Tier 1 Status and Recent Major WLCG Service Incidents

Harry.Renshall@cern.ch

LCG-LHCC Referees Meeting

22 September 2008

Overview

- Per Site:
  - Deployment of 2008 Resources
  - Forward look to 2009 procurements
  - Status of 24 by 7 Services Support – the sites provide 24x7 support to users as standard operations: completed by all sites
  - Status of VO-boxes Support (note all ATLAS VOboxes are at CERN except for BNL)

- Major WLCG service incidents in the last 6 months
  - Where an MoU/WLCG quality of service downtime was exceeded triggering a postmortem investigation and report. There were seven such major incidents (some with multiple downtimes) and a similar number of major but shorter incidents.

Tier 1 Status (1/4)

- ASGC: CPU at 70%, disk at 80%, tape at 60%
  - Will meet 2008 pledges November (and exceed in CPU).
  - 2009 CPU (most of) already in Nov 2008, storage on April target.
  - VObox SLA defined and in place and needs to be signed-off by the experiments.
- CC-IN2P3: CPU at 100%, disk at 60%, tape at 100%
  - Expecting rest of 2008 disk and half of 2009 disk to be delivered last week.
  - 2009 launching the purchasing framework. Part will be available by the end of this year, remaining orders to be made when budget is known.
  - VObox SLA ready to go to experiments for approval before implementation. Will be completed in next few weeks.
- CERN: All 2008 pledges available
  - 2009 pledges on track for April 2009 availability.
  - VObox support is defacto available but needs to be officially approved.

Tier 1 Status (2/4)

- DE-KIT: CPU at 80%, disk at 80%, tape at 70%
  - Remainder of 2008 pledge in October as synchronised with experiments (mostly ALICE) requirements.
  - 2009 CPU orders sent and disk order to be made soon. Tape libraries big enough for 2009 – will order media and drives. Expect to meet April target.
  - VObox support complete.
- INFN/CNAF: CPU at 60%, disk at 40%, tape at 40%
  - Rest of 2008 pledges in place and starting to be installed.
  - 2009 tenders start end September for end February delivery so will be tight to have all ready for April deadline.
  - VObox support complete.
- NDGF: CPU at 120%, disk at 80%, tape at 35%
  - Have all the storage and should be installed this month.
  - 2009 will have 2 more sites and expect to be on target.
  - Only ALICE have VOboxes in NDGF. These need detailed description of their many functions for the SLA.

Tier 1 Status (3/4)

- PIC: All 2008 pledges available
  - 2009 planning awaited
  - VObox SLA in place and the CMS contact persons need to be defined.
- RAL: All 2008 pledges available
  - 2009 Disk tender closing now and CPU tender in October. New building available September so on target for April 2009.
  - VObox support complete.
- NL-T1: CPU at 60%, disk at 20%, tape at 10%
  - 2008 disk finally passed acceptance and being configured. All 2008 pledges by November (tape media quickly added as/when needed).
  - 2009 framework tenders in place. Possible to be ready by April but needs additional space/power/cooling currently under acquisition to be on time.
  - SLA document not finalised yet. Needs both NIKHEF and SARA to approve.

Tier 1 Status (4/4)

- TRIUMF: All 2008 pledges available
  - 2009 procurements to be made early October for February delivery so should be on target for April 2009.
  - No local VOboxes.
- US-ATLAS: CPU at 100%, disk at 70%, tape at 100%
  - New power and cooling nearly ready so should fulfill 2008 disk pledge in October/November (adding 1 PB).
  - 2009 specifications are ready. Next comes discussions with US-ATLAS funding agency.
  - VObox support complete.
- US-CMS: All 2008 pledges available.
  - Need to verify 2009 disk order will be in place then will be on target for April 2009.
  - VObox support complete.

CERN-PROD SRM v22 blockage on Saturday morning 24/5/2008: 5 hours

- **Nearly complete outage of the 5 CERN-PROD CASTOR2 SRM endpoints: srm-alice, srm-atlas, srm-cms, srm-dteam and srm-lhcb from 06.30-11.30. Impact was failure of all Srm2 transfers CERN to and from Tier1s.**
- **The outage was reported by ATLAS to the atlas-operator-alarm list. The 24/7 operator observed a high load on the SRM db service due to a high number of oracle sessions and rebooted the DB server at ~11:30 after which the endpoints slowly recovered. In parallel the data service standby service manager saw the ATLAS report and informed the SRM service experts.**
- **Problem was due to a significant slowdown in the request processing on the shared (non-LHC) stager, castorpublic. The castorpublic problem has not been understood but the service recovered when the service manager noticed the degradation (in CERN SLS) and restarted the daemons just before 9am. The SRM service did however not recover because its own database was already blocked with too many sessions.**
- **Operator procedure for handling mails to <VO>-operator-alarm list was improved as a result.**
- **The number of oracle sessions on the shared database server was capped to avoid overload and the SRM server and daemon thread pool sizes were reduced to match the maximum number of sessions**

CNAF Electrical power problems 20-30 June

- **On Friday June 20, 2008 the INFN Tier-1 had a building infrastructure problem which required its UPS system to be temporarily by-passed. The UPS batteries had to be removed, and within 40 minutes we switched off all the Tier1 (without major incidents). Power was available again on Sunday June 22, 2008 and the full set of services was available on Tuesday June 24, 2008.**

- **On Thursday June 26, 2008, with the UPS still unavailable for the reasons mentioned above, there was an outage caused by the local power supplier; this brought down the entire INFN Tier-1. On June 27, 2008, the INFN Tier-1 was available again, with all subsystems restored.**

- **On Saturday June 28, 2008, an electrical problem probably caused by the local power supplier caused the opening, with some mechanical damage, of the two main electrical switches of the INFN Tier-1 causing yet another downtime. Power was restored in the afternoon of June 28, and all services were up again on June 30, 2008, when the UPS was put back in operation.**

- **In summary, in about 9 days there were three electrical outages of the INFN Tier-1 and which caused considerable damage to several components.**

IN2P3 Cooling problem reducing number of worker nodes 21 June

- A cooling compressor broke down at 14.20 on Saturday 21 June causing a rapid machine room temperature rise triggering an SMS message to the staff on standby duty.
- It was necessary to stop 260 batch workers (done at 15.30) then later (17.45) an additional 80 – about half the total batch capacity at the time.
- Repairs were made by 19.30 and the worker nodes were restarted at about 22.00.
- Long term actions will be to install additional cooling to add redundancy, improve communications for this type of incident and automate such stopping of worker nodes as a function of the machine room temperature.

CNAF network switch problems 5-8 July

- **On Saturday July 5, 2008 at 03:29 a 10 Gigabit/s interface on one of the Tier-1 core switches started flapping. The effect was intermittent connectivity failures to various sets of computers across the Tier1. What rendered detection of the fault not immediately obvious was that no traces of the flapping were recorded in the log files of the core switch actually exhibiting the problem.**

- **On Monday July 7, 2008 the priority for the replacement of the faulty network card was escalated to the highest possible level to the switch vendor. At 17:00 the faulty network card was replaced, and the network was operational again. A fallout of the network problem was that several systems were stuck and had to be rebooted.**

- **On Tuesday July 8, 2008 at 11:00, during certification of all INFN Tier-1 subsystems and services, some other network problems were detected. At 12:30 the cause of these problems was identified through log messages in a faulty core switch management card, which caused among other problems random packet loss. Another ticket was opened with the switch vendor, and in the afternoon a replacement management card was received. The replacement of the card and a related operating system upgrade to the core switches finished at 22:00.**

- **In the morning of Wednesday July 9, 2008 all network, storage and farm subsystems and services were checked and certified as ready for operation. But since the INFN Tier-1 was still in downtime, the decision was taken to replace the the broken component of the electrical switch from the June incident and this was done on Thursday July 10, 2008.**

CERN ATLAS conditions data capture failure from 26-30 July

- **ATLAS conditions capture stuck from 26th July 2-3 a.m. due to a known bug in Oracle streams when dropping tables with referential constraints.**
- **Problem fixed around 15:15 on 30th of July (missing logfiles identified, copied and registered, capture & propagation successfully restarted)**
- **new logfiles found missing on the 31st of July (due to ATLDSC restarts, when fixing the streams setup) - missing logfiles identified, copied and registered, capture & propagation successfully restarted**
- **ATLAS conditions OFFLINE -> Tier1s replication was not working properly from 26th of July till 30th of July afternoon**
- **Improvements in the monitoring will be needed to spot similar problems (assigned to development)**

RAL CASTOR failures 14 August – 27 August

- **The incident was first detected by ATLAS and was also seen on OPS tests on 2008-08-14, following an upgrade to CASTOR 2.1.7 and migration to a new ORACLE RAC configuration on 13th August. It manifested itself as a unique constraint violation in the CASTOR request handler. At this point RAL were unsure whether other tables were affected by the problem so took the decision to close the ATLAS instance while looking into the impact on other tables, since propagating the corruption would have led to a worse situation.**
- **14/08/08 onward. Various related and unrelated problems impacted the service in the days just after the 13th August upgrade and various interventions were carried out which may have resolved some issues, but the unique constraint violation remained masked under much other activity and there were five long unscheduled CASTOR ATLAS downtimes during this period (CMS and LHCb were also affected).**
- **19/08/08 Identified problem relates to unique constraint violation in Oracle**
- **20/08/08 Experiments/CASTOR team liaison meeting concluded no resolution was likely soon and 7 day downtime should be set**
- **27/08/08 Conclude that database (or data) corruption is unlikely and that simply restarting request handler clears problem. Commence new policy of manual/automated restart of CASTOR when automation detects problem and open CASTOR instances.**
- **27/08/08 Apply kks_use_mutex_pin=false on advice from CERN. Fault effectively resolved from this point**
- **1/09/08 RAL revalidated as an ATLAS Tier-1**

CERN ATLAS transfers/access to first beam data stuck 10 September

- **On 10 September at 18.10 first attempt at exporting ATLAS data from CASTOR to all T1s failed.**
- **19:34 - GGUS ALARM TICKET submitted by ATLAS shifter was received by CERN Computer Centre operator routed from GGUS. Automated SMS mechanism not triggered.**
- **19:45 - CASTOR expert called and rapidly identified that the problem is due to a disk hotspot. The resolution was to enable internal replication in the 'default' pool and temporarily remove the hot server from production. This forced a better load balancing of the files over the other servers in the pool and ATLAS were informed.**
- **10 September 2008 20:47 - CASTOR re-enabled the diskserver after having confirmed that the requests were better load balanced over all servers in pool.**
- **10 September 2008 20:57 - ATLAS confirms that situation is back to normal. Although this did not exceed the Mou limits it was happened at a particularly sensitive time.**

Conclusions

- These major incidents were largely unpredictable
- Not all of them can be avoided (or limited) by adding redundancy or improving monitoring.
- There were 4 incidents a month (counting RAL as 6 separate CASTOR ATLAS downtimes) with an average impact time of 30 hours (skewed by the long CNAF and RAL incidents).
- Six can be attributed to hardware and 9 to software.
- We must expect similar incidents during data taking and be prepared how to respond to them both at the sites and global LCG service levels.