GRID OPERATIONS: EVOLUTION OF OPERATIONAL MODEL OVER THE FIRST YEAR

H. Cordier, G. Mathieu, F. Schaer, IN2P3, Lyon, France
M. Schulz, P. Nyczyk, J. Novak, CERN, Geneva, Switzerland
M.H. Tsai, ASGC, Taipei, Taiwan

Abstract

The paper reports on the evolution of operational model set up in the **Enabling Grids for E-sciencE (EGEE)** project, and on the implications of Grid Operations in LHC Computing Grid (LCG).

The primary tasks of Grid Operations cover monitoring of resources and services, notification of failures to the relevant contacts and problem tracking through a ticketing system. Moreover, an escalation procedure is enforced to urge the responsible bodies to address and solve the problems. An extensive amount of knowledge has been collected, documented and published in a way which facilitates a rapid resolution to the common problems.

The number of sites in production quickly expanded from 60 to 170 in less than a year. At the same time, the operations model evolved from one single person at CERN to a distributed model involving more and more geographically scattered teams. The evolution of both procedures and workflow requires steady refinement of the associated tools as ticketing system, knowledge database and integration platform. Since EGEE/LCG production infrastructure relies on the availability of robust operations mechanisms, it is essential to gradually improve the operational procedures and to track the progress of the tools' on-going development.

CIC-ON-DUTY: LOAD SHARING ON INFRASTRUCTURE MANAGEMENT

EGEE, along with its sister project LCG, manages with over 180 sites the world's largest Grid infrastructure. When EGEE was first launched, it was managed centrally from the Operations Centre at CERN. While this worked quite well, it also had disadvantages.

EGEE came up with a scheme where the Core Infrastructure Centres shared the load. Dubbed CIC-on-duty (COD), this new system began in October 2004. In this system, responsibility for managing the infrastructure passes around the globe on a weekly basis.

Already the CIC-on-duty scheme is proving worthwhile. Since the initial deployment of COD, the percentage of "good" sites has increased dramatically. At the same time the number of sites has multiplied and criterion for "good" sites has become more stringent, making the COD accomplishment even more difficult.

The approach taken in this scheme reflects the collaborative nature of the EGEE project in general, and the successes that such an approach can provide [13].

BEGINNING OF GRID OPERATIONS : NEEDS AND MOTIVATIONS

At the very beginning a single person was responsible for approximately sixty sites. The work was labor intensive involving the use of several existing tools and internally developed tests such as the TestZone Tests which later became the *Site Functional Tests* (**SFT**) [6].

Routine operations tasks would start with the operator proactively searching for faults using available test suites. The incident would then be recorded in the *Savannah* [2] system to assist in problem tracking. This is followed by searching in the *Grid Operations Centre Database* (**GOCDB2**) [12] which holds site metadata for contact information of the affected site. Finally the operator would use an email client to send out a notification for the problem detected.

At first the site adminisrator was contacted and the operator would provide full support and expertise for problem resolution to a given site. No clear procedures nor recommendations existed at the time.

As the number of sites increased rapidly, the amount of time spent with administrative tasks increased proportionally. This meant that less time was available for operators to provide technical diagnosis and troubleshooting support afforded to sites before.

The first draft of the *Operations Procedure Manual* [10] was written by the person monitoring the sites when the workload increased to a point where the job had to be distributed. This person had then to design a "cookbook" to keep track of his experience and to train a small team at CERN. At that moment, the need for an initial *escalation procedure* as unofficial documentation arose and mid 2004 the first draft of this cookbook was widely advertised.

The operations procedure has evolved into a document that describes such characteristics from a site as entering/exiting Grid production infrastructure, scheduling downtimes, and their status down to a node level. Moreover the escalation procedure validated by ROCs and applied by CODs is now well established and has been crucial for the Grid stablility and availibility (Figure 1).

Indeed, the whole process of contacting sites and helping them to solve their problems turned to be out so time consuming that the scope of the job of a operator would gradually change from then on.

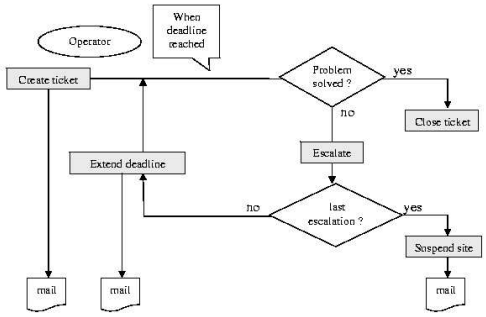

Figure 1: Escalation procedure

To further distribute the operations duties, four institutions CNAF (IT), RAL (UK), IN2P3 (FR) joined the small team at CERN in running the COD. Then Russia and Taiwan entered the new geographically distributed structure in 2005. Now six teams are part of the rotating shifts.

The set of tools has evolved into an integrated *dashboard*, which enables a given operator to have an overall view of site failures at sites, their respective severity as well as the status and the history of the corrective actions. Instructions and recepies for the occuring problems were collected on the GOC wiki pages [8], producing a common knowledge-base.

Over a year later, the main mandate for operators in the COD team now is to check the monitoring tests, detect failures at sites and diagnose the problem as precisely as possible. They contact the site and its ROC to ask them to solve the problem. As time passed operations experience has accumulated at ROCs and site, so the operators' expertise is called upon only when the ROCs and sites cannot solve the problem by themselves.

The COD helps maintain a high level of Grid infrastructure availability by assigning tickets and sending email to relevant parties. Afterwards, the COD provide follow-up support until the resolution of the problem according to the escalation procedure. The COD teams operate COD shifts and at the same time, they drive the Grid expertise in operations as most of them are involved in the developement of new monitoring tools.

De facto, the COD activity has made their actual mandate expand in scope and span.

TOOLS USED IN OPERATIONS

Monitoring Tools (SFT-Gstat)

Several tools are used in order to detect problems at sites. There is a need to separate tools that comes from the Grid architecture : some malfunctions can be detected from any (Grid) node while other problems can only be detected while a job is running at a site.

Gstat [7] is a *Grid Information System monitoring* application. Its primary goal is to detect faults, verify the validity and display useful data from the Information System. Gstat can be decomposed into modular agents and filters. Agents are responsible for making queries and collecting raw data for further analysis by filter components. Filters then create the processed data which is then used to generate a web based user interface.

SFT [6] stands for *Sites Functional Tests*. Its aim is to provide information on potential site-specific failures. To do so, a Grid job is sent every 4 hours on every site, each job containing a set of tests that make sure that all main midleware features are working on a given site/node. Each individual test output and results is then published and, given a set of requirements, the site is displayed as healthy or not on a web page, which is then used by CIC on duty people and site admins to fix problems.

Ticketing System

Savannah as an instance at CERN ([1], [2]) was used in the beginning as the problem tracking tool for operational issues. Then, the initial need of tracking problems was completed by the need of assigning them to specific Support Units like regional support. Such support units were already in place in the EGEE/LCG user support system, the *Global Grid User Support* (**GGUS**) [3], [4].

It was agreed that Operations support would take benefit to use the same ticketing system as User Support, since the people having to solve the problems are often the same. The decision to use GGUS for Grid operations was taken in January 2005.

For each problem detected on a Grid site, a ticket is created and assigned to the *Responsible Unit* of the EGEE Federation the site belongs to. Problem date, affected node, problem description and link to useful documentation are provided in the ticket. A mail is also sent to the Responsible Unit and the affected site's administrators. The deadline depending on the severity of the problem is set between one and three days, depending on the chosen priority.

Although they have to be solved by Regional Responsible Units, tickets are supervised by the operators on duty. If the deadline is reached and the problem remains, ticket is escalated following a well defined procedure (see below).

Currently, each week around thirty tickets are created, and operators handle another seventy (modification, escalation, closing).

Administration Tools - GOCDB2 and ROC Weekly Reports

In order to discover which Grid services should be tested, monitoring applications reference a central site information respository. This database called **GOCDB2** [12]

is developed and maintained by Rutherford Appleton Laboratory (RAL).

This database contains a site ID-card with registration data (name, location, contact information, administrator contact, security contact, etc.) and site status (candidate, uncertified, production, suspended, etc.). Keeping this information up to date is a shared responsibility between the site and the ROC. This repository is used by the Grid monitoring services.

The COD only monitors sites that are in ”production” status. It is worthwile to mention that local instances of Sites Functional Tests may be used to ”certify” the sites that are candidates to the production Grid infrastructure.Indeed, Site functional Tests on demand (a.k.a SFT2) has been used now for several months by several federations for domestic monitoring and certification purposes.

Based on these statistics and on the site reports dubbed ”ROC weekly reports” that all federations validate through weekly operations meetings, it appears that the visibility and the reactivity of sites and federations have increased drastically.

Integration

The tools described above are developped by different teams and hosted by different institutions. This situation is advantageous for sharing workload and responsibilities. However, distributed development also makes high level views difficult to realize.

The main difficulty is to cross reference to make a clear diagnosis: for a given site where a problem has been detected, we need to know for example what are the results according to some other monitoring tools, if there is a ticket already reporting this problem, whom is to be contacted, or if this problem had already occurred.

Going through this process using all the available tools one by one is really time demanding, and sometimes not efficient enough. Therefore, operators need a synthetic entry point with synoptic summaries as well as quick access to all the results of these tools, so that they can work from the begining to the end in the same layer.

To answer this need, we set up a specific section in the CIC Portal [5]: the **CIC-on-duty Dashboard**. The first version was put in production in november 2004, and was a simple webpage grouping links to the used tools.

Since then, a lot of improvements have been made, and the dashboard switched from a static links page to a fully dynamic integration tool. It is now designed as a set of webpages giving access to all the tools the operators need to perform their daily work: site information, monitoring results and ticketing system (Figure 2). A list of sites having problems is built from SFT and Gstat results. To associate the problems with corresponding tickets, each problem is cross reference with all open GGUS tickets. In addition a mailing tool with email templates was developed to streamline the notification process.

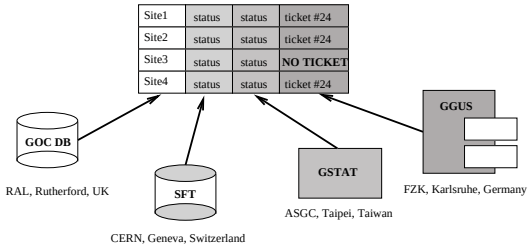

Figure 2: Pillars of the daily operations

Data from GOCDB2, Gstat and SFT are retrieved by MySQL queries, parsing flat files and querying R-GMA respectively. Interaction with the GGUS system is done by web services as described in [4].

Building an integrated tool eventually remains a great challenge. Main issues are the distributed data sources, the heterougenious data formats and that they are accessed through different technologies.

At first, cron mechanisms were used to get the data, then dedicated code to parse and display them. The CIC-on-duty dashboard is now based on the **Lavoisier** service [9]: all different data sources are presented as Lavoisier ”Data Views”. This has enabled us to improve the efficiency of the tool and to lower and to lower overall development time of new features.

DOCUMENTATION AND TRAINING

The operations procedure is designed to describe the daily work of the COD people. The mandate of a COD people is to detect, diagnose a specific problem at a site hosting resources and services that are crucial for the Grid infrastructure availability.

When one of the specific tests is failing, a given COD is then to open a ticket assigned to the sites’s ROC as the ROC is the responsible entity within the EGEE/LCG project. At this stage, the COD follow up the life-cycle of a given problem. If the failure is not attended at on time, and if the ROC does not provide some relevant info, the current COD is responsible in ”escalating” the ticket. The steps of this escalation procedure are precisely described within the COD Operational Manual [10] until the ticket can be closed.

The operations procedure is undergoing constant evolution, taking benefits from the experience of over a year of activity and from the feedback of the new teams joining the daily operations. All specific tips needed to operate but dependant on specific middleware for example were taken out of the operations procedure to a wiki [11] for the teams to update. This is more scalable now that the teams have gone from one to six in less than a year.

This wiki is designed to keep track of some COD topics: development and improvement of the monitoring tests and integration of the tools. Work Topics such as tests framework, tests development, integration, metrics and failover

procedure are discussed at each face-to-face quaterly meetings and topics' leaders keep the debate focused in some smaller parallel sessions.

Moreover, the procedure describes a number of processes including the description of the steps a new team should go through in order to officially enter the planning of the CODs. Once the new team feels comfortable with the COD tools, they volunteer to replace once a regular team as the lead team. Of course, the backup team would be a very experienced one. It is advised that a new team assists duty twice before entering officially the planning.

The flexibility of these shifts planned by pairs since early 2006 enable the CODs to cope with national holidays, trainings of new teams and work overload.

ASSESSMENT OF COD ACTIVITY AND INDUCED BENEFITS

The COD activity can be assessed by the number of tickets handled per week by a given COD, eventhough it is not an absolute value. Indeed, the number of SFTs has increased over the months and the tests themselves have become stricter. The results of the tests are categorizied as critical and non critical and problems at sites are handled by problem severity and by size of site in CPU number.

VOs wish lists are taken into account by the implementation of site white listing. Indeed, VOs can point out their own SFT set as well as their custom tests suite according to the Freedom of Choice for Resources tool [15]. This will enable operations to be VO-oriented.

Operations and ticketing system interfacing, COD teams have assigned more than one thousand five hundred tickets to federations. Federations act as first support unit to sites and are the entity responsible for closing the tickets in GGUS. Their response time increased over the months and the life time of a given ticket has decreased drastically since mid 2005.

Statistics over time for site availability and analysis of "ROC Weekly Report" completed by sites and federations help reveal useful trends on the health of the Grid infrastructure and also provide feedback for further improvements.

EVOLUTION OF THE YEAR TO COME

COD activity has been crucial in stabilizing sites [13]. The structure is going to have an extended mandate based on a applications-based approach. Indeed, as some sites'sanity is crucial to an operational Grid infrastructure, some others' is crucial to operational environment.

Also, some specific services are to be closely monitored to meet LHC experiments deadlines. Consequently, CODs are working on the monitoring of sets of core services relevant to Grid infrastructure and to each Virtual Organization. These requirements have lead to the development of the **SAME** framework: *Service Availability Monitoring Environment* [14].

The SAME project is designed to coordinate individual sensors that monitor Grid core services. This project defines how sensor results will be published and stored. SAME also includes plans for the development of analysis tools that leverages the published data to produce metric and alarms to better understand quality of Grid services and improve the responsiveness to faults.

Eventually, with ten federations involved in the shifts in the coming months is another challenge to be met by the early stages of EGEE-II. We will then have to work on getting an extendend daily coverage for the monitoring of the Grid infrastructure using time zones.

REFERENCES

[1] **Savannah Ticketing System**, http://savannah.gnu.org/

[2] **CERN Savannah**, http://savannah.cern.ch/

[3] **GGUS Ticketing System**, http://ggus.org

[4] **Global Grid User Support: the model and experience in the Worldwide LHC Computing Grid**
T.Antoni, F.Donno, H.Dres, G.Mathieu, P.Strange, D.Spence, M.H.Tsai, M.Verlato
Proceedings of Computing in High Energy and Nuclear Physics (CHEP06), Mumbai, India, February 2006

[5] **CIC Portal**, http://cic.in2p3.fr/

[6] **Sites Functional Tests (SFT)**
https://lcg-sft.cern.ch:9443/sft/lastreport.cgi

[7] **Information System Monitoring (GSTAT)**
http://goc.grid.sinica.edu.tw/gstat/

[8] **GOC Wiki Pages**
http://goc.grid.sinica.edu.tw/gocwiki/FrontPage

[9] **Lavoisier:**
A Data Aggregation and Unification Service
S. Reynaud, G. Mathieu, P. Girard, F. Hernandez, O. Aidel
Proceedings of Computing in High Energy and Nuclear Physics (CHEP06), Mumbai, India, February 2006

[10] **COD Operational Manual**
https://edms.cern.ch/fi le/701575

[11] **Operations Tools Wiki Page**
http://goc.grid.sinica.edu.tw/gocwiki/OpDocs

[12] **Grid Operations Center Database (GOCDB2)**
http://goc.grid-support.ac.uk/gridsite/gocdb

[13] **EGEE Newsletter 9.**
http://egee-intranet.web.cern.ch/egee-intranet/newsletter/Sept-2005.html

[14] **Service Availibility Monitoring (SAME)**
http://goc.grid.sinica.edu.tw/gocwiki/Service_Availability_Monitoring_Environment

[15] **Freedom of Choice for Resources (FCR)**
https://goc.grid-support.ac.uk/gridsite/bdii/site-apps/FCR-cgi/fcr.cgi