ATLAS report on summer issues

S. Jézéquel

8 September 2010

Support on ATLAS operation

Significant and recognized

contribution from ATLAS cloud supports

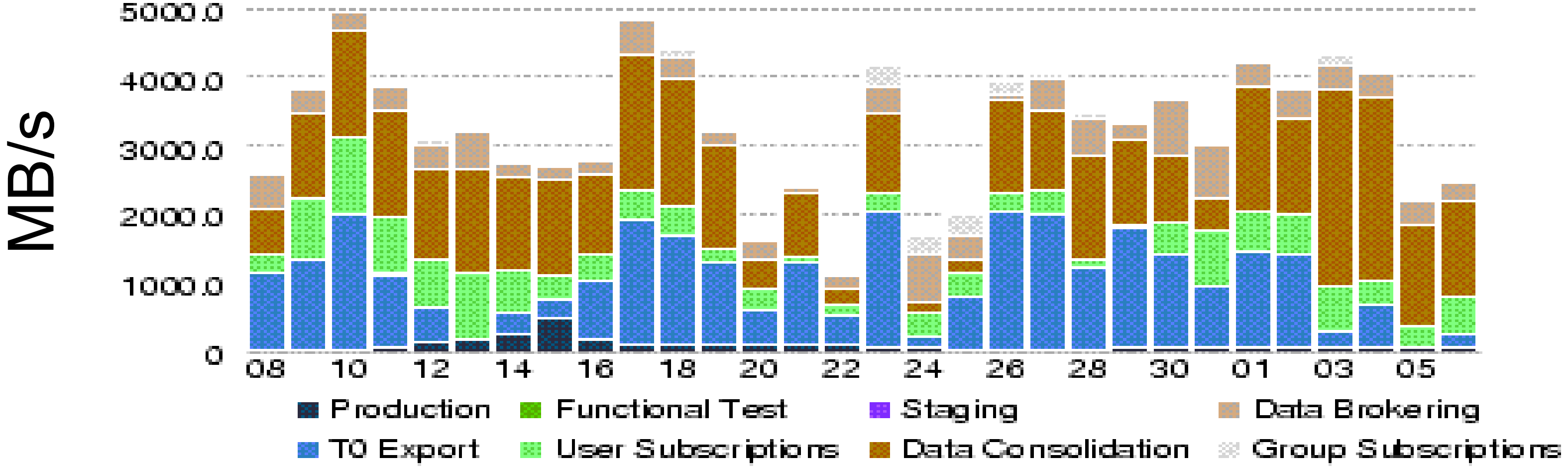
to run

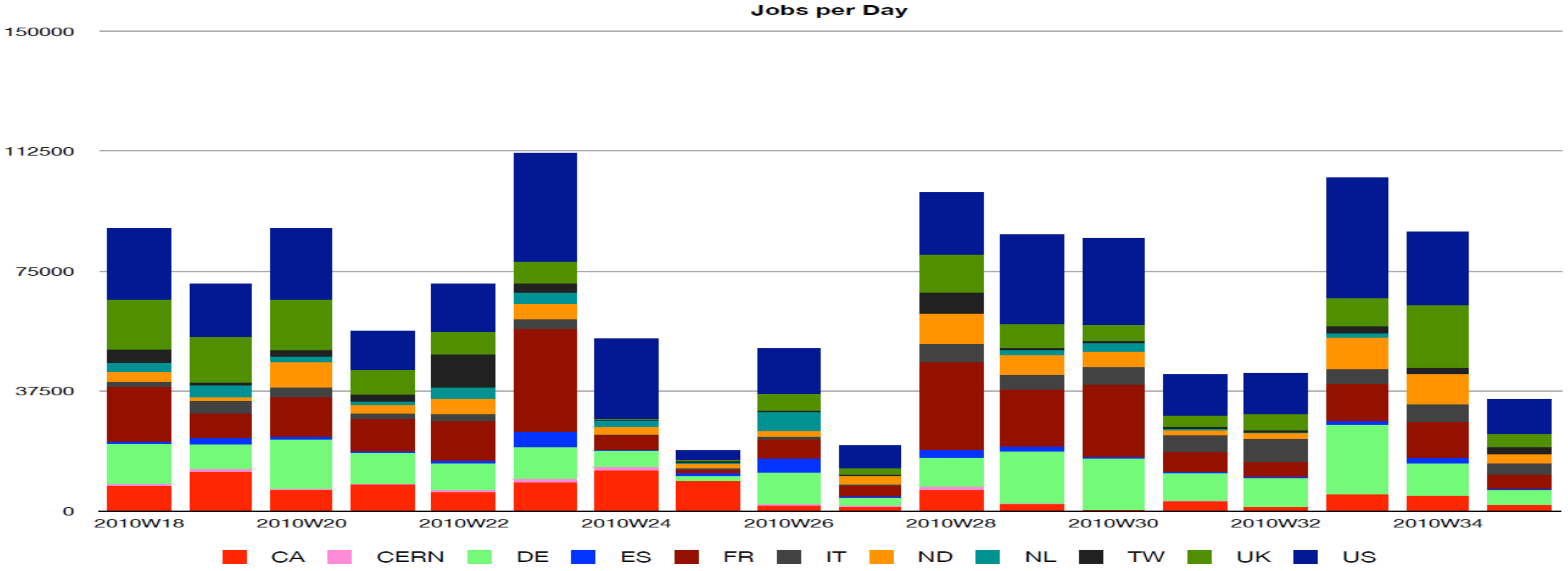
ATLAS Manager On Duty (AMOD)

- Share the load with ATLAS team based at CERN
- Improve the documentation
- After his/her period, AMOD reports:
  - A lot of work
  - Improved knowledge of ATLAS Computing Operation

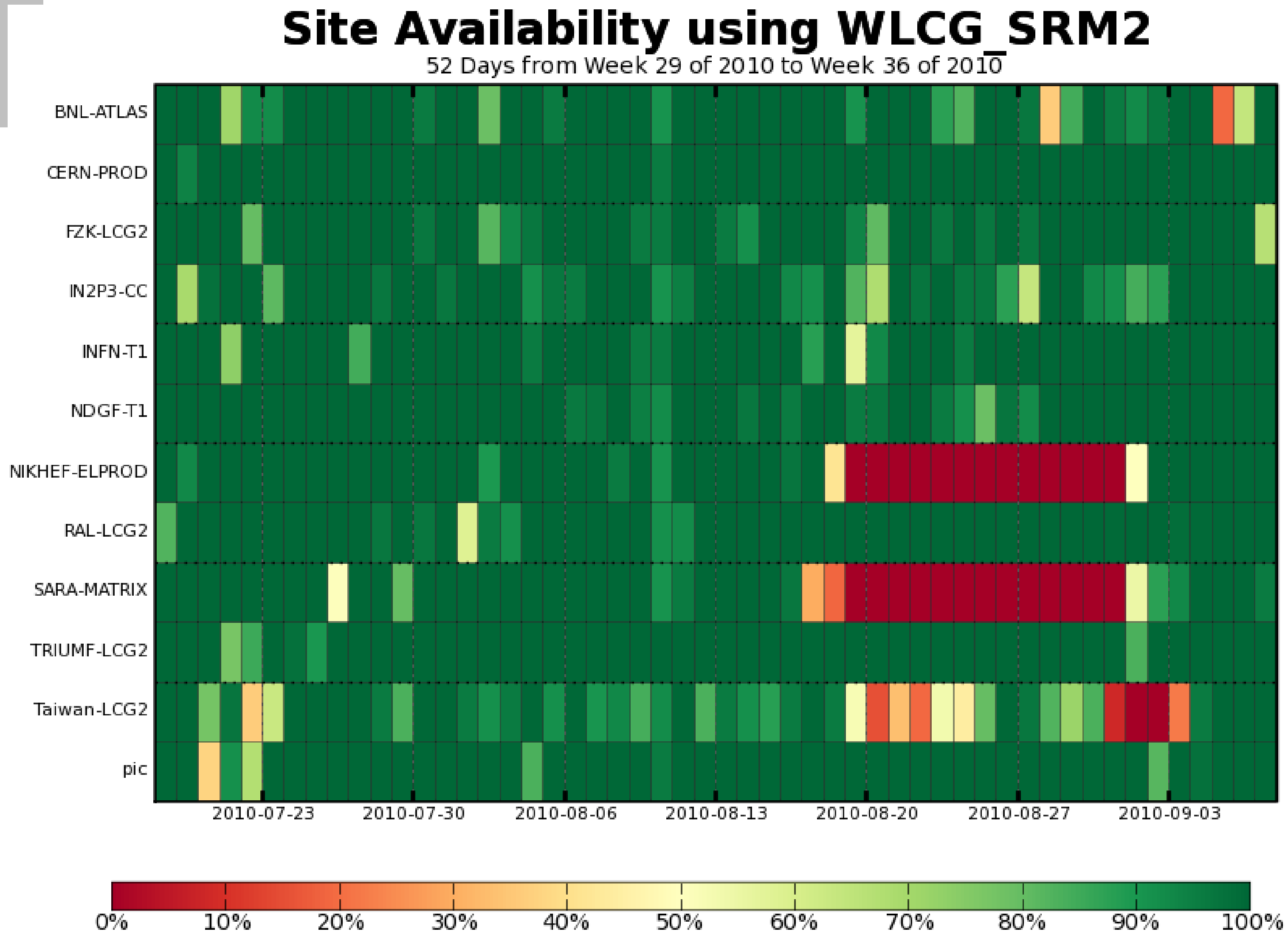
Contribution from other inactive clouds is welcomed

ATLAS activity during summer





T1 reliability



T1-T1 transfer issues

- FZK-NDGF : GGUS 60437 (24 July – 26 August)
  - Network hardware problem in FZK
- NDGF-RAL : GGUS 61306 (Started 19th August)
- BNL-CNAF : GGUS 61440 (Started 23 August)
  - Issue adressed by CNAF since mid- August

- Always difficult for ATLAS to decide who to push first
- One month delay : Summer effect ?

Checksum saga

- Castor/Storm: Share same code to compute checksum during transfers
- Many iterations to converge on reliable checksum computation
- Now OK in CERN/INFN-T1

- Castor outside CERN should/will include checksum
- Time scale to deploy checksum in Storm+GPFS/Lustre in T2s ?

RAL : Unstable storage

- In last 6 months, ATLAS has experienced storage server outage:
  - Files not unaccessible for > 24 hours
  - Most of the time, all files accessible after intervention
  - Blocking production (nasty when could be run somewhere else)
  - According to RAL, happened 6 times in last 6 months
  - ATLAS noticed 4 and usually pointed out by transfers

- Is it more often than somewhere else ?
  - Example FZK : once in 6 months (unavailability > 1 day) according to GGUS
  - Any detailed comparison between T1s foreseen ?

RAL : Broken disk server

- Beginning of August: a disk server down and never fully recovered
- Detailed report from A. Dewhurst : ADC weekly (16th August)
- Disk server expected to be recovered but failed again (after few hours)
  - Each reconstruction attempt took few days
- After discussion with RAL/ATLAS UK,
    - ATLAS provided a list of important data (to be automated): ~4000
      - Single replica useful for further activity (not log files)
    - Action on RAL was to recover some files while disk server was running
    - With list of lost files in disk server (few 10k),
      - ATLAS identified permanently lost ones and recovered others
      - In the mean time, UK cloud was blacklisted for production and data export

Unavailable disk servers: Action

- A week was lost expecting to get back the disk server
- Document to define actions for AMOD and sites under validation
  - Example : Cloud will not receive any new production task
- Automation of the recovery procedure for lost files to be deployed soon

SARA : LFC unavailability

- LFC: one of the key component for ATLAS located outside CERN
  - LFC down →
    - Unable to find path for existing files
    - Unable to register new files → No Grid computing activity
    - Affects all sites within the cloud

- Untill now, LFC complete unavailability was rare or during a short period
  - Exception : CNAF power cut → Datagard copy in ROMA

- ATLAS lacked SARA LFC catalog during 2 weeks :
  - SARA LFC not accessible in read or write mode : 18 August-2 September
  - All details reported by H.C. Lee in ADC weekly meeting (backup slides)

  - Took time to trace the problem to a hardware issue on the Oracle rack
  - Oracle backup (missing 'only' 1-2 hours) restored on a 'recycled' setup
  - Still in process to ramp up activities one after the other (ready in few days)

LFC unavailability : Actions

- Files transfered during the 1-2 hour hole were identified (ATLAS) and will be transfered again
- Even though current Oracle setup can sustain the current load, expect to get back the Oracle production setup

- Discussions started to setup a live replication of LFC catalogs
  - hosted outside the site
  - could be used for monitoring purpose
  - If master copy is unavailable, transform the replica as master LFC
  - SARA will probably be the first candidate

LFC unavailability : Actions (2)

- ATLAS will setup an action list which will be more aggressive (LFC, storage)
  - Example of failure : dedicated meeting with NL-T1 only after 10 days

Warning : ASGC

- ASGC availability has decreased recently:
  - Instabilities in Castor

- 12 GGUS tickets in the last 3 weeks
- Schedulded downtime on 31st August extended until 3rd

ATLAS already requested to help this site

BACKUP

SARA Oracle recovery

ADC weekly report (6 Sept.): H.C. Lee

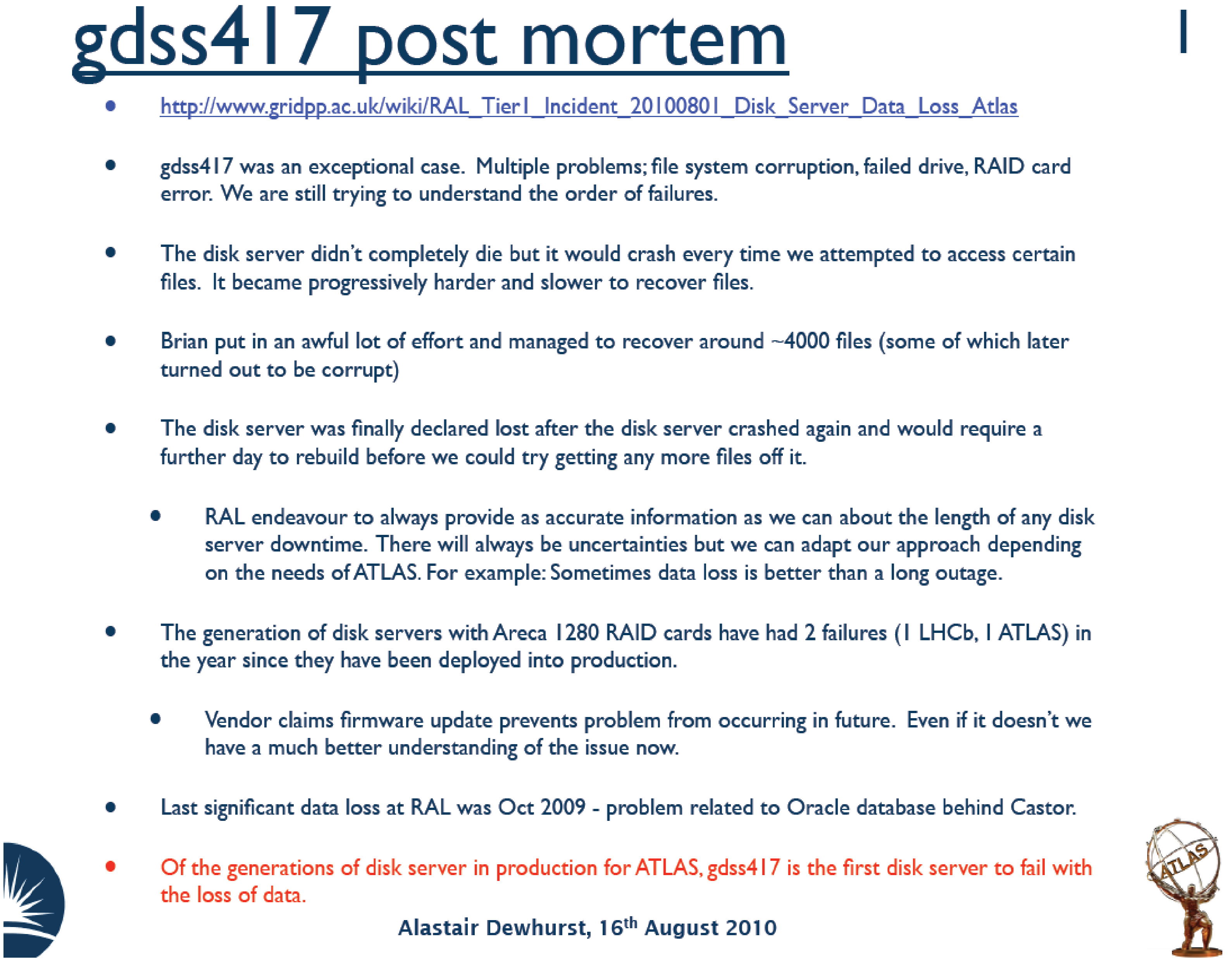
- Aug 18, 11:00 am CEST - SAN hiccup caused by a bug in the firmware of the raid controller brought down the Oracle database, affecting LFC, FTS, 3D and entire ADC activities in NL
  - neither Oracle nor storage hardware gave any addition errors except the hiccup itself
  - SARA decided to firstly bring the Oracle back online and plan for a firmware update as soon as the services are back. Usually it should just take few hours to recover ... but ...
- the following database recovery failed (logical block corruptions in backups)
  - making SARA to think there have been already corruptions in recent backups so SARA began to try earlier backups (that have been archived in tape) to save the data as much as possible ... but none of the attempts succeeded.
- 23 Aug - tried RMAN block recovery and made an Oracle service request
  - ATLAS was informed to prepare the lost of data on the corrupted logical blocks
  - database recovery still failed after RMAN block recovery
  - investigation so-far led SARA to think it's perhaps something wrong with the Oracle ASM
- 26 Aug - recalled there was a kernel upgrade on Aug 16 on the Oracle RAC that might give the trouble, tried to restore the old kernel ... the initial small scale db restore looked fine ... however, the following full scale restore failed again.

SARA Oracle recovery

ADC weekly report (6 Sept.): H.C. Lee

- weekends of 28, 29 Aug - preparing a system to restore Oracle directly on a filesystem (bypassing Oracle ASM)
- 30 Aug - database rollback directly to filesystem failed again, ruling out the kernel and Oracle ASM suspicious ... leading SARA to think it might be something wrong in even lower level: the hardware (however, there was again no indication of any hardware error)
- 31 Aug - parallel approaches for checking the hardware
  - power cycling on the Oracle RAC
  - restore Oracle database on a different hardware (a big dCache node: 16 CPUs, 96 GB memory, 20 Gb network link)
- 1 Sept - Oracle restore on the big dCache machine passed the validation. The database was rollback to about 2 hours before the crash on 18 Aug.
- 2 Sept - FTS and LFC came online and hardware issue was escalated to the Sun (Storagetek 6540)
- 3 Sept - ATLAS deletion service for NL cloud and functional test to SARA-MATRIX_DATADISK resumed. Both worked stable during the entire weekend.
- 6 Sept - full scale functional tests restored

ADC weekly report (16 Aug.): A. Dewhurst



gdss417 post mortem

I

- http://www.gridpp.ac.uk/wiki/RAL_Tier1_Incident_20100801_Disk_Server_Data_Loss_Atlas
- gdss417 was an exceptional case. Multiple problems; file system corruption, failed drive, RAID card error. We are still trying to understand the order of failures.
- The disk server didn't completely die but it would crash every time we attempted to access certain files. It became progressively harder and slower to recover files.
- Brian put in an awful lot of effort and managed to recover around ~4000 files (some of which later turned out to be corrupt)
- The disk server was finally declared lost after the disk server crashed again and would require a further day to rebuild before we could try getting any more files off it.
  - RAL endeavour to always provide as accurate information as we can about the length of any disk server downtime. There will always be uncertainties but we can adapt our approach depending on the needs of ATLAS. For example: Sometimes data loss is better than a long outage.
- The generation of disk servers with Areca 1280 RAID cards have had 2 failures (1 LHCb, 1 ATLAS) in the year since they have been deployed into production.
  - Vendor claims firmware update prevents problem from occurring in future. Even if it doesn't we have a much better understanding of the issue now.
- Last significant data loss at RAL was Oct 2009 - problem related to Oracle database behind Castor.
- Of the generations of disk server in production for ATLAS, gdss417 is the first disk server to fail with the loss of data.

Alastair Dewhurst, 16th August 2010