



LCG Grid Operations Service

LCG Resource Administrators' Guide



<i>Date:</i>	16 February 2004
<i>Version:</i>	0.2
<i>Status:</i>	Final
<i>Author:</i>	Trevor Daniels



Document Log			
Issue	Date	Author	Comment
0.1	10 Nov 2003	Trevor Daniels	First draft for discussion in GOC Steering Group and LCG Security Group
0.2	16 Feb 2004	Trevor Daniels	Incorporating comments and general tidying up
0.2	8 June 2004	Ian Neilson	Moved to status FINAL after approval by GDB on 18 May 2004



1 Introduction

This document provides guidance to Resource and Service Administrators (hereafter called ‘Resource Administrators’) for meeting the requirements of the LCG Security and Availability Policy with respect to operating Resources and Services on the LHC Computing Grid.

The adopted Security and Availability Policy places four responsibilities on Resource Administrators:

Site Policy

Resource Administrators must ensure their implementations of LCG services comply with both their site policies and the LCG Policy.

Notifying Site Personnel

Resource Administrators are responsible for ensuring that all appropriate personnel concerned with security or system management on their site are notified of and accept the requirements of the LCG Policy before implementing any LCG services.

Resource Administration

The Resource Administrators are responsible for the installation and maintenance of Resources assigned to them, and subsequently for the quality of the operational service provided by those Resources. This quality will be defined by the Service Level Agreement (SLA) for each Service as published by the Administrator of that Service.

Service Level Agreement

The Administrator of each Service instance must maintain an assessment of the risks inherent in their particular Service design or resulting from local services or operational practice which might affect that Service’s Availability, Reliability or Performance, and publish the expected values of these service parameters in accordance with the GOC Procedures for Resource Administrators. The publication of this information, together with other details described in the LCG Service Level Agreement Guide, constitutes the SLA with the user community for that service.

This Guide defines in more detail the requirements of the last two of these four responsibilities, and should be read in conjunction with the companion document “*LCG Service Level Agreement Guide*”.

2 Definitions

LCG is the LHC Computing Grid project which was formed to build the computing environment to support the scientific exploitation of the Large Hadron Collider (LHC) at CERN.

An **LCG Service** is one of the production Grid services which are made available to a remote and general community of LCG users via the Internet. LCG Services include but might not be limited to User Interfaces, Computing Elements, Information Services, Logging and Bookkeeping, Resource Brokers, Replica Catalogues, the Security Infrastructure and Storage Elements.

A **Site** is an institute which is providing one or more **LCG Services** to the LHC Computing Grid.

The **LCG Resources** at a site are the hardware, software, data and supporting infrastructure required to provide the LCG Services operated by that site.



The **Resource Administrator** of an LCG Service at a Site is the person responsible for providing and maintaining an LCG Resource or LCG Service at that Site. (Although the singular is used in this document several individuals may actually be involved, including system programmers, operators, user support staff, networking staff and security personnel.)

3 Service Quality

3.1 Operational Cover

During the early stages of LCG the most likely cause of service failure will be due to immature software. Most sites will need to recover from this type of failure by manual action, which will usually entail investigating and recording the circumstances of the failure to aid subsequent diagnosis followed by stopping and restarting relevant service processes.

The time required to recover services which fail in this way is determined by the time to detect the failure plus the time taken by the support staff responsible for operating that service to respond and take the appropriate action. The local arrangements for detecting Service failures, raising alerts and calling out staff to attend to them are therefore key parameters in determining Service Availability.

Part of the SLA requires these arrangements to be recorded and it is the responsibility of the Resource Administrator to ensure that the published SLA reflects these working practices accurately. The information required is:

- Brief description of Service monitoring and alerting arrangements
eg all processes monitored by Nagios; alert raised on screen and by pager
- Number of staff (full-time equivalents - ftes) available and competent to support this Service, ie able to recover a failing service by determining the nature of the failure, stopping/restarting appropriate processes and/or booting appropriate hosts.
eg 4 ftes
- Hours of on-site cover provided by at least one of those staff per week
eg 75 hours per week in two overlapping shifts from 7am to 10 pm Mon-Fri
- Schedule of off-site cover provided by those staff by call-out and estimated call-out response time
eg 1 hour response time outside on-site hours from Mon evening to Fri morning. No cover Fri evening or at week-ends.

3.2 Risks

Each Resource Administrator should consider the risks which might affect the quality of all LCG Services under his/her control with a view to

- a) minimising the risks and/or their effects
- b) estimating the effect the residual risks might have on the LCG Services

The risks which should be considered include instabilities of the essential services and environment (power, cooling, flooding, fire, etc), failure of the essential networking infrastructure, failure of some essential hardware component, inherent software failures, administrative and human errors, malicious attacks or the effects of untargeted but disruptive hacking. A more comprehensive list of possible risks is taken from the LCG Risk Analysis conducted by the LCG Security Group and reproduced at Appendix A.



The risk assessment methodology in standard use at the Site should be employed in order to calculate an estimate of the Availability and Reliability of the Service, taking all significant risks into account. Alternatively the methodology described in Appendix B may be used.

The estimated or targeted values for the Availability and Reliability are then published by the Service Administrator in accordance with procedures described in the LCG Service Level Agreement Guide.

3.3 Physical Security

There are no special requirements imposed on the physical security of the resources used to provide Grid services other than those imposed on services of this nature by the host site. The exception to this is the provision of a Certificate Authority service, which is covered by document "Approval of LCG Certificate Authorities" referenced from the Security and Availability Policy.

3.4 Change Control

All LCG Service-related software is distributed from CERN under the control of the Grid Deployment Team. Announcements of new releases and updates to LCG software are made on the LCG_Rollout mailing list, to which the system administrators of all Grid services must subscribe.

Resource Administrators **must** ensure they or a deputy are always able to respond promptly to announcements on this list, since software changes required to meet new or developing security threats will be made by this means. Resource Administrators are required to respond to security-related announcements of required changes within 24 hours and to other releases or upgrades of software within 3 working days.

Scheduled changes to operational Services which are due to local reasons must be announced in advance in the appropriate site-related page on the GOC website.

4 Accounting

[This section will be expanded later when the details of the mechanism are known]

All sites offering a CE Service must provide accounting information to the Grid Operations Service to enable LCG-wide accounting statistics to be maintained.

Each site is required to preserve accounting information locally for all jobs submitted under any of the LCG VOs until the required information has been transmitted to the GOC and an acknowledgement of safe receipt obtained. The required information is held in the batch and gatekeeper logs. Each site must convert their local accounting data into the format prescribed by the GOC before transmitting it either by using one of the supplied filters or by writing their own using one of the supplied filters as a model. Accounting information must be transmitted within 5 working days of the end of each calendar month. *[The details of how this is to be done will be inserted later]*

The GOC will collate all accounting data and prepare written and online reports in agreed formats which will be available from the GOC website. *[Details to be decided]*

5 Fault Investigation, Tracking and Rectification

Fault analysis and rectification on a computing structure as diverse and complex as LCG involves many people: grid users, the several experiment support teams, the grid user support call centre, the grid operations centre, the grid deployment team and the local support staff at each of the sites offering grid services as well as the software developers and hardware maintenance personnel.

The responsibility for rectifying operational faults which affect a production grid service lies exclusively with the local support staff at the site which is providing that service. However, before a fault can be rectified in this way it must first be localised to one (or perhaps occasionally a few) sites,



and this process of fault analysis and isolation may well involve several of the other groups of people mentioned above.

5.1 *Fault investigation*

A **user** who believes (s)he has encountered a problem or fault with a grid service will first approach either his local experiment support staff if the fault appears more likely to lie with experiment-specific software or the Global Grid User Support (GGUS) centre at FZK, Karlsruhe (<https://gus.fzk.de/>) (or one of the User Support outstations) otherwise. Experiment support staff will forward problems which on investigation are found to be unrelated to the experiment's application software to GGUS.

GGUS will deal with grid-related problems of a general nature which can be resolved by user action and will forward problems of an operational nature to the Grid Operations Centre (GOC) at RAL (<http://www.grid-support.ac.uk/GOC/>) (or one of the GOC outstations) for further investigation.

The **GOC** will investigate all problems of an operational nature which are passed to it in order to determine which site should take responsibility for further investigation and eventual resolution. It will often be immediately clear which site this is, but occasionally a complex fault may require extensive investigation involving several sites before the fault can be isolated and localised to a particular site. For example, the failure of job to access an imported file may indicate a problem with an RB, a CE, a WN or any part of the RC mechanism as well as user or application error. This preliminary investigation, called first-level fault analysis, will be carried out by the GOC and the responsibility for progressing the fault remains with the GOC until the fault is either resolved or it is passed to the local support staff at a site.

The **local support staff** are responsible for assisting the GOC in carrying out first-level fault analysis and for progressing (with the help of their agents) all faults passed to them to full resolution or to a point where they can demonstrate to the GOC's satisfaction that they are not responsible for the fault condition.

GOC, GGUS and the local support teams are jointly responsible for investigating and isolating defects in grid software which come to their attention and for notifying the LCG deployment team of them via the lcg-rollout list (or its subsequent replacement).

5.2 *Fault tracking*

Faults which concern application software and those that are resolved completely by one of the experiment support teams are tracked by experiment-specific help desk software, currently believed to be based on Savannah. Neither the GOC, GGUS nor the Service Administrators will normally be involved in investigating these faults.

Faults reported to GGUS will be entered and tracked in the GGUS fault tracking system, currently based on an existing Remedy system. An interface from Savannah to Remedy is being developed by GGUS to enable tickets to be transferred from one system to the other.

The GOC will also use the Remedy system at GGUS to track faults which are its responsibility. This enables responsibility for investigating faults to be easily transferred between GGUS and GOC. Registered users will be able to enter fault reports directly into this Remedy system.

The GOC will transfer tickets from Remedy to local support staff by direct email, and will continue to update the Remedy system as the fault is investigated by local support staff. The lcg-rollout list will be used for reporting the investigation of faults which are of relevance to the general community of Resource Administrators.

Faults which are traced to defective software will be notified to the Grid Deployment Team at CERN either via the lcg-rollout list or by direct entry by GOC or the Deployment Team into a deployment fault tracking system.



5.3 *Fault Rectification*

The Resource Administrator, together with his/her local supporting staff are responsible for all installation and operational aspects of the grid Services they are running. They alone will be responsible for diagnosing and rectifying operational faults, ie other than those arising from defective software distributed by the Deployment Team, on the services under their control.

Faults arising from defective software will be passed by the Deployment Team for rectification to the appropriate software developers.

5.4 *Fault Escalation*

The GOC will monitor the progress with investigating and rectifying all faults of which it is aware, and will take appropriate steps to escalate those that do not appear to be making satisfactory progress. Escalation procedures will normally involve contacting senior personnel at the site carrying the responsibility for that service.

Users affected by a fault may request escalation via GGUS.



Appendix A

[Extract from the Risk Analysis conducted by the LCG Security Group - tba]



Appendix B

A Simple Risk Assessment Methodology

The numbers involved in risk assessments are necessarily imprecise estimates, and only approximate order-of-magnitude calculations are required. Nevertheless, these will indicate clearly enough which risks are significant and which operational practices are effective and justified.

Note that this simple approach is suitable only for estimating the Availability and Reliability of instant short-term Services. Where the longer term storage of data, for example, is involved the effect of a risk beyond simply causing a break in service must also be considered.

1. Identify the Risks

List all the risks which could cause the Service to fail during its anticipated lifetime, using Appendix A as a guide.

2. Estimate the Failure Rates

Estimate the number of service failures per annum which are expected to result from each risk. For unlikely risks this will be fractional. An order-of-magnitude estimate is adequate.

3. Estimate the Recovery Times

Estimate the average length of time required to recover the Service following each risk. Note this must cover the total down-time of the Service, including the time taken to detect the failure, the time to respond to it (for example, by call-out), the time to diagnose the problem and the time to rectify it. Take sensible averages. If a service has on-site operational cover for 75 hours a week, 1-hour call-out cover for a further 36 hours overnight and no cover for the remaining 57 hours at the weekend the time to respond to a fault on average would be

$$(75 \times 10 \text{ mins} + 36 \times 1 \text{ hour} + 57 \times 57 \text{ hours} / 2) / 168 = 10 \text{ hours}$$

If the site policy is to simply restart failing services as quickly as possible the estimated total recovery time might then be 10 hours and 10 minutes.

4. Estimate the Reliability of the Service

The Reliability of the Service (mean time to failure - MTTF) is estimated by taking the inverse of the failure rates aggregated over all risks. So, for example, a total failure rate of 7 per annum from all risks would yield a MTTF of $365/7 = 52$ days.

5. Estimate the Availability of the Service

Calculate the estimated unscheduled downtime of the service in hours per annum. This is given by the sum over all the risks of the failure rate times the recovery time. The Availability is then given by the unscheduled downtime per annum divided by the number of scheduled service hours per annum, usually $52 \times 168 = 8736$.

6. Take Mitigating Action

If the estimates of Reliability and Availability are below the acceptable values for the Service consider taking mitigating action to improve them. The table of risks showing their rates of occurrence and times to recover from them will indicate what actions would be most effective.