Feedback from the Tier1s

GDB, September 2008

CNAF 24x7 support

- *On-call person for all critical infrastructural services (cooling, power etc..)*
- *Manager on shift looking after all services (e.g. WMS, CEs, SRMs)*
- *SMS triggered by infrastructural and service alarms*
- *Starting from this October all scheduled interventions requiring an outage will be done on the third Tuesday of each month*

CC-IN2P3
Support during out-of-office hours

- Operation of critical services are supported during out-of-office hours during all the year
  - Really urgent interventions can be triggered at any time by the on-call engineer
    - Improvements still needed to detect that an intervention is considered urgent
  - We defer major (re)configurations of services until when experts are available on site
    - For instance, avoid reconfiguration during holiday periods
  - For a particular service, whenever possible experts organize their vacation periods to make sure that there is always one expert available
    - level of expertise is not equal between all the members of a team
  - Details of the organisation were given in a presentation to the Management Board meeting in August 2006:
    - http://indico.cern.ch/conferenceDisplay.py?confId=a063099

CC-IN2P3 Outage planning

- One planned outage per quarter, scheduled one year in advance
  - Planning available at web site
    - Current plan goes until december 2009
  - Details of the services and experiments impacted for each outage are published 10 days in advance
    - the outage may be cancelled if it is not really needed. This is also announced at least 10 days before
  - Interventions include power, cooling, hardware and software updates
  - Please note that there are service updates that do not need the service to be down, or only partially (impactingonly a limited set of experiments)
    - These are announced to the impacted experiments, but are not necessarily planned long in advance

24x7 at RAL

- Out of Hours: 17:00-09:00 Monday-Friday, weekends all day and public holidays.
  - Primary on-call guarantees response within 2 hours to pager call. Over 95% of calls responded within target time – most much faster.
  - Pager call may originate from:
    - Internal exception monitored by nagios
    - Multiple consecutive SAM test failures (about 2 hours)
    - Email alarm
  - Primary on-call may fix fault (documentation) or call second line support (for Fabric, CASTOR, Grid, Databases, Network or Machine room.
    - Second line guarantees response within 2 hours to phone call (usually much faster but no guarantee).
    - Second line does NOT guarantee to work all weekend to fix fault. Will do what is reasonable/achievable given personal circumstances.
- System runs throughout the year (skeleton service over Xmass). No changes planned.

RAL Downtime

- Team coordinates downtime requirements each Monday afternoon at a formal review meeting. Historically
  - Downtimes scheduled as required by component managers
  - Subject to constraints imposed by WLCG/EGEE rules
  - Merge downtimes wherever possible
- No requirement for routine downtimes for machine room/power/cooling. We can maintain service (at risk) during routine preventive maintenance.
  - Require other (non-routine) scheduled disruptions to give 3 months notice.
- Up until now have taken the view that our main priority is development ahead of data taking. During data taking periods we intend to minimise downtime requirements but have no formal policy on what that should mean in practice.
  - Open to guidance here but note the tension between need to address operational issues and experiments' need for stability.

24x7 at NDGF.

- General setup:
- First line handled by Operator on Duty in office hours / day time weekend, and by the NORDUnet NUNOC team outside office hours.
- Second line support is handled by Operator on Duty (365 days a year -8 hours a day)
- Third line support (on the sites, which are all redundant) handled by local staff, in business hours.
-
- All critical services are redundant, except for a very few (SRM door).
- Semi-Automatic setup in place for moving service - i.e. on call 24x7 staff can orchestrate a move if a critical box goes down.
- Setup finalized, though some final tweaks will be implemented over the next short time, better documentation of procedures etc.
- We plan outages well in advance (months) and will coordinate this at the weekly WLCG operation meetings.

NDGF 24x7 Service

Last Update: Nov 9th, 2007

Introduction

Our mission is to: **Provide a high performance and high reliability service for the Nordic storage and production grid, especially the Nordic WLCG Tier-1.** To further help our service reliability, we need to be able to react to problems on a 24x7 basis. This document states the goal and objectives for our 24x7 service and also describes the tools, procedures, organization and workflows that are needed to achieve these objectives.

Services

The primary goals are to provide high availability services to achieve **99%** availability to our critical services. To meet this level of services the **Level 1 Service** class is created. This level of availability is required for the operation of the WLCG Tier-1 and it is also expected that other e-Science projects hosted by NDGF will request this kind of high availability. Many of the critical services can, however, be split into several redundant services and hence the requirement on uptime will be replace by redundancy. I.e. such services are defined moved to the **Level 2 Service** class.

A third service class, **Level 3 Service**, are ran on best effort. For Level 3 best effort services, engineers are notified of faults that occur, but are not required to respond immediately. These service faults can be addressed on a best effort basis or the on the next business day.

Level 1 – critical services:

These services are critical single point of failures that will, if non-functional, prevent crucial tasks from being done. Crucial tasks are dictated by e-Science projects and include:

- The CERN Project:
  - Data taking on a 24x7 basis – 99% uptime at 100 specified days a year, maximum downtime is 12 hours

Some level 1 critical service can fully or partly through redundancy be reduced to lower level services.

Requirements:
Average annual availability: 99%
Max time to respond: 4 hrs
Backup: continuously
Hot-spare with heartbeat monitoring

In general Level 1 services are continuously backed up and a hot spare is ready with automatic failover to the backup service.
Level 1 services are only run at the NORDUnet high availability hosting facility in Örestaden, DK.

Level 1 Service list:

- dCache admin interface
- dCache gridftp door

TRIUMF 24x7 support

• TRIUMF's main control room operates 24x7x365 for cyclotron operations. Cyclotron operators are not to deal with the Tier-1 operations. They are only p;rovided with a simple procedure to contact Tier-1 staff in case of emergencies: site-wide power outage, etc. and whenever CERN calls.

• Tier-1 staff are not on-site 24x7 (not necessary). Some Tier-1 staff are only 10 minutes away from TRIUMF in case immediate physical presence is required during off hours..

• System fully monitored with various nagios sensors. In case of problems a page is issued if:

- SAM tests failure (2 consecutives) (subject to some false alarms)
- Power (UPS), Cooling, temperature alarms
- Experiment alarm (special email with authorized DN's)

• Pager is rotated among Tier-1 personnel on a weekly basis.

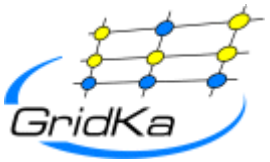
• To contact the pager we use two methods:
- via network/email (primary)
- dedicated phone line with a modem (as a backup)

• Paging system is also monitored: a test page is sent everyday 5 minutes prior to noon (lunch time). This is to test that nagios sensors and the paging mechanism is functioning properly.

• Downtimes are kept to a minimum and are scheduled when upgrades are truely necessary (gLite, dCache) (always within 4 hours in the last year). We tend to combine various upgrades all together to minimize downtime.

• We have been providing a true 24x7 support for more than a year and we will continue to do so.

24x7 on-call service at GridKa

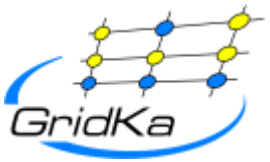


- **Until August 2008:**
  - On-call service only for infrastructure (power, cooling) and network.
  - No 24x7 service for middleware, databases and storage.
    - Weekend / public holiday coverage on voluntary basis.

- **Starting now (September 2008):**
  - 24x7 on-call service also for middleware, storage (dCache, file servers) and Oracle databases.
  - Currently two additional on-call groups: middleware / storage+DBs
  - One person per group is always on-call outside business hours.
  - Many system experts volunteer to be called by person who is on-call, but are not standby all the time.
  - Person on-call is not necessarily an expert for all systems that might fail.
    - Will try standard recovery procedures where possible.
    - Will call system expert (if available) when standard recovery procedures fail.

  - On-call people receive alarm on mobile phones issued from Nagios.
  - Sensors include SAM, internal functional tests of middleware + SRM, processes, file systems, pings, ...

24x7 on-call service at GridKa



- Nagios setup and sensors will be continually improved.
  (expecting steep learning curve.)
- 24x7 coverage for lower level services will be introduced
  (e.g., server hardware, operating systems)
- On-call service runs all year.

Mainenance / downtime planning

- Some updates, e.g. of dCache, or larger network maintenances cannot be avoided and will also in the future lead to outages of the whole site or services for several hours.
- Most network restructuring at GridKa has been done already before the LHC start.
  (e.g. to improve redundance of WAN connections)
- No major maintenance windows are *currently* planned.
- No regularily occuring (e.g. quarterly) downtimes planned for GridKa.
- Major maintenances / downtimes are coordinated with all experiments (LHC and non-LHC) in the GridKa Technical Advisory Board.
- Short downtimes of single services are coordinated with the affected experiments.

NL-T1

- At the dutch T1, we have a voluntary mechanism for after hours support. we are staffed from 08.00 til 18.00 on normal working days. outside of this, for certain classes of alarm tickets, "special" communication paths are used to contact experts; however there is no obligation by the expert to respond outside of working hours.
-
- we try to make the underlying system as robust as possible via redundancy and proactive monitoring.
-
- Our 24x7 document is undergoing the final edits as we speak, this is the gist of it.

24x7 at PIC

- Out-of-hours hours support of critical services:
  - Manager on Duty (MoD) weekly shifts, 24x7. System in place for the whole year. For the moment, no difference between data taking period and other.
  - MoD regularly checks monitoring system. Each alarm has an associated recovery procedure that the MoD will follow in case of failure.
  - Critical alarms generate SMS to MoD mobile phone.

- Scheduled downtimes:
  - Once a year, complete site outage (2-3 days) for electrical maintenance of the building.
    - Up to now, downtime scheduled inside Easter week. If this has to change, need to coordinate with other T1s and experiments with as much time as possible (one year notice).
  - Would like to set up a schema of regular scheduled maintenances.
    - In the last months, we have tested monthly scheduled maintenances (2nd Tuesday of every month). Very positive experience. Allows to plan and notify experiments with enough time.
    - During data taking, the frequency of these scheduled maintenances might be longer (quarterly?). Open to discuss and coordinate this with experiments and other Tier1s.