



# LHC COMPUTING GRID

## LCG-2 MIDDLEWARE OVERVIEW

---

**Document identifier:** CERN-LCG-GDEIS-498079

**EDMS id:** 498079

**Version:** 0.1

**Date:** October 1, 2004

**Section:** LCG Experiment Integration and Support

**Document status:** PUBLIC

**Author(s):** Simone Campana, Maarten Litmaath, Andrea Sciabà

**File:** LCG-mw

---

*Abstract: This document contains an overview of the LCG-2 services and functionalities provided by the middleware to Grid users.*

---



---

## Document Change Record

Issue	Item	Reason for Change
30/09/04	v0.1	First Draft

## Files

Software Products	User files
PDF	<a href="https://edms.cern.ch/file/498079/0.1/LCG-2_Overview.pdf">https://edms.cern.ch/file/498079/0.1/LCG-2_Overview.pdf</a>
HTML	<a href="https://edms.cern.ch/file/498079/0.1/LCG-2_Overview.html">https://edms.cern.ch/file/498079/0.1/LCG-2_Overview.html</a>



---

## CONTENTS

<b>1</b>	<b>INTRODUCTION.....</b>	<b>5</b>
1.1	OBJECTIVES OF THIS DOCUMENT .....	5
1.2	APPLICATION AREA.....	5
1.3	DOCUMENT EVOLUTION PROCEDURE .....	5
1.4	TERMINOLOGY .....	7
1.5	ACKNOWLEDGEMENTS .....	8
<b>2</b>	<b>EXECUTIVE SUMMARY.....</b>	<b>9</b>
<b>3</b>	<b>THE LCG-2 ARCHITECTURE.....</b>	<b>10</b>
3.1	GENERAL VIEW.....	10
3.2	THE WORKLOAD MANAGEMENT SYSTEM .....	11
3.2.1	User Interface . . . . .	11
3.2.2	Resource Broker . . . . .	11
3.2.3	Computing Element . . . . .	13
3.2.4	Worker Node . . . . .	13
3.2.5	Proxy Server . . . . .	13
3.3	DATA MANAGEMENT SYSTEM .....	14
3.3.1	Replica Location Service . . . . .	14
3.3.2	Storage Element . . . . .	14
3.4	INFORMATION SYSTEM.....	15
3.5	MONITORING SYSTEMS.....	16
3.5.1	GridICE . . . . .	16



---

3.5.2	R-GMA . . . . .	17
3.6	INSTALLATION SYSTEM.....	17
3.6.1	Middleware installation . . . . .	17
3.6.2	VO software installation . . . . .	18
3.7	AUTHENTICATION AND AUTHORISATION SYSTEM.....	19



---

# 1. INTRODUCTION

## 1.1. OBJECTIVES OF THIS DOCUMENT

The LHC Computing Grid Project (LCG) has deployed a worldwide computational Grid service to provide the LHC experiments with the computing and data storage resources they need. The LCG employs middleware from various sources, like the Virtual Data Toolkit (VDT), the DataGrid Project and the DataTAG Project. The main tasks of LCG are to integrate all the middleware components to make it work coherently, to develop or to modify the middleware when needed, to prepare a consistent way to distribute the middleware and to provide central guidance to the deployment and the management of Grid resources across the sites involved in the project.

The purpose of this document is to give a brief but complete description of the middleware services used in the version 2 of the LCG (to which we will refer henceforth as LCG-2) and of their functionalities.

## 1.2. APPLICATION AREA

This document is intended as a general overview of the LCG-2 middleware, more detailed than the one presented in the LCG-2 User Guide. It can be of interest for both users and developers which are not familiar with the LCG-2 middleware and want a concise introduction to it.

This document can be considered as a companion to the LCG-2 User Guide, to which the reader is advised to refer for a more basic introduction to Grid concepts.

## 1.3. DOCUMENT EVOLUTION PROCEDURE

This document will change as needed to reflect any significant changes in the LCG-2 middleware.

### APPLICABLE DOCUMENTS

- [A1] LCG-2 User Guide  
<https://edms.cern.ch/document/454439/1/>



---

## REFERENCES

- [R1] The Globus Alliance  
<http://www.globus.org/>
- [R2] The DataGrid Project  
<http://www.edg.org/>
- [R3] The Virtual Data Toolkit  
<http://www.cs.wisc.edu/vdt/>
- [R4] The Condor Project  
<http://www.cs.wisc.edu/condor/>
- [R5] CASTOR  
<http://cern.ch/castor/>
- [R6] Storage Resource Management (SRM) Middleware Project  
<http://sdm.lbl.gov/srm/>
- [R7] DCache  
<http://www.dcache.org/>
- [R8] S. Andreozzi, M. Sgaravatto, C. Vistoli, "Sharing a conceptual model of Grid resources and services"  
Proceedings of the Conference on Computing in High Energy and Nuclear Physics (CHEP 2003),  
March 2003, La Jolla CA, USA  
GLUE schemas activity  
<http://www.cnaf.infn.it/~sergio/datatag/glue/>
- [R9] The DataTAG project  
<http://cern.ch/datatag/>
- [R10] International Virtual Data Grid Laboratory  
<http://www.ivdgl.org/>
- [R11] Lemon, Fabric Monitoring Kit  
<http://cern.ch/lemon/>
- [R12] Nagios  
<http://www.nagios.org/>
- [R13] Local Configuration System  
<http://www.lcfg.org/>
- [R14] GridICE  
<http://infnforge.cnaf.infn.it/gridice/>



[R15] R-GMA: Relational Grid Monitoring Architecture  
<http://www.r-gma.org/>

[R16] Virtual Organization Membership Service  
<http://infnforgc.cnaf.infn.it/projects/voms/>

#### 1.4. TERMINOLOGY

<b>API:</b>	Application Programming Interface
<b>BDII:</b>	Berkeley Database Information Index
<b>CE:</b>	Computing Element
<b>CERN:</b>	European Laboratory for Particle Physics
<b>DMS:</b>	Data Management System
<b>EDG:</b>	DataGrid Project
<b>EDT:</b>	DataTAG Project
<b>ESM:</b>	Experiment Software Manager
<b>GASS:</b>	Global Access to Secondary Storage
<b>GIIS:</b>	Grid Information Index Service
<b>GOC:</b>	Grid Operations Centre
<b>GRAM:</b>	Globus Resource Allocation Manager
<b>GRIS:</b>	Grid Resource Information Service
<b>GSI:</b>	Grid Security Infrastructure
<b>GUID:</b>	Globally Unique Identifier
<b>IC:</b>	Information Catalogue
<b>IS:</b>	Information System
<b>JC:</b>	Job Controller
<b>LB:</b>	Logging and Bookkeeping
<b>LCAS:</b>	Local Centre Autorisation System
<b>LCG:</b>	LHC Computing Grid
<b>LCMAPS:</b>	Local Credential MAPPING Service
<b>LFN:</b>	Logical File Name
<b>LHC:</b>	Large Hadron Collider
<b>LM:</b>	Log Monitor
<b>LRC:</b>	Local Replica Catalog
<b>MDS:</b>	Metadata Directory Service
<b>MSS:</b>	Mass Storage System
<b>NS:</b>	Network Server
<b>PFN:</b>	Physical File Name
<b>PRS:</b>	Proxy Renewal Service
<b>PS:</b>	Proxy Server
<b>RB:</b>	Resource Broker
<b>RFIO:</b>	Remote File Input/Output Protocol



---

<b><i>RLS:</i></b>	Replica Location Service
<b><i>RMC:</i></b>	Replica Metadata Catalog
<b><i>SE:</i></b>	Storage Element
<b><i>UI:</i></b>	User Interface
<b><i>VDT:</i></b>	Virtual Data Toolkit
<b><i>VO:</i></b>	Virtual Organisation
<b><i>WM:</i></b>	Workload Manager
<b><i>WMS:</i></b>	Workload Management System
<b><i>WN:</i></b>	Worker Node

## **1.5. ACKNOWLEDGEMENTS**

This work received support from the following institutions:

- Istituto Nazionale di Fisica Nucleare, Roma, Italy.
- Ministerio de Educación y Ciencia, Madrid, Spain.



---

## 2. EXECUTIVE SUMMARY

This document is a pragmatic approach to a description of the middleware services in use in LCG-2 and their functionalities.

After the introduction and this summary, the architecture of the LCG-2 is presented. Every subsystem (workload management, data management, etc.) is analysed from the point of view of the services that compose it, their functions and their interactions. This analysis is intentionally concise and many technical details are not covered. The focus is more on the reality of the service deployment in LCG-2 than on the pure architectural side.



---

## 3. THE LCG-2 ARCHITECTURE

### 3.1. GENERAL VIEW

The LCG is a collection of geographically distributed resources and services with the purpose of allowing people working in LHC experiments, organised in Virtual Organisations (VO), to efficiently run their programs for event generation, reconstruction and analysis, without the need to know where to run a particular job, where to get its input data and where to store its output data. The LCG is composed of a *Workload Management System (WMS)*, a *Data Management System (DMS)*, an *Information System (IS)*, an *Authorisation and Authentication System*, an *Accounting System* (not yet operational), various *monitoring services* and *installation services*.

The WMS is responsible for the management of the jobs submitted by users; it matches the job requirements to the available resources and schedules the job for execution on an appropriate computing cluster; then, it tracks the job status and allows the user to retrieve the job output when ready.

The DMS allows users to move files in and out of the Grid, to replicate them among different locations, and to locate them. This is achieved transferring data via a number of protocols (GridFTP is the most commonly used) and by interacting with a central file catalog, the Replica Location Service (RLS).

The IS provides information about the LCG-2 resources and their statuses. The information is published by the individual resources and copied into central databases; it is used by the WMS to match the resources against the job requirements and to rank them, by the DMS to choose storage resources, and by the monitoring systems.

The Authentication and Authorisation system contains the list of all the people authorised to use LCG-2, divided by virtual organisation. This list is normally downloaded to all machines running Grid services in order to map the LCG-2 users to the local users of the machine.

In addition, in LCG-2 there is a number of monitoring services: GridICE monitors the usage of Grid resources (e.g. the number of jobs running, the storage space available, etc.); R-GMA allows users to retrieve custom information from running jobs and store it in a relational database; finally, there are monitoring systems to check the status of Grid services (e.g. whether they are running) and their functionality: they are more intended for the LCG-2 operations staff than for users.

At some sites, dedicated fabric management services, like LCFGng, are in use to manage the installation, the upgrade and the maintenance of the local Grid services.

In the next sections, the LCG-2 services will be described in greater detail.



## 3.2. THE WORKLOAD MANAGEMENT SYSTEM

In this section the LCG-2 workload management system is described (Figure 1). We assume the reader is familiar with the key concepts of the Globus Toolkit [R1], particularly the Grid Security Infrastructure (GSI), the Globus Resource Allocation Manager (GRAM) and the Global Access to Secondary Storage (GASS).

The WMS middleware was developed by EDG [R2], VDT [R3] (which includes Condor-G and Globus), and has several custom modifications by LCG. The WMS also relies on a regular batch system, like OpenPBS or LSF, to manage the worker nodes.

The LCG-2 WMS is deployed on five kinds of machines: the User Interface (UI), the Resource Broker (RB), the Computing Element (CE), the Worker Node (WN) and the Proxy Server (PS).

### 3.2.1. User Interface

The *User Interface (UI)* is the component that allows users to access the functionalities of the WMS, and it is usually identified with the host on which it is installed: in this meaning, it is the host on which a user has to log in to access and use the Grid. It provides a command line interface, a graphical interface and a C++ programming interface to the WMS. The basic functionalities are: **a)** list the computing resources compatible with a given set of job requirements, **b)** submit a job, **c)** get the job status, **d)** cancel a job, **e)** retrieve the logging information of a job and **f)** retrieve the output of a job.

### 3.2.2. Resource Broker

The following services, unless noted otherwise, usually run on the same machine, designated as *Resource Broker*.

The *Network Server (NS)* accepts incoming requests from a UI, authenticates the user, copies the input and output sandbox between the UI and the RB, optionally registers the user proxy for periodic renewal by the Proxy Renewal Service, and forwards the requests to the Workload Manager.

When a job is submitted, the *Workload Manager (WM)* calls the *Matchmaker* to find the resource which best matches the job requirements, interacting with the IS and the RLS. The *Job Controller (JC)* is then called to submit the job to Condor-G.

The *Condor-G* component [R4] submits the job to the CE; in addition, it submits an extra job (the *grid monitor*) per CE and per user to monitor the user jobs. A Condor service, called *GAHP server*, acts as a GRAM client for all jobs, and as GASS server for the results from the grid monitor jobs. The *Log Monitor (LM)* continuously parses the Condor-G log files looking for events concerning active jobs. If a job fails to be fully executed by the batch system, the LM will inform the WM for optional resubmission of the job to another CE.

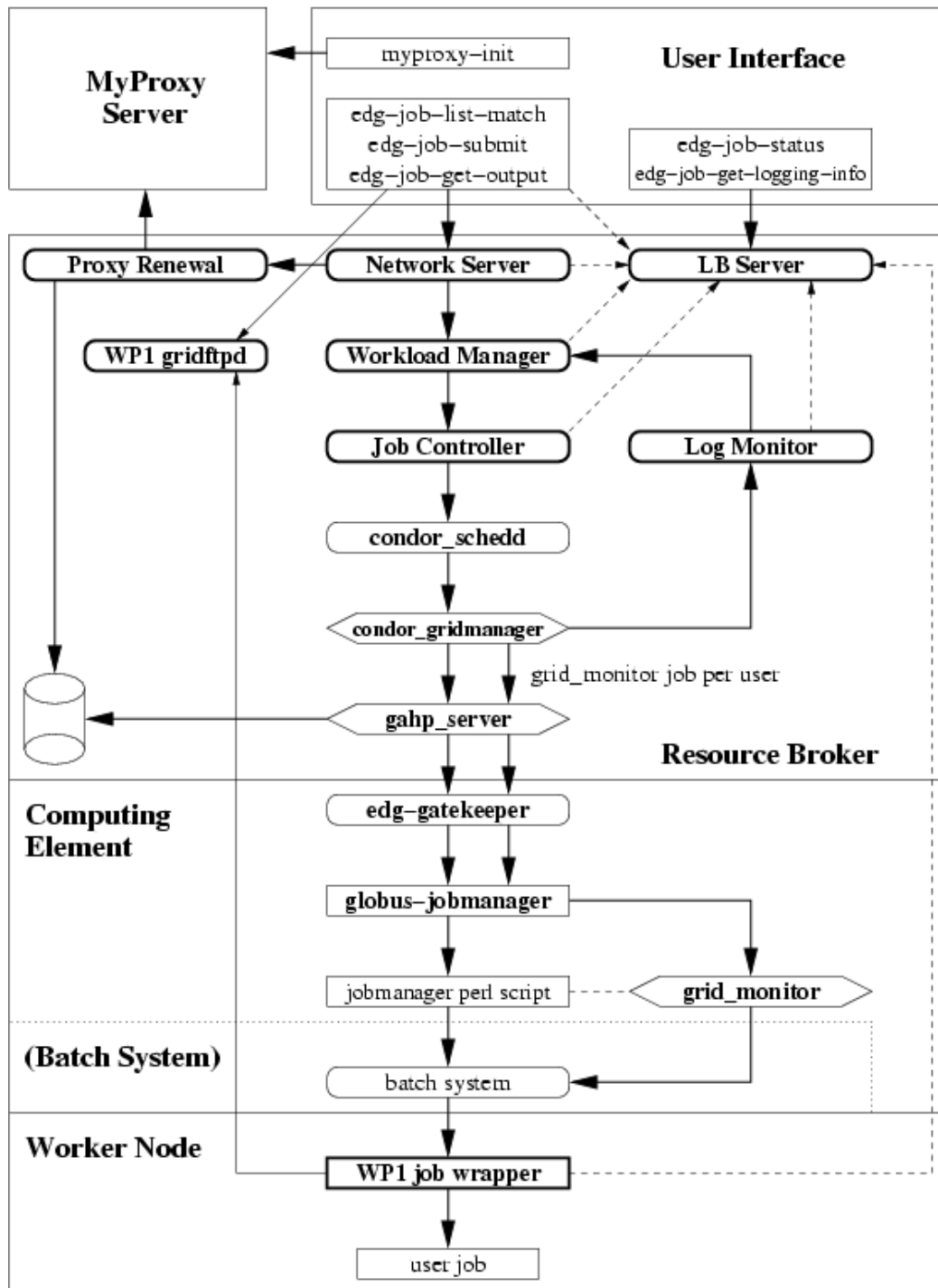


Figure 1: LCG-2 job submission chain.



---

The *Logging and Bookkeeping (LB)* service collects and stores in a database the job status information supplied by the different components of the WMS. The collection is done by **LB local-loggers**, which run on the RB and on the CE, while the **LB server**, which normally runs on the RB, saves the collected information in the database. The database can be queried by the user from the UI, and by RB services themselves.

The RB machine also runs accessory services, like a MySQL server for the LB, and a GridFTP server.

### 3.2.3. Computing Element

The *Computing Element (CE)* is the Grid interface to a computing cluster. The same word is used to refer to the machine where the Grid services run, and to the Grid identifier of a queue of the local batch system, defined as `<hostname> : <port> / <batchsystem> - <queue name>`. We will refer to the former meaning.

The CE runs a *gatekeeper*, which accepts requests from Condor-G and creates a *job manager* (a generic interface to the basic batch system functions) per job. Job managers are used only to submit or to cancel jobs, while the job status is queried by a *grid monitor* through a query process, a single instance of which runs on the CE per user. The batch system is the last element of the chain, and often has a server running on the CE.

### 3.2.4. Worker Node

The *Worker Node (WN)* is a host to execute jobs; a set of WNs managed by a single CE constitute a computing cluster. There are no LCG-2 services running on a WN and it requires only a minimal amount of middleware to be Grid-aware, i.e. the LCG-2 commands and libraries that a job may need to invoke. What is run on the WN is not exactly the user's executable or script, but a script wrapped around it by the WMS, which, among other things, copies the input/output sandbox from/to the RB.

### 3.2.5. Proxy Server

In LCG-2, users authenticate themselves using temporary credentials called *proxy certificates*, which contain also the corresponding private key. Proxy certificates do not represent a significant security risk only if they are reasonably short-lived (by default, a dozen hours). For longer jobs, a proxy renewal system is used, consisting of a *Proxy Renewal Service (PRS)* on the RB and a *Proxy Server (PS)* on a dedicated host. A PS stores long-lived user proxies (with a lifetime of several days, usually) which it uses to generate, on request of the PRS, short-lived proxies for jobs whose proxies are about to expire.



### 3.3. DATA MANAGEMENT SYSTEM

The DMS relies on two kinds of services: the *Replica Location Service (RLS)* and the *Storage Element (SE)*.

#### 3.3.1. Replica Location Service

This service is the official file and metadata catalog of LCG-2, and is composed of two parts: the *Local Replica Catalog (LRC)* contains the mappings between globally unique file identifiers (GUID) and physical file names (PFN), where both GUIDs and PFNs can have attributes, while the *Replica Metadata Catalog (RMC)* contains the mappings between logical file names (LFN) and GUID, where LFNs can have attributes (also referred to as *metadata*).

This architecture is somewhat different from the original RLS concept as developed by Globus and EDG, where the RMC was a separate component and the RLS foresaw a distributed rather than centralized structure, with an LRC at each site and an overall Replica Location Index (RLI) system, a service showing which LRCs hold information on a given GUID.

The RLS is deployed as a set of central servers, one for each VO, consisting of a database backend (Oracle or MySQL) and an application server, which accepts requests from clients through the RLS API. Among the RLS clients we count the data management command line tools to copy, register, replicate or delete files, and the RB Matchmaker when the job description specifies input files.

#### 3.3.2. Storage Element

A *Storage Element (SE)* provides uniform access to large storage spaces. The Storage Element may control large disk arrays, or a hierarchical mass storage system (MSS) with a tape robot back-end. Each LCG-2 site provides one or more SEs.

In the current LCG-2 release, the SE consists of a disk server or a front-end to an MSS, with a GridFTP server and an RFIO server. The GSIFTP protocol offers the functionality of FTP, but enhanced to use GSI security. It is responsible for secure, fast and efficient file transfer to/from the Storage Element and it is used to access data at remote sites. The Remote File Input/Output protocol (RFIO) [R5] is a POSIX-compliant file access protocol which can be used to access byte ranges in files stored on a local SE. A GSI-enabled version of RFIO is foreseen to provide WAN access.

The simple GridFTP SE is known as a *Classic SE*. A second type of SE, not yet widely deployed, uses a *Storage Resource Manager (SRM)* [R6] as interface; SRM is currently supported by dCache [R7] and CASTOR [R5]. An SRM-based SE typically uses a disk pool manager, allowing files to be transparently relocated to different disks and allowing the pool to grow by just adding another disk. Classic SEs will be phased out in favour of SRM SEs.

### 3.4. INFORMATION SYSTEM

The LCG-2 information system is based on the Globus *Metadata Directory Service (MDS)* with modifications from LCG which considerably improve the scalability and the robustness of the system. The information itself is organised following the *GLUE schema* [R8], developed jointly by the DataTAG [R9] and iVDGL [R10] projects.

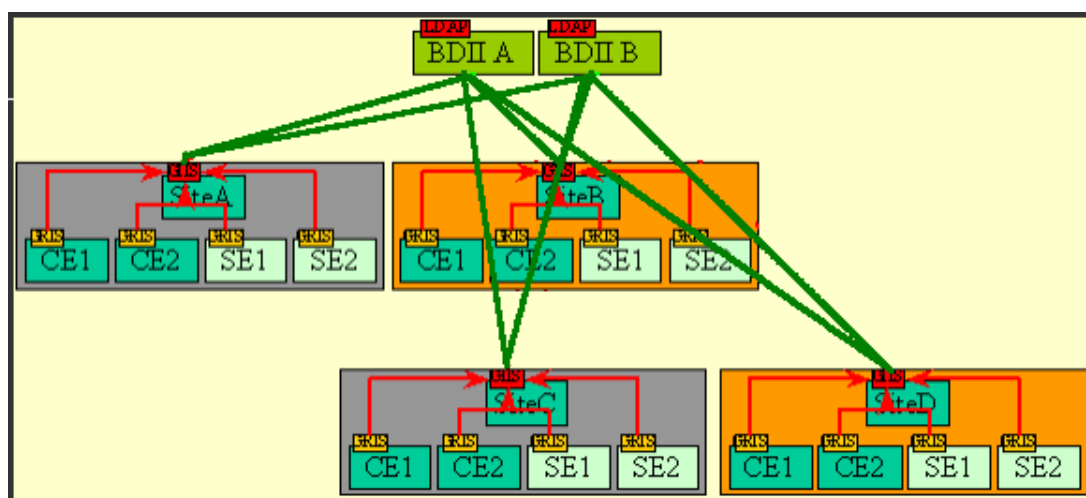


Figure 2: The Information System in LCG-2.

The system uses OpenLDAP databases to cache and publish the information, and has a hierarchical structure, depicted in Figure 2. Each Grid resource is served by a *Grid Resource Information Service (GRIS)*, which usually runs on the same machine. The GRIS calls, when queried, an *information provider*, a script running on the Grid resource which obtains both static and dynamic information on the resource; the information is then read into an OpenLDAP database and an OpenLDAP server returns the result of the query.

Higher in the hierarchy lies the Grid Information Index Service (GIIS), one being deployed at each site, to which the local GRISes are registered. When a GIIS is queried, it queries in turn each registered GRIS and returns the aggregated result. A cache mechanism prevents the information provider from running at each query. A GIIS can register to another GIIS, although this does not happen in LCG-2<sup>1</sup>.

The top level of the IS is the *Berkeley Database Information Index (BDII)*. The BDII periodically queries a list of GIISes and/or GRISes and executes the information providers listed in its configuration file, if any. Each VO can configure its BDII to query only those sites that are relevant to the VO.

In order to improve the scalability, the BDII uses two databases, one read-only and one write-only, which are switched when an update is completed; as a consequence, an increase in the number of sites

<sup>1</sup>Regional GIISes used to be deployed in LCG-1, but they were found to worsen the performances of the IS.

leads to a proportional increase in the time needed to update the database, and to less up-to-date information, but not to a degradation in the response time of the IS. All sites are queried in parallel, but the database has to be updated sequentially.

### 3.5. MONITORING SYSTEMS

#### 3.5.1. GridICE

GridICE [R14] (Figure 3) is a fairly complete Grid monitoring service to collect, store and visualise the status of the LCG-2 resources: in particular low level information (CPU load, memory and disk usage, etc.), status of services (gatekeepers, GridFTP servers, etc.) and Grid information (number of running and queued jobs, free CPUs, free space on SEs, etc.).

The monitored information is described by an extended version of the GLUE schema, and is collected by the *Measurement Service*, partly from the IS and partly using *sensors* running on the monitored hosts; the sensors are based on Lemon [R11]. The information from the sensors is published by a special GRIS that does not register to any GIIS (but whose existence is published by the local GIIS). The *Data*

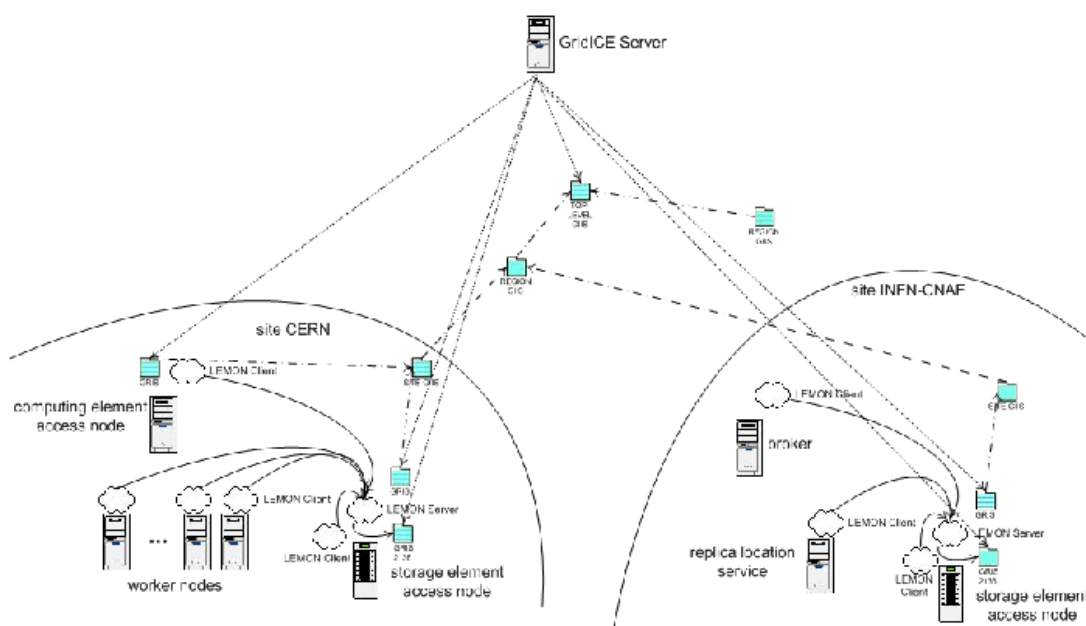


Figure 3: Typical GridICE deployment.

*Collector Service*, based on Nagios [R12], is used to discover new resources from the IS and store all the information to be monitored on persistent storage (a PostgreSQL database). Finally, GridICE provides a web interface.



In LCG-2, GridICE is deployed as a central server at the Grid Operations Centre (GOC) and Lemon servers and special GRISes at each site (normally installed on the local SE). Worker nodes are monitored only in a small number of sites.

### 3.5.2. R-GMA

**R-GMA** [R15] is a monitoring and information management service for distributed resources; it uses SQL as a query language and gives an RDBMS-like view of VO-specific information.

In LCG-2, it consists of a central *Information Catalogue (IC)*, one *MON box* per site, and clients installed on the WNs, the CE, the SE, the RB and the UI.

The MON box is the R-GMA server and is installed on a dedicated machine only at large sites; otherwise it is usually installed on the SE. The IC is installed at the GOC and contains a list, called *Registry*, of registered R-GMA servers and tables.

The model is not dissimilar to the one used for the IS: R-GMA clients at a site can push the desired information to the local MON box, which has full WAN connectivity and pushes in turn the information to the IC, where it is stored in user-defined tables of a central database. One of the purposes of R-GMA is in fact to allow users to monitor their jobs at run time.

A second application of R-GMA in LCG-2 is *accounting*, that is the ability to keep track of the usage of the LCG-2 resources per VO in a given lapse of time, down that the job level. At the moment accounting is not done in LCG-2 using R-GMA, though, but by manually sending the CE log files to the GOC, where they are further processed.

## 3.6. INSTALLATION SYSTEM

### 3.6.1. Middleware installation

The LCG-2 middleware is distributed with support for *LCFGng* [R13], which can be employed to install and manage the configuration of Grid farms. Figure 4 gives a schematic view of the main constituents and their interactions.

The LCFGng server at a site contains all the information about the configuration of the site. This configuration is described in text files (source files) which are compiled (*mkxprof*) into XML fragments (called *profiles*), and published on a web server.

Each profile describes the configuration parameters relative to a host in the farm. When a profile changes, the LCFGng client on the corresponding host (*rdxprof*) is notified. The client acknowledges the server, retrieves the profile via HTTP and caches the parameters in a DBM file. Informations about the statuses of all hosts are collected by the server and published on a web page.

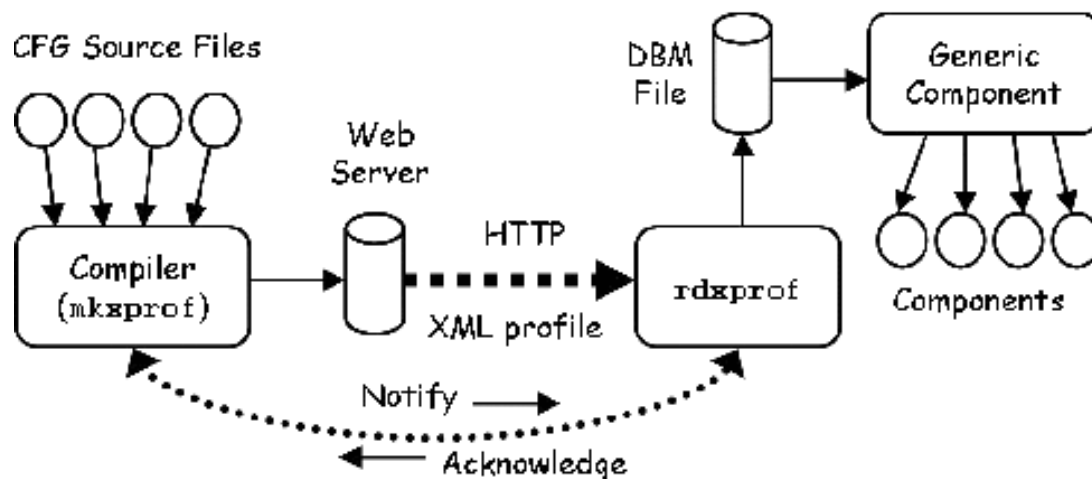


Figure 4: Schema of the LCFGng installation system.

On each host, *component* scripts read the configuration parameters, modify the host configuration when needed and notify any associated daemons. Components are derived from a “generic component” (which implements default functions and performs standard operations) and redefine the necessary methods. One component in particular (updaterpms) is responsible to maintain the software packages according to the central specification. For this purpose, each client host needs to access a filesystem (usually remote) containing both the RPM files and the list of those which should be installed.

LCFGng also provides the capability to install new machines with a limited manual intervention: a minimal operating system (PXE) is loaded from the network (usually via a dhcp daemon on the LCFGng server); this program partitions the system and installs a copy of the real operating system.

It is not compulsory to use LCFGng to install Grid Resources: also a completely manual installation procedure is supported. This allows integrating the LCG-2 middleware installation with other cluster management systems.

### 3.6.2. VO software installation

The installation of software specific to VOs is managed with a mechanism that can be used by VO users with special privileges, called *Experiment Software Managers (ESM)*, and that does not require superuser privileges. Presently, this mechanism requires a file system to be shared among all the WNs of a CE: only jobs of ESMs of a given VO have write permission to the VO’s shared area and can use it to install the VO software; read permissions are granted for all jobs from the members of the VO. In addition, the ESMs can modify the value of an attribute published by the GRIS of a CE to advertise, for example, the availability of a given software version on that CE. This can be exploited at job submission to select only CEs that publish a certain attribute value.



---

It is worth pointing out that LCG sites are not required to provide such file system: in that case, users will have to install the software they need “on the fly” for each job, and it will be removed automatically at the end of the job. To overcome this limitation, a service is foreseen to manage the local installation of VO software on WNs. This service will consist of a server (*Tank*), which acts as software repository and manages the installation/removal of software on/from individual WNs, and a client (*Spark*) running on each WN.

### 3.7. AUTHENTICATION AND AUTHORISATION SYSTEM

In LCG-2, user authentication is based on central databases, one per VO, containing the certificate subjects of all LGC-2 users. These databases are accessed by the RBs, the CEs and the SEs to locally build a list of authorised users. When a request to run a job arrives at a CE, the *Local Centre Authorization System (LCAS)* verifies that the request can be accepted based on the local policy. This functionality is not yet exploited in LCG-2, though. Finally, the *Local Credential MAPPING Service (LCMAPS)* maps the user’s credentials to local credentials (namely, a local UNIX user).

It is foreseen to replace the current VO management system, based on LDAP databases, with a *Virtual Organization Membership Service (VOMS)* [R16]. The VOMS allows extending proxy certificates with information about the user’s role and capabilities, and the groups he belongs to, which may correspond to different privileges for jobs. This contrasts with the current system, where essentially all the users of a VO share the same privileges<sup>2</sup>.

---

<sup>2</sup>ESMs have special privileges, but these are fixed in the site configuration: it is impossible to dynamically create new groups or roles.