GridKa/KIT Service Incident Report 2011-07-28

Type of Incident: data loss
Location: GridKa/KIT (FZK-LCG2)
Reported by: Andreas Petzold (KIT)
Date: July 28, 2011

Description

The content of a file system was lost after a dirty GPFS file system was mounted and written to. The dirty file system was caused by a power outage in a storage rack. After the power outage GPFS did not complain about the dirty file system and online file system check showed no problems. 87% of the original 103007 files were recovered intact, after repeated attempts to run file system checks had failed.

Impact

11225 files of the ATLAS VO have been lost. A list of lost and recovered file names has been reported to the VO. ATLAS reported that some unique data stored only at GridKa has been lost. Even though the majority of files was recovered, ATLAS was forced to declare lost 43599 files, in order to speed up recovery during the critical time before the EPS-HEP conference.

Analysis

After the power outage of the storage rack, the file system was left in a dirty state. The second offline file system check silently failed and the file system could be mounted, even though it was still dirty. This is the original cause of the problems that occurred afterwards. From this point on, any write operation to the file system likely caused data loss.

Further online file system checks reported no problems and operators enabled the affected dCache pool in r/w mode, which immediately caused further loss of data. Location information regarding some of the affected files was lost from the dCache system, due to the pool checking its current inventory, which hampered later recovery efforts.

It was not possible to bring the dCache pool online with the file system being mounted in r/o mode. If possible, this would have made $>$80000 files immediately available to ATLAS again. It would be very helpful, if in the future dCache provided the possibility to bring a pool back online in an emergency mode on a r/o file system. This feature request is already being implemented by the developers.

GPFS support has been involved, but the problems cannot be diagnosed with the version of GPFS that is currently installed on that specific GPFS cluster. An update of GPFS will be scheduled to continue the investigation of the cause of the failing file system check.

Timeline

12.07.2011
- blown PSU in storage rack causes circuit breaker trip
- rack off, since rack is not on UPS, power restored
- communication with DDN
- first offline file system check reports errors, second offline check silently fails and file system can be mounted
- online GPFS file system checks report no problem
- dCache pools restarted
- f01-032-105_1D_atlas reports incorrect free space (4TB) - first indication of trouble

13.07.2011
- dCache team investigates f01-032-105_1D_atlas pool problems: inconsistencies between pool inventory and file system

14.07.2011
- I/O errors discovered on file system

15.07.2011
- offline file system check is run, reports problems, 1.6TB lost
- DDN support excludes any hardware problems

18.07.2011
- GPFS support involved
- further attempts for file system checks crash GPFS cluster

19.07.2011
- ATLAS declares lost and re-imports affected files with certain file name pattern to help recover stuck jobs

20.07.2011
- attempts to bring pool back online in r/o mode fail
- start to copy accessible files to new file system

21.07.2011
- ATLAS declares lost more files

22.07.2011
- file migration to new file system finished; 400 I/O errors
- start check sum validation on new pool

25.07.2011
- check sum validation finished; 1 bad file
- new pool online
- ATLAS keeps on deleting files

26.07.2011
- more files recovered from recently decommissioned h/w
- final numbers: files lost 11221, files recovered 89399 (43599 declared lost and re-imported by ATLAS)