WLCG Service Report

Jamie.Shiers@cern.ch

~ ~ ~

WLCG Management Board, 4th November 2008

Overview

- Since last week the template for the daily WLCG operations meeting has been updated (one file per week) so that it starts with a summary of Service Incident Reports from the previous week(s)
  - **Received; Open; Due.**
- **Hopefully this will simplify follow-up…**
- Unfortunately, the # of incidents that fall into the “degraded service” / “service or site down” continues to be at a (much) higher level than 1 per week...
  - **Tip of the iceberg? Statistical fluctuations? Half-term?**
- And there are quite a few incidents that should probably also be included
  - “Harvesting” EGEE broadcasts is both labour intensive and not very informative – no analysis of problem; not scalable or sustainable;
  - And its not clear that this would catch all major incidents in any case...

Site “down” times

- Even though not necessarily visible through standard monitoring, events which make one (or more) sites effectively unusable are **not that uncommon**
- At least a fraction (naturally…) of problems occur shortly before or during weekends – and are often not fully resolved until the **following week**

💣 The chance of two or more Tier1 sites being “down” at the same time (weekend) – as well as the consequences (see example) – means that this would be **extremely painful** during data taking and / or reprocessing

➢ **We need to address this issue with priority – there are a number of chronic issues – some “trivial” (physicist sense) that are not being addressed…**

Shifter Report (Sites Issues)

- Sat-Sun (Oct 25-26) – worst possible situation I can recall
  - 3 Tier-1s had problems (ASGC, IN2P3-CC,SARA)
    - 28 Tier2s didn't get data for 24+h
      - 21 Tier2s couldn't get/ship data, participate in production, because T1 was down
      - 2 Tier2s had problems with DQ2 SS (SS were stopped and not restarted)
  - IN2P3-CC – dCache problem
  - SARA dCache problem, bug reported to the developers
    - Data transfer backlog from T2s to T1
    - No new MC tasks will be submitted until produced data will be shipped to T1
  - ASGC is out of production for 10+ days
    - No cosmics data replication
    - No MC Production
    - No Functional Test
    - Site downtime isn't reported

Suggested Actions

- **Sites** should **spontaneously** provide information on service incidents – we have no news e.g. from ASGC about the prolonged downtime of the CASTOR services; there have been several incidents at SARA in the past days (or more?) affecting the storage services; quite a few sites rarely or **never** attend the operations meetings
- **Services** – are sites deploying services using the techniques that have been repeatedly described and / or with adequate resources? Have we documented sufficiently the service and its operation?
- **Experiments** – are there changes that can be made that expose experiments less to the inevitable problems? The situation will probably take **at least** months before any significant improvement can be seen – i.e. we may have to live with this for 2009!

☝ **Remember – things are currently very calm with respect to what we must expect when the machine is running (and after, when we have data to reprocess…)**

💣 **If we don't get this under control in the next months "sustainable operations" may be a contradiction in terms…**

Weekly Update – Experiments (1/2)

- ALICE – on-going work on WMS migration – essentially out of RB now; Continuing with the implementation of the WMS into ALICE s/w. A pilot version of the submission module of ALICE has been implemented in Torino and at CERN to ensure load balancing among different WMS per site. Presented at ALICE TF meeting;

- ATLAS – problems with file registration continued and then was solved – on ATLAS side (not LFC) ☺ Backlog drained away rapidly...; **Conditions DB access & related stress tests – on-going discussions within ATLAS and with service providers on way forward, access to conditions from Tier2s most likely requires revision to current production model – look at FroNTier / Squid, which might have (minimal?) service impact;** still some cosmic data being collected but not automatically distributed – sites have to ask; ATLASMCDISK space in CASTOR at the RAL Tier1 (see later...)

Weekly Update – Experiments (2/2)

- CMS - run "CRAFT" on-going - magnet still on since yesterday afternoon. Everything basically fine except backlog of queued transfers to CASTOR tapes over w/e every time a new run started. From CASTOR point of view "nothing wrong" - CMS trying a different way(s) of patterns of copying out data to CASTOR. CAF: problem with low free disk space. CASTOR team gave +150TB to CAF. CMS will still make an effort to delete as much data as possible. Problem with CASTOR at ASGC still not solved, in contact with Oracle global support. The global cosmics run continues until 11 November.

Sounds familiar, or just déjà vu?

- It is not uncommon for an experiment to run into a problem previously seen, analyzed, resolved by another

- More sharing of “solution” would mean less pain and a better use of the available effort

- A case in point: it seems CMS saw the ***pnfs*** overloads already and modified their applications to limit directory entries below 1000
  - Which is curiously reminiscent of DBL3/HEPDB for which directories were typically partitioned above this limit

Site Issues

- There has been a multiple disk failure in a server that forms part of the ATLASMCDISK space in Castor at the RAL Tier1. It is probable that data has been lost from this server. Approximately 4 Terabytes of the disk capacity was used on this system, and there are some 72,000 entries in the nameserver for files on this disk.

  The disk server (gdss154) has a RAID5 array with a hot spare. It suffered a double disk failure during the night of Friday 17th October. Replacement disks were ordered on the Monday for delivery the next working day. The disks didn't turn up, and the server was overlooked. There was a further disk failure on Monday (27th October) which led to failing file transfers which were noticed yesterday. Work is ongoing to see if data can be recovered from the server, but this is rather hopeful. We are reviewing our procedures to learn from this.

  The disk server forms part of the ATLASMCDISK are. A first analysis of the data on the disk shows that 90% is MC data for which this is a secondary copy. Of the remaining 10% we expect that the bulk has already been copied elsewhere. Simone - from site point of view action will be to provide list of files declared unrecoverable. Some might be in other Tier1s, some might have to be processed or produced again.

- Friday update: we lost 58732 files that were stored on the ATLASMCDISK tokens server at RAL. The loss is shared almost equally between production data and AOD replication. It is planned start cleaning the LFC by 2 pm GMT this afternoon.

Service Incident Reports

Site	Date	Duration	Service	Impact	Assigned to	Status
NDGF	18-20 Oct	2 days	Streams		Input from NDGF pending	Received
CERN	24 Oct	3-4 hours	FTS	Channels down or degraded	Gavin	Received
RAL	18 Oct	55 hours	CASTOR	down	Andrew Sansum	Received
ASGC	25 Oct	Days	CASTOR	Down	ASGC	Due
SARA	28 Oct	7 hours	SE/SRM/ tape b/e	Down	SARA	Due

- NIKHEF power cut? Other SARA storage incidents?

Things to be validated…

- *Service validation if software is changed/upgraded*
- *Specific tests (e.g. throughput) to ensure that no problems have been introduced*
- *Tests of functions not yet tested (e.g. Reprocessing/data recall at Tier 1s)*
- "Simulated" downtime of 1-3 Tier1s for up to – or exceeding – 5 days (window?) to understand how system handles export including recall from tape
- Extensive concurrent batch load – do shares match expectations?
- Extensive overlapping "functional blocks" – concurrent production & analysis activities (inter & intra – VO)
- Reprocessing and analysis use cases (Tier1 & Tier2) and conditions "DB" load - validation of current deployment(s)

71 people have registered so far – without counting speakers.

Already close to usable capacity of IT amphitheatre…

Personal remarks

- “Tier-1” issues need to be addressed
- Very prompt sites (T1s,T2s) response to GGUS tickets and emails (even to eLog entries).
  - I didn’t place any alarm tickets, but even ‘less urgent’ cases were answered by experts within several hours.

Post Script

- **We know sites are being asked to do a lot – probably too much**

- **And then we come along and say "make the services more reliable too!"**

IMHO, this second issue should be the priority – when the services are (more) reliable you'll have less stress & more time, which will allow enhancements / upgrades to be planned and scheduled

- **What are the real killers? AFAIK DM & DB…**