Report of the

LCG Operations

GDB Working Group 4

Date:	6th December 2002
Version:	0.5
Status:	Draft

Document Log

Issue	Date	Comment	Author
0.0		First draft	Federico Ruggieri
0.1	31/10/02	Reformatted and various comments	Tony Cass
0.2	15/11/02	New Comments	Federico Ruggieri
0.3	22/11/02	Radically rewritten for wider audience	Tony Cass
0.4	27/11/02	Incorporates comments following Working Group meeting at CERN	Tony Cass
0.5	06/12/02	Some changements	Federico Ruggieri

Document Change Record

Issue	Item	Reason for Change

1 Introduction

The LCG Operations Working Group[1] was asked to

1. Recommend an LCG model of operations; including defining the responsibilities of the LCG teams, the experiments and the participating site operations teams.
2. Recommend a standard LCG operating environment; define the mechanism for implementing standard operating environments that satisfy the needs of the experiments; for example OS level, compilers, etc. How should the upgrade process be handled? What tools are required to implement this? What is the certification process?
3. Agree on an LCG policy for deployment of middleware including the integration, testing, and certification process for release candidates of the software.
4. Recommend a common accounting and reporting process for resource usage so that aggregate resource usage LCG-wide can be monitored. What are the requirements for this process? What reporting is needed back to the participating sites or to funding agencies?
5. Recommend configuration management policies – what are the system administration requirements for grid services at a site?

Our proposals in these areas are detailed in the rest of this document. Note that our proposals are for the *initial* LCG service. The ultimate answers to these questions will be the result of a long process of generally agreed proposals that show their worth in worldwide trials or tests. We do not pretend to define here the "final" solutions for LHC production but, rather, a preliminary model that can reasonably be introduced during the first phase of LCG. Nevertheless, we include a final section pointing to various issues we feel should be investigated during the LCG-1 service as part of the preparation for follow on services.

Given the nature of the questions to the Working Group, no simple executive summary is possible. However, the group was asked one specific question:

Is LCG a single resource or several — one per experiment?

Our answer is clear. Whatever the underlying implementation, **the LCG service should be presented to users as a single resource.** It is essential that any local restrictions (e.g. that no ATLAS jobs be run at Fermilab) be handled by the Grid middleware and not exposed to users.

[1] Members: John Gordon/UK, Klaus-Peter Mickel/Germany,
Federico Ruggieri/Italy (Chair), Hiroshi Sakamoto/Japan, Tony Wildish/CMS
Technical advisers: Tony Cass/CERN, Luca Dell'Agnello/CNAF, Luciano Gaido/INFN

2 LCG Operation Areas

2.1 LCG Operations Model

The basic model is that ***sites*** are responsible for providing computing resources that meet the requirements of the ***experiments***—who, in turn, have the duty to collaborate with site managers to ensure efficient resource usage. ***LCG Teams***, especially ***middleware suppliers***, are responsible for providing the software to allow transparent use of resources across sites but must do so in a way that is compatible with the general operation of large computing centres—including, or especially, computer centres that do not cater solely to LHC (or HEP) users.

For computing resources to be usable, three types of software are required.

Basic Operating System	Selection of the OS version is agreed between ***sites*** based on general Linux developments as experiment software is expected to be relatively independent of the Linux version. However, • the compiler dependency is clearly recognised, and • on average, new systems should not be introduced more than once per year.
Grid Middleware	Developed by the middleware suppliers and upgraded as necessary to improve service delivered to experiments. The middleware architecture and delivery format are to be agreed with sites. On average, new middleware versions should not be introduced more than three times per year—although the acceptable frequency will likely depend on the transparency of the upgrade process for system administrators and experiments. Implementations of the mapping of grid level resource commands (such as job submit, query, cancel) shall be provided for the LSF and PBS local scheduling systems together with a precise explanation of the expected behaviour of these commands to allow implementation of the mapping for other local resource managers.
Experiment specific software	This software is expected to "sit on top of" the basic operating system—i.e. no root privileges are necessary for installation, only write permission to the relevant directory tree.

Specific team responsibilities are defined in the following sections.

2.1.1 Site Operation Team responsibilities

Sites shall identify one or more people to fill the following roles.

System Administrators	Installation and upgrading of the basic OS, compilers and applications as defined in the standard environment and grid middleware as defined by Working Group 1. At least initially, ensure that approved experiment software releases are installed on nodes and advertised correctly. Long term, this should be handled more automatically—see section 3 for more details.

	Ensure full implementation of grid level resource scheduling commands in the interface to the local resource scheduling system. Ensure correct configuration of the interface between grid middleware and the local environment—including the correct advertisement of local capabilities.
Security Officer	To coordinate security issues with other sites and the Grid operations team. We expect the detailed responsibilities to be defined by Working Group 3.
Network Specialist	To coordinate networking connectivity, bandwidth and use issues with other sites and the Grid operations team.

2.1.2 LCG Team responsibilities

Middleware suppliers	Middleware suppliers must take due account of system operations needs when designing and packaging middleware. In particular • Software required on worker nodes should be delivered in libraries or, at most, as kernel extensions. Requiring daemons to run on each worker node introduces operational complexity. We expect that this will reasonably be limited to well defined scopes. • Software running on worker nodes must not assume any direct external network connectivity. Such connectivity can be assumed only for the GRID elements (Compute, Storage, Gatekeeper, etc. in EDG terms). On the other end present implementations do require some in and out-bound connectivity (Sandboxes, Interactive Jobs, etc.). There should be an exaustive discussion on the realistic requirements in this field. • Middleware must be packaged in RPM format. Other formats may be provided, but RPM format is required. Apart from the use of RPMs, no assumptions may be made about, or placed on, the actual method of installation.
Grid Deployment	Provide support to sites for the installation of grid middleware, experiment software and other components such that they can function effectively as part of the LCG-1 service. Develop, or foster development of, tools to ensure that this team is not needed for future LCG services (see sections 2.2.2 and 3 for a discussion of desirable tools).
Grid Operations	General coordination between, and monitoring of, sites including coordinating effort • find common solutions to similar problems, • maintain definitions of minimal requirements, and to • support centres in the migration to new operating environments.
Grid Support	We expect the detailed responsibilities and roles of this team to be defined by Working Group 5.

2.1.3 Experiment Responsibilities

Experiments	Provide and support "job prologue procedures" that avoid the installation of shareable software for each job. In principle, if resources are

advertised and requested correctly then this should not be needed. Experiments should have some means to check that the software under the agreed root (see section 2.2.1) is as expected. If so, this should be used. Software should only be untarred in the sandbox if it is not already available on the node.

See section 2.2.3 for a discussion of the requirements on experiment software.

2.2 Standard LCG Operating Environment

All the existing Regional Centres have already implemented their own Model of Operations, but working together will require a new level of coordination and organization with few clear rules to be obeyed. Define results, not mechanism.

In this first phase a minimum set of requests that can define the operating environment of LCG-1 can be:

2.2.1 Basic Operating System

All LCG-1 systems provided for running experiment jobs ("worker nodes") shall run the RedHat 7.3 Linux distribution. Middleware suppliers are required to make known their requirements and constraints on the kernel configuration. A definitive list of required kernel patches must be available at least two months before initial LCG-1 production. Each participating site must publish a list of their local additional patches, again at least two months before initial LCG-1 production. Changes to this list shall be published as they occur but are not expected to be frequent—being provoked, for example, by a requirement to install a patch for a security exposure.

Middleware that runs on non-worker nodes shall be compatible with the basic operating system defined for the worker nodes.

Whatever compiler is used for the basic system installation, gcc 3.2 and the relevant libraries shall be available for use by the experiments (boost ?).

In principle, additional software tools, e.g. Perl, Python and so on, shall be installed as provided in the RedHat 7.3 distribution. However, it is possible that the grid middleware components place some requirements on these tools. The precise details of the additional software tools will therefore have to be defined in consultancy with the experiments once the Working Group 1 recommendations are known. Nevertheless, this working group believes that there must be extremely good reasons to require deviations from the default application versions of the base OS distribution.

In addition to those delivered with the RedHat 7.3 distribution, the following PERL modules shall be available on worker nodes: Mail::Mailer and SNMP_Session. These modules shall be taken from …

Each worker node shall have a directory `/opt/hep/experiment` below which experiments will expect to find software. This directory shall be read only for batch jobs and shall provide at least 2GB of space per experiment (sufficient to support up to 3 versions of the experiment specific software). Beyond these two requirements, the implementation mechanism (local, mounted from some distributed file system) may be chosen freely at each site. Site system administrators should, however, understand the implications for the local installation of experiment software in this directory.

We do not consider it feasible to require all sites to maintain a complete list of VO members. Middleware tools and the site configuration must therefore allow the mapping of grid certificates to generic rather than individual local identifiers. The mapping process may or may not attempt to use an individual local identifier first. If the mapping is to a generic local identifier then

- middleware tools must permit the establishment of an audit trail recording the DN of the certificate to which a generic local identifier corresponds at all times;
- there must be no concurrent allocation of generic local identifiers to distinct DNs, and
- jobs running with a generic local identifier may have fewer privileges than jobs running as a locally identified user but must in all cases have full access to the resources of the appropriate VO.
- mapping with Kerberos tokens/id's should also be available (under discussion).

2.2.2 Installation Mechanism and Certification Process

It is expected that most, if not all, of the sites participating in the LCG-1 service will be major computer centres with mature services—and, perhaps, services for communities other than the LHC experiments or the HEP community. It is therefore likely that most centres will have well established installation and operation procedures. In this environment, we feel it is inappropriate to mandate the use of any particular installation tool. Sites are therefore free to choose any overall installation procedures provided that they can handle the installation of software delivered in RPM format. However, sites should clearly review any fabric management tools associated with the middleware selected by Working Group 1 (e.g. LCFG from EDG).

Tools to verify that the installation at a given site is correct—from the basic worker node configuration through the local grid environment to the connection to the wider world—are desirable. Unfortunately, we do not consider that tools to enable the full certification of a site will be available in the near future. Given this, we believe that considerable system administration effort will be needed at each LCG-1 site to install, understand and configure the required software correctly. We believe that, in the short term, the Grid Deployment team should assist sites with these tasks. In the longer term, we consider that the Grid Deployment team must ensure that tools are available to automate at least the certification process.

Ideally, the Grid Deployment team would set a framework for site certification by the time of the LCG-1 service. With such a framework in place, package developers will understand how their software will be verified—by both local site administrators and the Grid Operations team—and, in some cases, may even be able to provide compliant verification procedures.

2.2.3 Experiment Software

As noted above, it is expected that experiment software be installed below the `/opt/hep/experiment` directory and that no particular privileges (other than write access to this directory) are required for installation.

Experiment software must be versioned. In order to allow software installation in parallel with job execution, individual components of a particular software release may

not be replaced directly. New copies of changed components must be provided in a separate location.

We find it hard to prescribe any mechanism by which experiment software is installed in the `/opt/hep/experiment` directory. Long term, it is clear we need a mechanism for the automated installation of experiment software at relevant HEP sites. This installation should be launched across all sites once the relevant software librarian certifies a new release.

Automated installation of experiment software should not be triggered on demand if a particular job requires a newer version than that available on the node. In these cases the job should install the required software in the sandbox and the mismatch in software versions should be reported to the site manager and the Grid Operations team.

2.3 Middleware Deployment Policy

Middleware, and, indeed, any tools or packages required to be installed at LCG-1 sites, must be fully tested and certified by the Grid Deployment team before local system administrators are asked to install them. We consider the steps in the certification process to be as follows.

1. Initial delivery: Package developers must deliver code that compiles and runs correctly. The package must be delivered in the correct install format (RPM) and, eventually, with a verification procedure compatible with the agreed framework (see section 2.2.2).
2. Integration: Updated packages are tested in conjunction with other LCG-1 components by the Grid Deployment team. It is highly desirable that middleware and other software packages be modular in order that any changes in behaviour can easily be traced to a given component.
3. Testing: Extensive tests should be performed, including tests at one or more production sites in collaboration with experiment experts. Where necessary, verification procedures for the overall LCG-1 environment should be updated or developed.
4. Certification: Packages will only be certified once they have been shown to be compatible with the rest of the LCG environment and to work correctly in the test environment over a long period.

2.4 LCG Accounting

Activity in this field is already under way (see http://server11.infn.it/workload-grid/docs/DataGrid-01-TED-0126-0_8.pdf). It will be very difficult to implement consistent accounting across all LCG-1 sites during this first phase, but a well defined accounting model should be available for the second release and tested during the LCG-1.

The requirements for this system are as follows:

- the accounted resources must include CPU time, job statistics (elapsed time vs. cpu usage, job completion codes—success, timeout, user error), Mass Storage use (disk and tape as appropriate) and network transfers at the file level.
- accounting details shall be gathered by user and aggregated to produce summaries by experiment (VO) and centre.

- weekly and monthly reports must be supplied to the Grid Operations team. Tier1 centres must be capable of providing daily reports if requested.

Notes:

1. CPU usage needs to be accounted in some standard unit. The only realistic basic unit at present is the SPECInt2000-second. Unfortunately, the SPEC consortium changes their processor benchmark regularly and so conversion factors must be stored along with the historical accounting data. The alternative would be to create a new HEP benchmark but we do not see this happening in the near future—and, anyway, we need to report capacity and usage to the wider world in some generally understandable unit.
2. The requirement to account resource utilisation by user has implications for the mechanism used to map grid certificates to local identifiers. If there is no one-to-one mapping of certificates to local ids then it must be possible to appropriately combine resources used by multiple local ids to give the correct resource usage figures for the different grid users concerned.

2.5 Configuration Management

Each site should have at least a GRID system Administrator for the day by day management.

Management should be coordinated among all the RC with a very light organisation. Planning should be done twice per year together with the review of what was achieved in the previous periods. A GRID Management Board will be probably needed in the end, but due to the large number of countries and RC involved worldwide, a complete representative committee is excluded.

3 LCG Service Evolution

As has been noted elsewhere in this report, we expect experience gained during the LCG-1 service to lead to developments which will improve the operational procedures for future services. In no particular order, we present here areas that we feel must be developed and/or explored.

- As the number of sites participating in LCG services expands, fewer and fewer will have full-time dedicated system administrators. In addition to providing installation tools to minimise the effort required for each site to participate—and keep up –to-date—tools are required to allow sites to verify that the local installation is correct and to allow the grid operations team to verify that a site is correctly (re-)configured to participate in the overall environment. There should be provision for different levels of grid certification. These should range from verifying that a simple "hello world" job will run correctly to verifying that sample production jobs provided by each experiment produce the expected results.
- It is not expected that most experimental code will require WAN connectivity on worker nodes—jobs should interface to Grid services at the local site which provide the required connectivity. However,
    - are there specific use cases that do require WAN access? If so, at what level and how will the resource requirements be specified? Will this limit execution of such code to certain sites?
    - Is there any requirement to install and run experiment specific code on non-worker nodes? Or proxies?
- Some mechanism is required to allow an authorised experiment representative to launch the installation of a new version of experiment software at all sites. We assume that experiment software will be treated and transferred between sites as a simple data file. The installation tools will have to take account of the varied ways in which the /opt/hep/*experiment* directory can be implemented. Points to note here are
    - Must sites guarantee that N versions of the software will be available?
    - Can older versions be deleted automatically as newer software is installed?
    - In the (very) long term, can grid middleware identify which software is being requested to tune any automatic deletion (it may be appropriate to have X.5 and X.3 available, and not X.5 and X.4)?
- It is unlikely that all worker nodes for future LCG services will be running the same Linux distribution, let alone the same distribution version. We believe, however, that it is reasonable to require Linux distributions that are LSB compliant and support an RPM based installation process. Heterogeneity should be introduced carefully at first, perhaps as RedHat version 8 becomes available. We note here that the inevitable future heterogeneity is another reason for grid middleware and experiment specific software to be only loosely coupled to the base operating system.

In addition to the Grid Deployment Board as the ultimate decision body we see HEPiX as the appropriate forum for technical discussion in these areas and strongly encourage greater participation by the LHC experiments in future HEPiX meetings.

Finally, any tools provided must be such that all sites can run jobs from any (LHC) experiment. Site policies may then exclude jobs from some experiments, but such restrictions must not be imposed as a side effect of using any particular tool.