CCRC'08 Weekly Update

Jamie Shiers

~ ~ ~

WLCG MB, 19th February 2008

Agenda

- Do not plan to repeat talk from today's meeting with LHCC referees

- Focus on recommendations and key issues

1. Scope & Timeline of the Challenge(s)

2. **Problem Tracking, Reporting & Resolution**

Scope & Timeline

- We will not achieve sustained exports from ATLAS+CMS(+others) at nominal 2008 rates for 2 weeks by end February 2009

- There are also aspects of individual experiments' work-plans that will not fit into Feb 4-29 slot

- **Need to continue thru March, April & beyond**

- After all, the WLCG Computing Service is in full production mode & this is its purpose!

- **Need to get away from mind-set of "challenge" then "relax" – its full production, all the time!**

Handling Problems…

- Need to clarify current procedures for handling problems – some mismatch of expectations with reality
  - e.g. no GGUS TPMs on weekends / holidays / nights…
    - ➢ **c.f. problem submitted with max. priority at 18:34 on Friday…**
  - Use of on-call services & expert call out as appropriate
    - {alice-,atlas-}grid-alarm; {cms-,lhcb-}operator-alarm;
  - Contacts are needed on all sides – sites, services & experiments
    - e.g. who do we call in case of problems?

- Complete & open reporting in case of problems is essential!
  - Only this way can we learn and improve!
  - ☛ **It should not require Columbo to figure out what happened…**

- Trigger post-mortems when MoU targets not met
  - This should be a light-weight operation that clarifies what happened and identifies what needs to be improved for the future
  - Once again, the problem is at least partly about communication!

FTS “corrupted proxies” issue

- The proxy is only delegated if required
  - The condition is lifetime < 4 hours.
- The delegation is performed by the glite-transfer-submit CLI. The first submit client that sees that the proxy needs to be redelegated is the one that does it - the proxy then stays on the server for ~8 hours or so
  - Default lifetime is 12 hours.
- **We found a race condition in the delegation - if two clients (as is likely) detect at the same time that the proxy needs to be renewed, they both try to do it and this can result in the delegation requests being mixed up - so that that what finally ends up in the DB is the certificate from one request and the key from the other.**
- We don’t detect this and the proxy remains invalid for the next ~8 hours.

- The real fix requires a server side update (ongoing).

- The quick fix. There are two options: ... **[being deployed]**

ATLAS CCRC'08 Problems 14-18 Feb

- There seem to have been 4 unrelated problems causing full or partial interruption to the Tier0 to Tier1 exports of ATLAS.

1. *On Thursday 14th evening the Castor CMS instance developed a problem which built up an excessive load on the server hosting the srm.cern.ch request pool. This is the SRM v1 request spool node shared between all endpoints. By 03:00 the server was at 100% cpu load. It recovered at 06:00 and processed requests till 08:10 when it stopped processing requests until 10:50. There were 2 service outings totalling 4:40 hours. S.Campana entered in the CCRC08 elog the complete failure of ATLAS exports at 10:17, in the second failure time window, and also reported the overnight failures as being from 03:30 to 05:30. This was replied to by J.Eldik at 16:50 as a 'site fixed' notification with the above explanation asking SC for confirmation from their Atlas monitoring. This was confirmed by SC in the elog at 18:30. During the early morning of 15th the operator log received several high load alarms for the server followed by a 'no contact' at 06:30. This lead to a standard ticket being opened. The server is on contract type D with importance 60. It was followed by a sysadmin at 08:30 who were able to connect via the serial console but not receive a prompt and lemon monitoring showed the high load. They requested advice on whether to reboot or not to the castor.support workflow. This was replied to at 11:16 with the diagnosis of a problem of the monitoring because of a pile-up of rfiod processes.*

- **SRM v1.1 deployment at CERN coupled the experiments – this is not the case for SRM v2.2!**

ATLAS problems cont

2. Another srm problem was observed by S.Campana around 18:30 on Friday.

- He observed connection timed out errors from srm.cern.ch for some files. He made an entry in the elog, submitted a ggus ticket and sent an email to castor.support hence generating a remedy ticket. ggus tickets are not followed at the weekend nor are castor.support tickets which are handled by the weekly service manager on duty during working hours. The elog is not part of the standard operations workflow. A reply to the castor ticket was made at 10:30 on Monday 18th asking if the problem was still being seen. At this time SC replied he was unable to tell as a new problem, the failure of delegated credentials to FTS, had started. An elog entry that this problem was 'site fixed' was made at 16:50 on the 18th with the information that there was a problem on a disk server (hardware) which made several thousand files unavailable till Saturday. Apparently the server failure did not trigger its removal from Castor as it should have. This was done by hand on Saturday evening by one of the team doing regular checks. The files would then have been restaged from tape.
- The ggus ticket also arrived at CERN on Monday. (to be followed)

ATLAS problems – end.

3. There was a castoratlas interruption at 23.00 on Saturday 16 Feb. This triggered an SMS to a castor support member (not the piquet) who restored the service by midnight. There is an elog entry made at 16:52 on Monday. At the time there was no operator log alarm as the repair pre-empted this.
4. For several days there have been frequent failures of FTS transfers due to corrupt delegated proxies. This has been seen at CERN and several Tier 1. It is thought to be bug that came in with a recent gLite release. This stopped ATLAS transfers on the Monday morning. The workaround is to delete the delegated proxy and its database entry. The next transfer will recreate them. This is being automated at CERN by a cron job that looks for such corrupted proxies. It is not yet clear how much this affected ATLAS during the weekend. The lemon monitoring shows that ATLAS stopped, or reduced, the load generator about midday on Sunday.

Some (Informal) Observations (HRR)

- *The CCRC'08 elog is for internal information and problem solving but does not replace, and is not part of, existing operational procedures.*
- *Outside of normal working hours ggus and CERN remedy tickets are not looked at. Currently the procedure for ATLAS to raise critical operations issues themselves is to send an email to the list atlas-grid-alarm. This is seen by the 24 hour operator who may escalate to the sysadmin piquet who can in turn escalate to the FIO piquet. Users who can submit to this list are K.Bos, S.Campana, M.Branco and A.Nairz. It would be good for IT operations to know what to expect from ATLAS operations when something changes. This may be already in the dashboard pages.*

- (Formal follow-up to come...)

Monitoring, Logging & Reporting

- Need to follow-up on:
  - Accurate & meaningful presentation of status of experiments' productions wrt stated goals
  - "Critical Services" – need input from the experiments on "check-lists" for these services, as well as additional tests
  - MoU targets – what can we realistically measure & achieve?

- **The various views that are required need to be taken into account**
  - e.g. sites, depending on VOs supported, overall service coordination, production managers, project management & oversite

- **March / April F2Fs plus collaboration workshop, review during June CCRC'08 "post-mortem"**

Supporting the Experiments

- Need to focus our activities so that we support the experiments in as efficient & systematic manner as possible
- Where should we focus this effort to have maximum effect?
- What “best practices” and opportunities for “cross fertilization” can we find?

- The bottom line: it is in everybody’s interest that the services run as smoothly and reliably as possible and that the experiments maximize the scientific potential of the LHC and their detectors...

- **Steady, systematic improvements with clear monitoring, logging & reporting against “SMART” metrics seems to be the best approach to achieving these goals**

Draft List of SRM v2.2 Issues

Priorities to be discussed & agreed:

- Protecting spaces from (mis-)usage by generic users
  - Concerns dCache, CASTOR
- Tokens for *PrepareToGet/BringOnline/srmCopy* (input)
  - Concerns dCache, DPM, StoRM
- Implementations fully VOMS-aware
  - Concerns dCache, CASTOR
- Correct implementation of *GetSpaceMetaData*
  - Concerns dCache, CASTOR
    - Correct size to be returned at least for T1D1
- Selecting tape sets
  - Concerns dCache, CASTOR, StoRM
    - ¿ by means of tokens, directory paths, ??

Feedback by Friday 22nd February!

Service Summary

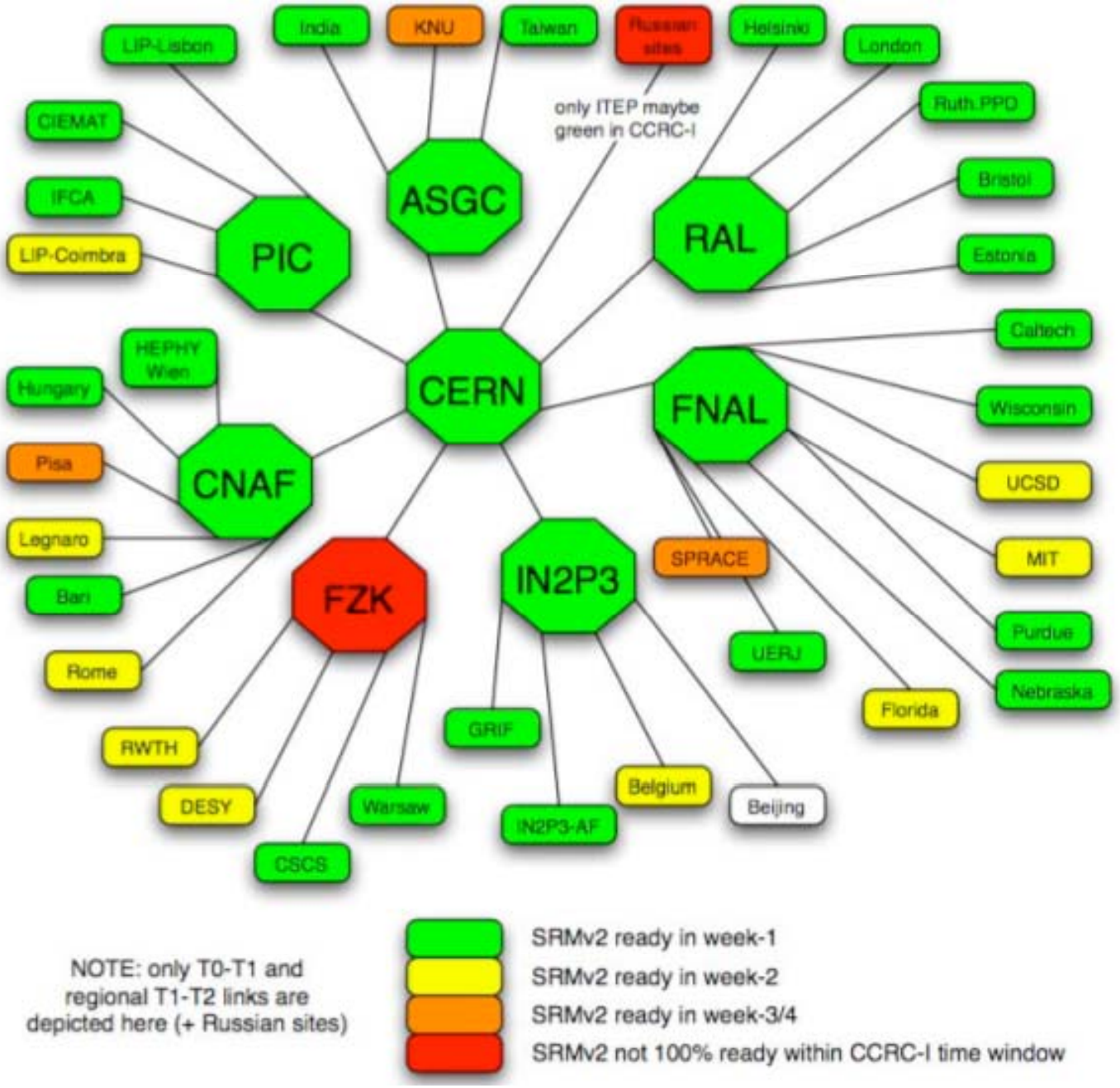
- From a service point of view, things are running reasonably smoothly and progressing (reasonably) well

➢ **There are issues that need to be followed up (e.g. post-mortems in case of "MoU-scale" problems, problem tracking in general…) but these are both relatively few and reasonably well understood**

☛ **<u>But we need to hit all aspects of the service as hard as is required for 2008 production to ensure that it can handle the load!</u>**

- And resolve any problems that this reveals...

BACKUP SLIDES


India
KNU
Taiwan
Russian sites
Helsinki
London
LIP-Lisbon
only ITEP maybe green in CCRC-I
Ruth.PPD
CIEMAT
ASGC
Bristol
IFCA
RAL
PIC
LIP-Coimbra
Estonia
Caltech
HEPHY Wien
CERN
FNAL
Hungary
Wisconsin
Pisa
CNAF
UCSD
Legnaro
SPRACE
MIT
FZK
IN2P3
Bari
Purdue
UERJ
Rome
Florida
Nebraska
GRIF
RWTH
Belgium
DESY
Warsaw
Beijing
IN2P3-AF
CSCS
NOTE: only T0-T1 and regional T1-T2 links are depicted here (+ Russian sites)
SRMv2 ready in week-1
SRMv2 ready in week-2
SRMv2 ready in week-3/4
SRMv2 not 100% ready within CCRC-I time window