File Catalogue, containing information about all files in the Grid. This last one can have multiple read-only instances de-localized at main computing centres all holding the same information.

LFC was developed to overcome serious performance and security issues of the old EDG-RLS catalogues [55]; it also adds some new functionalities such as transactions, sessions, bulk queries and a hierarchical namespace for

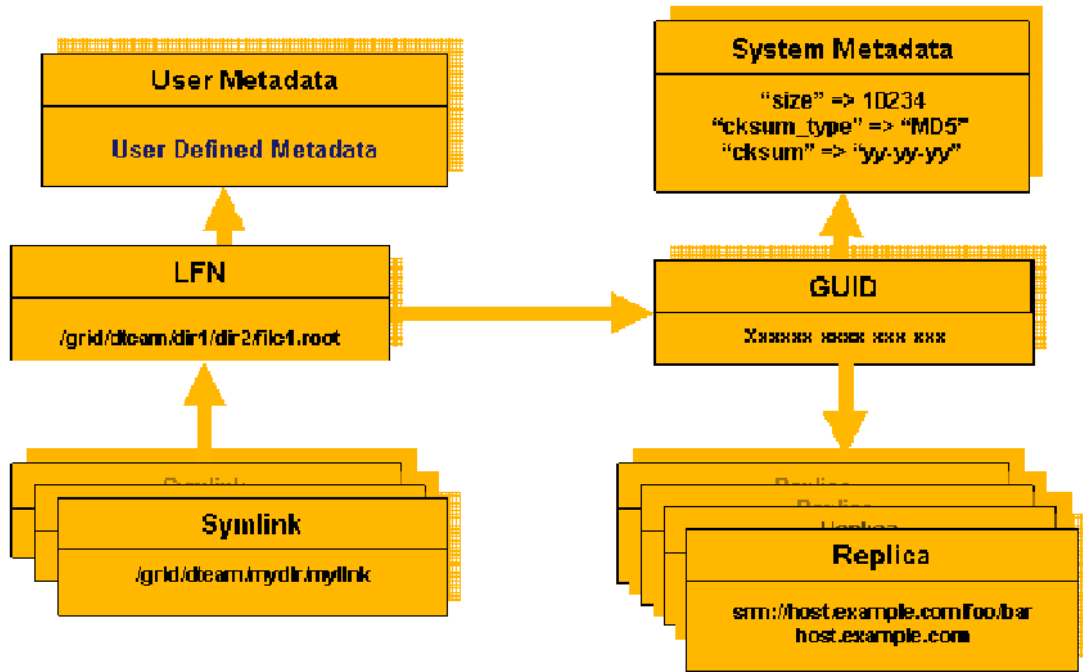


Figure 11: Architecture of the LFC

LFNs. It consists of a unique catalogue, where the LFN is the main key (Figure 11). Further LFNs can be added as symlinks to the main LFN. System metadata are supported, while for user metadata only a single string entry is available (rather a description field than real metadata).

**Note:** a file is considered to be a *Grid file* if it is **both** physically present in a SE **and** registered in the file catalogue. In this chapter several tools will be described. In general high level tools like `lcg_util` (see Sec. 7.5.1) will ensure consistency between files in the SEs and entries in the file catalogue. However, usage of low level Data Management tools could create inconsistencies between SEs physical files and catalogue entries resulting in corruption of GRID files. This is why the usage of low level tools is strongly discouraged unless really necessary.

#### 7.4.1. LFC Commands

In general the user should interact with the file catalogue through high level utilities (`lcg_util`, see Section 7.5.1). The CLIs and APIs that are available for catalogue interaction provide further functionality and more fine-grained control for catalogue operations; in some situations, they represent the only possible way to achieve the desired result.

In gLite 3 the environment variable `LFC_HOST` can be set to hold the host name of the LFC server. This is mandatory for the LFC CLIs and APIs; for GFAL and `lcg_util` (see later) such variable, if set, supersedes the endpoint definition published in the Information System.

The directory structure of the LFC namespace has the form:

/grid/<VO>/<subpaths>

Users of a given VO will have read and write permissions only under the corresponding <VO> subdirectory. More restrictive access patterns on deeper subpaths of the directory tree can be enforced by the VO.

Once the correct environment has been set, the following commands can be used:

lfc-chmod	Change access mode of a LFC file/directory
lfc-chown	Change owner and group of a LFC file/directory
lfc-delcomment	Delete the comment associated with a file/directory
lfc-getacl	Get file/directory access control lists
lfc-ln	Make a symbolic link to a file/directory
lfc-ls	List file/directory entries in a directory
lfc-mkdir	Create directory
lfc-rename	Rename a file/directory
lfc-rm	Remove a file/directory
lfc-setacl	Set file/directory access control lists
lfc-setcomment	Add/replace a comment
lfc-entergrpmap	Defines a new group entry in the Virtual ID table
lfc-enterusmap	Defines a new user entry in Virtual ID table
lfc-modifygrpmap	Modifies a group entry corresponding to a given virtual gid
lfc-modifyusmap	Modifies a user entry corresponding to a given virtual uid
lfc-rmgrpmap	Suppresses group entry corresponding to a given virtual gid or group name
lfc-rmusmap	Suppresses user entry corresponding to a given virtual uid or user name.

Man pages are available for all the commands. Most of them work in a very similar way to their Unix equivalents, but operating on directories and files of the catalogue namespace. Where the path of a file/directory is required, an absolute path can be specified (starting by /) or, otherwise, the path is prefixed by the contents of the LFC\_HOME environment variable.

Users should use these commands carefully, keeping in mind that the operations they are performing affect the catalogue, but not the physical files that the entries represent.

#### **Example 7.4.1.1** (Listing the entries of a LFC directory)

The `lfc-ls` command lists the LFNs in a given directory.

**Attention!** The `-R` option, for recursive listing, is available for the command, but **it should not be used extensively**. It is a very expensive operation on the catalogue and should be avoided as much as possible.

In the following example content of the directory `/grid/dteam/MyExample` is listed:

```

$ lfc-ls /grid/dteam/MyExample

/grid/dteam/MyExample:
day1
day2

```

day3  
day4

#### **Example 7.4.1.2 (Creating directories in the LFC)**

The `lfc-mkdir` creates a directory in the LFN namespace:

```

$ lfc-mkdir /grid/lhcb/test_doe/MyTest
$ lfc-ls -l /grid/lhcb/test_doe
drwxrwxr-x  0 doe z5                0 Feb 21 16:50 MyTest
  
```

#### **Example 7.4.1.3 (Creation of symbolic links)**

The `lfc-ln` command can be used to create a symbolic link to a file. In this way two different LFNs will point to the same file.

In the following example, we create a symbolic link `/grid/lhcb/test_doe/MyTest/newname` to the original file `/grid/lhcb/test_doe/testfile`:

```
$ lfc-ln -s /grid/lhcb/test_doe/testfile /grid/lhcb/test_doe/MyTest/newname
```

And check that the new alias exists:

```

$ lfc-ls -l /grid/lhcb/test_doe/MyTest/newname
lrwxrwxrwx  1 doe  z5      0 Feb 21 16:54 /grid/lhcb/test_doe/MyTest/newname
-> /grid/lhcb/test_doe/testfile
  
```

Remember that links created with `lfc-ln` are soft. If the LFN they are pointing to is removed, the links themselves are not deleted, but will still exist as broken links.

#### **Example 7.4.1.4 (Adding metadata information to LFC entries)**

The `lfc-setcomment` and `lfc-delcomment` commands allow the user to associate a comment with a catalogue entry and delete such comment. This is the only user-defined metadata that can be associated with catalogue entries. The comments for the files may be listed using the `--comment` option of the `lfc-ls` command. This is shown in the following example:

```

$ lfc-setcomment /grid/cms/MyFiles/file1 "Most promising measure"

$ lfc-ls --comment /grid/cms/MyFiles/file1
/grid/dteam/MyFiles/file1 Most promising measure
  
```



### Example 7.4.1.5 (Removing LFNs from the LFC)

The `lfc-rm` command can be used to remove files and directories from the LFN namespace, but with two basic limitations:

- a file can be removed only if there are no SURLs associated to it. If SURLs exist, the `lcg_util` commands should be used instead (Section 7.5.1);
- a directory can be removed (`-r` option) only if it is empty.

In the next example, the directory `trash` is removed:

```
$ lfc-ls -l -d /grid/dteam/MyExample/trash
drwxr-xrwx  0 dteam004 cg                0 Jul 06 11:13 /grid/dteam/MyExample/trash

$ lfc-rm -r /grid/dteam/MyExample/trash

$ lfc-ls -l -d /grid/dteam/MyExample/trash
> /grid/dteam/MyExample/trash: No such file or directory
```

## 7.4.2. Access Control Lists

LFC allows to attach to a file or directory an *access control list (ACL)*, a list of permissions which specify who is allowed to access or modify it. The permissions are very much like those of a UNIX file system: read (`r`), write (`w`) and execute (`x`). A combination of these permissions can be associated to these entities:

- a user (`user`);
- a group of users (`group`);
- any other user (`other`);
- the maximum permissions granted to specific users or groups (`mask`).

Permissions for multiple users and groups can be defined. If this is the case, a mask must be defined and the “effective” permissions are the logical AND of the user or group permissions and the mask.

In LFC, users and groups are internally identified as numerical *virtual uids* and *virtual gids*, which are virtual in the sense that they exist only in the LFC namespace.

A user can be specified as a name, as a virtual uid or as a Distinguished Name. A group can be specified as a name, as a virtual gid or as a VOMS FQAN.

In addition, a directory in LFC has also a *default ACL*, which is the ACL associated to any file or directory being created under that directory. After creation, the ACLs can be freely changed. When creating a sub-directory, its default ACL is inherited from the parent directory’s default ACL.

**Note:** it is worth to know that, internally, LFC maps different VOMS FQANs to different virtual groups, and a user cannot be in more than one group at any given time. This means that, for example, it might be possible for a user to be allowed to delete an LFC file when his VOMS proxy has as FQAN `/atlas`, but not when his VOMS proxy has as FQAN `/atlas/Role=production`. This goes against the “expected” behaviour, because having the production role should imply having more privileges, not less. A future version of LFC will solve this issue.

#### **Example 7.4.2.1**    *(Print the ACL of a directory)*

In the following example, the ACL for a given directory is displayed:

```

$ lfc-getacl /grid/atlas/UserGuide

# file: /grid/atlas/UserGuide
# owner: /C=CH/O=CERN/OU=GRID/CN=John Doe
# group: atlas
user::rwx
group::rwx          #effective:rwx
other::r-x
default:user::rwx
default:group::rwx
default:other::r-x
  
```

The output prints the DN and the group of the owner of the directory, followed by the ACL and the default ACL. In this example, the owner and all users in the `atlas` group have full privileges to the directory, while other users cannot write into it.

#### **Example 7.4.2.2**    *(Modify the ACL of a directory)*

Suppose that we want to associate a set of permissions to a given FQAN for the LFC directory from the previous example. This could be done by doing:

```

$ lfc-setacl -m g:/atlas/Role=production:rwx,m:rwx,d:g:/atlas/Role=production:rwx,d:m:rwx \
/grid/atlas/UserGuide
  
```

The `-m` option means that we are modifying the existing ACL.

The added ACL is specified as a comma-separated list of entries, where each entry is a colon-separated list of fields: an ACL type (`user`, `group`, `other`, `mask`, or these preceded by `default`), a user or group, and a permission set. Notice that ACL types can be abbreviated using their first letter.

In this example, we have set a permission set for a group (the `/atlas/Role=production` FQAN), the mask and the same for the default ACL.

If now we print again the ACL for the directory:

```

$ lfc-getacl /grid/atlas/UserGuide

# file: /grid/atlas/UserGuide
# owner: /C=CH/O=CERN/OU=GRID/CN=John Doe
# group: atlas
user::rwx
group::rwx          #effective:rwx
group:/atlas/Role=production:rwx      #effective:rwx
mask::rwx
other::r-x
default:user::rwx
default:group::rwx
default:group:/atlas/Role=production:rwx
default:mask::rwx
default:other::r-x
  
```

we see now permissions for both the owner's group (atlas) and the /atlas/Role=production FQAN, and the same for the default ACL. The effective permissions for the owner's group and the VOMS FQAN take into account the mask.

Other options of `lfc-setacl` are `-d` to remove ACL entries, and `-s` to replace the complete set of ACL entries.

## 7.5. FILE AND REPLICA MANAGEMENT CLIENT TOOLS

The gLite 3 middleware offers a variety of data management client tools to upload/download files to/from the Grid, replicate data and interact with the file catalogues. Every user should deal with data management through the LCG Data Management tools (usually referred to as *lcg\_util* or `lcg-*` commands). They provide a high level interface (both command line and APIs) to the basic DM functionality, hiding the complexities of catalogue and SEs interaction. Furthermore, such high level tools minimize the risk of grid files corruption.

Some lower level tools (like `edg-gridftp-*` commands, `globus-url-copy` and `srm-*` dedicated commands) are also available. These low level tools are quite helpful in some particular cases (see examples for more details). Their usage, however, is strongly discouraged for non expert users, since such tools do not ensure consistency between physical files in the SE and entries in the file catalogue and their usage might be dangerous.

### 7.5.1. LCG Data Management Client Tools

The LCG Data Management tools (*lcg\_util*) allow users to copy files between UI, CE, WN and a SE, to register entries in the file catalogue and replicate files between SEs. The name and functionality overview of the available commands is shown in the following table.

#### Replica Management

<code>lcg-cp</code>	Copies a Grid file to a local destination (download)
<code>lcg-cr</code>	Copies a file to a SE and registers the file in the catalogue (upload)
<code>lcg-del</code>	Deletes one file (either one replica or all replicas)
<code>lcg-rep</code>	Copies a file from one SE to another SE and registers it in the catalogue (replicate)
<code>lcg-gt</code>	Gets the TURL for a given SURL and transfer protocol
<code>lcg-sd</code>	Sets file status to "Done" for a given SURL in an SRM's request

### File Catalogue Interaction

<code>lcg-aa</code>	Adds an alias in the catalogue for a given GUID
<code>lcg-ra</code>	Removes an alias in the catalogue for a given GUID
<code>lcg-rf</code>	Registers in the catalogue a file residing on an SE
<code>lcg-uf</code>	Unregisters in the the catalogue a file residing on an SE
<code>lcg-la</code>	Lists the aliases for a given LFN, GUID or SURL
<code>lcg-lg</code>	Gets the GUID for a given LFN or SURL
<code>lcg-lr</code>	Lists the replicas for a given LFN, GUID or SURL

The `--vo <vo_name>` option, to specify the virtual organisation of the user, is present in all commands, except for `lcg-gt` and `lcg-sd`. Its usage is mandatory unless the variable `LCG_GFAL_VO` is set, see below. The `--config <file>` option (to specify a configuration file) and the `-i` option (to connect insecurely to the file catalogue) are currently ignored.

### Timeouts

The commands `lcg-cr`, `lcg-del`, `lcg-gt`, `lcg-rf`, `lcg-sd` and `lcg-rep` all have timeouts implemented. By using the option `-t`, the user can specify a number of seconds for the tool to time out. The default is 0 seconds, that is no timeout. If a tool times out in the middle of an operation, all actions performed till that moment are rolled back, so no broken files are left on a SE and no existing files are not registered in the catalogues.

### Environment variables

- For all `lcg-*` commands to work, the environment variable `LCG_GFAL_INFOSYS` must be set to point to a top BDII in the format `<hostname>:<port>`, so that the commands can retrieve the necessary information. Remember that the BDII read port is 2170;
- the endpoint(s) for the catalogues can also be specified (taking precedence over that published in the IS) through the environment variable `LFC_HOST`. If no endpoints are specified, the ones published in the Information System are taken;
- if the variable `LCG_GFAL_VO` is set to indicate the user VO, the `--vo` option is not required. However, if the VO name is specified in the command option, the `LCG_GFAL_VO` variable is ignored;
- The `VO_<VO>_DEFAULT_SE` variable specifies the default SE for the VO `<VO>`.

The user must hold a valid proxy and be authorized on the SE in order to use `lcg-cr`, `lcg-cp`, `lcg-rep` and `lcg-del`. While access to resources (SEs and LFCs) is authenticated, the data channel is not encrypted.

**Note:** The user will often need to gather information on the existing Grid resources in order to perform DM operations. For instance, in order to specify the destination SE for the upload of a file, the information about the available SEs must be retrieved in advance. There are several ways to retrieve information about the resources on the Grid, which are described in Chapter 5.

In what follows, some examples are given. Most commands can run in verbose mode (`-v` or `--verbose` option). For details on the options of each command, refer to the man pages of the commands.

### **Example 7.5.1.1** (Uploading a file to the Grid)

In order to upload a file to the Grid, that is to transfer it from the local machine to a Storage Element and register it in the catalogue, the `lcg-cr` command (which stands for *copy&register*) can be used:

```
$ lcg-cr --vo dteam -d lxb0710.cern.ch file:/home/does/file1
guid:6ac491ea-684c-11d8-8f12-9c97cebf582a
```

where the only argument is the local file to be uploaded (a fully qualified URI) and the `-d <destination>` option indicates the SE used as the destination for the file. The command returns the file GUID. If no destination is given, the SE specified by the `VO_<VO>_DEFAULT_SE` environmental variable is taken. Such variable is set in all WNs and UIs.

The `-P` option allows the user to specify a relative path name for the file in the SE. The absolute path is built appending the relative path to a root directory which is VO- and SE-specific and is published in the Information System. If no `-P` option is given, the relative path is automatically generated.

It is also possible to specify the destination as a complete SURL, including SE hostname, the path, and a chosen filename. The action will only be allowed if the specified path falls under the user's VO directory.

The following are examples of the different ways to specify a destination:

```
-d lxb0710.cern.ch
-d sfn://lxb0710.cern.ch/data/dteam/my_file
-d lxb0710.cern.ch -P my_dir/my_file
```

The option `-l <lfn>` can be used to specify a LFN:

```
$ lcg-cr --vo dteam -d lxb0710.cern.ch -l lfn:my_alias1 file:/home/does/file1
guid:db7ddbc5-613e-423f-9501-3c0c00a0ae24
```

**Note:** LFNs in LFC are organized in a hierarchical namespace (like UNIX directory trees). So the LFN will take the form `lfn:/grid/<vo>/<dir1>/...`. Subdirectories in the namespace are **not** created automatically by `lcg-cr` and the user should manage himself their creation through the `lfc-mkdir` and `lfc-rmdir` command line tools described in the previous section.

The `-g` option allows to specify a GUID (otherwise automatically created):

```
$ lcg-cr --vo dteam -d lxb0710.cern.ch \
```

```
-g guid:baddb707-0cb5-4d9a-8141-a046659d243b file:`pwd`/file2
```

```
guid:baddb707-0cb5-4d9a-8141-a046659d243b
```

**Attention!** This option should not be used except for expert users and in very particular cases. Because the specification of an existing GUID is also allowed, a misuse of the tool may end up in a corrupted GRID file in which replicas of the same file are in fact different from each other.

Finally, in this and other commands, the `-n <#streams>` options can be used to specify the number of parallel streams to be used in the transfer (default is one).

**Attention!** When multiple streams are requested, the GridFTP protocol establishes that the GridFTP server must open a new connection back to the client (the original connection, and only one in the case of one stream, is opened from the client to the server). This may become a problem when a file is requested from a WN and this WN is firewalled to disable inbound connections (which is usually the case). The connection will in this case fail and the error message returned (in the logging information of the job performing the data access) will be "425 can't open data connection".

#### **Example 7.5.1.2**    *(Replicating a file)*

Once a file is stored on an SE and registered in the catalogue, the file can be replicated using the `lcg-rep` command, as in:

```
$ lcg-rep -v --vo dteam -d lxb0707.cern.ch guid:db7ddbc5-613e-423f-9501-3c0c00a0ae24

Source URL: sfn://lxb0710.cern.ch/data/dteam/does/file1
File size: 30
Destination specified: lxb0707.cern.ch
Source URL for copy: gsiftp://lxb0710.cern.ch/data/dteam/does/file1
Destination URL for copy: gsiftp://lxb0707.cern.ch/data/dteam/generated/2004-07-09/
file50c0752c-f61f-4bc3-b48e-af3f22924b57
# streams: 1
Transfer took 2040 ms
Destination URL registered in LRC: sfn://lxb0707.cern.ch/data/dteam/generated/2004-07-09/
file50c0752c-f61f-4bc3-b48e-af3f22924b57
```

where the file to be replicated can be specified using a LFN, GUID or even a SURL, and the `-d` option is used to specify the SE where the new replica will be stored. This destination can be either an SE hostname or a complete SURL, and it is expressed in the same format as with `lcg-cr`. The command also admits the `-P` option to add a relative path to the destination (as with `lcg-cr`).

For one GUID, there can be only one replica per SE. If the user tries to use the `lcg-rep` command with a destination SE that already holds a replica, the command will exit successfully, but no new replica will be created.

#### **Example 7.5.1.3**    *(Listing replicas, GUIDs and aliases)*

The `lcg-lr` (*list replicas*) command allows users to list all the replicas of a file registered in the file catalogue:

```

$ lcg-lr --vo dteam lfn:/grid/dteam/does/my_alias1

sfn://lxb0707.cern.ch/data/dteam/generated/2004-07-09/file79aee616-6cd7-4b75-8848-f091
sfn://lxb0710.cern.ch/data/dteam/generated/2004-07-08/file0dcabb46-2214-4db8-9ee8-2930
  
```

Again, a LFN, the GUID or a SURL can be used to specify the file. The SURLs of all the replicas are returned.

The `lcg-lg` (*list GUID*) returns the GUID associated with a specified LFN or SURL:

```

$ lcg-lg --vo dteam sfn://lxb0707.cern.ch/data/dteam/does/file1

guid:db7ddbc5-613e-423f-9501-3c0c00a0ae24
  
```

The `lcg-la` (*list aliases*) can be used to list the LFNs associated with a particular file, which can be identified by its GUID, any of its LFNs, or the SURL of one of its replicas:

```

$ lcg-la --vo dteam guid:baddb707-0cb5-4d9a-8141-a046659d243b

lfn:my_alias1
  
```

#### **Example 7.5.1.4 (Copying files out of the Grid)**

The `lcg-cp` command can be used to copy a Grid file to a non-grid storage resource. The first argument (source file) can be a LFN, GUID or SURL of a valid Grid file, the second argument (destination file) must be a local filename or a valid TURL. In the following example, the verbose mode is used and a timeout of 100 seconds is specified:

```

$ lcg-cp --vo dteam -t 100 -v lfn:/grid/dteam/does/myfile file:/tmp/myfile

Source URL: lfn:/grid/dteam/does/myfile
File size: 104857600
Source URL for copy:
gsiftp://lxb2036.cern.ch/storage/dteam/generated/2005-07-17/fileea15c9c9-abcd-4e9b-8724-1ad60c5afe5b
Destination URL: file:///tmp/myfile
# streams: 1
# set timeout to 100 (seconds)
      85983232 bytes   8396.77 KB/sec avg   9216.11
Transfer took 12040 ms
  
```

Notice that although this command is designed to copy files from a SE to non-grid resources, if the proper TURL is used (using the GSIFTP protocol), a file could be transferred from one SE to another, or from out of the Grid to a SE. **This should not be done**, since it has the same effect as using `lcg-rep` BUT **skipping the file registration**, making in this way this replica invisible to Grid users.

Be aware that in the case of a MSS, the file may be not present on disk but only stored on tape. For this reason, `lcg-cp` on such a file could time out, waiting for the file stage-in on a disk buffer.

### Example 7.5.1.5 (Obtaining a TURL for a replica)

The `lcg-gt` allows to get a TURL from a SURL and a supported protocol. The command behaves very differently if the Storage Element exposes an SRM interface or not. The command always returns three lines of output: the first is always the TURL of the file, the last two are meaningful only in case of SRM interface.

- For a classic SE (no SRM interface), the command obtains the TURL by simple string manipulation of the SURL and the protocol (checking in the Information System if it is supported by the Storage Element). No direct interaction with the SE is involved. The last two lines of output are always zeroes:

```
$ lcg-gt sf://lxb0710.cern.ch/data/dteam/generated/2004-07-08/file0dcabb4
6-2214-4db8-9ee8-2930de1a6bef gsiftp

gsiftp://lxb0710.cern.ch/data/dteam/generated/2004-07-08/file0dcabb46-22
14-4db8-9ee8-2930de1a6bef
0
0
```

- In the case of a SRM interface, the TURL is returned to `lcg-gt` by the SRM itself. For a MSS, the file will be staged on disk (if not present already) before a valid TURL is returned. It could take `lcg-gt` quite a long time to return the TURL (depending on the conditions of the stager) but a successive `lcg-cp` of such TURL will start copying the file immediately. This is one of the reasons for which a SRM interface is desirable for all MSS.

The second and third lines of output represent the *requestID* and *fileID* for the `srm-put` request (hidden to the user) which will remain open unless explicitly closed (at least with SRM 1). It is important to know that some SRM SEs are limited in the maximum number of open requests. Further requests will fail, once this limit has been reached. It is therefore good practice to close the request once the TURL is not needed anymore. This can be done with the `lcg-sd` command which needs as arguments the TURL of the file, the `requestID` and `fileID`.

```
$ lcg-gt srm://srm.cern.ch/castor/cern.ch/grid/dteam/generated/2005-04-12/file
fadle7fb-9d83-4050-af51-4c9af7bb095c gsiftp

gsiftp://srm.cern.ch:2811//shift/lxfsrk4705/data02/cg/stage/filefadle7fb-9d
83-4050-af51-4c9af7bb095c.43309
-337722383
0

[ ... do something with the TURL ... ]
```



```
$ lcg-sd gsiftp://srm.cern.ch:2811//shift/lxfsrk4705/data02/cg/stage/filefad1
e7fb-9d83-4050-af51-4c9af7bb095c.43309 -337722383 0
```

### **Example 7.5.1.6 (Deleting replicas)**

A file stored on a SE and registered in LFC can be deleted using the `lcg-del` command. If a SURL is provided as argument, then that particular replica will be deleted. If a LFN or GUID is given instead, then the `-s <SE>` option must be used to indicate which one of the replicas must be erased, unless the `-a` option is used, in which case all replicas of the file will be deleted and unregistered (on a best-effort basis). If all the replicas of a file are removed, the corresponding GUID-LFN mappings are removed as well.

```
$ lcg-lr --vo dteam guid:91b89dfe-ff95-4614-bad2-c538bfa28fac

sfn://lxb0707.cern.ch/data/dteam/generated/2004-07-12/file78ef5a13-166f-4701-
8059-e70e397dd2ca
sfn://lxb0710.cern.ch/data/dteam/generated/2004-07-12/file21658bfb-6eac-409b-
9177-88c07bb1a57c

$ lcg-del --vo dteam -s lxb0707.cern.ch guid:91b89dfe-ff95-4614-bad2-c538bfa28fac
$ lcg-lr --vo dteam guid:91b89dfe-ff95-4614-bad2-c538bfa28fac

sfn://lxb0710.cern.ch/data/dteam/generated/2004-07-12/file21658bfb-6eac-409b-
9177-88c07bb1a57c

$ lcg-del --vo dteam -a guid:91b89dfe-ff95-4614-bad2-c538bfa28fac

$ lcg-lr --vo dteam guid:91b89dfe-ff95-4614-bad2-c538bfa28fac

lcg_lr: No such file or directory
```

The last error indicates that the GUID is no longer registered within the catalogue, as the last replica was deleted.

### **Example 7.5.1.7 (Registering and unregistering Grid files)**

The `lcg-rf` (*register file*) command allows to register a file physically present in a SE, creating a GUID-SURL mapping in the catalogue. The `-g <GUID>` allows to specify a GUID (otherwise automatically created).

```
$ lcg-rf -v --vo dteam -g guid:baddb707-0cb5-4d9a-8141-a046659d243b \
sfn://lxb0710.cern.ch/data/dteam/generated/2004-07-08/file0dcabb46-2214-4db8-9ee8-2930de1
guid:baddb707-0cb5-4d9a-8141-a046659d243b
```

Likewise, `lcg-uf` (*unregister file*) allows to delete a GUID-SURL mapping (respectively the first and second argument of the command) from the catalogue:

```
$ lcg-uf --vo dteam guid:baddb707-0cb5-4d9a-8141-a046659d243b \
sfn://lxb0710.cern.ch/data/dteam/generated/2004-07-08/file0dcabb46-2214-4db8-9ee8-2930de1
```

If the last replica of a file is unregistered, the corresponding GUID-LFN mapping is also removed.

**Attention!** `lcg-uf` just removes entries from the catalogue, it does not remove any physical replica from the SE. Watch out for consistency.

### **Example 7.5.1.8** (Managing aliases)

The `lcg-aa` (*add alias*) command allows the user to add a new LFN to an existing GUID:

```
$ lcg-la --vo dteam guid:baddb707-0cb5-4d9a-8141-a046659d243b
lfn:/grid/dteam/does/my_alias1

$ lcg-aa --vo dteam guid:baddb707-0cb5-4d9a-8141-a046659d243b lfn:/grid/dteam/does/new_alias

$ lcg-la --vo dteam guid:baddb707-0cb5-4d9a-8141-a046659d243b
lfn:/grid/dteam/does/my_alias1
lfn:/grid/dteam/does/new_alias
```

Correspondingly, the `lcg-ra` (*remove alias*) command allows a user to remove an LFN from an existing GUID:

```
$ lcg-ra --vo dteam guid:baddb707-0cb5-4d9a-8141-a046659d243b lfn:/grid/dteam/does/my_alias1

$ lcg-la --vo dteam guid:baddb707-0cb5-4d9a-8141-a046659d243b
lfn:/grid/dteam/does/new_alias
```

## **7.6. FILE TRANSFER SERVICE**

The *File Transfer Service (FTS)* is the gLite 3 low level data movement service. The user can schedule asynchronous and reliable file replication from source to destination (point-to-point, i.e. no file routing via intermediate storage) while participant sites can control the network usage. The FTS handles internally the SRM negotiation between the source and destination SEs and the management of the underlying GridFTP transfers.

### **7.6.1. Basic Concepts**

- **Transfer Job:** a set of files to be transferred in a source/destination pair format. A job may contain optional parameters for the underlying transport layer (GridFTP). Finally, the job carries along a cyphered pass

phrase to decrypt user credentials from the MyProxy server;

- **File:** a source/destination SURL pair to be transferred;
- **Job State:** a function of the individual file states constituting the Job;
- **File State:** the state of an individual file transfer;
- **Channel:** a specific network pipe used for file transfers. **Production channels** are high bandwidth, dedicated network pipe between Tier-0, Tier-1's and other major Tier-2's centers. **Non-production channels** are assigned typically to open networks and do not guarantee a minimum throughput as production channels do.

The transfer jobs are processed asynchronously (batch mode). Upon submission, a job identifier is returned to the user. This identifier can be used to query the status of the job as it progresses through the system or cancel the job. Once a job has been submitted to the system it is assigned to a transfer channel based on the SEs containing the source and the destination. Finally, FTS accepts only SURLs as source and destination. Logical entries like LFNs or GUIDs are at the moment not supported.

### 7.6.2. Transfer job states

The possible states a job can assume are:

- **Submitted:** the job has been submitted to FTS but not yet assigned to a channel
- **Pending:** the job has been assigned to a channel and files are waiting for being transferred
- **Active:** the transfer for some of the job's files is ongoing
- **Canceling:** the job is being canceled
- **Done:** all files in a job were successfully transferred
- **Failed:** some file transfers in a job have failed
- **Canceled:** the job has been canceled
- **Hold:** the job has aborted and requires manual interventions (moving it to **Pending** or **Failed**)

The final states for jobs are **Done**, **Canceled** and **Failed**. The possible job status transitions are depicted in Figure 12.

### 7.6.3. Individual file states

The possible states for individual files are the following:

- **Submitted:** the status of all files in a job which is in **Submitted** status
- **Pending:** the status of all files in a job which is in **Pending** status
- **Active:** the transfer of the file is ongoing

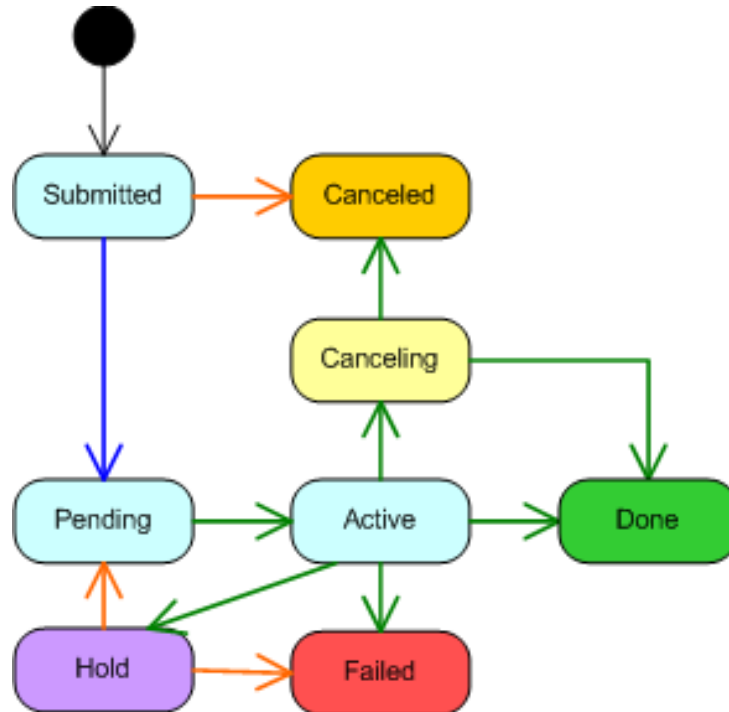


Figure 12: FTS transfer job states.

- **Canceling:** the transfer of the file is being canceled
- **Waiting:** the transfer of the file has failed; depending on the VO policies, it will then go to *Pending*, *Failed* or *Hold* status
- **Done:** the transfer of the file has finished correctly
- **Failed:** the transfer of the file has failed
- **Canceled:** the transfer of the file has been canceled
- **Hold:** the transfer of the file has failed and requires manual interventions (moving it to *Pending* or *Failed*)

The final states for files are *Done*, *Canceled* and *Failed*. The possible file status transitions are depicted in Figure 13.

#### 7.6.4. FTS Commands

Before submitting a job, the user is expected to upload an appropriate password-protected long-term proxy to the MyProxy server used by FTS.

```
$ myproxy-init -s myproxy-fts.cern.ch -d
```

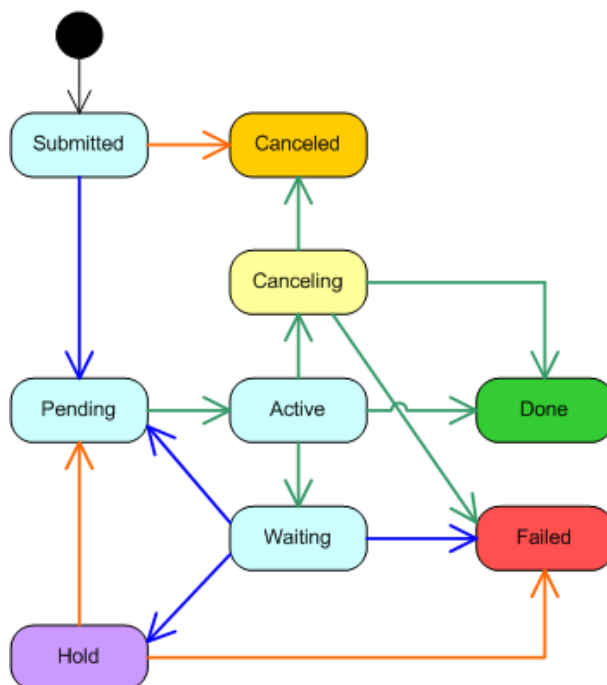


Figure 13: FTS individual file states.

**Attention!** This is a different usage of MyProxy with respect to the WMS Proxy Renewal. In the latest case the `-n` option must be used while here it must be omitted. In addition, the same MyProxy server can not be *simultaneously* used for WMS Proxy Renewal and FTS authentication, because they require a different configuration of the MyProxy server.

The same password is passed to FTS at job submission time. The following user-level commands for submitting, querying and canceling jobs are described here:

<code>glite-transfer-submit</code>	Submits a transfer job
<code>glite-transfer-status</code>	Displays the status of an ongoing transfer job
<code>glite-transfer-list</code>	Lists all submitted transfer jobs owned by the user
<code>glite-transfer-cancel</code>	Cancel a transfer job

For completeness the following administrative commands are also briefly described. Only FTS service administrators are allowed to use them.

<code>glite-transfer-channel-add</code>	Creates a new channel with defined parameters on FTS
<code>glite-transfer-channel-list</code>	Displays details of a given channel defined on FTS
<code>glite-transfer-channel-set</code>	Allows administrators to set a channel Active or Inactive
<code>glite-transfer-channel-signal</code>	Changes status of all transfers in a given job or channel

**Example 7.6.4.1** (Submitting a job to FTS)

Once a user has successfully registered a long-term proxy to a MyProxy server, he can submit a transfer job. He can do it either by specifying the source-destination pair in the command line:

```

$ glite-transfer-submit -m myproxy-fts.cern.ch \
-s https://fts.cnaf.infn.it:8443/sc3/glite-data-transfer-fts/services/FileTransfer \
srm://srm.sara.nl/pnfs/srm.sara.nl/data/lhcb/doi/zz_zz.f \
srm://srm.cnaf.infn.it/castor/cnaf.infn.it/grid/lcg/lhcb/test/SARA_1.25354

Enter MyProxy password:

Enter MyProxy password again:

c2e2cdb1-a145-11da-954d-944f2354a08b
  
```

or by specifying all source-destination pairs in an input file (bulk submission). The `-m` option specifies the MyProxy server to use; the `-s` option specifies the FTS service endpoint to be contacted. If the service starts with **http://**, **https://** or **httpg://** it is taken as a direct service endpoint URL; otherwise is taken as a service instance name and Service Discovery is invoked to look up the endpoints. If not specified the first available transfer service from the Service Discovery will be used. This is true for all subsequent examples.

```

$ glite-transfer-submit -m "myproxy-fts.cern.ch" \
-s https://fts.cr.cnaf.infn.it:8443/sc3/glite-data-transfer-fts/services/FileTransfer \
SARA-CNAF.in -p $passwd
  
```

where the input file `SARA-CNAF.in` looks like:

```

$ cat SARA-CNAF.in

srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/lhcb/test/doi/zz_zz.f \
srm://sc.cr.cnaf.infn.it/castor/cnaf.infn.it/grid/lcg/lhcb/test/doi/SARA_1.25354
srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/lhcb/test/doi/zz_zz.f \
srm://sc.cr.cnaf.infn.it/castor/cnaf.infn.it/grid/lcg/lhcb/test/doi/SARA_2.25354
srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/lhcb/test/doi/zz_zz.f \
srm://sc.cr.cnaf.infn.it/castor/cnaf.infn.it/grid/lcg/lhcb/test/doi/SARA_3.25354
.....
  
```

The `$passwd` in the example is an environment variable set to the value of the password to be passed to FTS.

**Attention!** The transfers handled by FTS within a single job bulk submission must be all assigned to the same channel, otherwise FTS will not process such transfers and will return the message: *Inconsistent channel*.

#### **Example 7.6.4.2** (Querying the status of a job)

The following example shows a query to FTS to infer information about the state of a transfer job:

```

$ glite-transfer-status \
-s https://fts.cnaf.infn.it:8443/sc3/glite-data-transfer-fts/services/FileTransfer \
-l c2e2cdb1-a145-11da-954d-944f2354a08b

```

```

Pending
  Source:      srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/lhcb/test/doi/zz_zz.f
  Destination: srm://sc.cr.cnaf.infn.it/castor/cnaf.infn.it/grid/lcg/lhcb/test/doi/
SARA_1.25354
  State:      Pending
  Retries:    0
  Reason:     (null)
  Duration:   0

```

**Attention!** The verbosity level of the status of a given job is set with the `-v` option; the status of individual files is however available through the option `-l`.

#### **Example 7.6.4.3** *(Listing ongoing data transfers)*

The following example allows to query all ongoing data transfers in the specified (intermediate) state in a defined FTS service. In order to list only the transfer jobs relative to a channel, this must be specified with the `-c` option.

```

$ glite-transfer-list \
-s https://fts.cnaf.infn.it:8443/sc3/glite-data-transfer-fts/services/FileTransfer Pending
...
c2e2cdb1-a145-11da-954d-944f2354a08b Pending
...

```

#### **Example 7.6.4.4** *(Canceling a job)*

An example of cancellation of a previously submitted data transfer job is shown here:

```

$ glite-transfer-cancel \
-s https://fts.cnaf.infn.it:8443/sc3/glite-data-transfer-fts/services/FileTransfer \
c2e2cdb1-a145-11da-954d-944f2354a08b

```

## **7.7. LOW LEVEL DATA MANAGEMENT TOOLS**

In this section some details on lower level data management tools are given.

### 7.7.1. GSIFTP

The following low level tools can be used to interact with GSIFTP servers on SEs:

<code>edg-gridftp-exists TURL</code>	Checks the existence of a file or directory on a SE
<code>edg-gridftp-ls TURL</code>	Lists a directory on a SE
<code>edg-gridftp-mkdir TURL</code>	Creates a directory on a SE
<code>edg-gridftp-rename sourceTURL destTURL</code>	Renames a file on a SE
<code>edg-gridftp-rm TURL</code>	Removes a file from a SE
<code>edg-gridftp-rmdir TURL</code>	Removes a directory on a SE
<code>globus-url-copy sourceTURL destTURL</code>	Copies files between SEs

**Attention!** The commands `edg-gridftp-rename`, `edg-gridftp-rm`, and `edg-gridftp-rmdir` should be used with extreme care. In fact, these commands do not interact with any of the catalogues and therefore they can compromise the consistency/coherence of the information contained in the Grid.

All the `edg-gridftp-*` commands accept `gsiftp` as the only valid protocol for the TURL.

Some examples are shown. To obtain help on these commands use the option `--usage` or `--help`. More information on the GSIFTP protocol is available in [41].

#### *Example 7.7.1.1 (Listing and checking the existence of Grid files)*

The `edg-gridftp-exists` and `edg-gridftp-ls` commands can be useful in order to check if a file is physically in a SE, regardless of its presence in the Grid catalogues.

```

$ lcg-lr --vo dteam guid:27523374-6f60-44af-b311-baa3d29f841a

sfn://lxb0710.cern.ch/data/dteam/generated/2004-07-13/file42ff7086-8063-414d-9000-
75c459b71296

$ edg-gridftp-exists \
gsiftp://lxb0710.cern.ch/data/dteam/generated/2004-07-13/file42ff7086-8063-414d-9000-
75c459b71296

$ edg-gridftp-exists \
gsiftp://lxb0710.cern.ch/data/dteam/generated/2004-07-13/my_fake_file

error gsiftp://lxb0710.cern.ch/data/dteam/generated/2004-07-13/my_fake_file
does not exist

$ edg-gridftp-ls \
gsiftp://lxb0710.cern.ch/data/dteam/generated/2004-07-13/file42ff7086-8063-414d-9000-
75c459b71296

/data/dteam/generated/2004-07-13/file42ff7086-8063-414d-9000-75c459b71296

```



### *Example 7.7.1.2 (Copying a file with globus-url-copy)*

The `globus-url-copy` command can be used to copy files between any two Grid resources, and from/to a non-grid resource. Its functionality is similar to that of `lcg-cp`, but source and destination must be specified as URLs.

```

globus-url-copy \
gsiftp://lxb0710.cern.ch/data/dteam/generated/2004-07-13/file42ff7086-8063-414d-9000-
75c459b71296 file://`pwd`/my_file
  
```

## 7.7.2. CASTOR and RFIO

Direct access to the CASTOR Mass Storage System (not via SRM) can be obtained through its native CLI. The clients are available in every gLite 3 UI or WN and described below, divided into two logical subgroups. Remember however that CASTOR supports insecure RFIO access only and therefore such commands must be used from a UI or WN in the same LAN of the CASTOR SE.

- Clients interacting with the CASTOR namespace:

<code>nsls</code>	list directories/files in CASTOR
<code>nsfind</code>	search for files in CASTOR
<code>nsmkdir</code>	create a directory in CASTOR
<code>nsrm</code>	remove directories/files from CASTOR
<code>nschmod</code>	change access mode of a directory/file in CASTOR
<code>nschown</code>	change owner and group of a directory/file in CASTOR
<code>nsrename</code>	rename a directory/file in CASTOR
<code>nsln</code>	create a link to a file in CASTOR
<code>nstouch</code>	change the filestamp of a file in CASTOR
<code>nssetcomment</code>	add/replace a comment associated with directory/file in CASTOR
<code>nsdelcomment</code>	delete the comment associated with a directory/file in CASTOR
<code>nssetchecksum</code>	set or reset the checksum for a tape segment
<code>nssetacl</code>	set the ACL for a directory/file in CASTOR
<code>rfcp</code>	Remote file copy
<code>rfstat</code>	Display remote file or filesystem status
<code>rfcat</code>	Remote file concatenation to standard output

- Clients managing CASTOR file classes:

<code>nschclass</code>	change the class of a CASTOR directory in the name server
<code>nsdeleteclass</code>	delete a file class definition
<code>nsenterclass</code>	define a new file class
<code>nslistclass</code>	query the CASTOR Name Server about a given class or list all existing classes
<code>nsmodifyclass</code>	modify an existing file class

File classes reflect into different service classes exposed to the user. For example, the CASTOR file class attribute defines whether a file is permanent or not: all permanent files belong to a class with an associated tape pool; all scratch files belong to a class with no associated tape pool. Setting the file class attributes requires administrator privileges.

### 7.7.3. dCache and DCAP

Analogously to CASTOR and the RFIO protocol, CLIs for direct access to dCache storage systems are available:

`dccp` allows to copy files from/to dCache SEs via the native dcap protocol

## 7.8. JOB SERVICES AND DATA MANAGEMENT

With both the LCG-2 and gLite WMS, some specific JDL attributes allow the user to specify requirements on the input data (see Chapter 6 for information on how to define and manage Grid jobs).

### *Example 7.8.1 (Specifying input data in a job)*

If a job requires one or more input files stored in a SE, the `InputData` JDL attribute can be used. Files can be specified by both by LFN and GUID.

An example of JDL specifying input data looks like:

```
Executable = "/bin/hostname";
StdOutput = "sim.out";
StdError = "sim.err";
DataCatalog = "http://lfc.cern.ch:8085";
InputData = {"lfn:/grid/dteam/does/fileA"};
DataAccessProtocol = {"rfio", "gsiftp", "gsidcap"};
OutputSandbox = {"sim.err", "sim.out"};
```

The `InputData` field may also be specified through GUIDs. This attribute is used only during the match-making process, to find an appropriate CE to run the job. It has nothing to do with the real access to files that the job can do while running. However, it is reasonable to list in `InputData` the files that will be accessed by the job and vice-versa.

The `DataAccessProtocol` attribute is used to specify the protocols that the application can use to access the file and is mandatory if `InputData` is present. Only data in SEs which support one or more of the listed protocols are considered. The WMS will schedule the job to a CE *close* to the SE holding the largest number of input files requested. In case several CEs are suitable, they will be ranked according to the ranking expression.

**Note:** a SE or a list of SEs is published as “close” to a given CE via the `GlueCESEBindGroupSEUniqueID` attribute of the `GlueCESEBindGroup` object class. For more information see Appendix G.

### *Example 7.8.2 (Specifying a Storage Element)*

With the LCG-2 WMS the user can ask the job to run close a specific SE using the attribute `OutputSE`. For example:

```
OutputSE = "srm.cern.ch";
```

The WMS will not submit the job if there is no CE close to the `OutputSE` specified by the user.

**Note:** the same is not true with the gLite WMS. Even if there are CEs close to a given `OutputSE` specified by the user, no resources get matched when this field is defined on the JDL.

### *Example 7.8.3 (Automatic upload and registration of output files)*

In the LCG-2 WMS (but not in the gLite WMS), the `OutputData` attribute allows the user to automatically upload and register files produced by the job on the WN. For each file, three attributes can be set:

- the `OutputFile` attribute is mandatory and specifies the name of the generated file to be uploaded to the Grid;
- the `StorageElement` attribute is an optional string indicating the SE where the file should be stored, if possible. If not specified, the WMS automatically chooses a SE defined as close to the CE;
- the `LogicalFileName` attribute (also optional) represents a LFN the user wants to associate to the output file.

The following code shows an example of JDL requiring explicitly an `OutputData` attribute:

```
Executable      = "test.sh";
StdOutput       = "std.out";
StdError        = "std.err";
InputSandbox    = {"test.sh"};
OutputSandbox   = {"std.out", "std.err"};
OutputData = {
  [
    OutputFile="my_file";
    LogicalFileName="lfn:/grid/dteam/does/my_file";
    StorageElement = "castorsrm.pic.es";
  ]
};
```

Once the job is terminated and the user retrieves its output, the `Output Sandbox` downloaded will contain a further file, automatically generated by the `JobWrapper`, containing the logs of the output upload.

```
$ cat DSUpload_ZboHMYWoBsLVax-nUCmtaA.out
#
# Autogenerated by JobWrapper!
#
# The file contains the results of the upload and registration
# process in the following format:
```

```
# <outputfile> <lfn|guid|Error>
```

```
my_file    guid:2a14e544-1800-4257-afdd-7031a6892ef7
```

#### ***Example 7.8.4*** (Selecting the file catalogue to use for match making)

For the WMS to select CEs that are close to the files required by a job (by the `InputData` attribute), it has to locate the SEs where these files are stored. To do this, the WMS uses the Data Location Interface service, which acts as interface to a file catalogue. Since it is possible that several different DLI services exist on the Grid, the user has the possibility to select which one he wants to talk to by using the JDL attribute `DataCatalog`. The user will specify the attribute and the endpoint of the DLI as value, as in this example:

```
DataCatalog = "http://lfc-lhcb-ro.cern.ch:8085/";
```

If no value is specified, then the first DLI service that is found in the information system is used (which should probably be the right choice for the normal user).

## APPENDIX A THE GRID MIDDLEWARE

The only operating system currently supported by gLite 3 is Scientific Linux 3[45] and the supported architecture is IA32. It is foreseen to have soon support for Scientific Linux 4 and the x86\_64 and IA64 architectures.

The gLite 3 middleware layer uses components from several Grid projects, including EGEE, Datagrid (EDG), DataTag (EDT), DataGrid (EDG), INFN-GRID, Globus and Condor. In some cases, patches have been applied to some components, so the final software used is not exactly the same as the one distributed by the original project.

The components which are currently used in gLite 3 are listed in table 1.

Component	EGEE	EDG	EDT	INFN Grid	Globus	Condor	Other
<b>Basic middleware</b>							
Globus Toolkit 2.4.3 ClassAds 0.9.7					✓	✓	
<b>Authentication and Authorisation</b>							
MyProxy 0.6.1 VOMS VOMRS LCAS/LCMAPS	✓	✓					✓  ✓
<b>Workload management</b>							
Condor-G 6.6.7 EDG WMS gLite WMS	✓	✓				✓	
<b>Data management</b>							
LFC DPM FTS GFAL LCG DM tools	✓ ✓ ✓ ✓ ✓						
<b>Fabric management</b>							
Quattor YAIM	✓						✓
<b>Monitoring</b>							
GridICE				✓			
<b>Information system</b>							
MDS Glue Schema BDII R-GMA LCG Information tools	✓ ✓ ✓	✓	✓		✓		

Table 1: Software components of gLite 3 and projects that contributed to them.

## APPENDIX B ENVIRONMENT VARIABLES AND CONFIGURATION FILES

Some of the configuration files and environmental variables that may be of interest for the Grid user are listed in the following tables. Unless explicitly stated, they are all located/defined in the User Interface.

### Environment variables

Variable	Definition	UI	WN
EDG_LOCATION	EDG middleware installation directory	✓	✓
EDG_WL_JOBID	JobID (defined for a running job)		✓
EDG_WL_LOCATION	LCG-2 WMS UI installation directory	✓	
EDG_WL_RB_BROKERINFO	Location of the .BrokerInfo file		✓
EDG_WL_UI_CONFIG_VAR	Non-standard location of the LCG-2 WMS UI configuration file	✓	
EDG_WL_UI_CONFIG_VO	Non-standard location of the VO-specific LCG-2 WMS UI configuration file	✓	
GLITE_LOCATION	gLite middleware installation directory	✓	✓
GLITE_SD_PLUGIN	Sets the type of service discovery implementation to be used (file, bdi, rgma)	✓	✓
GLITE_SD_SITE	Sets the local site where to find services	✓	✓
GLITE_SD_VO	Sets the default VO for which to find services	✓	✓
GLITE_WMS_CLIENT_CONFIG	Non-standard location of the gLite WMProxy UI configuration file	✓	
GLITE_WMSUI_CONFIG_VAR	Non-standard location of the gLite WMS UI configuration file	✓	
GLITE_WMSUI_CONFIG_VO	Non-standard location of the VO-specific gLite WMS UI configuration file	✓	
GLOBUS_LOCATION	Globus middleware installation directory	✓	✓
LCG_CATALOG_TYPE	Type of file catalogue used by lcg.util and GFAL (it should be lfc)	✓	✓
LCG_GFAL_INFOSYS	BDII contact string for lcg.utils and GFAL (<hostname>:<port>)	✓	✓
LCG_GFAL_VO	User's VO for lcg.utils and GFAL	✓	✓
LFC_HOST	Location of the LFC catalogue	✓	✓
LCG_LOCATION	LCG middleware installation directory	✓	✓
LCG_RFIO_TYPE	Type of RFIO for GFAL (dpm or castor)	✓	✓
VO_<VO>_DEFAULT_SE	Default SE for the VO <VO>	✓	✓
VO_<VO>_SW_DIR	<VO>'s software installation directory		✓
X509_CERT_DIR	Directory containing the CA certificates	✓	✓
X509_USER_CERT	User's certificate file	✓	
X509_USER_KEY	User's private key file	✓	
X509_USER_PROXY	User's proxy certificate file	✓	✓
X509_VOMS_DIR	Directory containing the certificates of the VOMS servers	✓	✓

### Configuration files

Configuration File	Notes
<code>\$GLITE_LOCATION/etc/vomses</code>	System-level configuration of the VOMS CLI
<code>\$HOME/.glite/vomses</code>	User-level configuration of the VOMS CLI
<code>\$GLITE_LOCATION/etc/&lt;vo&gt;/glite_wms.conf</code>	Configuration file for the WMProxy CLI for the VO <VO>
<code>\$GLITE_LOCATION/etc/glite_wmsui_cmd_var.conf</code>	Generic configuration file for the gLite WMS CLI via NS
<code>\$GLITE_LOCATION/etc/&lt;vo&gt;/glite_wmsui.conf</code>	VO-specific configuration file for the gLite WMS CLI via NS for the VO <VO>
<code>\$EDG_WL_LOCATION/etc/edg_wl_ui_cmd_var.conf</code>	Generic configuration file for the LCG-2 WMS
<code>\$EDG_WL_LOCATION/etc/&lt;vo&gt;/edg_wl_ui.conf</code>	VO-specific configuration file for the LCG-2 WMS



## APPENDIX C JOB STATUS DEFINITION

As it was already mentioned in Chapter 6, a job can find itself in one of several possible states. Also, only some transitions between states are allowed. These transitions are depicted in Figure 14. For completeness, also the DAG states are described in 15.

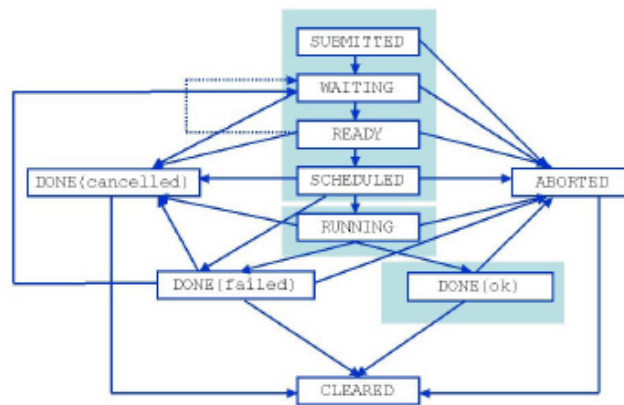


Figure 14: Possible job states in gLite 3

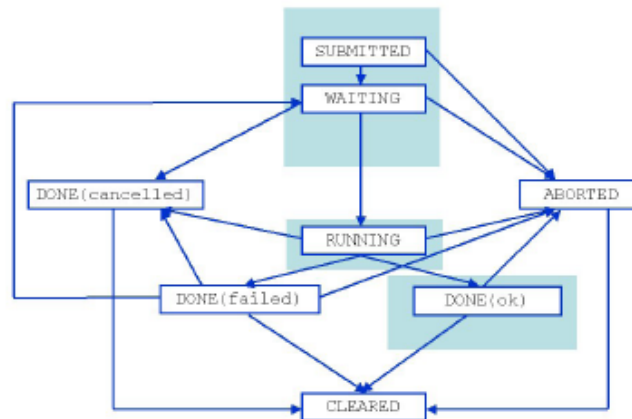


Figure 15: Possible DAG states in gLite 3

And the definition of the different states is given in this table.

<b>Status</b>	<b>Definition</b>
SUBMITTED	The job has been submitted by the user but not yet processed by the Network Server or WMPProxy
WAITING	The job has been accepted by the Network Server or WMPProxy but not yet processed by the Workload Manager
READY	The job has been assigned to a Computing Element but not yet transferred to it
SCHEDULED	The job is waiting in the Computing Element's queue
RUNNING	The job is running
DONE	The job has finished
ABORTED	The job has been aborted by the WMS (e.g. because it was too long, or the proxy certificated expired, etc.)
CANCELED	The job has been canceled by the user
CLEARED	The Output Sandbox has been transferred to the User Interface

## APPENDIX D USER TOOLS

### D.1. INTRODUCTION

This section introduces some tools that are not really part of the gLite 3 middleware stack, but that can be useful for Grid users nonetheless. Rather than a full description, an introduction to the functionality of some of them and pointers to more detailed sources of documentation will be provided in this Appendix.

### D.2. JOB MANAGEMENT FRAMEWORK

When submitting very large numbers of jobs to the WLCG/EGEE Grid, the management and the monitoring of these jobs can be a cumbersome task, if done manually. A simple framework to automatically submit and manage large numbers of jobs is available to assist users in developing more sophisticated job tracking tools.

The framework consists mainly of two commands:

- `submitter_general.pl`: it performs the automatic job submission
- `get_output.pl`: it retrieves and handles the corresponding outputs

More information on this tool can be found in the User level tools Wiki[46].

### D.3. JOB MONITORING

The `lcg-job-monitor` command can be used to monitor from the UI the progress of a job currently running on a WN. This tool provides some information and statistics about a given jobID, like memory usage, swap usage, CPU time, user DN, etc.

The information is retrieved by querying the `JobMonitor` table in R-GMA. The command can return information either for a single job (given the jobID), for a user (given the DN) or for a whole VO. This command currently works only with the LCG-2 WMS.

The command syntax is:

```
lcg-job-monitor [-j <jobID>] [-v <VO>] [-u <DN>] [-q <query_type>]
```

where the `<query_type>` can be `LATEST`, `HISTORY` or `CONTINUOUS`.

More information on this tool can be found in the User level tools Wiki[46].

## D.4. JOB STATUS MONITORING

The `lcg-job-status.py` command allows to recover the logging information of a job from R-GMA. The information is retrieved by querying the `JobStatusRaw` table in R-GMA. This command currently works only with the LCG-2 WMS.

The command syntax is:

```
lcg-job-status.py [-j <jobID>] [-q <type>]
```

where the query type can be either `LATEST`, `CONTINUOUS` or `HISTORY`.

For `LATEST` and `HISTORY` queries, an output is printed and the command exits. In the case of `CONTINUOUS` queries, the status is checked every 5 seconds until the program is interrupted via `Ctrl-C` or a status `Done` or `Aborted` is reached.

More information on this tool can be found in the User level tools Wiki[46].

## D.5. TIME LEFT UTILITY

These commands and API can be invoked from a running job to know:

- how much CPU time or wall clock time the job has consumed;
- how long the job can still run before reaching the CPU time or the wall clock time limit of the batch system queue; this is done by directly querying the batch system whenever possible.

The results returned by the tools are not always accurate, particularly if a site has not set a time limit, or if the batch system is not supported by the tools. Therefore, they should be used with some care.

**Attention!** These commands can generate a significant load on the CE, and therefore they should not be configured to run too often, in particular if there is a large number of concurrent jobs on the CE. A reasonable time interval would be one hour.

The following files should be present in the WN (either already in the release or shipped with the job in a tarball):

- `lcg-getJobStats`: a wrapper bash script around the corresponding Python script;
- `lcg-getJobTimes`: a wrapper bash script around the corresponding Python script;
- `lcg-getJobStats.py`: a Python script;
- `lcg-getJobTimes.py`: a Python script;
- `lcg_jobConsumedTimes.py`: a Python module;
- `lcg_jobStats.py`: a Python module.

The only commands the user should normally use are `lcg-getJobStats` or `lcg_jobStats.py`. The other commands, `lcg-getJobTimes` or `lcg_jobConsumedTimes.py`, are used to estimate the used CPU time and wall clock time without querying the batch system, but by parsing the proc filesystem; they are internally called by `lcg-getJobStats` when it cannot get the information from the batch system.

More information on this tool can be found in the User level tools Wiki[46].

## APPENDIX E VO-WIDE UTILITIES

### E.1. INTRODUCTION

This section describes some administrative tools that are only relevant to selected VO members (VO managers, VO software managers, etc.). Links to other sources of documentation are provided when available.

### E.2. FREEDOM OF CHOICE FOR RESOURCES

The *Freedom of Choice for Resources (FCR)* is a tool for VO Software Managers to set up selection rules for Computing and Storage Elements, which will determine whether a particular resource will be available or not to the VO.

The FCR interface[49] allows the VO manager to decide if a resource supporting his VO should be visible at any time, invisible at any time or visible only if it passes periodic tests run by the WLCG/EGEE operations team, or by the VO itself (see next section). The VO manager can also decide what tests are to be considered critical for its VO: only the failure of critical tests determines the exclusion of a resource.

### E.3. SERVICE AVAILABILITY MONITORING

The *Service Availability Monitoring (SAM)*[50] is a framework to provide a centralized and uniform monitoring tool for all Grid services.

It is based on the concept of running periodic tests on all known Grid services to determine whether they are working properly. These tests can both be directly run from a User Interface, or sent out as Grid jobs (for example, to test a Computing Element). It provides detailed information about the overall status of the service, and about the outcome of the individual tests, and keeps a historical record of the information.

The SAM is a very useful tool both for the WLCG/EGEE operations team and for the VO members. It also allows Virtual Organisations to complement the standard tests with custom tests covering VO-specific functionalities.

The user can view the results of the SAM tests on the production WLCG/EGEE infrastructure from the SAM Web interface [51]. From this page, the user can choose the service type (CE, LFC, WMS, etc.), a VO, a region and the specific tests he is interested in.

### E.4. THE VO BOX

The *VO box*[52] is a type of node, which is deployed at many sites, where the VO can run specific agents and services. The access to the VO box is restricted to the VO software manager of the VO.

The VO box offers basically two main features to the users:

- VOs can run their own services from this node;
- it provides direct access to the software area of each VO, also accessible from all WNs of the site.

Each VO should negotiate with the site the setup of the VO box depending on the services which are run inside that node.

## E.5. VO SOFTWARE INSTALLATION

Authorized users can install VO-specific software on the computing resources of WLCG/EGEE. The availability of such software can be advertised in the Information System[47].

The *VO Software Manager* is the member of the VO with the privileges to install VO-specific software on the different sites. The software manager can install, validate or remove VO-specific software on a site at any time through a normal Grid job. Normally, the software manager privileges are expressed by a special VOMS role, which must be taken when creating the VOMS proxy used to submit the software installation job.

The VO software manager can also modify the VO-specific information for the CEs using the command `lcg-ManageVOTag`, available from the UI and the WN.

Each site should provide a dedicated space where each supported VO can install or remove software. The amount of available space must be negotiated between the VO and the site, as well as any special priority for software installation jobs.

## APPENDIX F DATA MANAGEMENT AND FILE ACCESS THROUGH AN APPLICATION PROGRAMMING INTERFACE

In this section, an overview of the available Data management API will be given, and some details on the most high-level API will be provided.

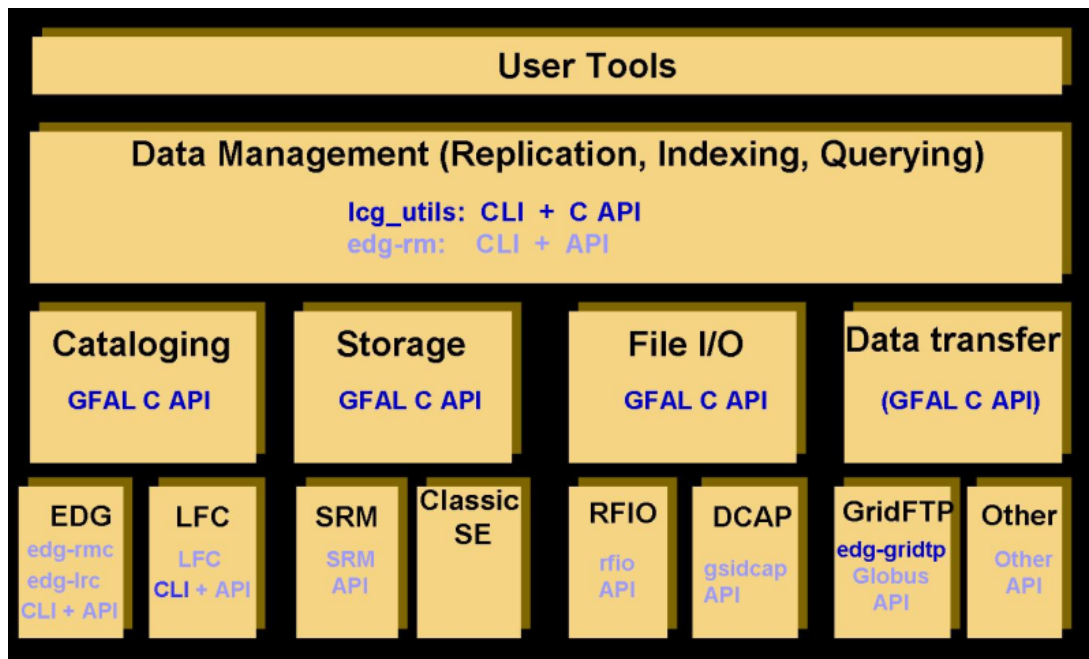


Figure 16: Layered view of the Data Management APIs and CLIs

Figure 16 shows a layered view of the different gLite Data Management API and of the CLI which were described earlier. The CLI and API whose use is discouraged are shadowed.

Just below the user tools, we find the `lcg_util` API. This is a C API that provides the same functionality as the `lcg-*` commands for Data Management we have already seen: in fact, the commands are just wrappers around the C calls. This layer should cover most of the basic needs of user applications. It is independent from the underlying technology, since it will transparently interact with the LFC catalogue and will use the correct protocol (GSIFTP, RFIO or `gsidcap`) for file transfer.

In the following table, all the available methods and a short description are listed.



Method name	Description
lcg_aa	add an alias in the LFC for a given GUID
lcg_cp	copy a Grid file to a local destination
lcg_cr	copy and register a file
lcg_del	delete one file (either one replica or all replicas)
lcg_gt	get the TURL given the SURL and the transfer protocol
lcg_la	get the list of aliases for a given LFN, GUID or SURL
lcg_lg	get the GUID for a given LFN or SURL
lcg_lr	get the list of replicas for a given LFN, GUID or SURL
lcg_ra	remove an alias in LFC for a given GUID
lcg_rep	copy a file from one SE to another and register it in LFC
lcg_rf	register in LFC a file residing on a SE
lcg_sd	set file status to “Done” for a given SURL in a specified request
lcg_uf	unregister in LFC a file residing on a SE

Apart from the basic calls `lcg_cp`, `lcg_cr`, etc., there are other calls that enhance them with a buffer for complete error messages (`lcg_cpx`, `lcg_crx`, ...), that include timeouts (`lcg_cpt`, `lcg_crt`, ...), and both (`lcg_cpxt`, `lcg_crxt`). Actually, all calls use the most complete version (i.e. `lcg_cpxt`...) with default values for the arguments that were not provided.

Below the `lcg_util` API, we find *the Grid File Access Library (GFAL)*. GFAL provides a POSIX-like interface for I/O operations on Grid files, effectively hiding the interactions with the LFC, the SEs and SRM. The function names are obtained by prepending `gfal_` to the POSIX names, for example `gfal_open`, `gfal_read`, `gfal_close`.

GFAL accepts GUIDs, LFNs, SURLs and TURLs as file names. It will automatically select the most appropriate transfer protocol, depending on the kind of SE the file is located on (if a TURL is used, the protocol is already implicitly specified).

**Note:** In the case where LFNs or GUIDs are used, GFAL (and, as a consequence, `lcg_util`) needs to contact the LFC to obtain the corresponding TURL. For GFAL to be able to discover the LFC endpoints and to find out information about the Storage Elements, the user must set the environment variables `LCG_GFAL_VO` and `LCG_GFAL_INFOSYS` to the VO name and the BDII hostname and port. For example:

```
export LCG_GFAL_VO=cms
export LCG_GFAL_INFOSYS=exp-bdii.cern.ch:2170
```

The endpoint of the catalogue may also be directly specified by setting the environment variable `LFC_HOST`. For example:

```
export LFC_HOST=prod-lfc-cms-central.cern.ch
```

In Figure 17, it is shown the flow diagram of a `gfal_open` call. This call will locate a Grid file and return a remote file descriptor so that the caller can read or write file remotely, as it would do for a local file. As shown in the figure, first, if a GUID is provided, GFAL will contact a file catalogue to retrieve the corresponding SURL. Then, it will access the SRM interface of the SE that the SURL indicates, it will get a valid TURL and also pin the file so that it is there for the subsequent access. Finally, with the TURL and using the appropriate protocol, GFAL will open the file and return a filehandle to the caller.

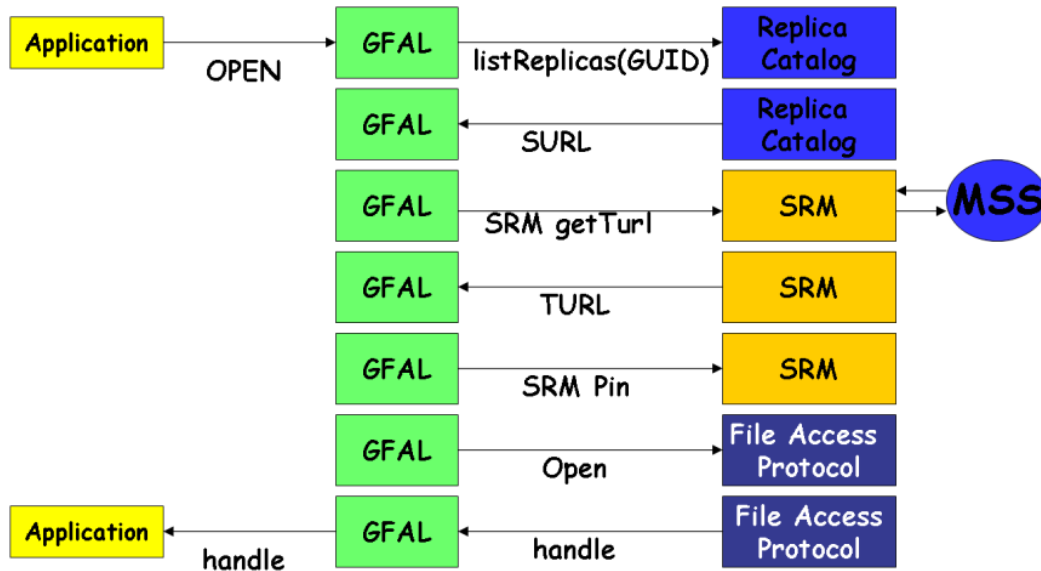


Figure 17: Flow diagram of a GFAL call

It is important to notice that if a file is created with GFAL naming it by SURL, for it to be used as a Grid file, it should be manually registered with a LFN using `lcg-rf`.

In addition, GFAL may expose functionality applicable only to a specific underlying technology (or protocol), if this is considered useful. A good example of this is the exposed SRM interface that GFAL provides. Some code exploiting this functionality is shown later.

For more information on GFAL, refer to the manpages of the library (`gfal`) and of the different calls (`gfal_open`, `gfal_write`...).

Finally, below GFAL, we find some other CLI and API which are technology dependent. Their direct use is in general discouraged (except for the mentioned cases of the LFC client tools and the `edg-gridftp-*` commands). Nonetheless, some notes on the RFIO API are given later on.

**Example F.0.1** (Using `lcg_util` API to transfer a file)

The following example copies a file from a SE to the local host. The file can be then accessed locally with normal file I/O calls.

The source code follows (`lcg_cp_example.cpp`):

```
#include <iostream>

extern "C"{
```

```

#include "lcg_util.h"
}

using namespace std;

int main(int argc, char **argv){

/* Check syntax (there must be 2 arguments) */
if (argc != 3) {
    cerr << "Usage: " << argv[0]<< " source destination\n";
    exit (1);
}
char * src_file=argv[1];
char * my_file=argv[2];
char * dest_file=new char[200];
char * vo=getenv("LCG_GFAL_VO");
int nbstreams=1;
int verbose=1;
int timeout=180;

/* Form the name of the destination file */
char * pwd=getenv("PWD");
strcpy(dest_file,"file:");
strcat(dest_file,pwd);
strcat(dest_file,"/");
strcat(dest_file,my_file);

/* The lcg_cp call itself */
if(lcg_cpt(src_file, dest_file, vo, nbstreams, 0, 0, verbose, timeout)==0){
    cout << "File correctly copied to local filesystem " << endl;
}
else{
    perror("Error with lcg_cp!");
}

/* That was it */
cout << endl;
return 0;

} //end of main

```

The code can be compiled with the following command:

```

$ c++ -I$LCG_LOCATION/include -L$LCG_LOCATION/lib -L$GLOBUS_LOCATION/lib \
    -llcg_util -lgfal -lglobus_gass_copy_gcc32 -o lcg_cp_example lcg_cp_example.cpp

```

**Note:** The link with libglobus\_gass\_copy\_gcc32.so should not be necessary, and also the one with libgfal.so

should be done transparently when linking `liblcg_util.so`. Nevertheless, their explicit link as shown in the example was necessary for the program to compile in the moment that this guide was written.

The resulting executable will take two arguments: the name of a Grid file and a local name for the file in the current directory. For example:

```

$ ./lcg_cp_example lfn:/grid/cms/does/gridfile localfile
Using grid catalog type: lfc
Using grid catalog : prod-lfc-cms-central.cern.ch
Source URL: lfn:/grid/cms/does/gridfile
File size: 7840
VO name: cms
Source URL for copy: gsiftp://lxfsr4601.cern.ch//castor/cern.ch/grid/cms/generated/
2006-11-14/fileee903ced-b61a-4443-b9d2-a4b0758721a8
Destination URL: file:/home/does/localfile
# streams: 1
# set timeout to 180 (seconds)
          0 bytes      0.00 KB/sec avg      0.00 KB/sec inst
Transfer took 6080 ms
File correctly copied to local filesystem
  
```

### ***Example F.0.2 (Using GFAL to access a file)***

The following C++ code uses GFAL to access a Grid file. The program opens the file, writes a set of numbers into it, and closes it. Afterwards, the file is opened again, and the previously written numbers are read and shown to the user. The source code (`gfal_example.cpp`) follows:

```

#include<iostream>
#include <fcntl.h>
#include <stdio.h>
extern "C" {
#include "/opt/lcg/include/gfal_api.h"
}

using namespace std;

/* Include the gfal functions (are C and not C++, therefore are 'extern') */
extern "C" {
  int gfal_open(const char*, int, mode_t);
  int gfal_write(int, const void*, size_t);
  int gfal_close(int);
  int gfal_read(int, void*, size_t);
}

/***** MAIN *****/
main(int argc, char **argv)
  
```

```

{
int fd; // file descriptor
int rc; // error codes
size_t INTBLOCK=40; // how many bytes we will write each time (40 = 10 int a time)

/* Check syntax (there must be 2 arguments) */
if (argc != 2) {
    cerr << "Usage: " << argv[0]<< "filename\n";
    exit (1);
}

/* Declare and initialize the array of input values (to be written in the file) */
int* original = new int[10];
for (int i=0; i<10; i++) original[i]=i*10; // just: 0, 10, 20, 30...

/* Declare and give size for the array that will store the values read from the file */
int* readValues = new int[10];

/* Create the file for writing with the given name */
cout << "\nCreating file " << argv[1] << endl;
if ((fd = gfal_open (argv[1], O_WRONLY | O_CREAT, 0644)) < 0) {
    perror ("gfal_open");
    exit (1);
}
cout << " ... Open successful ... " ;

/* Write into the file (reading the 10 integers at once from the int array) */
if ((rc = gfal_write (fd, original, INTBLOCK )) != INTBLOCK) {
    if (rc < 0)    perror ("gfal_write");
    else cerr << "gfal_write returns " << rc << endl;
    (void) gfal_close (fd);
    exit (1);
}
cout << "Write successful ... ";

/* Close the file */
if ((rc = gfal_close (fd)) < 0) {
    perror ("gfal_close");
    exit (1);
}
cout << "Close successful" << endl;

/* Reopen the file for reading */
cout << "\nReading back " << argv[1] << endl;
if ((fd = gfal_open (argv[1], O_RDONLY, 0)) < 0) {
    perror ("gfal_open");
    exit (1);
}
cout << " ... Open successful ... ";

```

```

/* Read the file (40 bytes directly into the readValues array) */
if ((rc = gfal_read (fd, readValues, INTBLOCK) != INTBLOCK) {
  if (rc < 0) perror ("gfal_read");
  else cerr << "gfal_read returns " << rc << endl;
  (void) gfal_close (fd);
  exit (1);
}
cout << "Read successful ...";

/* Show what has been read */
for(int i=0; i<10; i++)
  cout << "\n\tValue of readValues[" << i << "] = " << readValues[i];

/* Close the file */
if ((rc = gfal_close (fd)) < 0) {
  perror ("gfal_close");
  exit (1);
}
cout << "\n ... Close successful";
cout << "\n\nDone" << endl;

} //end of main

```

The command used to compile and link the previous code (it may be different in your machine) is:

```
$ c++ -I$LCG_LOCATION/include -L$LCG_LOCATION/lib -l gfal -o gfal_example gfal_example.cpp
```

As temporary file, we may specify one in our local filesystem, by using the file:// prefix. In that case we get the following output:

```

$ ./gfal_example file://`pwd`/test.txt

Creating file file:///afs/cern.ch/user/d/does/gfal/test.txt
... Open successful ... Write successful ... Close successful

Reading back file:///afs/cern.ch/user/d/does/gfal/test.txt
... Open successful ... Read successful ...
  Value of readValues[0] = 0
  Value of readValues[1] = 10
  Value of readValues[2] = 20
  Value of readValues[3] = 30
  Value of readValues[4] = 40
  Value of readValues[5] = 50
  Value of readValues[6] = 60
  Value of readValues[7] = 70
  Value of readValues[8] = 80

```

```

    Value of readValues[9] = 90
    ... Close successful
  
```

Done

This example will not work in all cases from a UI, though. Due to the limitations of the insecure RFIO protocol, GFAL can work with a classic SE or a CASTOR SE only from a worker node at the same site. The reason is that insecure RFIO does not handle Grid certificates, and while the local UNIX user ID to which a user job is mapped on the WN will be allowed to access a file in the local SE, the UNIX user ID of the user on the UI will be normally different, and will not be allowed to perform that access.

In opposition to the insecure RFIO, the secure version, also called *gsirfio*, includes all the usual GSI security, and so it can deal with certificates rather than with UNIX user IDs. For this reason, it can be used with no problem to access files from UIs or in remote SEs, just as *gsidcap* can. As a consequence, the example will work without any problem from the UI with DPM and dCache.

**Attention:** Some SEs support only insecure RFIO (classic SEs and CASTOR), while others support only secure RFIO (DPM), but they all publish `rfio` as the supported protocol in the Information System. The result is that currently GFAL has to figure out which one of the two RFIO versions it uses based on the environment variable `LCG_RFIO_TYPE`. If its value is `dpm`, the secure version of RFIO will be used; otherwise insecure RFIO will be the used. Therefore, the user must correctly define the indicated variable depending on the SE he wants to talk to.

Another important issue is that of the names used to access files. For classic SEs, SURLS and TURLs must include a double slash between the hostname of the SE and the path of the file. This is a known limitation in GFAL and insecure RFIO. For example:

```

sfn://lxb0710.cern.ch//flatfiles/SE00/dteam/my_file
rfio://lxb0710.cern.ch//flatfiles/SE00/dteam/my_file
  
```

As seen in previous examples, the `lcg-*` commands will work with SURLS and TURLs registered in the catalogues, even if they do not follow this rules. Therefore, it is always better to use LFNs or GUIDs when dealing with files, not to have to deal with SURL and TURL naming details.

In addition to GFAL, there is also the possibility to use the RFIO C and C++ API, which also allows to remotely open and read a file. Nevertheless, this is not recommended, as it can work only with classic SEs and CASTOR SEs located in the same local area network, and RFIO does not understand LFNs, GUIDs or SURLS. More information on RFIO and its API can be found in [42].

### *Example F.0.3 (Explicit interaction with the SRM using GFAL)*

The following example program can be useful for copying a file that is stored in a MSS. It asks for the file to be staged from tape to disk first, and only tries to copy it after the file has been migrated.

The program uses both the `lcg_util` and the GFAL API. From `lcg_util`, just the `lcg_cp` call is used. From GFAL, `srm_get`, which requests a file to be staged from tape to disk, and `srm_get_status`, which checks the status of the previous request, are used.

The source code follows:

```

#include <stdio.h>
#include <stdlib.h>
#include <sys/types.h>
#include <iostream>
#include <sstream> // for the integer to string conversion
#include <unistd.h> // for the sleep function
#include <fstream> // for the local file access
extern "C"{
    #include "gfal_api.h"
    #include "lcg_util.h"
}

using namespace std;

main(int argc, char ** argv){
/* Check arguments */
if ((argc < 2) || (argc > 2)) {
    cerr << "Usage: " << argv[0] << " SURL\n";
    exit (1);
}

/*
 * Try to get the file (stage in)
 * int srm_get (int nbfiles, char **surls, int nbprotocols, char **protocols, int *reqid,
 *             char **token, struct srm_filestatus **filestatuses, int timeout);
 *
 * struct srm_filestatus{
 *   char *surl;
 *   char *turl;
 *   int fileid;
 *   int status;};
 */
int nbreplies; //number of replies returned
int nbfiles=1; // number of files
char **surls; // array of SURLs
int nbprotocols; // number of bytes of the protocol array
char * protocols[] = {"rfio"}; // protocols
int reqid; // request ID
//char **token=0; // unused
struct srm_filestatus *filestatuses; // status of the files
int timeout=100;

/* Set the SURL and the nbprotocols */
surls = &argv[1];
nbprotocols = sizeof(protocols) / sizeof(char *);

/* Make the call */

```



```

if ((nbreplies = srm_get (nbfiles, surls, nbprotocols, protocols,
    &reqid, 0, &filestatuses, timeout)) < 0) {
    perror ("Error in srm_get");
    exit (-1);
}

/* Show the retrieved information */
cout << "\nThe status of the file is: " << endl;
cout << endl << filestatuses[0].status << " -- " << filestatuses[0].surl;
free(filestatuses[0].surl);
if(filestatuses[0].status == 1){
    cout << " (" << filestatuses[0].turl << ")" << endl;
    free(filestatuses[0].turl);
}
else {cout << endl;}
free(filestatuses);

if(filestatuses[0].status == -1){
    cout << endl << "Error when trying to stage the file. Not waiting..." << endl;
    exit(-1);
}

/*
 * Now watch the status until it gets to STAGED (1)
 * int srm_getstatus (int nbfiles, char **surls, int reqid, char **token,
 *                    struct srm_filestatus **filestatuses, int timeout);
 */
cout << "\nWaiting for the file to be staged in..." << endl;
int numiter=1;
int filesleft=1;

char * destfile = new char[200];
while((numiter<50) && (filesleft>0)){
    //sleep longer each iteration
    sleep(numiter++);
    cout << "#"; // just to show we are waiting and not dead
    cout.flush();

    if ((nbreplies = srm_getstatus (nbfiles, surls, reqid, NULL, &filestatuses, timeout))
        < 0) {
        perror ("srm_getstatus");
        exit (-1);
    }

    if (filestatuses[0].status == 1){
        cout << "\nREADY -- " << filestatuses[0].surl << endl;
        filesleft--;
        // Create a name for the file to be saved

```

```

    strcpy(destfile, "file:/tmp/srm_gfal_retrieved");
    cout << "\nCopying " << filestatuses[0].surl << " to " << destfile << "... \n";
    // Copy the file to the local filesystem
    if(lcg_cp(filestatuses[0].surl, destfile, "dteam", 1, 0, 0 , 1)!=0){
        perror("Error in lcg_cp");
    }
}
free(filestatuses[0].surl);
if(filestatuses[0].status == 1) free(filestatuses[0].turl);
free(filestatuses);
}

if(numiter>49){
    cout << "\nThe file did not reach the READY status. It could not be copied." << endl;
}

/* Cleaning */
delete [] destfile;

/* That was all */
cout << endl;
return reqid; // return the reqid, so that it can be used by the caller
} //end of main

```

The `srm_get` function is called once to request the staging of the file. In this call, we retrieve the corresponding TURL and some numbers identifying the request. If a LFN was provided, several TURLs (from several replicas) could be retrieved. In this case, only one TURL will be returned (stored in the first position of the `filestatuses` array).

The second part of the program is a loop that will repeatedly call `srm_getstatus` in order to get the current status of the previous request, until the status is equal to 1 (ready). There is a `sleep` call to let the program wait some time (time increasing with each iteration) for the file staging. Also a maximum number of iterations is set (50), so that the program does not wait forever, but rather ends finally with an aborting message.

When the file is ready, it is copied using `lcg_cp` in the same way as seen in a previous example.

A possible output of this program is the following:

The status of the file is:

```
0 -- srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1
```

Waiting for the file to be staged in...

```
#####
```

```
READY -- srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1
```

```
Copying srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1 to
file:/tmp/srm_gfal_retrieved...
Source URL: srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1
File size: 2331
Source URL for copy:
gsiftp://castorgrid.cern.ch:2811//shift/lxfs5614/data03/cg/stage/test_1.172962
Destination URL: file:/tmp/srm_gfal_retrieved
# streams: 1
Transfer took 590 ms
```

where the 0 file status means that the file exists but it lies on the tape (not staged yet), the hash marks show the iterations in the looping and finally the READY indicates that the file has been staged in and it can be copied (what it is done afterwards as shown by the normal verbose output).

If the same program were run a second time, passing the same SURL as argument, it would return almost immediately, since the file has been already staged. This is shown in the following output:

The status of the file is:

```
1 -- srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1
(rfio://lxfs5614//shift/lxfs5614/data03/cg/stage/test_1.172962)
```

Waiting for the file to be staged in...

```
#
READY -- srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1
```

```
Copying srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1 to
file:/tmp/srm_gfal_retrieved...
Source URL: srm://castorsrm.cern.ch/castor/cern.ch/grid/dteam/testSRM/test_1
File size: 2331
Source URL for copy:
gsiftp://castorgrid.cern.ch:2811//shift/lxfs5614/data03/cg/stage/test_1.172962
Destination URL: file:/tmp/srm_gfal_retrieved
# streams: 1
Transfer took 550 ms
```

where the 1 file status means that the file is already in disk.

## APPENDIX G THE GLUE SCHEMA

The GLUE information schema [21] provides a standardised description of a Grid computing system, to enable resources and services to be presented to users and external services in a uniform way. The schema has been defined as a joint project between a number of Grids. It does not attempt to model the real systems in any detail, but rather to provide a set of attributes to facilitate key use cases. The intended uses are resource discovery (“what is out there?”), selection (“what are the properties?”) and monitoring (“what is the state of the system?”). Inevitably, when real systems are mapped on to the standard schema some details are lost and some features may not entirely match the assumptions in the schema.

The schema has evolved as experience has been gained and new features have required schema support. The current schema version deployed with gLite 3 is 1.2, but version 1.3 has now been defined and is expected to be deployed during 2007. A working group is also being created in the context of the Open Grid Forum to define a major revision (2.0) to be deployed in 2008.

### G.1. BASIC CONCEPTS

The schema itself is defined in an abstract way, using a simplified UML description, in terms of *objects* which have *attributes* and *relations* to other objects. Some attributes may be *multivalued*, i.e. there may be many instances of the same attribute for one object instance, and most attributes are *optional*. Most objects have an *identifier*, which is either globally unique (*UniqueID*) or unique in its local context (*LocalID*). Many objects also have a human-readable *Name* to provide a more user-friendly description (i.e. the Name is an attribute of an object instance, as opposed to the name of the abstract object itself). Neither the UniqueID nor the LocalID should have any semantics, i.e. they should not be interpreted or decoded by the middleware.

Attributes have *types*. In many cases these are simply `int32` or `string`. However, where possible more restrictive types are used, notably for URLs (or the more general URIs) and for enumerated lists, where the allowed values must be taken from a defined set. Other restrictions on values may be implicit, e.g. for latitude and longitude. In most cases, if an attribute is not available or is irrelevant it is simply omitted, rather than taking a special value.

The schema is split into a number of high-level pieces. *Site* defines some information about an entire Grid site. *Service* provides a general abstraction for any Grid service. *CE* and *SE* provide detailed information about Computing and Storage Elements, as these are the most important components of the Grid. In GLUE terms a CE represents a queue in a batch system; the schema also describes a *Cluster* which represents the hardware (Worker Nodes) which can be accessed via a CE.

Finally, *CESEBind* allows relationships between “close” CEs and SEs to be defined. Historically there was also a *Host* concept to allow individual Worker Nodes to be represented in detail, but in a Grid context this has not proved to be useful.

These high-level abstractions are in most cases broken down further into a number of objects representing those details of the system which need to be exposed to the outside world.

## G.2. MAPPINGS

The abstract schema has a set of *mappings* to specific Information System technologies, currently LDAP, R-GMA (relational) and XML, of which the first two are currently in use in WLCG/EGEE. These technologies are quite different and therefore the mappings have significantly different structures, although the basic objects and attributes are preserved. For example, relational tables cannot support multivalued attributes directly (a table cell can only hold a single value), hence these are split into separate tables with an extra key attribute. Similarly the LDAP mapping introduces extra attributes (*ForeignKey* and *ChunkKey*) to allow relations between objects to be expressed in a way which is usable in LDAP queries.

These mappings are to some extent a matter of judgement, hence they are defined by hand rather than being automated. Details can be found on the GLUE web site [21], or simply by exploring the LDAP or R-GMA structures with a browser.

A further mapping is to the Classad language used in JDL files, which is the principle way in which most users interact with the schema. The WMS is able to convert from both LDAP and R-GMA to this form. In most respects this “flattens” the schema and simply provides a list of attributes which can be used to construct the `Rank` and `Requirements` expressions.

## G.3. INFORMATION PROVIDERS

The information presented in the schema is produced by programs known as *Information Providers*. These are structured as a framework *Generic Information Provider* with a set of plugins for specific parts of the schema, which may vary depending on the circumstances, e.g. to cope with different batch systems or SE technologies. The same providers are used for both LDAP and R-GMA.

Broadly speaking these divide into *static* and *dynamic* providers. Dynamic information (e.g. the number of running jobs) changes on a short timescale and therefore has to be collected from some other system (e.g. a batch system) every time the provider is run. By contrast, static information (e.g. the maximum CPU time for a queue) changes infrequently, and the providers therefore read it from a configuration file which is updated only when the system is reconfigured.

To reduce the load on the underlying systems the dynamic information is normally cached for a short time. In addition there are usually further caches and propagation delays in the wider Information Systems. It should therefore be assumed that dynamic information is always somewhat out of date, typically by a few minutes.

## G.4. GLUE ATTRIBUTES

The rest of this appendix provides some information about those schema attributes which are the most important in WLCG/EGEE. For a full list of attributes see the GLUE documentation. Attributes not mentioned here are generally either not defined in WLCG/EGEE, not usable in practice, or are standard elements like `UniqueID` or `Name`.

It should be emphasised that the Information Providers in WLCG/EGEE do not provide everything defined in the schema (most schema attributes are optional), and in addition WLCG/EGEE imposes some extra constraints

on the attributes which are not part of the schema itself, hence some of the comments here do not apply to the schema in general. Attributes defined in the 1.3 schema are included here for completeness, but it should be borne in mind that this schema is not expected to be fully deployed for some time (the LDAP mapping includes the schema version as attributes, `GlueSchemaVersionMajor` and `GlueSchemaVersionMinor`). Some attributes are deprecated; these are mentioned here if they are still in common use.

### G.4.1. Site information

This provides information about a grid site as a whole. Most of the information is intended to be human-readable, as opposed to being used by the middleware. Most entries are set by hand by the system managers and hence may vary from site to site, although some are configured in a standard way by the WLCG/EGEE tools.

- **GlueSite object**

- `GlueSiteUniqueID`: This is the unique name for the site, as defined in the GOC database. This is more or less human-readable, and the Name is currently the same string in most cases.
- `GlueSiteDescription`: A general description of the site.
- `GlueSiteEmailContact`: A `mailto:` URL defining a general email contact address for the site; however, note that in WLCG/EGEE users should normally contact sites via GGUS. Separate attributes may define specific contacts for user support, security and system management.
- `GlueSiteLocation`: The geographical location of the site as a string, normally in the form City, State, Country.
- `GlueSiteLatitude`, `GlueSiteLongitude`: The map reference for the site, in degrees. The resolution usually locates the site to within 100m.
- `GlueSiteWeb`: A URL pointing to a web page relating to the site.
- `GlueSiteSponsor`: The organisation(s) providing funding for the site.
- `GlueSiteOtherInfo`: A multivalued string which may contain any further information the site considers useful; in WLCG/EGEE this generally includes the Tier affiliation, in the form TIER-n.

### G.4.2. Service information

This provides a general abstraction of a Grid service (not necessarily a web service, and perhaps not even something with an externally visible endpoint). This information is also available via the *Service Discovery* API (see Section 5.3). At present the Service information is not directly related to the CE and SE information described below, but it is likely that the proposed major revision of the GLUE schema will describe everything as a specialisation of a service concept.

- **GlueService object**

- `GlueServiceType`: The service type, taken from a defined list which can be found on the Glue web site [21].
- `GlueServiceVersion`: The version of the service, in the form major.minor.patch.

- `GlueServiceEndpoint`: The network endpoint for the service.
- `GlueServiceStatus`: The status of the service (one of OK, Warning, Critical, Unknown, Other).
- `GlueServiceStatusInfo`: A textual explanation of the Status.
- `GlueServiceWSDL`: For web services this is a URL pointing to the WSDL definition of the service.
- `GlueServiceSemantics`: This is a URL which would typically point to a web page explaining how to use the service.
- `GlueServiceOwner`: The service owner, if any; typically a VO name.
- `AccessControlBaseRule`: A set of ACLs defining who is allowed access to the service.

### G.4.3. Attributes for the Computing Element

These are attributes that give information about the computing system (batch queues and Worker Nodes). These are mostly available for use in the JDL, and consequently are the most important for most users.

Note that the term CE is overloaded; in different contexts it may refer to the front-end machine through which jobs are submitted, or to the entire set of computing hardware at a site. However, in the GLUE schema a CE is a single queue in a batch system, and there are typically many CEs at a site submitting jobs to the same set of WNs. This means that attributes published per-CE, e.g. the total number of available CPUs (or job slots), cannot be summed in a simple way.

The original schema concept was to represent the computing hardware as a Cluster consisting of one or more SubClusters, where each SubCluster would describe a set of identical WNs. However, limitations in the WMS mean that currently only a single SubCluster per Cluster (and hence per CE) is supported. Most Grid sites have WNs which are heterogeneous in various ways (processor speed, memory size etc), hence they have to publish the attributes of a representative WN which may not always match the hardware on which a job actually runs. However, the variations are generally not too large, and nodes should not differ in more vital attributes like the OS version. In some cases sites may publish more than one Cluster, e.g. to make a set of nodes with very large memory available via a separate set of queues.

Many sites use scheduling policies which give jobs priority according to who submits them, often to give specific VOs a higher priority. This was not representable in the original schema, e.g. a queue shown with a large number of queued jobs might in fact execute a job from a particular VO immediately. As a result most sites have configured separate queues for each VO. However, this increases the management effort, and can also result in a very large number of CEs being published. The 1.2 schema revision therefore introduced the concept of a VOView, which allows a subset of the CE information to be published separately for each VO (or subgroup) for which scheduling policies are defined. This is supported by the latest version of the WMS, so it is likely that the Grid sites will gradually move back to a single set of generic queues.

#### • GlueCE object

- `GlueCEUniqueID`: The unique identifier for the CE. This is constructed from various information including a host name, batch system type and queue name, but for most purposes it should simply be treated as an identifier. The constituent information is available in other attributes if needed.
- `GlueCECapability`: Introduced in version 1.3 of the schema, this will enable sites to advertise any features not represented by specific attributes.

- `GlueCEInfoTotalCPUs`: The total number of CPUs on all WNs available via the CE, which is usually the maximum number of jobs which can run. This attribute is deprecated in favour of `MaxRunningJobs`.
- `GlueCEInfoApplicationDir`: The path of a directory in which application software is installed; normally each VO has a subdirectory within this directory.
- `GlueCEInfoDefaultSE`: The unique identifier of an SE which should be used by default to store data.
- `GlueCEStateStatus`: The queue status: one of `Queueing` (jobs are accepted but not run), `Production` (jobs are accepted and run), `Closed` (jobs are neither accepted nor run), or `Draining` (jobs are not accepted but those already in the queue are run). The JDL normally has a default `Requirement` for the `Status` to be `Production`.
- `GlueCEStateTotalJobs`: The total number of jobs in this queue (running + waiting).
- `GlueCEStateRunningJobs`: The number of running jobs in this queue.
- `GlueCEStateWaitingJobs`: The number of jobs in this queue waiting for execution.
- `GlueCEStateWorstResponseTime`: The worst-case time between the submission of a new job and the start of its execution, in seconds.
- `GlueCEStateEstimatedResponseTime`: An estimate of the likely time between the submission of a new job and the start of its execution, in seconds. This is usually the default `Rank` in the JDL, i.e. jobs will be submitted to the queue with the shortest estimated time to execution. However, note that the estimate may not always be very accurate, and that all queues which currently have free execution slots will have an `EstimatedResponseTime` of 0 (or close to 0).
- `GlueCEStateFreeCPUs`: The number of CPUs not currently running a job. This is deprecated in favour of `FreeJobSlots`, since the relationship between CPUs and jobs is not always one-to-one.
- `GlueCEStateFreeJobSlots`: The number of jobs that could start immediately if submitted to this queue.
- `GlueCEPolicyMaxWallClockTime`: The maximum wall clock time (i.e. real time as opposed to CPU time) allowed for jobs submitted to this queue, in minutes. Jobs will usually be killed automatically after this time. Specify a `Requirement` on this attribute for jobs which are expected to spend a significant time waiting for I/O.
- `GlueCEPolicyMaxCPUtime`: The maximum CPU time available to jobs submitted to this queue, in minutes. Jobs will usually be killed after this time. Note that this value should be scaled according to the `SI00` (`SpecInt`) rating, published as a `SubCluster` attribute (see below). All jobs should have a suitable `Requirement` on this value, otherwise they may be killed before they finish.
- `GlueCEPolicyMaxTotalJobs`: The maximum allowed total number of jobs in this queue. Jobs which exceed this limit will be rejected if the WMS attempts to submit them.
- `GlueCEPolicyMaxRunningJobs`: The maximum number of jobs in this queue allowed to execute simultaneously.
- `GlueCEPolicyMaxWaitingJobs`: The maximum allowed number of waiting jobs in this queue. Jobs which exceed this limit will be rejected if the WMS attempts to submit them. This attribute is new in version 1.3 of the schema.
- `GlueCEPolicyAssignedJobSlots`: The number of job execution slots assigned to this queue. This will normally be the same as `MaxRunningJobs`.
- `GlueCEPolicyMaxSlotsPerJob`: The maximum number of job slots which can be occupied by a multi-processor job. A value of 1 means that the CE does not accept multi-processor jobs. This attribute is new in version 1.3 of the schema.



- `GlueCEPolicyPreemption`: This flag is `TRUE` if jobs may be pre-empted, i.e. suspended after they start executing. This attribute is new in version 1.3 of the schema.
- `GlueCEAccessControlBaseRule`: This defines a set of rules which specify who can submit a job to this queue. This is usually of the form `VO:<vo>`, but may also specify VOMS roles or groups. This is taken into account automatically by the WMS.

- **GlueVOView object**

- The `VOView` object overloads a subset of the CE attributes for users defined by the `AccessControlBaseRule`. Some attributes are only defined in the 1.3 schema version.
- `GlueCECapability`: As for CE. New in 1.3.
- `GlueCEInfoTotalCPUs`: As for CE. Deprecated.
- `GlueCEInfoApplicationDir`: As for CE, but points to a VO-specific location.
- `GlueCEInfoDefaultSE`: As for CE.
- `GlueCEStateRunningJobs`: As for CE.
- `GlueCEStateWaitingJobs`: As for CE.
- `GlueCEStateTotalJobs`: As for CE.
- `GlueCEStateEstimatedResponseTime`: As for CE.
- `GlueCEStateWorstResponseTime`: As for CE.
- `GlueCEStateFreeJobSlots`: As for CE.
- `GlueCEStateFreeCPUs`: As for CE. Deprecated.
- `GlueCEPolicyMaxWallClockTime`: As for CE.
- `GlueCEPolicyMaxCPUTime`: As for CE.
- `GlueCEPolicyMaxTotalJobs`: As for CE.
- `GlueCEPolicyMaxRunningJobs`: As for CE.
- `GlueCEPolicyMaxWaitingJobs`: As for CE. New in 1.3.
- `GlueCEPolicyAssignedJobSlots`: As for CE.
- `GlueCEPolicyMaxSlotsPerJobs`: As for CE. New in 1.3.
- `GlueCEPolicyPreemption`: As for CE. New in 1.3.
- `GlueCEAccessControlBaseRule`: As for CE. This defines the set of users for which this `VOView` is valid.

- **GlueSubCluster object**

- `GlueSubClusterTmpDir`: This should be the name of a scratch directory which is shared across all WNs, e.g. via NFS. However, in practice this is not currently reliable at most sites.
- `GlueSubClusterWNTmpDir`: This should similarly be a scratch directory on a disk local to the WN. However, again this is not currently set reliably.
- `GlueSubClusterPhysicalCPUs`: The total number of real CPUs on all nodes in the subcluster. Currently this value is often not set.
- `GlueSubClusterLogicalCPUs`: The total number of logical CPUs on all nodes in the subcluster (e.g. including the effect of hyperthreading). Currently this value is often not set.

- `GlueHostOperatingSystemName`: This is the name of the OS installed on the WNs. The convention for the OS Name, Release and Version in WLCG/EGEE can be found at: [http://goc.grid.sinica.edu.tw/gocwiki/How\\_to\\_publish\\_the\\_OS\\_name](http://goc.grid.sinica.edu.tw/gocwiki/How_to_publish_the_OS_name).
- `GlueHostOperatingSystemRelease`: The name of the OS release installed on the WNs.
- `GlueHostOperatingSystemVersion`: The version of the OS installed on the WNs.
- `GlueHostProcessorModel`: The CPU model name as defined by the vendor.
- `GlueHostProcessorVendor`: The name of the CPU vendor.
- `GlueHostProcessorClockSpeed`: The CPU clock speed in MHz.
- `GlueHostMainMemoryRAMSize`: The amount of physical memory in the WNs, in MB.
- `GlueHostMainMemoryVirtualSize`: The total amount of memory (RAM + swap space) on the WNs, in MB.
- `GlueHostNetworkAdapterOutboundIP`: TRUE if outbound network connections are allowed from a WN. This is normally the case in WLCG/EGEE.
- `GlueHostNetworkAdapterInboundIP`: TRUE if inbound network connections are allowed to a WN. This is not normally the case in WLCG/EGEE.
- `GlueHostArchitectureSMPSize`: The number of CPUs per WN.
- `GlueHostBenchmarkSI00`: The nominal SpecInt2000 speed rating for the CPU on a WN. This should be used to scale any requested time limit.
- `GlueHostApplicationSoftwareRunTimeEnvironment`: This is a multivalued string which allows the presence of specialised installed software to be advertised. VO-specific software uses the format VO-<vo>-<sw\_name\_version>.

- **GlueLocation object**

- The Location object was defined to advertise the location of installed software. However, in version 1.3 of the schema it is replaced by a new Software object.
- `GlueLocationName`: The name of the software.
- `GlueLocationPath`: The name of the directory where the software is installed.
- `GlueLocationVersion`: The software version number.

- **GlueSoftware object**

- This object is new in version 1.3 of the schema.
- `GlueSoftwareName`: The name of the software.
- `GlueSoftwareVersion`: The software version number.
- `GlueSoftwareInstalledRoot`: The name of the directory where the software is installed.
- `GlueSoftwareEnvironmentSetup`: The fully-qualified path name of a script with which to set up the application environment.
- `GlueSoftwareModuleName`: The name of the module with which to set up the application environment using a module management tool.
- `GlueSoftwareDataKey`: The name of any additional information item.
- `GlueSoftwareDataValue`: The value associated with the Key.

#### G.4.4. Attributes for the Storage Element

The part of the schema relating to the Storage Element has been evolving rapidly in line with the development of the SRM protocol, hence many of the attributes are new in the 1.3 schema version. Also, even for the current (1.2) schema the attributes are not always filled correctly by the information providers, or supported correctly by the middleware. This is expected to improve during 2007 as the data management software matures.

In addition to overall SE information, the schema introduces the concept of a *Storage Area* (SA). Originally this referred to a specific area of disk space in which files could be stored, but the SRM has a somewhat more abstract view of an SA as something which collects files which share some attributes. There is also a *Storage Library* (SL) which was introduced to represent the physical storage hardware, but this has been found not to be useful and is now deprecated, and hence not described here.

Auxiliary concepts are the *Control Protocol* and *Access Protocol*. The former relates to the protocol used to manage the SE; in WLCG/EGEE this currently means some version of the SRM protocol. The latter specifies the protocols used for data transfer; a typical SE will support several of these.

Storage systems have many variations, and are evolving rapidly. To allow some flexibility to publish information not otherwise represented in the schema, *Capability* attributes can be used to publish extra information, either as a simple identifier or as key=value pairs.

- **GlueSE object**

- **GlueSEUniqueID**: The unique identifier for the SE. This is usually the SE hostname, but it should be emphasised that this should not be assumed; the SE should be contacted using the endpoint(s) specified in the Protocol objects.
- **GlueSESizeTotal**: The total size of the SE storage in GB. In the 1.3 schema version this is deprecated and split into online and nearline components.
- **GlueSESizeFree**: The total amount of space available to store new files, in GB. In the 1.3 schema version this is deprecated and split into online and nearline components.
- **GlueSETotalOnlineSize**: The total amount of online (disk) storage, in GB. New in the 1.3 schema.
- **GlueSETotalNearlineSize**: The total amount of nearline (tape) storage, in GB. New in the 1.3 schema.
- **GlueSEUsedOnlineSize**: The total amount of online (disk) storage available to store new files, in GB. New in the 1.3 schema.
- **GlueSEUsedNearlineSize**: The total amount of nearline (tape) storage available to store new files, in GB. New in the 1.3 schema.
- **GlueSEArchitecture**: This describes the general hardware architecture of the SE. The value is one of: *tape* (a system including a tape storage robot), *disk* (simple disk storage), *multidisk* (a disk array, e.g. RAID) and *other*.
- **GlueSEImplementationName**: The name of the underlying software implementation, e.g. dCache or DPM. New in the 1.3 schema.
- **GlueSEImplementationVersion**: The version number of the software implementation. New in the 1.3 schema.
- **GlueSEStatus**: The current operational status of the whole SE. Values can be *Queuing* (the SE can accept new requests but they will be kept on hold); *Production* (the SE processes requests normally);

Closed (the SE will not accept new requests and will not process existing ones); and Draining (the SE will not accept new requests, but will still process existing ones). New in the 1.3 schema.

- **GlueSEAccessProtocol object**

- `GlueSEAccessProtocolType`: The protocol type, e.g. `gsiftp` or `rfio`. See the GLUE web site [21] for the full list of types.
- `GlueSEAccessProtocolVersion`: The protocol version.
- `GlueSEAccessProtocolEndpoint`: A URL specifying the endpoint for this protocol. Note that with an SRM the endpoint is normally obtained dynamically.
- `GlueSEAccessProtocolCapability`: A multivalued string allowing arbitrary capabilities to be advertised.
- `GlueSEAccessProtocolMaxStreams`: The maximum number of data streams allowed for a single transfer using this protocol. New in the 1.3 schema.

- **GlueSEControlProtocol object**

- `GlueSEControlProtocolType`: The protocol type (usually SRM in WLCG/EGEE).
- `GlueSEControlProtocolVersion`: The protocol version.
- `GlueSEControlProtocolEndpoint`: A URL specifying the endpoint for this protocol.
- `GlueSEControlProtocolCapability`: A multivalued string allowing arbitrary capabilities to be advertised.

- **GlueSA object**

- `GlueSAPath`: This defines a path name to the root directory for this area. If specified this should be prefixed to the name (SURL) used to store the file.
- `GlueSAType`: This specifies a guarantee on the lifetime of files in the storage area. Values can be `permanent` (files will not be deleted automatically), `durable` (files may be purged after notification of the owner), `volatile` (files may be purged automatically after the expiration of a lifetime), or other. Currently WLCG/EGEE only supports the `permanent` type, but `volatile` (scratch) files may be supported in future.
- `GlueSAStateAvailableSpace`: The total space available for new files in this SA. Note that the units are kB, which with modern storage systems is too small. In the 1.3 schema this attribute is deprecated.
- `GlueSAStateUsedSpace`: The space used by files in this SA. Note that the units are kB, which with modern storage systems is too small. In the 1.3 schema this attribute is deprecated.
- `GlueSAStateTotalOnlineSize`: The total online (disk) space in GB for this SA. New in 1.3.
- `GlueSAStateUsedOnlineSize`: The online (disk) space in GB used by files stored in this SA. New in 1.3.
- `GlueSAStateFreeOnlineSize`: The online (disk) space in GB available for new files in this SA. New in 1.3.
- `GlueSAStateReservedOnlineSize`: The online (disk) space in GB which has been reserved for a specific purpose but not yet used. New in 1.3.
- `GlueSAStateTotalNearlineSize`: The total nearline (tape) space in GB for this SA. New in 1.3.
- `GlueSAStateUsedNearlineSize`: The nearline (tape) space in GB used by files stored in this SA. New in 1.3.

- `GlueSAStateFreeNearlineSize`: The nearline (tape) space in GB available for new files in this SA. New in 1.3.
- `GlueSAStateReservedNearlineSize`: The nearline (tape) space in GB which has been reserved for a specific purpose but not yet used. New in 1.3.
- `GlueSAAccessControlBaseRule`: This defines a set of rules which specify who can store files in this SA. This is usually of the form `VO:<vo>` (or for historical reasons simply the VO name), but may also specify VOMS roles or groups.
- `GlueSARetentionPolicy`: This specifies the quality of storage for files in this SA. Values can be `custodial` (high quality, typically on tape), `output` (medium quality, typically redundant disk storage), or `replica` (low quality, files may be lost if a disk fails). New in 1.3.
- `GlueSAAccessLatency`: This specifies how quickly files stored in this SA are guaranteed to be available for use. Values can be `online` (files are always on disk and can be read immediately), `nearline` (files may not be immediately accessible, e.g. on tape, and may need to be staged in before access), or `offline` (files may need manual intervention to make them accessible). New in 1.3.
- `GlueSAExpirationMode`: The policy for expiration of files in this SA. Values can be `neverExpire`, `warnWhenExpired` (a warning is generated when the lifetime is exceeded), or `releaseWhenExpired` (files will be deleted automatically when the lifetime is exceeded). Note that currently WLCG/EGEE does not expect files to expire. New in 1.3.
- `GlueSACapability`: A multivalued string allowing arbitrary capabilities/properties to be advertised. New in 1.3.

- **GlueSAVOInfo object**

- The `GlueSAVOInfo` object allows the specification of VO-specific information for an SA which supports multiple VOs. This may also be used for subgroups or roles within a VO. New in 1.3.
- `GlueSAVOInfoPath`: A VO-specific path which supercedes the `GlueSAPath` if present.
- `GlueSAVOInfoTag`: A VO-defined string which allows an SA to be selected according to the type of file being stored.
- `GlueSAVOInfoAccessControlBaseRule`: This defines a subset of the users specified by the `GlueSAAccessControlBaseRule` for whom this `VOInfo` object applies.

#### G.4.5. Attributes for the CE-SE Binding

The CE-SE binding schema represents a means for advertising relationships between a CE and one or more SEs. This is typically for CEs and SEs at the same site, but this is not required. In any case the relationship is defined and published by the site hosting the CE. The implication is that files on the SE(s) can be accessed quickly from WNs composing that CE compared with general file access over the WAN. It may also imply access for protocols like `rfile` which restrict access using host-based authorisation. Among other things, the WMS uses the `CESEBind` information to schedule jobs with input files specified in the JDL, to ensure that the jobs go to CEs from which the files are rapidly accessible.

Historically the `CESEBind` was also used to advertise an NFS mount point from which files on an SE were directly visible from WNs. However, this is not currently supported in WLCG/EGEE.

- **GlueCESEBind object**

- GlueCESEBindCEUniqueID: The unique ID for the CE.
- GlueCESEBindSEUniqueID: The unique ID for the SE.
- GlueCESEBindWeight: If multiple SEs are bound to a CE this allows a preference order to be expressed (larger numbers are preferred). This is not generally used in WLCG/EGEE at present.