# WLCG Service Incident Reports

WLCG Service Coordination Team

~~~

WLCG Collaboration Workshop, 7th July 2010
~~~

# Overview

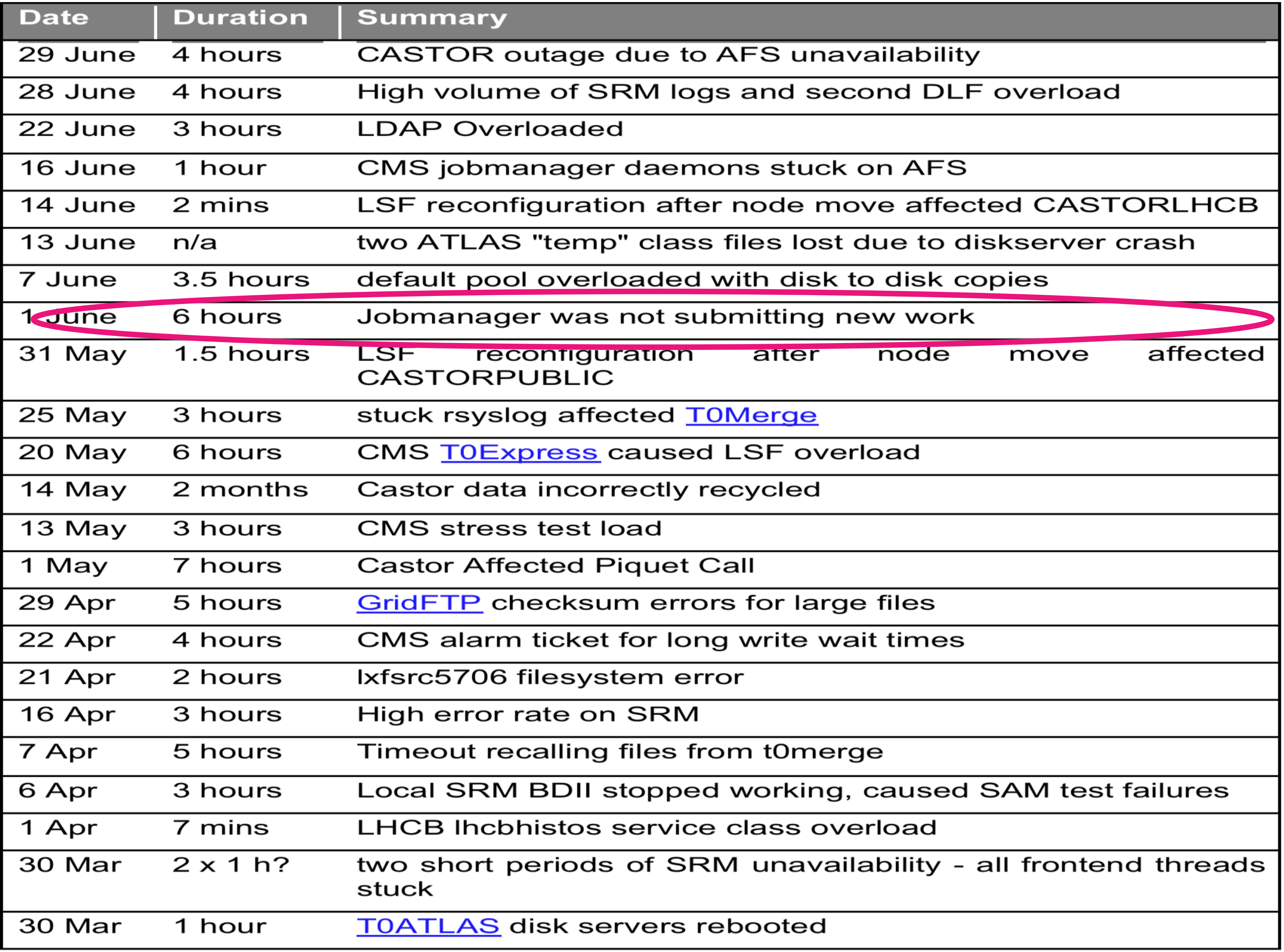
- Brief reminder of Service Incident Reports (introduced in February run of CCRC'08)…
- Summary of SIRs since LHC restart (Q2 – 1 day)
- Drill-down into 1 specific SIR
- Discussion

# Service Incident Reports: When?

- Degradation goes beyond some MoU target for any service classified as critical for at least one of the VOs !
- SCOD asks for it !
- When it’s useful for your own purpose
  - Tracking of incidents and the restoration → your knowledgebase for ***when*** it happens next time

# SIRs – Categorization

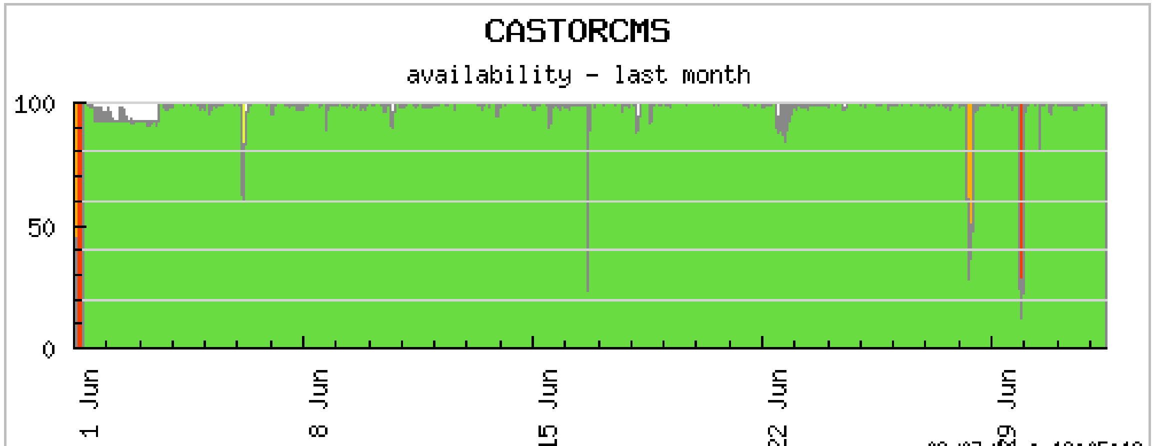
- Since the LHC restart on 30 March 2010 there have been some 20 CERN-CASTOR related SIRs, 6 (8?) DB-related and 9 others
- Quickly review these by category…



| Date | Duration | Summary |
|---|---|---|
| 29 June | 4 hours | CASTOR outage due to AFS unavailability |
| 28 June | 4 hours | High volume of SRM logs and second DLF overload |
| 22 June | 3 hours | LDAP Overloaded |
| 16 June | 1 hour | CMS jobmanager daemons stuck on AFS |
| 14 June | 2 mins | LSF reconfiguration after node move affected CASTORLHCB |
| 13 June | n/a | two ATLAS "temp" class files lost due to diskserver crash |
| 7 June | 3.5 hours | default pool overloaded with disk to disk copies |
| 1 June | 6 hours | Jobmanager was not submitting new work |
| 31 May | 1.5 hours | LSF reconfiguration after node move affected CASTORPUBLIC |
| 25 May | 3 hours | stuck rsyslog affected T0Merge |
| 20 May | 6 hours | CMS T0Express caused LSF overload |
| 14 May | 2 months | Castor data incorrectly recycled |
| 13 May | 3 hours | CMS stress test load |
| 1 May | 7 hours | Castor Affected Piquet Call |
| 29 Apr | 5 hours | GridFTP checksum errors for large files |
| 22 Apr | 4 hours | CMS alarm ticket for long write wait times |
| 21 Apr | 2 hours | lxfsrc5706 filesystem error |
| 16 Apr | 3 hours | High error rate on SRM |
| 7 Apr | 5 hours | Timeout recalling files from t0merge |
| 6 Apr | 3 hours | Local SRM BDII stopped working, caused SAM test failures |
| 1 Apr | 7 mins | LHCB lhcbhistos service class overload |
| 30 Mar | 2 x 1 h? | two short periods of SRM unavailability - all frontend threads stuck |
| 30 Mar | 1 hour | T0ATLAS disk servers rebooted |

# CASTOR SIRs - Comments

- More quantitative statements regarding service degradation would be useful – particularly true in incident summaries, but also in text body
- Analysis and follow-up: who ensures that this happens and to which bodies is this reported?
- How is this information shared?



# DB SIRs

| Site | Date | Duration | Summary |
|---|---|---|---|
| ASGC | 29 June | ~15 hours | Streams LCRs not applied from central 3D DB for 15 hours |
| CERN | 26 June | 1 hour | ATLAS offline DB (ATLR) – 9 Oracle services did not failover properly after a node eviction |
| CERN, PIC + T1s | 24, 25 June | 10 hours | LHCb streaming to PIC not working for 10 hours, to other sites not working for 40 minutes |
| CERN | 2 June | ? | ATLAS and LHCb online and offline databases – access and QoS compromised |
| CERN | 31 May, 1 June | ? | CMS online, LCGR and ATLAS offline databases – services unavailable during patching |
| CERN | 26 May | ? | CMS offline database – h/w failure affecting one node |

# DB SIRs – Comments

- Two formats used – 1 for Streams and 1 for others
  - A single format – closer to the "standard" – would be beneficial

- Several are not very conclusive in their analysis or recommendations, others still "open"
  - e.g. "Next time the problem happens detailed hang analysis/tracing will be performed"

| Site | Service | Date | Duration | Summary |
|---|---|---|---|---|
| RAL | SE | 30 June | N/A | 1083 CMS files lost |
| CERN | AFS | 29 June | 5 hours | Complete FC disk array affected CASTOR and als LHC! |
| KIT | CMS dCache | 22 June | 3 hours | Service down |
| CERN | CREAM CE | 7 June | 3 hours | Job submission failure |
| PIC | Power | 21 May | 19 hours | Whole site out |
| CERN to ASGC | OPN | 12 – 15 May | Days | Reduced bandwidth |
| CNAF | StoRM | 28 & 29 April | 9/12 hours | SRM blockage (h/w) followe by MCDISK full and StoR bug |
| IN2P3 | AFS | 26 April | 17.5 hours | AFS crashed after serv overload. Batch also affecte |
| IN2P3 | Batch | 24 April | 17 hours | Service location servic stopped responding blockin most batch syste commands |
| IN2P3 | Grid downtime notification | 20 April | 9 hours & 5 days | Grid downtime notificatio impossible after two separa incidents |

# Summary

- More consistency in formats and wording, e.g. based on that in MoU
  - Service Interuption;
  - Degradation by > 20%, > 50%

- More consistency in follow-up: introduce high-level reviews at T1SCM
  - Is information considered correct and complete?
  - Has follow-up been done?

- Try to identify areas where real improvements can be made: monitor against realistic metrics