Frédéric Hemmer
January 31, 2005

Planning for the first major release of the gLite middleware

Version 1.12

1. Introduction

This document presents the plan for the first major gLite software release to be produced by the EGEE project by the end of March 2005.

The goal of this first major gLite release is to provide the minimum functionality necessary to satisfy the immediate needs of the pilot user groups (LHC and Bio-medical) and address the principal short-comings identified with the existing LCG-2 production middleware. The contents are heavily influenced by the prototypes assembled and deployed on the prototype testbed since May 2004 and offers equivalent functionality.

Below is a plan from now until the end of March to achieve the above goal. In brief, the plan is to provide a set of high level modules described below by the end of January to grid operations groups (SA1), while focusing on testing, bug fixing, common logging and configuration and simple user interfaces during February and March.

The rest of the document elaborates on the functionality, responsibilities, status and plans of the modules of the proposed gLite software stack for this first release (release 1)

2. Workload Management System (WMS)

2.1. Functionality

The Workload Management System (WMS) operates via the following components and functional blocks:

- The Workload Manager (WM) itself, responsible of accepting and satisfying job management requests coming from its clients. The WM will pass job submission requests to appropriate Computing Elements for execution, taking into account requirements and preferences expressed in the Job Description. The decision of which resource should be used is the outcome of a matchmaking process between submission requests and available resources. This not only depends on the state of resources, but also on policies that sites or VO administrators have put in place (on the Computing Elements).
- The Computing Element (CE), which handles Job submission (including staging of required files), cancellation, suspension and resume (subject to support by the Local Resource Management System – LRMS), job status inquiry and notification. The CE is able to work in a push model (where a job is pushed to a Computing Element CE for its execution) or in a pull model (where a CE asks a known WM – or a set of WM's – for jobs).

- The Logging & Bookkeeping services (LB), which tracks jobs during their lifetime in term of events (important points of job life, such as submission, starting execution, etc.) gathered from the WM's and the CE's (they are instrumented with LB calls). The events are first passed to a local logger then to bookkeeping servers.
- The Accounting Services collect information about usage of Grid resources by users, group of users (including VO). This information can be used to generate reports/billing but also to implement resources quotas. Access to the accounting information is protected by ACL's.
- Job Provenance Services, whose role is to keep track of submitted jobs (completed or failed), including execution conditions and environment, and important points of the job life cycle for longs periods (months to years). This information can then be reprocessed for debugging, post-mortem analysis, and comparison of job execution and re-execution of jobs.
- Interfaces to Data Management allowing the WMS to locate sites suitable for job submission (or to schedule data transfer to an appropriate site) are available for LCG RLS, the Data Location Interface (DLI) and the StorageIndex interface (allowing for querying catalogs exposing this interface - a set of two methods listing SEs for a given LFN or GUID).

Compared to the AliEn workload management (TaskQueue) the WMS is missing the job optimizer that can trigger file replication. With the introduction of the File Placement Service, the WMS could be modified to provide this functionality.

2.2. Improvements to EDG/LCG

The WM components above have been defined to address the short-comings of the existing WM in LCG-2, notably in the areas of robustness and overall efficiency.

The first improvement to the WM is that it now has a repository of resource information made available in read-only mode. In effect it provides a cache of the Information System. This repository, dubbed Information Super Market (ISM), is updated either by active polling of resources (CE in push mode) or by arrival of notifications (CE in pull mode) or by a combination of both. The ISM can be configured so that certain notifications can trigger the matchmaking process. Also, the existence of the ISM reduces the need to be in permanent contact with an information system and hence increases overall robustness.

Another improvement is the possibility for the WM to keep submission requests in a Task Queue (keeping a list of jobs to be submitted together with their requirements) if no resources are available that match job requirements. Non-matching requests will be retried periodically or as soon as notifications of available resources appear in the ISM. In particular, this allows for a CE to signal (using the CE Monitor component) that it is ready for accepting jobs for a particular VO.

A third improvement is the use of a new Condor development (Condor-C), allowing for reliable job submission between the WM and the CE (even in case of network failures), reusing a large part of the Condor code. At the CE level, once instantiated through the Globus gatekeeper, Condor is interfacing to the LRMS through a new

layer called Batch Local Ascii Helper protocol (BLAHP), somewhat similar to the Globus Ascii Helper Protocol (GAHP) of Condor-G.

The Job Provenance Service is a new development that did not exist in EDG/LCG but the ability to track large sets of jobs has been requested by the LHC experiments.

Accounting services as such have not been used in LCG, rather ad-hoc collection of accounting data have been merged and processed to provide consistent reports.

Finally, improvements on local account management (Dynamic Account System) and CE head node monitoring (not to be confused with CE monitor component, CE head node monitoring will monitor CE resources such as number of open files/sockets, CPU, processes, etc.) are being developed by the Globus team together with JRA3 (Security).

DAG based jobs, i.e. multiple jobs grouped in an acyclic graph, are supported as well as bulk job submissions (as requested by the LHC experiments).

2.3. Responsibility

The Workload Management system, the matchmaking process, the Logging and Bookkeeping Services, the Accounting Services and the Job Provenance Services are the responsibility of the IT/CZ cluster.

The CE is the responsibility of ANL (Gatekeeper, setuid services, DAS and monitoring of resources consumption on the CE head node), Univ. of Madison (Condor and improvements), and IT/CZ (BLAHP to LSF/PBS and CE Monitor). A new Web Services based CE is also being implemented by IT/CZ but will not be part of the first release.

JRA3 retains an overall responsibility for all security architecture and implementation of related libraries, in particular for VOMS, LCAS and LCMAPS. The VOMS server is however evolved in IT/CZ with JRA3 coordinating the work.

2.4. Current Status and Plans

The WMS comprising the Workload Manager, CE, CE Monitor, Logging and Bookkeeping Services, Accounting Services and interfaces to Data Management for RLS and StorageIndex have been released to SA1 as part of Integration Build I20041217 on December 17, 2004. The new CE Monitor part (pull model) and interface to StorageIndex have not yet been exercised by SA1 however.
Web Services components to WMS will be available after release 1. The WMS supports the following job types: batch, mpi, DAG, interactive, checkpointable and partitionable. Job splitting (as recommended by LCG) and bulk job submission can be achieved through the submission of DAGs without dependencies. Support for bulk job submission without DAGs specification will become available with the Web Services based component (coming after release 1, although CE Monitor and LB query are already providing Web Services).

CE Monitor should be tested in January 2005.

Interfaces to Data Management using the StorageIndex and DLI interfaces are being tested by the developers and should be available in January.

The DAS Services (renamed in the meanwhile Workspace Management Service) is available as a tech preview. Setuid services are being implemented as a workaround by Madison waiting for final delivery by ANL. CE head node monitoring will not be available for release 1.

A prototype of the Job Provenance exists, but there is still a long way to achieve a 'one-button' resubmission of jobs with access to data.

Accounting Services are available, accounting sensors for the CE are being tested.

3. Data Management

3.1. Functionality

It is understood that LCG and Biomedical applications does not have an *immediate* need for enhanced functionalities such as distributed catalogs with lazy updates using a messaging system. LCG specifically identified a non-distributed catalog, gLiteI/O for data transfer and an SRM service with Castor and dCache backends as key data management components. It is therefore proposed to refocus on the simple Data Management functions as described below.

- The File and Replica Catalog (FireMan catalog) presents a hierarchical view of a logical file name space. The two functions of the catalog are to resolve Logical File Name to Storage URL translation (via a GUID) and to locate the site at which a given file resides. The catalog provides Access Control List support (ACL) that will be able to be mapped onto VOMS roles(For release 1 ACLs will support lists of individual Distinguished Names – DNs); file access is secured through these ACLs. It provides Web Services interfaces with full WSDL available. Bulk operations are supported. Storage Index interface for use by Workload Management is available. The client interface is currently limited in the sense that it has to be used through WSDL (therefore requiring programming and the use of tools such as gSOAP). Catalog partitioning is not supported, in other words it is not possible to store part of a hierarchy in one site and another part in another site. Some metadata capabilities are supported if restricted to key/value pair associated to directories. A more general Metadata Interface for application specific Metadata has been proposed (allowing application specific Metadata catalog implementations to be performed, while retaining a common user interface).
- The gLite I/O which is a POSIX-like I/O for access to grid files via their Logical Name. This provides open/read/write/close style of calls to access files while interfacing to the above mentioned catalog. It enforces the file ACLs specified in the catalog.
- The File Placement Service (FPS) takes data movement requests and executes them based on policy. It maintains a persistent transfer queue thus providing

reliable data transfer even in the case of network outage and interacts fully with the FireMan catalog. Proper authorization is ensured by checking the file(s) ACLs in the catalog. The File Placement Service can be used without the interaction with the FireMan catalog (File Transfer Service).

3.2. Improvements to EDG/LCG

The major concerns with the existing data management tools of LCG-2 that gLite tries to address are the poor performance of RLS, a non-distributed catalog, lack of consistent grid-storage interfaces and an unreliable data transfer layer.

The FireMan catalog exposes a hierarchical name space, bulk operations, ACLs and offers Web Services Interfaces. It should scale better than the EDG RLS (but remains to be proven).

The equivalent of gLite I/O in LCG is GFAL. gLite I/O provides ACL support which was not available from GFAL, as well as support for the FireMan catalog. In addition, the catalog interaction was moved from the client (GFAL) to the server (gLite I/O).

There was no equivalent to the File Transfer Service in LCG, until recently when LCG Service Challenges started. The LCG Service Challenge code has been developed in collaboration with JRA1 (with a different focus). The database schema and service logic are common to the LCG code.

3.3. Responsibilities

The Data Management Services is the responsibility of CERN.

3.4. Status and Plans

The File and Replica Catalog is currently in the build system, but is not fully integrated yet (meaning deployment modules and/or documentation are not ready). This is scheduled during the course of January 2005 at which point it will be available for testing.

The gLite I/O component (replacing GFAL) has been available since October in the first Integration Build (I20041020). Usage by ARDA has however discovered some serious deficiencies (reliability, corruption, performance) which have been or are in the process of being fixed.

The File Placement Service are currently in the build system, but is not fully integrated yet (meaning deployment modules and/or documentation are not ready). This is scheduled during the course of January 2005 at which point it will be available for testing.

In addition, it is proposed to converge the File Placement Service with the current code used for the data movement parts of the LCG Service Challenges.

It is also proposed to postpone all other efforts in Data Management (e.g. catalog distributed updates through messaging systems) until a simple File Catalog has demonstrated enough functionality, usability and performance and gLite I/O has been fixed. In particular the lcg-utils command line tools and libraries should now be

adapted to be able to use the FireMan catalog (and therefore providing backward compatibility with LCG-2 at user level).

Testing of all Data Management components is scheduled to start at the beginning of February, when Data Management components are available from the build system and fully integrated.

4. Information System

4.1. Functionality

The Information & Monitoring system is R-GMA. As stated by LCG, an information system is required for the WMS to work in push-mode. It offers refactored consumer/producers services with Java, C, C++ and Python API's allowing users to manipulate data using a SQL like statements on (virtual) tables. It also proposes a prototype Service Discovery API for use by other middleware providers such as the WMS so as to simplify large-scale deployment. R-GMA currently is VO agnostic.

4.2. Improvements to EDG/LCG

R-GMA has been little used in LCG so far. The gLite version of R-GMA is different than the EDG/LCG one, as R-GMA has been significantly reworked to cope with Web Services. It also now offers registry replication, eliminating single point of failure.

4.3. Responsibilities

The responsibility for R-GMA entirely sits in the UK cluster. There are however discussions going on in using R-GMA interfaces for accounting data (IT/CZ cluster).

4.4. Status and Plans

R-GMA has been released as part of Integration Build I20041217 on December 17, 2004 and is being tested. Multi-VO support and Web Services version have been delayed after release 1.

5. Other Services

The following section describes additional services of which only VOMS will be included in release 1 while the others will be addressed later.

5.1. VOMS

VOMS provides support for group membership, roles and capabilities. It is in particular used by the Workload management System and the FireMan catalog for ACL support to provide the functionality identified by LCG. The main evolution from EDG/LCG is support for SLC3, bug fixes and better conformance to IETF RFCs. VOMS is currently being integrated and will be available for testing in January. The overall responsibility for all security packages is JRA3, however the VOMS server is the responsibility of IT/CZ, JRA3 coordinating the work.

5.2. Grid Access Service

A prototype of a Grid Access Service (GAS) has been demonstrated for AliEn Data Management functions and the gLite metadata catalog interface. It provides an entry point to authentication and an indirection layer to Grid Services API's allowing applications to run unmodified even if Services API's do change as well as for VO to express their preference for the Grid Services they want to use (e.g. to specify they want to use their own metadata catalog implementation). This is an optional component. The work was so far performed by CERN but responsibilities are unclear here as the GAS was not originally foreseen.. GAS is not considered to be a high-priority component for release 1 since access to various services is independent of the GAS and are still possible without it.

5.3. Package Manager

A prototype of a Package Manager, a component that allows for on demand installation, configuration, upgrade and removal of application software suites has been demonstrated for AliEn. The work was so far performed by CERN but responsibilities are unclear here as the Package Manager was not originally foreseen. The Package Manager is not considered a high-priority component for release 1 and hence work on this component will be resumed later.

6. Monitoring the progress

A weekly meeting which involves the JRA1 management, cluster heads (IT/CZ, UK, CERN, integration & testing), SA1 and NA4 (ARDA) representatives and dedicated to ensuring the delivery of the gLite software according to this plan will be organised. An updated version of the summary table given below will be produced at each of these meetings.
An important milestone is the end of January when all of the software components for release 1 should have completed the integration step (see table below). The majority of the testing by the different groups will be performed during the months of February and March.

7. Summary of components for release 1

The following table summarises the high-priority components of gLite that will be included in the first major release (release 1). It will be kept up to date as a means of monitoring the progress of gLite development. Future dates indicate the week in which the milestone is planned to be completed.

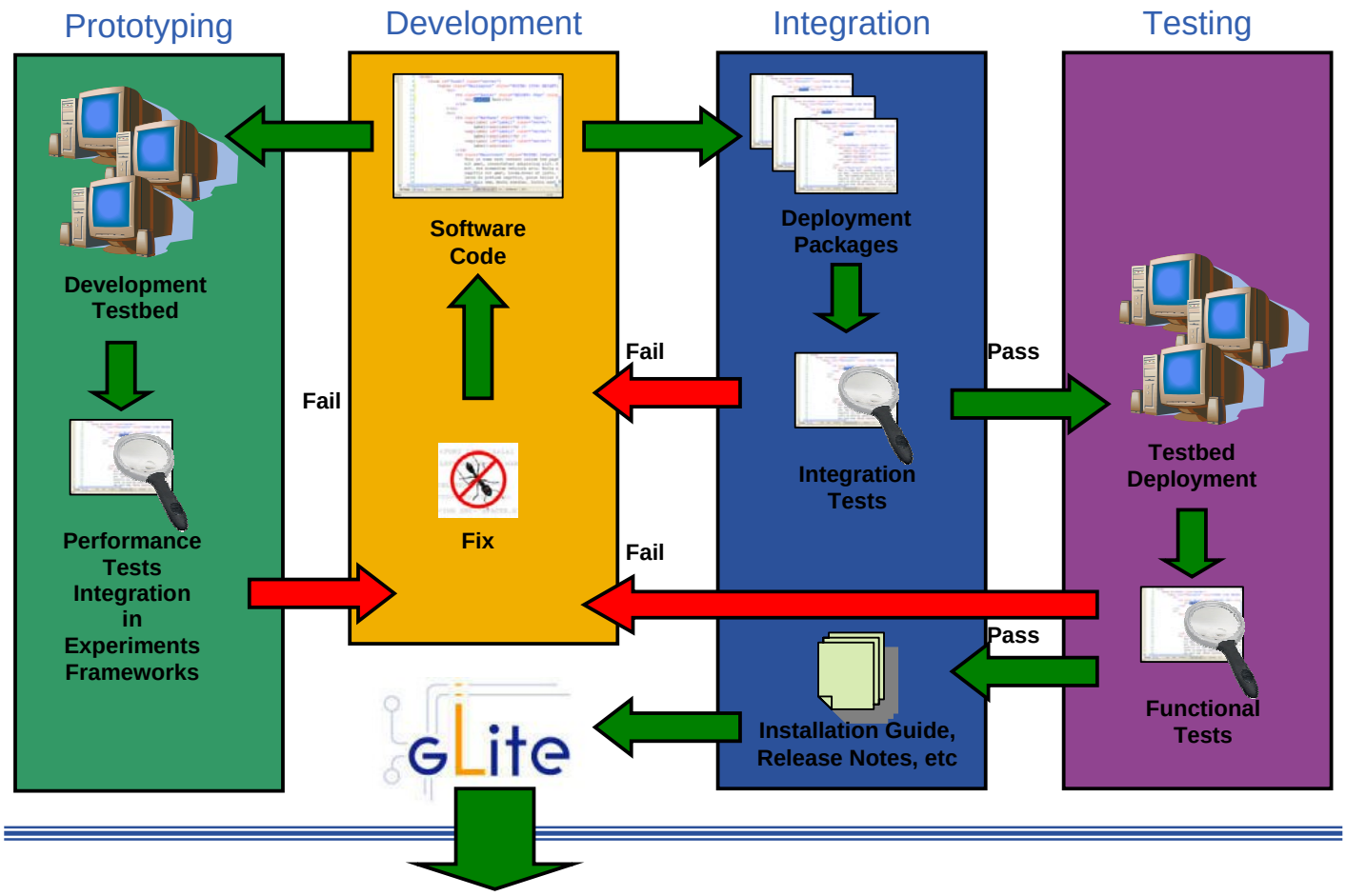
Services	Component	Responsible	Remaining development work to be done and expected dated [1]	Integration Completion Date[2]	Date available for ARDA/Bio testing[3]	Date for feedback from ARDA/Bio testing	Date available for JRA1 testing[4]	Date for feedback from JRA1 testing	Date available for SA1 testing[5]	Date for feedback from SA1 testing[6]
WMS	WM	IT/CZ	none	December 10, 2004 (push-only & RLS interface functionality exposed)	October 15, 2004	Under test since (first feedback at ARDA workshop 4 days after being available) extensive tests in December 2004. Iterations still going on	November 2 2004	December 10 2004 Installation, Configuration, LCG test suite run once successfully	February 2, 2005	Feedback provided based on initial examinationin January 2005
	CE (Condor-C based)	IT/CZ, ANL, Wisconsin,	none	December 10, 2004	See WM	Not directly exposed	See WM	See WM	See WM	See WM

[1] Further development work required to meet the release 1 goals defined for this module.
[2] Integration complete means the component is part of an integration build that includes code, deployment modules & documentation.
[3] ARDA and Biomed perform functionality & performance tests and prepares for integration with experiment frameworks. These tests are performed on the gLite development testbed before or in parallel to the JRA1 tests. Final assessment will be performed on the pre-production service.
[4] JRA1 performs installation, configuration and basic functionality tests on a dedicated testbed.
[5] The date in this column is when the module is expected to be available with JRA1 testing suite. Where the date given is earlier than the date for feedback from JRA1 testing, SA1 may take the components after integration is completed for initial examination and to allow testing to proceed in parallel.
[6] SA1 performs tests according to the certification and validation process put in place for the LCG-2.

Services	**Component**	**Responsible**	**Remaining development work to be done and expected dated** [1]	**Integration Completion Date**[2]	**Date available for ARDA/Bio testing**[3]	**Date for feedback from ARDA/Bio testing**	**Date available for JRA1 testing**[4]	**Date for feedback from JRA1 testing**	**Date available for SA1 testing**[5]	**Date for feedback from SA1 testing**[6]
		JRA3								
	CE Monitor (pull mode operation)	IT/CZ	Demonstrate CE Monitor (pull mode) Jan. 2005, then finish integration and testing	February 2, 2005	February 2, 2005	February 23, 2005	February 7, 2005	February 18 2005 Installation, configuration, basic functionality	February 18, 2005	March 9 2005
	Logging & Bookkeeping	IT/CZ	none	December 10, 2004	February 2, 2005	March 2, 2005	See WM	See WM	See WM	See WM
	StorageIndex Interface	IT/CZ, CERN	Demonstrate interface to StorageIndex Jan. 2005, then finish integration and testing	February 2, 2005	February 2, 2005 (pending on Fireman catalog availability)	March 2, 2005	February 7, 2005 (pending on Fireman catalog availability)	February 18 2005 Basic functionality	February 7, 2005 (pending on Fireman catalog availability)	March 9 2005

Services	Component	Responsible	Remaining development work to be done and expected dated	Integration Completion Date	Date available for ARDA/Bio testing	Date for feedback from ARDA/Bio testing	Date available for JRA1 testing	Date for feedback from JRA1 testing	Date available for SA1 testing	Date for feedback from SA1 testing
DM	gLite I/O	CERN	Fix gLite I/O reported problems. February 2004 (2 bugs pending)	November, 2004 (first iteration)	October 20, 2004	November 2004	October 20, 2004	October 26, 2004: Basic gLite I/O functionality.	October 26, 2004	Initial feedback provided in November 2004
				Second version correcting problems reported by ARDA available on January 2005.	January 2005	Since the "testing" part has been given to JRA1, we look for a more complete system using gliteIO + we are suggesting tests within JRA1 (joint JRA1/ARDA testing effort)		Enhanced functional test suite, some stress and regression tests in collaboration with ARDA		
	FireMan catalog (Global catalog equivalent to AliEn catalog)	CERN	finish integration and testing	February 2, 2005	November 20, 2004	See gLiteIO	February 7, 2005	February 25 2005 Installation, configuration and Basic functionality in collaboration with ARDA	February 7, 2005	March 9 2005
					Second version correcting problems reported by ARDA: February 2, 2005					

	FTS/FPS	CERN (Wisconsin)	Remove Stork dependency as agreed with LCG, then finish integration and testing.	February 2, 2005	February 2, 2005	March 2, 2005	February 7, 2005	February 25 2005: Installation, configuration and Basic functionality in collaboration with ARDA	February 7, 2005	March 2 2005
	User Interface	CERN	Port lcg-utils to use FireMan February 2005 (finish integration and testing). Need to identify manpower for this task by Feb 8	February 25 2005	See WM and R-GMA for these parts; lcg-utils in February 2005	March 16, 2005	February 25 2005	March 4 2005 Testing on best effort	February 25 2005: Existing CLIs (WMS, R-GMA) may be installed earlier; data mgmt tools can be accessed through their WSDL, but deployment modules will only be available then.	March 16 2005
IS	R-GMA	UK	Service Discovery to be finalized and integration and testing to be finalised by February 25, 2005 i	December 17, 2004	February 2, 2005	March 29 2005 (Application monitor)	December 17, 2004	Currently no effort available to test (best effort from RAL)	December 17, 2004	March 29 2005 (depend on establishing a migration plan from the LCG2 version)
VOMS		JRA3+IT/CZ	Finish integration, Testing	February 2, 2005	Will not test directly (VOMS is used for managing the prototype VO but ARDA does not do any VOMS specific tests).		February 7, 2005	Currently no effort available to test (best effort from NIKHEF)	February 7, 2005	March 2 2005


Prototyping
Development
Integration
Testing
Development Testbed
Performance Tests Integration in Experiments Frameworks
Software Code
Fix
Deployment Packages
Integration Tests
Testbed Deployment
Functional Tests
Installation Guide, Release Notes, etc
Fail
Fail
Fail
Pass
Pass
gLite
Operations (SA1)