RAL Tier 1: From Development to Production

S de Witt, G Smith, A Sansum

STFC Rutherford Appleton Laboratory, R89, Harwell, Didcot, Oxfordshire, OX11 0QX, UK

shaun.de-witt@stfc.ac.uk

Abstract. In this paper we present the steps in the evolution of the Rutherford Appleton Laboratory Tier 1 Facility from a development system to the high quality entity in production today, and how this improvement will continue into the future. We look at the processes and tools we have in place to monitor performance, ensure changes do not impact operations and the planning procedures to ensure smooth running.

1. Background

The history of the UK Tier 1 facility dates back to 2000 when Rutherford Appleton Laboratory (RAL), together with 15 other academic institutes around the United Kingdom requested funds to set up a UK Grid Infrastructure, which became the GridPP project – part of the UK contribution to the LHC Grid Project. RAL was selected to host the Tier 1 based on the fact that it already hosted the UK BaBar computing centre and had a long history in delivering large scale storage and computing facilities to the particle physics community. The first prototype Tier 1 was delivered in March 2002, consisting of 156 dual processor CPUs each with 30GB of storage. The development from this prototype system to the first version to be used in real production took considerable effort and was fraught with difficulties. There were periods when the development system would be down for weeks at a time. During this evolution of the three main components of the Tier 1 at RAL, the storage, grid services and fabric teams, were somewhat disjoint and it was evident that the software in use was not as robust as expected, requiring significant extra effort in testing before moving to production.

2. Test Systems

Test systems have always been an important part of the Tier 1 Infrastructure. At RAL, we have implemented two test systems with distinct purposes. Initially, in 2006 the first storage group test system was set up for the express purposes of testing 'edge cases', those cases which should never occur in production, but could if something goes wrong. This was deliberately set up as a small scale system which allowed us to test the behavior of the system when disk servers fill up, allowed for profiling of transfer rates and analysis of the best network tuning parameters. This was a small scale system which was independent of other developments and allowed basic functional tests of all of the main operations of the hierarchical storage management system, but was not scaled for stress testing.

By 2009 a requirement for a larger scale pre-production set up which can be interfaced to the grid tools and accessed by all Virtual Organisations (VOs) was realised. This system consisted of multiple space tokens and was set up to mimic the production systems of the Large Hadron Collider (LHC)

VO's to allow them access to the test system before significant upgrades were rolled out to production. This system is scaled to the size of a small production system, but is used concurrently by all experiments. This allows full end-to-end testing of any change with minimal effort by the end users.

In addition to these we also have smaller scale test setup for individual components including the File Transfer Service (FTS) and the Berkley Database Information Index (BDII), either of which can be used with the pre-production setup.

3. The Role of the Production Team

The production team within the Tier 1 was created when it was realised an independent team was required to take a lead in coordinating the work of the three existing teams. Their initial roles were to perform not only this coordination role but also to assist in the routine operations, formalise the already existent procedures, and to develop a strategic plan for moving the development facility into a fully operational one ready for LHC data taking. To this end, the team formalised and enhanced procedures and, where applicable, formalised the interactions between the teams. Much of this co-ordination was achieved through enhanced use of the already existing ticketing system and regular liaison involving the all the teams involved including the production team itself during any process.

To assist with monitoring of the many systems in the Tier 1, they also undertook a review of the existing monitoring put in place during the development phase, added additional monitoring where appropriate and removed redundant alerts. This made the monitoring system far more useable since prior to this many alerts were put in place which proved to be inappropriate, but were never actually removed due to time constraints, leading to a flood of false alerts. The Tier 1 now reviews alerts via a regular weekly meeting to discuss whether they are still appropriate or whether new monitoring needs to be put in place in response to operational issues.

The team, supported by members of the grid services group, also put in place the procedures and tools needed to run out-of-hours support. This support is designed to allow problems to be addressed outside of the normal working day (defined as the hours between 17:00 and 08:30 hours) and over weekends and holidays. This was clearly not as simple as putting together a rota of on-call support from the various teams, but also developing the infrastructure such that staff were notified whenever there was a significant problem, as well as actually categorizing errors to assess whether to call out or whether a problem could be safely left to working hours. Within the out-of-hours we use a role *primary on-call* whose purpose is to triage problems. In most cases the primary on-call person can take appropriate action, but in the event of unexpected or more significant errors we also have second-line on call staff who are experts in storage management, grid services, fabric or databases who can be called on (similar to the 'expert on-call role at CERN). In the event of a major problem out-of-hours the primary on-call also acts as a conduit for information and takes the role of incident coordinator. During these out-of hours incidents, we make use of internet 'chat-rooms' where several staff can communicate with each other.

A disaster management plan has also been developed to mitigate against major threats to the Tier 1. This is used at an early stage in any unfolding scenario to take control of the situation and minimize any adverse impact. The process consists of several levels, ranging from the disaster management team merely being aware of an issue and needing to keep a watch on it, such as delayed delivery of hardware, to levels where the emergency team and support staff are required to attend outside of normal working hours on-site, for example when there has been a complete failure of the database under the mass storage system. Further, based on the experience with the generic plan, we have also evolved fifteen specific plans for well understood potential issues.

4. Change Control and Planning

One of the most significant changes brought in to ensure smooth operations of the Tier 1 has been the introduction of a change control process. This was developed as a lightweight process to ensure changes made are both well thought out, well tested and well planned. In order to keep the process lightweight it allows for various levels of approval within the process. The process is shown diagrammatically in figure 1. It should be noted that while this system was created and has evolved to meet the needs of the Tier 1, the final procedure maps closely onto the IT Infrastructure Library (ITIL) framework for change control.

Three categories of changes have been identified and each has a different path through the change control process. The formality of the change control process is largely dependent on the type of change, ranging from a simple post implementation review to a fully documented change request. To foster a culture of continuous improvement, most changes are reviewed post implementation to assess how successful the change was based on criteria such as whether it had the desired effect, whether any problems were encountered and whether the implementation overran. Details of the procedures to follow for each type of change are given in the following subsections.

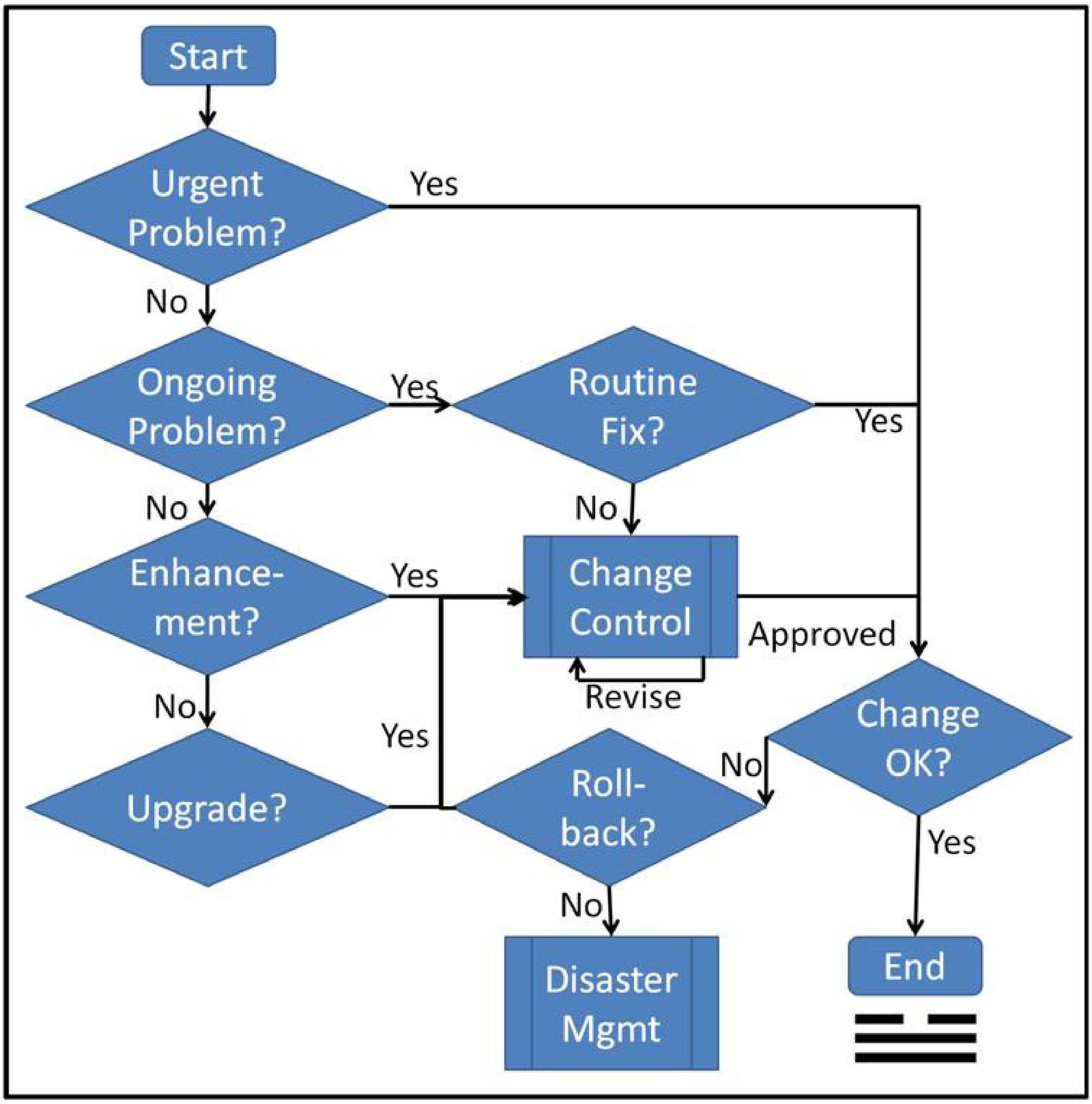

Figure 1: The RAL Change Control Process

4.1. Immediate operational Problems

Changes which are necessary to correct immediate operational problems do not need to go through a formal change review process prior to implementation since the priority is to restore the service to a fully operational mode. Once the affected service has been restored, details of the procedure are recorded in an e-logger or, for more serious incidents, by the preparation of a formal incident review. Incidents are also logged in the ticketing system so that the whole problem can be reviewed with a view to identifying improvements to procedures or new developments which could be implemented to prevent a recurrence.

4.2. Routine Changes

Routine changes which are performed on a regular basis and have well understood impacts on the Tier 1. These represent a set of approved exceptions which have been agreed between the team leads and the production team and do not require a formal review process unless the change has an unexpected impact. Examples of such changes are adding new disk servers, altering the FTS channel settings and restarting services. Experience has demonstrated that this is the most common type of change and the lightweight nature of the process helps ensure the Tier 1 remains responsive to user requests.

4.3. Significant Changes

Non urgent problems, service enhancements and significant upgrades all need to go through a formal change control procedure. For these, requestors of the change are required to fill a simple form which gathers information such as the reason for the change, details of any testing performed, risk assessment of the change, whether users need to be notified of the change and an implementation plan, including any required downtime for a service. This document is initially reviewed by the team leader of the team which will control the change who will score it based on perceived risk and impact of failure. Based on this score, the team lead can either approve the change or push it up to the change control board where all of the team leads, including at least one member of the management team will review the change and either approve the change, or request modifications or additional testing or suggest alternative approaches to consider before it is approved. In some cases changes may even be rejected if the perceived risk outweighs the benefit or an incompatibility is uncovered during the review process, but this is extremely rare.

It is expected that most changes which fall into this category will have undergone some testing on one of the testbeds at the Tier 1. Depending on the complexity of the change, basic functional testing may suffice, or significant stress and end-to-end testing may also be required.

5. Tools

The RAL Tier 1 uses a large number of both open source and custom designed tools to allow us to monitor the system and ensure the smooth operational running of the facility. This monitoring covers many levels, from the basic hardware operations and disk verification to overall system monitoring such as looking at the number of unmigrated files (those waiting to be written to tape), the success and failure rates of jobs running on the Tier 1 batch farm and transfer statistics of the FTS.

For hardware monitoring of the system we use a number of tools. The main ones are *Nagios* [1] for active monitoring and *Ganglia* [2] for passive monitoring. In addition we use the CERN provided *fsprobe* tool [3] to monitor the health of every file system. For an overview of the systems as a whole we use the CERN *Service Level Status* tool (SLS) [4].

To help facilitate communication and coordination between teams we use *Request Tracker* [5] for ticket management and *Numara's FootPrints* [6] for Incident tracking. In addition, the Tier 1 has an internal e-logger for recording changes and an internally developed status page [7] and blog [8] external communication.

Managing any large scale facility such as a Tier 1 requires a good fabric management tool. At the RAL Tier 1, we initially used the *puppet* tool ([9] within the mass storage team, but the whole Tier 1 has started to move to use *quattor* [10],[11] after a further review and evaluation. Both tools were shown to have advantages and disadvantages, but on balance it was felt in our case Quattor, with its pre-provided templates for various grid services, offered sufficient additional benefits to make it worth investing in the effort of moving between the two systems.

6. Summary

Since the first prototype, the RAL Tier 1 has evolved from an unstable, unreliable development system to a high quality production system which undergoes only minimal and essential downtime for maintenance. The addition of a production team has seen much better coordination between the different groups within the facility. They have brought a more cautious approach to changes and upgrades, while still maintaining flexibility to address immediate operational issues. The constant review of issues and responsive has brought about a culture of 'continuous improvement'. Even the change control process, which was initially perceived to be quite burdensome, has proved of enormous benefit since it forces implementers to consider the consequences of any change fully, and not just how it affects their team.

7. Reference

[1] http://www.nagios.org/
[2] http://ganglia.sourceforge.net/
[3] https://twiki.cern.ch/twiki/bin/view/FIOgroup/DiskRefFsprobe
[4] Lopienski S 2008 *J. Phys:Conf. Ser.* **119** 052025
[5] http://bestpractical.com/rt/
[6] http://www.numarasoftware.com/footprints/service_desk_software.aspx
[7] http://www.gridpp.rl.ac.uk/status/
[8] http://www.gridpp.rl.ac.uk/blog/
[9] http://projects.puppetlabs.com/
[10] Childs S ,Poleggi M E,Loomis C ,Mejías L F M, Jouvin M, Starink R, De Weirdt S, Meliá G C, 2008 *Proc. of the 22nd conf. on Large installation system administration conference (San Diego)*, pp.175-189
[11] http://quattor.sourceforge.net/