Enabling Grids for E-sciencE

# Feedback from SEE first COD shift

*Emanoil Atanassov*

*Todor Gurov*

Information Society

8th CIC on Duty meeting

Krakow 16-18 05/2006

# Outline

- Operational Tools
- Observations on site problems
- Other problems and issues
- Proposals for improvement
- Conclusions

# Operational Tools

- GSTAT – sometimes GSTAT gets stuck. It would be nice if it could give a warning, so that we can switch to the CNAF mirror. Downtimes are not consistent with GOC DB (difference of 1 hour or so). Please double-check if this is now fixed.
- GGUS – was down at the start of the shift, fixed in an hour or so. If GGUS is down, the CIC dashboard is also not available, which is bad.
- Certificate lifetime monitor – no problems, but it shows info only about CEs.

# Operational Tools(2)

- CIC portal – very useful. A new version was put in place during the shift, but it had some problems when showing a site with more than one ticket. Problems were reported, and resolved.
- SFTs – It takes some time to realize that SFTs are not updated. A warning could be helpful, so that new SFTs could be run manually.
- SFTs for the PPS on the gLite site failed throughout the shift, because of some middleware problems, that could not be resolved. It seems that PPS sites have good responsiveness to tickets, but the problems could not be solved easily.

# Observations on site problems

- JS problems – several sites with JS problems. We noticed that LHCB created several tickets with Subject "More than 1000 failed jobs at …" for such sites. Our impression is that many of these problems are caused by some, but not all, misconfigured worker nodes, and are not really resolve by the site administrators. "We did nothing, but the problem is resolved." Especially in the case of "Maradona problem" site administrators could be requested to identify the WN where the job failed.
- Sites with replica management problems – mostly sites using some other BDII, not lcg-bdii.cern.ch. Lots of tickets and lots of emails on such issues, but the problem is easily solved by changing two files – lcgenv.sh and lcgenv.csh on the WNs.
- R-GMA problems – no time to deal with them, because the number of sites affected is big. Many sites did not upgrade their MON boxes to secure R-GMA and consequently fail the R-GMA test. Is this test really a critical test? If it is, why gstat does not show the site in CT in such case?
- JL problems – several of the big sites with more than one CE had non-trivial JL problems. It appears they were caused by high load on the CE, which usually results from users submitting jobs directly to globus, or something similar. Seems to be a middleware problem.
- The current procedure does not deal efficiently with sites that mark the problem as "solved", but do not really do anything about their problem. Frequently it is obvious immediately that the problem is not solved, but what are we supposed to do?

# Other problems and issues

- Downtimes – CIC portal gives a warning if a ticket is going to be issued for a site that has been or is going to be in downtime during the same day. However, it is unclear if tickets should be issued for sites that are down outside their scheduled downtime – question about procedure. This lead to some confusion with one USA site. In the new GOC DB2 the downtimes are input in local timezone, which solves this problem. However, we still believe Gstat is not entirely in sync with that.
- The 4444 problem – many sites showed this problem. It appears it was a middleware problem, and it should only result in ticket if the site shows 4444 waiting jobs for extended period of time.
- If GIIS for a site is down, in CIC portal it appears as having 0 CPU, and it is shown at the bottom of the list, even if the site usually has 1000 CPUs. The number of CPUs should be taken as average over some period of time, to avoid this problem.

# Proposals for improvement

- **Operational tools**
  - CIC portal shows a site as down if it is down at GSTAT. However, it is very annoying to open the site's page, then we perform ldapsearch and see that the site is actually up, and nothing is to be done. The same applies about SFTs – one SFT failure is not enough. Therefore some history of failures should be visible in CIC portal.
  - The history of tickets issues for the sites should be easily visible, because maybe some sites have "replication failed" most of the time, and it is always in quarantine and this makes a vicious cycle of "1st email, 2nd email, 1st email, quarantine,1st email" and so on.
  - However, ticket reopening should be avoided, because it is not clear how relevant the ticket is to the new situation.

# Proposals for improvement(2)

- **Operational metrics**
  - Metrics are not ideal
  - A good metric for site’s operation is “number of tickets”
  - We propose to call downtime as measured by SFTs “SFT-downtime”, because sometimes a site is up, and running, but some test fails for some reason, perhaps just because of a network problem that lasts for 1 minute.
  - Another possible metric is “ticket downtime” – time, when the site has a critical ticket and its status is “1st email” or “2nd email” (or even quarantine). We believe that there are sites that are less than 50% available with respect to “ticket downtime”☺

# Proposals for improvement(2)

- **SA1 Operational procedures (in the document)**
  - Page 4, below, see point 2 in the subsection 3.2: “CIC-on-duty team regular tasks”. It is written: “….at the end of the day”, Maybe the correct phrase should be: “…at the end of the shift ( week).”
  - Subsection 3.3, the 2nd item: The correct mailing list is: project-eu-egee-sa1-cic-on-duty@cern.ch
  - All tables should be numbered.
  - (See page 10): There are 3 steps in the 1st table (subsection 5.4 ). But step 4 is mentioned in the paragraph under the table.
  - Page 15, from the top, point 4, in the 1st item: “…in the table 3”. Such table doesn’t exist.
  - In table 2: “Phone call to ROC” – Does this escalation step exist or not ?

# Conclusion

- **The Grid is not about solving tickets, the Grid is for users.**
- **It is extremely annoying to see 50% of your jobs being aborted at some sites.**
- **Example – a site that has 100 nodes and 1 WN misconfigured, will fail 1% of the SFTs and may not even get a ticket, but if a user submits 200 jobs to the site and the site is empty, the user will see 99 jobs ok and 101 jobs aborted.**
- **That is why we believe the monitoring should be more stringent.**