Follow-up Report KIT SIR Mar 2022

Timeline, impact and lessons learned

Steinbuch Center for Computing (SCC), Scientific Data Management (SDM)

Timeline – The incident struck 18th of March

09:30
- Reconfiguration of a switch pair severed connection of almost half the total disk storage from the GPFS cluster.

09:36
- dCache SEs break down due to IO errors from GPFS.

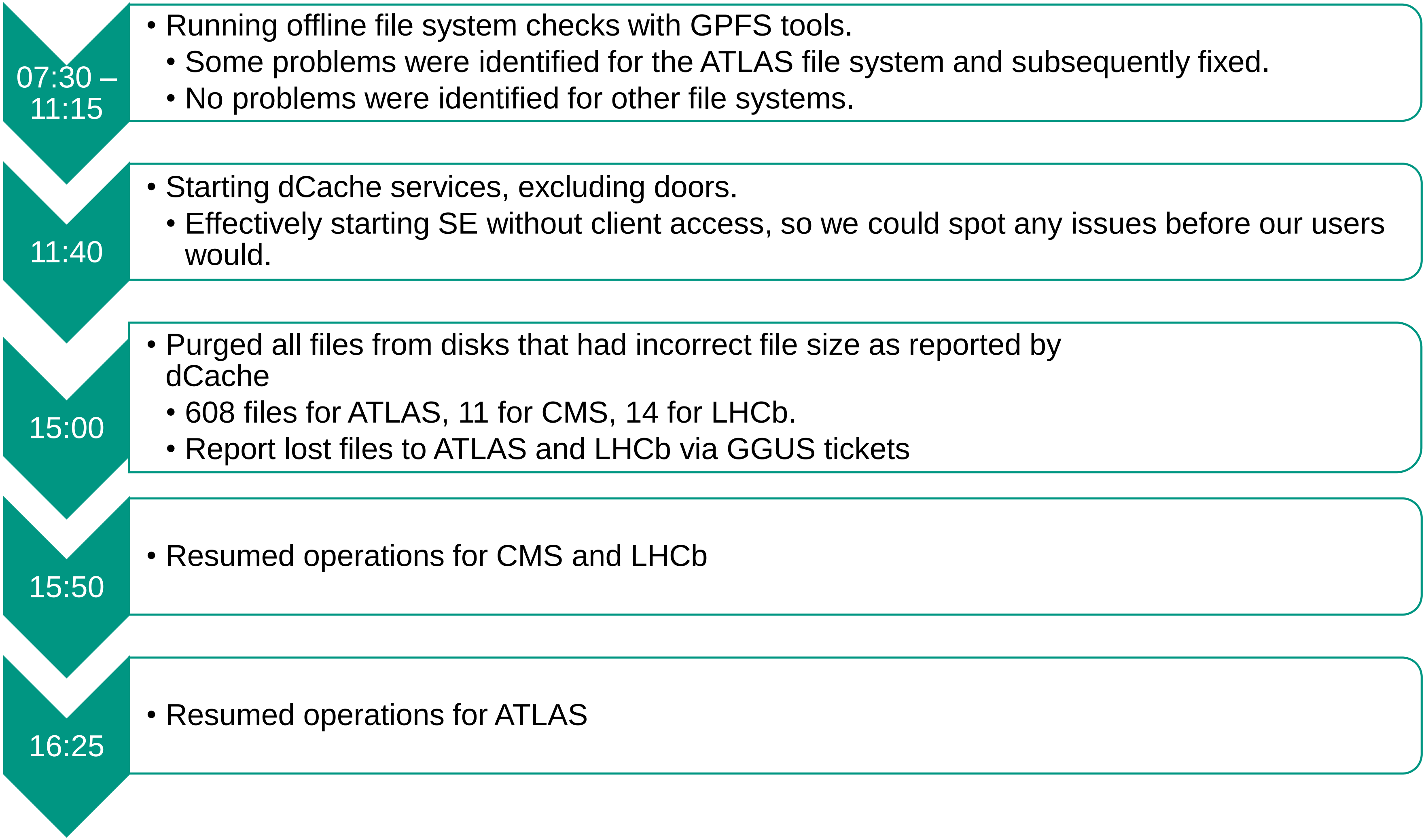
10:55
- First restart of SEs after switch configuration was fixed.

12:30
- Declared another downtime, since dCache pools and xrootd have obvious trouble with disk access.

14:20
- Second reboot of services

16:00
- Shut down SEs once more, since multiple files are found corrupted on disk.

Timeline – Slow recovery 19th of March



07:30 – 11:15

- Running offline file system checks with GPFS tools.
  - Some problems were identified for the ATLAS file system and subsequently fixed.
  - No problems were identified for other file systems.

11:40

- Starting dCache services, excluding doors.
  - Effectively starting SE without client access, so we could spot any issues before our users would.

15:00

- Purged all files from disks that had incorrect file size as reported by dCache
  - 608 files for ATLAS, 11 for CMS, 14 for LHCb.
  - Report lost files to ATLAS and LHCb via GGUS tickets

15:50

- Resumed operations for CMS and LHCb

16:25

- Resumed operations for ATLAS

Timeline – Cleanup of the aftermath

Karlsruhe Institute of Technology

21.03.

- Individual ATLAS and LHCb pools disabled themselves because more corrupted files were accessed by clients.
  - Such files were purged from disk and services resumed.

22.03.

- Focus on 30 min time frame around 09:30, 18.03. and validate file size and checksum for all newly written replicas.
  - Found another 138 files for ATLAS and 15 for LHCb.
- Started a **full consistency check** for all access latency ONLINE files

28.03.

- A checksum was calculated for all ATLAS and LHCb online files
  - Identified 5 and 28 damaged files for them respectively

Follow-up actions and lessons learned

- Because the same intervention was carried out on a different pair of switches some days before, we were less cautious this time.
  - The intervention was scheduled for a Friday and…
  - potential damage/problems in case of failure was severely underestimated and…
  - the actual changes were not reviewed by, nor coordinated with colleagues (one-man-show).
- During the last GridKa site downtime (20/21.06.2022) we ran a fire drill and demonstrated that for any of our current storage switch pairs, services will continue uninterrupted when one member is restarted.

Summerized impact

- ATLAS
  - 147 files were found corrupted due to this incident
  - 605 more (older) files were found damaged by the full consistency check, though all of them were considered „dark data“
- CMS
  - 0 files lost, 11 transfers were interrupted by the incident, which dCache correctly discarded as trashed data
- LHCb
  - 57 files were found corrupted due to this incident
  - 2 more (older) files were found damaged for unknown reasons
- Alice
  - xrootd has no database on site which we can get trustworthy meta data information from. So we only informed Alice about the incident. It is unknown to us whether Alice sustained significant data loss.