Cross-site problem resolution

Focus on reliable file transfer service

Gavin McCance

Service challenge technical meeting

15 September 2006

www.eu-egee.org
www.glite.org

Service issues

- **For the distributed transfer service there are two broad class of failure that cause service degradation**

1. Internal failures of software or hardware
   - FTS daemons lock-up, disk failure, kernel panics, file-system fills up…

2. Degradation or failure of an external service
   - MyProxy problems, information system problems
   - Castor problems
   - Tier-1 SRM problems
   - Networking problems

Internal failures

1. Internal failures of the FTS
   - FTS daemons lock-up, disk failure, kernel panics, file-system fills up…

- **Most of these problems can be handled by procedure, redundancy, etc**
  - RAID, redundant power supplies, operator procedures
  - Recovery procedures in case of server failure
- **There is already 24 x 7 support for known internal problems**
  - For previously unknown problems there is expert backup
    - Office-hours week-day only
    - Is this enough for software still under active development?
      - *(bearing in mind changes in behaviour of dependent software can also and do affect FTS)*

External failures

2. Degradation or failure of an external service
   - MyProxy problems, information system problems
   - Castor problems
   - Tier-1 SRM problems
   - Networking problems

- Two stages:
  - **Detection**
  - **Resolution**

External failures

- **Detection is 'easy'**
  - SAM tests are being improved to cover much of this
  - FTS 'knows' about all the transfers and can expose this information – the failure rate is measured
  - This needs work to integrate with current systems

- **Resolution: if the source of the problem is obvious:**
  - Obvious problems can be sent to the owner of the problem ~semi-automatically (FTS sends an alarm). e.g. 30% of transfers failed because your SRM didn't answer.
    - Appropriate for problems where the problem is obviously localised to one site
    - FTS knows where the problem is and sends an alarm to someone. This person with this role calls the right people using the appropriate system *and follows up.*

External failures

- **There are still many service degradations for which the cause is harder to determine**
  - “Transfer failure” (gridFTP isn’t always obvious about the cause).
  - Networking problems and firewall issues
  - Problems on SRM-copy channels (FTS doesn’t have much logging about what went wrong)
  - “This channel seems to be going slow for the last few hours” type problems

- **These require ‘expert’ involvement and investigation**
  - Local experts on FTS, Castor, networking
  - Remote experts on tier-1 site SRM, networking

- **Of course, the goal is to move as much of this as possible to the ‘automatic’ system**
  - Packaging ‘canned problems’ takes time and experience with the problem
  - Some things will never be moved

Who can we use?

- **For easy problems that require an alarm and follow-up we have CIC-on-duty**
  - Prerequisite is adequate monitoring
  - Can also handle problems that require a (small) bit of digging provided the tools and procedures are there
  - This needs to be our next development priority

- **… but CIC-on duty is office-hours week-day only**
    - (and moves time-zone)
  - We will not meet WLCG 99% service availability target with just this - two weekends downtime and you've failed to meet the target

- **For harder problems, we require an expert team**

Core hours proposal

- **Core hours = weekday office-hours**

- **Easy problems go to CIC-on-duty**
  - Alarms come from SAM and from FTS
  - Obvious alarms can be sent to correct site immediately
  - Procedures and tools are provided to dig (a little) deeper if the problem is not immediately obvious
  - The monitoring needs to be the next FTS development priority

- **Harder problems and problems requiring cross-site coordination go to an expert piquet team**
  - CIC-on-duty will get the alarm *detecting* service degradation
  - If the cause isn't obvious, call expert team to investigate
    - Send alarm ticket to site (incl. CERN)
    - Investigate with remote experts
  - CIC-on-duty should follow-up the issue

Out-of-hours proposal

- The WLCG expert piquet team extends the coverage provided by the CIC-on-duty
  - Proposal is 12-hour coverage including weekends
- The flow is the same – the team should make use of the same monitoring systems and alarm raising systems as CIC-on-duty
  - CIC-on-duty should perform follow-up during weekdays
- With this we accept up to 12 hours unattended service degradation
  - Recover using the transfer service catch-up
  - Maybe this needs review during accelerator running
- Need to resolve CIC-on-duty time-zone issues