C C L R C

RUTHERFORD APPLETON LABORATORY

Grid Operations Service for LHC - Proposal

T Daniels
JC Gordon
25 June 2003

BACKGROUND

An earlier proposal to establish a Grid Operations Centre for LCG was considered by GDB in June. Four main points emerged:

- The approach was too low-level, concentrating unduly on monitoring individual Grid components rather than Grid-related services.
- There was little recognition of the federal nature of the Grid, and of the important need to define service quality through Service Level Agreements.
- Consideration should be given to the relationship between the GOC and the proposed LCG Call Centre
- The issues which require investigation during the evolutionary phase should be more clearly identified.

The proposal has been re-written to address these points, with the central approach now being focussed on monitoring the quality of those critical Grid services which are being provided by the various centres. The many other functions of a GOC were not discussed in June and these have been left essentially unchanged.

The June meeting of GDB also accepted a suggestion that a GOC Group be established to assist in the drafting of this proposal and in the selection of monitoring tools to be deployed in the early stages of the GOC. The names for this group that have been proposed to date are Markus Schulz (CERN), Rolf Rumler (IN2P3), Cristina Vistoli (INFN), in addition to Trevor Daniels and John Gordon (both RAL). Vicky White is seeking some American participants. The group as constituted so far has assisted by commenting on earlier drafts of this proposal.

GDB is invited to consider and comment on this revised approach to a Grid Operations Service and its suitability for the operational needs of LCG-1.

Particular aspects which have not been elaborated in this proposal and on which the view of GDB is sought include

- Whether and to what extent is accounting a requirement which the GOC needs to address? This concerns not so much the recording of raw accounting data, but what is to be done with it; for example, whether action is to be taken to assist in ensuring users obtain a fair or agreed share of the available resources.

- Whether and to what extent is the GOC to exert any control over changes to the production Grid, and if any is required, how the authority for that control is to be provided?

- Whether and to what extent should the issue of operational control of a Grid within a federation of Grids be developed? It is far from obvious what monitoring such a federation of Grids means, or how the various GOCs might interact, or indeed whether the various services are related to separate Grids or to the whole federation. This, and related issues such as how the middleware in heterogeneous Grids might inter-work, are very much open questions at present. Such questions will be easier to approach after fully functional Grid Services become available, but it would be advantageous to prepare the way before then by exploring all the issues (including operations) concerned with a federation of Grids in, perhaps, an LCG Working Group.

1. INTRODUCTION

Since no global grid is yet in full production use the need for a Grid Operations Service (GOS) and the functions which it might provide cannot at this stage be clearly defined and the manner in which a GOS might best be operated to provide those functions is certainly not well understood, although experience of managing grid operations is emerging from both EDG and the GOC for iVDGL at Indiana. In these circumstances it is proposed that a GOS for serving LCG is established in an evolutionary manner, through a number of prototype stages operated during 2003. The experience gained during that time will then be available early in 2004 for specifying with confidence the requirement for and final form of the GOS for the main operational period of LCG.

The LHC Grid is a means by which the computing resources owned or being used by the main LHC experiments, and installed across many centres worldwide, may be shared, both within an experiment and between experiments. This permits peaks and troughs in the computing demands of regional groupings or individual experiments to be smoothed out, leading to an increased utilisation overall. At any particular time an experiment may choose to reserve its computing resources for its own use or to assign part of them to the LHC Grid for communal use (or indeed to assign them to any other Grid).

It is the operational problems arising from the dynamic configuration of the Grid and the organisational problems arising from the federal nature of its participants that make a Grid Operations Service a new and unknown entity.

This document sets out in Sections 2 and 3 the purpose, functions and activities of what is today seen as a possible, maybe idealised, Grid Operations Service - the LCG GOS Vision. The purpose of any vision is to provide targets, direction and guidance to the early stages of implementation and the investigations and trials being conducted. It is expected that this sketchy description will then be refined early in 2004 in the light of experience gained in implementing the early prototype service during 2003 to provide a clear specification of the final GOS for production deployment.

In the later sections we discuss the practical considerations of organisation, phasing of operational functions and facilities, the levels of staff required and the tools that will need to be procured or developed.

2. THE NATURE OF A GRID OPERATIONS SERVICE (GOS)

The overarching purpose of a Grid Operations Service is to carry out actions designed to ensure the Grid Infrastructure is operating and continues to operate in an optimal and reliable manner. That is, the operational aspects of the Grid Infrastructure are optimised for the users' benefit, the failures of critical Grid Services are detected and rectified promptly, and the overall performance is monitored to detect and understand longer-term trends.

This task is made particularly difficult because the Grid Infrastructure is distributed over several tens of computing centres worldwide, each of which may elect at any time to increase or decrease the computing capacity of the resources which they are prepared to commit to the Grid. Providing a reliable and guaranteed service in spite of this dynamic and distributed nature of the Grid Infrastructure and in spite of their being no overall point of organisational control is the goal of the GOS, and the main consideration of this proposal.

We approach this issue by first dividing operational responsibility clearly between the Grid Operations Centre (GOC) staff and the network, operations, systems and support staff (collectively called herein Local Support Groups) at individual centres.

> **Local Support Groups** have at all times the sole and exclusive responsibility for all aspects of operating the equipment within their computing centre.
>
> **The GOC** has responsibility for monitoring the operation of the Grid Infrastructure as a whole, devising and managing mechanisms and procedures which encourage optimal operation of the Grid, and working with Local Support Groups to assist them in providing the best possible service while their equipment is connected to the Grid.

The approach being proposed hinges on the publication by each computing centre participating in the Grid of the service level parameters associated with the services they are offering in a format to be agreed. In keeping with the ethos of the Grid the GOC does not specify what services must be provided or what the quality of those services must be (within some broad range), but the GOC will specify the minimum threshold of quality which must be declared (and subsequently achieved) for a centre to join (and remain in) the Grid. Achievement of this minimum standard and with the standard being advertised by the centre will then be monitored and published by the GOC.

A simplistic hypothetical example illustrates the principle. If a centre wishes to provide a Resource Broker service, then it would be required to publish a schedule showing the times the service will be available, and the level of availability and the level of reliability (see Appendix) which were targeted in designing the operational aspects of that service. The GOC will have already published the minimum acceptable levels of availability and reliability for that type of service, and the GOC will have the power to disconnect the service if it subsequently and persistently falls below that minimum standard. In effect the publishing of the targeted service level parameters by the centre hosting that service constitutes a service level agreement (SLA) between the Centre and the Grid community which the GOC then oversees and monitors. Other service level parameters would be published for all other offered persistent services in a similar fashion. The GOC would then measure and publish the service levels actually achieved in practice by monitoring that service. Action could then be taken if the actual quality of service fell 'persistently' (a parameter to be defined during the evolution of the GOS) below the quality being published by the centre as its target.

Being seen to achieve or surpass targeted high standards in comparison with other centres is a strong incentive for centres to strive for high quality. Furthermore, the higher the quality of service offered and achieved by a centre the more work would be attracted to that centre. This in itself could be advantageous to the centre's future funding and, depending on the accounting mechanisms chosen, could also be advantageous to any VO associated with that centre.

Operational aspects in which the GOS will play an important role include, in addition to devising the regime under which Grid services are delivered, the inter-working of Local Support Groups on Grid-related issues, coordinating Grid-related security precautions and incident responses, and the provision of operational information to users via the LCG Call Centres.

3. GRID OPERATIONS SERVICE: PROCESSES AND ACTIVITIES

We assume the responsibility for carrying out the processes and activities of the Grid Operations Service (GOS) will fall to one or more Grid Operations Centres (GOCs). The organisation of the GOCs is discussed later.

This Section provides a brief description of the functions of the GOS which are to be established by mid-2005 as they are foreseen today - the GOS Vision. These processes will be developed over the two years from mid-2003 to mid-2005, being modified as they evolve as indicated by the earlier experience gained.

In some areas the work proposed here may overlap with work which other groupings may be expecting to carry out, two examples being security and change control. While we have attempted to include only those aspects which would normally form part of the duties of a Grid-wide Operations Centre we seek the agreement of GDB that we have drawn the boundaries correctly.

3.1 Coordinating Grid Operations

The GOC will be responsible for coordinating the overall operation of the Grid. It will act as convenor of coordinating meetings of Local Support Groups, or Regional Centres, ensure that appropriate operational regimes are debated, agreed and followed, and generally provide the initiative, impetus and steer to operational developments.

3.2 Defining Service Level Parameters

The GOC will define the service level parameters associated with every persistent Grid-related Service. The target values for these must be published by a centre before that Service is made available on the Grid. These parameters will be chosen as far as is possible to meet two criteria: they should

a) accurately reflect the performance of that Service as perceived and valued by users, and

b) be capable of being measured automatically by suitable monitoring activities by both the centre and the GOC, insofar as this is practicable.

Typical parameters associated with most services will be their Availability and Reliability, or their equivalents. Other parameters will be defined for specific services where this is appropriate. These will in general be as high-level as possible, and as few as is consistent with adequate monitoring. For example, the capacity of a Compute Element as published by the Gatekeeper, the rate of processing work, the file transfer rates.

The Appendix gives an indication of the Grid Services and the nature of the service level parameters associated with each that might be monitored.

3.3 Monitoring Service Performance Levels

The GOC will monitor the actual service levels achieved and publish these alongside the 'contracted' or target service levels for each centre. Both passive and proactive monitoring is carried out on the published service level parameters of all operational services.

3.3.1 Passive Monitoring

This is the action of collecting together all the monitoring information logged by Grid components, for example, job submission and execution times from the Resource Brokers; data volumes transmitted through network interfaces from routers; transmission rates achieved from ftp servers. What precisely is actually available for monitoring is determined by the software developers : the GOC will simply take advantage of whatever is available, although it is expected that feedback to the developers will advise them of any missing operational information. Only information which is relevant to the performance of the Grid infrastructure, ie for monitoring the service level parameters, for determining long-term trends, for capacity planning or for measuring efficiency of operation, is of interest to the GOC. For consistency, the required statistics selected from those available and the format in which they are to be presented will be specified by the GOC.

3.3.2 Proactive Monitoring

Proactive Monitoring is carried out by the GOC itself using tools procured or developed specifically for the purpose. Its primary objective is to detect service failure promptly. Specific probes will be employed to replicate as closely as possible the activities of users, but in a controlled way so that unexpected behaviour is immediately apparent. A range of tools will be employed, most in constant operation. Examples include simple pings to measure the availability of specific hosts which are running critical services, network connectivity and congestion, heartbeat monitors running on all critical components to detect failures on sub-minute timescales. These feed the Alerting Process (see below) and the measurement of service level parameters.

In addition, short-period, directed, job submissions will provide overall tests of critical services, including Resource Brokers, Information Services, Computing Elements and intervening networks, and short jobs accessing data will test the operation of storage and replication facilities, both to detect the effect of increasing load on perceived performance (turnaround), to measure changes in performance following upgrades, to provide measures of performance more closely related to the performance seen by users and as a backstop to detect failures which elude the heartbeat monitors and other monitors of specific services.

When failures or unexpected responses are detected by any part of Active Monitoring an Alert is raised. The Alert will initially be presented to the part of the GOC which is at that time registered as being on duty. If an Alert is not acknowledged within a short time-out by the on-duty staff it will automatically escalate to other parts of the GOC, invoking automatic call-out of senior staff if necessary.

3.3.3 Publishing Operational Performance

Operational information gathered passively from various components in the Grid and proactively by directed monitoring is recorded and maintained in a consistent manner by the GOC. These, together with service failures which have been detected, are analysed periodically and published in regular reports in a manner which allows a direct comparison with the advertised target levels of service published by each centre or service provider.

3.4 First-Level Fault Analysis

The active monitoring process described above is intended to provide a means of detecting service failure as soon as possible after it occurs. It may or may not be immediately obvious which underpinning component is faulty. If it is not, GOC staff will assist in performing the first-level analysis of the fault to determine the underlying reason for failure as precisely as practicable by performing specific, directed, tests to diagnose more precisely the cause of the raised alert. These tests will usually be continued until the fault is localised to a specific, single Local Support Group.

3.5 Interacting with Local Support Groups

As soon as a fault is sufficiently localised, the responsibility for further investigation and rectification is passed to the appropriate Local Support Group, or Regional Centre. This may involve call-out in accordance with an agreed schedule if local staff are not immediately available. In addition, the Call Centres are informed. Monitoring of the failed service continues until normal operation is resumed, with regular reports being passed to the Call Centres.

Local Support Groups may well also need the assistance of the GOC to deal with local problems caused by Grid users or the Grid infrastructure. The advantage of involving the GOC is that they may have seen similar problems at other centres and be familiar with both the issue and the remedy.

3.6 Change Control

The GOC will provide a Change Control service. This is intended to ensure that major changes to the Grid are made in a coordinated fashion, to ensure that all interested parties are informed of prospective changes and have an opportunity to comment, to ensure that prospective changes have been adequately tested before entering production, and that a means of backing them out is available in the event that problems ensue.

The parties involved in change control are the Local Support Groups or the Regional Centres (usually the implementers of changes), the Grid Deployment Group at CERN (usually the initiators of changes), the GOC itself and the Call Centres (representing the users who are the intended beneficiaries or occasionally the victims of changes). The users themselves are not directly involved in change control.

Change Control applies to all changes which may result, deliberately or inadvertently, in changes to the normal operational regime as perceived by the user community, These include software upgrades, changes in the operational status of sites, scheduled downtime for major hardware changes, etc.

3.7 Coordinating Security Activities

By its nature the Grid provides close and intimate coupling between the LANs of many institutes. This presents unique security problems which as yet have not been fully resolved. Although such considerations are the prerogative of the Security Officers at the respective sites, the GOC could facilitate the resolution of such issues by playing a coordinating role, prompting discussion, organising meetings, etc.

It is essential that a Grid Incident Response Team (IRT), capable of responding to security incidents within the Grid environment with specialist advice to Local Support Groups, is established well before the Grid enters its full operational state in 2005. At present it is not clear how this is to be organised. One sensible and feasible model is that such a team is embedded within the GOC, and this is assumed in this proposal. Much of the information which is essential to an IRT (levels of deployed software, local contacts, network connectivities, firewall configurations, etc) will already be available within the GOC as part of their operational activities. Also, the area over which the IRT is required to operate is identical to that of the GOC, and many of the required skills are similar. It is therefore proposed that the GOC acts as a contact point for security incidents, operates an IRT for Grid-wide operations, and acts as a clearing house for incident resolution and eventual re-establishment of normal operations.

3.8 Operations Development

The tools required to carry out many of these activities will need active development, in some cases to create them from scratch, and in all cases to develop and improve them for use within the evolving Grid architecture. It will be appropriate for some of this development to be carried out by GOC staff, but the more esoteric features will need to be developed by appropriate technical staff within the Grid Deployment

Team or within local support teams. For example, heartbeat monitors are specific to individual processes within particular computer centres: these would need to be provided, perhaps to a specification issued by the GOC, by the systems staff local to each centre. At the other extreme, a system to submit simple jobs according to a schedule will be developed and maintained entirely by GOC staff. In general, operations tools which are generic or which act on the Grid as a whole will be the responsibility of GOC staff.

4. PHASING OF GOC FUNCTIONS

We envisage three Phases in which the full functions of a GOC are gradually established. These Phases will proceed iff approval is obtained from GDB for this proposal.

4.1 Phase 1 (Jul 03 - Oct 03)

a) Set up an initial monitoring centre by end Jul 03 using monitoring tools available for immediate deployment.

b) Develop Grid operations security policy in consultation with security officers.

c) Define the service level parameters which must be published and monitored for each of the critical Grid services.

d) Develop draft reporting formats and establish a monitoring regime for determining and presenting service level information.

e) Evaluate and select the tools which will be deployed in Phase 2. These will be capable of detecting fault conditions and raising alerts and may require the installation of local monitors.

4.2 Phase 2 (Nov 03 - May 04)

a) Select and establish a second monitoring centre; establish GOC coordination meetings.

b) Establish a Grid operations security coordination regime in consultation with security officers and local support groups.

c) Establish a simple change control regime.

d) Obtain, install and configure tools for detecting deviations from published service levels and presenting alerts. Set up automatic call-out mechanisms triggered by alerts.

e) In Feb 04 assess experience and write a proposal for the production GOC.

4.3 Phase 3 (Jun 04 - Jun 05)

a) Select and install the tools for first-level fault diagnosis and establish first-level fault analysis operating regime; establish fault tracking procedures.

b) Establish a Grid-specific IRT.

c) Select all the tools for the production GOC, deploy them in two operations centres and assess their suitability in production use.

d) Select and establish third monitoring centre and establish 24/7 working pattern.

5. PHASING OF PROCESSES AND ACTIVITIES

This section says when the various processes and activities described in Section 3 will become operational.

5.1 Coordinating Grid Operations

As there is only one monitoring centre during Phase 1, coordination within the centre will be internal and informal. During this time links will be established to contacts in all Local Support Groups.

A second operations centre will be established early in Phase 2 around Nov 03. During Nov and Dec 03 monitoring software which requires monitors to be installed in local centres will be deployed. To coordinate these activities a series of coordinating discussion groups and meetings involving GOC staff and Local Support Groups will be established during Nov 03, and continued thereafter to ensure the roll-out of operational facilities continues to be agreed by and coordinated across all centres.

Later, the same operations coordinating meetings will jointly develop the operating regimes for change control (Phase 2), alerting (Phase 2) and first-level fault diagnosis (Phase 3).

Towards the end of Phase 3 a third operations centre will be established and a 24/7 working pattern established.

5.2 Defining Service Level Parameters

During Phase 1, the service level parameters associated with the Resource Brokers, Job Submission Services and Information Services will be defined. This will require an extensive dialogue with Local Support Groups as part of the Grid coordination activity. An evaluation will be conducted to determine the way in which target values should be published by centres.

During Phase 2 a complete list of services and their service level parameters will be defined. In addition to the three services above this list will include the Data Replication Services, the Computing Element Services (Gatekeepers), the Storage Element Services, the Logging and Bookkeeping Services and the Grid Security Infrastructure.

5.3 Monitoring Service Performance Levels

5.3.1 Passive Monitoring

During Phase 1 the statistical information which could be made routinely available from each service and centre will be determined, and an automated means of

collating and presenting this information will be developed. Each centre will be required to publish the information they collect in a standard form. Initially the required information will consist of just the number of jobs run per day by the centre and an indication of their owner. This will provide the raw data for job accounting, and prompt the investigation of the questions of job owner, normalising factor, etc. Later in Phase 1, the number of grid-based bytes per day transmitted by the centre will be collected as an example of a network statistic. This minimal requirement will enable the passive monitoring mechanism to be established. A simple web-based reporting mechanism will be developed to present this information to all.

In Phase 2 the information gathered and presented will be extended in ways which will be determined by what is practical and what the community requires.

The most promising packages for collecting and presenting performance data are Ganglia, MonaLisa and R-GMA. Ganglia is already well-established, easy to install and will be implemented during Phase 1 at one or two centres and extended to most centres during Phase 2. MonaLisa and R-GMA are more sophisticated and flexible systems under development at CalTech and by WP3 of EDG respectively. They are not yet fully mature products. They will be installed and evaluated during Phase 2 and compared for operations use with Ganglia and other products.

5.3.2 Proactive Monitoring

During Phase 1 only existing and well-established products which do not require local sensing agents will be deployed. These include

- the GridPP Monitoring Map, which monitors the ability to submit a job to any centre via Globus, GridPP Resource Broker and the EDG RB at CERN and encodes the results on a map of the grid and provides minimal but useful job submission information to the end-user, and

- the MapCenter service developed by EDG WP7, which periodically tests communication with individual hosts at centres using icmp, a tcp port scan of selected ports and http Gets, and presents the results in a series of tables.

During Phase 2 a means of monitoring the defined service level parameters against the published targets and raising alerts when targets are likely to be in jeopardy is required. Key status indicators which will be useful in assessing actual performance include whether a CE is up or down, whether jobs are running, whether services are up, etc, but much of this activity in Phase 2 will be directed to determining exactly what are the critical services to monitor, how that is best achieved, how status information is best presented and when and how alerts should be raised. Products which provide this functionality include Nagios and possibly FMon from EDG WP4, although this latter currently has no alerting functionality. The most suitable appears to be GridICE from INFN, which integrates a number of these products around a Nagios core. All of these require sensing agents local to the services being monitored.

Products are readily available for automatically calling out off-duty staff via bleeper, mobile phone, etc whenever an alert is raised. One of these will be selected and installed during Phase 2.

5.3.3 Publishing Operational Performance

During Phase 1, target and actual performance levels will be published for a small number of high-level parameters relating to processing capacity, and the availability and reliability of job submission and the processing services of the operational computing centres.

During Phase 2 this will be extended gradually to other critical service components as the appropriate metrics are agreed and defined.

By the end of Phase 3 all the definitions will be completed and all target and actual service level parameters will be being published.

5.4 First-Level Fault Analysis

This activity begins in Phase 3 and will be targeted at identifying which component in the Grid is causing a general (rather than a specific) service failure and which support group carries the responsibility for rectifying it. It is likely that many of the high level faults will be reported by users via the Call Centre, but the main goal of active monitoring will be to detect such failures before the users do. Faults detected in this way will be tracked by the GOC until rectified, and a simple fault-tracking tool will be required for this purpose.

By the autumn of 2004 most of the monitoring information required for fault detection and diagnosis will be available via the selected tools. Additional tools will then be selected and deployed specifically to assist in the remote diagnosis of high-level faults. A wide variety of tools for first-level fault diagnosis exist and many of them will be useful. They range from simple and well-established network tools such as tracert to special-purpose tools developed for specific purposes. Initially, a tool set based on common existing practice will be assembled and extended as experience grows.

5.5 Interacting with Local Support Groups

The alerting regime will be established late in Phase 2 by the operations coordinating meetings, and used from early in Phase 3 by the GOC to ensure Local Support Groups are aware of conditions which are adversely affecting grid services as these are discovered by the diagnosis tools being deployed at that time. Call-out procedures and practices will be established to ensure trivial and serious faults are distinguished and support staff notified the next working day or called out as appropriate.

Similarly, an interface with Local Support Groups will be established during Phase 2 to assist Local Support Groups to deal with local problems caused by the Grid.

5.6 Change Control

During these evolutionary phases the main activity of change control will be to establish the necessary contacts with both developers and local systems staff, to establish procedures and mechanisms for collecting information about imminent and actual changes and to ensure that this is disseminated to all interested parties. No actual control will be exerted at this time.

A simple change information system will be established from Dec 03 during Phase 2. Its goal will be to ensure sufficient coordination is applied to prevent operational breaks or errors arising from incompatibilities, insufficient notice about changes or insufficient information about changes. Its operating regime will be devised and agreed by the operations coordination meetings (see 5.1), and will therefore have the support of the Local Support Groups of every centre.

An assessment will be made during Phase 3 of the type of change control which will be needed for the final operational GOC - in particular to what extent changes need to be controlled and coordinated. If it is determined that actual control is required then the tool to do this should be selected and purchased from those available off the shelf and deployed during the final stages of Phase 3.

5.7 Coordinating Security Activities

General security precautions taken by centres will continue at all times to be determined and implemented by the site security officers. The GOC will provide any advice or assistance required by those security officers but it is not expected that the GOC will be significantly involved in developing or implementing those general site-specific precautions. This is a matter for the local grid administrators and developers.

The GOC's interest in security is principally in setting out the Security Policy and Security Procedures for grid operations, which will be directed at containing any compromise to as small a number of grid hosts as possible, and in ensuring compromised grid hosts are both detected and restored to service rapidly. The Policy and Procedures will be developed in Phase 1 to provide a sound and agreed basis for the remaining activities.

During Phase 2 appropriate coordinating structures will be established with security officers. This will build on the security work initiated by the LCG Security Group, and carry it into an operational state.

During Phase 3 the Security Procedures will be developed further and supplemented by Incident Response Kits which may be used to rapidly determine the nature and extent of a suspected compromise. Expertise in recovering grid-specific software following a compromise will be developed and maintained, and be available to assist site-based administrators to investigate and recover from incidents.

At all stages the local site Security Officers will be kept informed and contact with them maintained.

5.8 Operations Development

No provision is made at this stage during the GOC establishment phases for extensive Operations Development. During these evolutionary phases only tools developed elsewhere will be selected and deployed. Any deficiencies in these tools that become apparent, and these are especially likely to be uncovered during the final deployment for the operational GOC in Phase 3, will be fed back to the appropriate development groups.

Some tools such as Nagios and its derivative, GridICE, require specific plug-ins and add-ons to monitor specific parameters, and if such tools are selected it may be that such additional software will need to be developed. If so, such development work should be carried out by appropriate development teams triggered by reports from the GOC.

What will be required within the GOC is a relatively straightforward development capability for operational aids. This will be at the application level in general - configuring tools, manipulating them and using them. This level of expertise will be expected of all GOC staff as part of their normal skill set.

6. ORGANISATION

Many of the activities required to run the GOS can be automated, and full advantage should be taken of this. It is not practical for staff to monitor such a large number of services by personal inspection, so all the active monitoring, as much as possible of the passive monitoring, and much of change control should be implemented to operate autonomously with the minimum of human involvement, with operations staff generally responding to exceptions signalled by raised alerts.

Other activities will need direct human attention. For example, GOC staff will be needed to deal with

a) coordinating operational activities,

b) some aspects of change control,

c) first-line fault analysis, alerting local operations staff to faults, responding to security incidents,

d) the analysis of monitoring statistics, the writing of the interpretive part of reports, and any follow-up actions associated with poor performance, and

e) the specification, some development and the deployment of operations tools.

Only (c) requires a commitment to work outside normal office hours; the rest may be carried out as part of a normal work-day.

The structure we recommend is a Regional GOC in the USA, a Regional GOC in Europe and a Regional GOC in Korea or Japan with roughly equal staffing, each with primary responsibility for the geographically local part of the grid as far as change control and the service commitments are concerned. These three Regional GOCs comprise the GOC. The responsibilities for coordination, reporting and tool development would fall to the GOC as a whole and be shared across the three Regional GOCs. Each of the three Regional GOCs would be on duty for fault detection, fault diagnosis and incident assistance for a ten-hour shift during the local daytime, which with a one-hour overlap at each end would provide 24/7 cover world-wide.

7. RELATION TO THE PROPOSED LCG CALL CENTRE

The proposed Grid Operations Centre and the proposed LCG Call Centre, which will provide direct support to end users, share a number of common characteristics: they both need to operate continuously to provide full 24/7 coverage around the world, they both need a comprehensive awareness of the extent of the Grid, a knowledge of the operational state of its current services, and a means of receiving and tracking faults. This suggests there is merit in considering whether these two functions should be merged into one.

However, there is a clear difference in the focus of the two activities. The GOC is concerned primarily with the availability and reliability of grid services, with security procedures and incidents, with matters of operational policy and practice. They interact mainly with Local Support Groups at centres. The LCG Call Centre (LCC) is concerned primarily with serving the users, addressing their concerns, ensuring they are kept informed, receiving and investigating faults and complaints. The GOC looks inward: the LCC looks outward.

These differences in focus and interest argue for keeping the organisations separate, operating under separate management structures, while the common activities suggest benefits could be gained by sharing information and to some extent facilities.

A suitable organisational structure which recognises these differences yet permits advantage to be taken of the similarities might be one in which the GOC and LCC were two arms of a single body, working closely together and housed within the same premises at each of the three locations.

8. STAFFING

The figures presented in this section are no more than an rough indication of the likely levels of staffing needed to carry out the work as described herein. There is considerable uncertainty in what will actually be required and the final figures may well deviate considerably from those presented here.

8.1 Prototype GOC Staffing

The program of an evolving GOS as outlined above is estimated to require the following staff numbers over the period Jul 03 to Jun 05. This covers the development work involved with setting up the various procedures, tools and processes, and for evaluating their efficacy. The effort required to carry out the actual operations tasks themselves is covered under the GOC operational staff below.

Quarter	Phase(s)	Staff (FTEs)
2003Q3	Phase 1	2
2003Q4	Phase 1 → Phase 2	3
2004Q1	Phase 2	4
2004Q2	Phase 2	4
2004Q3	Phase 3	3
2004Q4	Phase 3	3
2005Q1	Phase 3	2
2005Q2	Phase 3	2

This represents a total of c. 6 staff-years.

8.2 GOC Operational Staff

From mid-2005, when the production GOC starts operating, there should be round-the-clock cover. The way it is proposed to achieve this has been covered under the Section on Organisation above.

The staffing level for the three main GOC sites is determined by the activities under (a), (b), (d) and (e) in Section 6 above, ie all except (c). A minimum of 2 staff in total is required to ensure business continuity, 3-4 is probably optimum. It is hard to see a requirement for more than 5. The final figure will be determined after the experience gained during the operation of the prototype GOCs, particularly during Phase 3, but for planning purposes a figure of 3 staff in total, one at each Regional GOC, is suggested to cover these four activities. These would also act as the three Regional GOC Operations Managers.

Responding to Alerts and the Analysis of first-line faults will be the most demanding activities, even after the Grid has reached operational maturity, yet estimating the rate at which services might fail and the GOC effort which might be involved in detecting and initiating action to rectify them with any accuracy is not possible at this stage. The work to provide a better estimate of the rate of failures which require immediate attention will be carried out during Phase 2.

The recommended geographical distribution of the Regional GOCs ensures staff working normal office hours are available to give full 24/7 cover. Assuming one such person on duty would be adequate at any one time to respond to alerts and incidents world-wide would give a requirement, with shift overlap, for 70 hours/week at each of the three Regional GOCs. So to cover the working week and the two week-end shifts would require each Regional GOC to employ 2 technical staff engaged in monitoring, fault diagnosis and security duties, with these 2 staff each working one week-end in two on average. An alternative would be to employ 3 such staff at each Regional GOC, and this would then provide two people on duty during the working week.

In total, then, the estimated staffing requirement is for 3 (or 4 if double week-day cover is required) full-time staff in each Regional GOC, or around 9 (or 12) in total. Which level is most appropriate will be determined by the experience gained in Phases 1 and 2. Whichever is chosen, this position should be reached by June 2005. It is assumed that the Operations Manager and one of the technical staff of the first Regional GOC would be appointed by gradual redeployment from the staffing of the prototype GOC above. The second Regional GOC should be staffed with its Operations Manager and one technical assistant from 2003 Q3. In April 2005 these two should be brought up to full strength and the third Regional GOC should be set up and staffed to allow time for familiarisation and training in preparation for full production 24/7 operation from June 2005.

9. OPEN ISSUES

There are a number of issues which are both significant and relevant to a Grid Operations Service which are at present unclear, unresolved or even unconsidered. In this situation many aspects of the GOS functions cannot be defined or predicted at this time. In effect these issues comprise a research programme for the evolutionary stages of the GOS, to be carried out either by GOC staff or some other grouping.

In this document we simply list these issues and make some observations.

9.1 Service Level Definitions

Setting out the most appropriate performance metrics for each of the critical services is the key issue for the GOS, and this will form a significant part of its early work. In some respects this is similar to the work carried out by operations groups for decades, but there are a number of novel and complicating aspects. First is the need to monitor services at the connectivity, resource and collective layers rather than components at the fabric layer; second is the need to identify the small number of metrics which are service-related, which reflect aspects of performance important to users and which are susceptible to automatic monitoring; and third is the lack of any overall authority, which means progress has to be made by argument, consensus and agreement rather than prescription.

9.2 Accounting

It is not clear to what extent LCG wishes to account at the Grid level for the jobs running on LCG resources. By this we mean keeping a chronological record of the resources used by the various users or VOs across the Grid so that the aggregated resources used by the VOs may be compared one with another and with the agreements for resource provision made between VOs and computing centres.

A comprehensive accounting model has been developed by DataGrid, but it is unclear in the absence of discussion or guidance whether this might be appropriate to LCG. This model is based on the assumption that Grid resources are both provided by and consumed by the same set of VOs, and accounting is then a matter of debiting the account of the consuming VO and crediting the account of the providing VO whenever a job executes. A scheduling algorithm can then be devised to ensure that each VO stays roughly in balance over some period of time. This has a number of attractive features. It encourages centres to release resources to the community and offer as attractive a service as possible whenever they are underutilised in order to build up credit for their users, so encouraging high standards of service and performance, while, during times of pressure, through appropriate job scheduling algorithms, it ensures that VOs tend to return to balance.

This model is not immediately applicable to LCG since there is not, in general, a direct association of users and resources within single VOs; indeed, VOs as they are presently constituted contain only users, and some centres such as CERN are set up explicitly to serve all users. However, by generalising the model it is possible to devise a scheme which could be used by LCG, and which retains the desirable

features of load-balancing and providing incentives to computing centres to provide high quality services..

In this circumstance we simply raise the issue and seek guidance from GDB, but this could be a significant research area for the GOC if accounting is a serious requirement.

9.3 Incident Response Procedures

Experience with intrusions over recent years has lead to the development of sophisticated incident response procedures and kits which enable intrusions of single systems to be rapidly detected and rectified, while preventing the continuing propagation of the intrusion to other systems. However, these apply only to single centres. The Grid establishes a close coupling between the LANs and a large number of systems at many tens of centres, and no existing security procedure has been designed to be effective under these conditions.

Devising incident response kits and procedures specifically to safeguard the Grid is a major development area. Development topics, not necessarily for GOC staff, would include prompt detection of unauthorised modification of Grid software, and mechanisms for rapidly disconnecting systems which are suspected of being compromised.

9.4 Change Control

Consideration of the issue of change control begs the question of who has authority over the operation of the Grid. On the one hand the Grid is a loose federation of free agents who autonomously provide resources to the Grid (and autonomously remove them), and on the other hand is a resilient and reliable service for processing LHC data and simulations. The GOS has to reconcile these two aspects, and change control, particularly the 'control' aspect, is at the heart of the dilemma.

Traditionally, change control is applied to a critical service to minimise the likelihood that a change will cause an unexpected deterioration in that service. It achieves this by requiring an agent of change to satisfy some controlling body that all interested parties are aware of the change and its expected effect, that the change has been adequately tested and found satisfactory, and that a means of backing out the change has been identified and also tested.

The federal nature of the Grid as it is presently constituted does not easily admit the possibility of controlling change over the entire Grid in this way, and it falls to individual centres to implement change control locally. This, of course, may be entirely satisfactory, but at present that would need to be taken on trust. The question is, to what extent should the GOC be involved in the controlling aspect of change control (as opposed to the informing aspect, which is not contentious)? The view of GDB is sought.

APPENDIX

1. OPERATIONAL TERMS

Availability of a Service is defined to be the proportion of time that Service is effectively operational to the time it was scheduled to be operational expressed as a percentage, calculated over some agreed period of operation. A typical period of operation might be 4 weeks, a calendar month, a quarter or a year.

Reliability of a Service is expressed as the reciprocal of the rate of Failures of that Service, where a Failure is one which results in a break in that Service of sufficient magnitude and duration to be noticeable to users (to be defined for each service), calculated over some agreed period of operation, typically a quarter or year. It is closely related to mean time to failure - MTTF.

Performance of a Service is defined to be the rate of carrying out one or several typical or critical functions. Specific to each individual service. Chosen to indicate variation of overall performance and its dependence on loading.

2. COMMON GRID SERVICES

We list the primary Grid services which will be required for LCG as defined in EDG Release 2.0, and indicate some possible service level parameters which could be defined and monitored.

UI (User Interface) Enables preparation of JDL and submission of job to a RB. Not susceptible to monitoring by the GOC as it runs on individual client systems. The GOC will operate a number of client systems running the UI to enable tests to be carried out using specific clients.

RB (Resource Broker) Processes requests for resources, allocates resources, manages active jobs. To be provided by EDG-RB. Probably the most critical and difficult service. GOC will define minimum acceptable Availability, Reliability and Performance (rate of processing jobs (from JDL receipt to job submission)). Test each operational instance by periodic job submission and inspection of job status information.

IS (Information Services) The basic resource discovery mechanism. To be provided by either Globus MDS or EDG R-GMA - R-GMA assumed. GOC will define minimum acceptable Availability, Reliability and Performance (defined by rate of retrieving specific information). Test by publishing and retrieving specific information periodically through each operational instance.

JSS (Job Submission Services) Takes JDL from RB and submits it to the selected Gatekeeper. To be provided by Condor-G. GOC will define minimum acceptable Availability, Reliability. Performance is not likely to be an issue, but need to detect hangs. Test by directing test jobs through operational JSSs.

CE (Computing Element) The main service of the Grid. Interface provided by the Gatekeeper. GOC will define minimum acceptable Availability and Reliability of Gatekeeper, and Performance of entire CE (defined as ratio of available processing capacity to published processing capacity). Test by directed job submission to fast queue.

SE (Storage Element) Provides access to mass storage facilities, which must provide cost of retrieving specific items of data and be able to transfer that data on request to specific CEs. GOC will define minimum acceptable Availability, Reliability and Performance (defined by rate of retrieving specific information to the most local CE). Test by directed job submission to most local CE to read/write specific data item.

RC (Replica Catalog) Includes Replica Manager, Replication Metadata Catalog, Replica Location Service, Replica Optimisation Service. The most complex of the services to define and monitor, and how this is to be done requires further study.

LB (Logging and Bookkeeping) Needs to accept and log events from UI, RB, JSS, and CE and present status to UI. GOC will define minimum acceptable Availability and Reliability. Performance is not expected to be an issue. Test by observing status of jobs submitted in other tests.

GSI (Grid Security Infrastructure) The authentication mechanism, based on public key encryption, X.509 certificates, and SSL, plus extensions for single sign-on and delegation. Implemented by the Globus. Main services are Certificate Management and VOMS. How the performance of these is defined and monitored requires further study.