GridKa/KIT Service Incident Report 2011-06-05 (All times in CEST)

Type of Incident: data loss
Location: GridKa/KIT (FZK-LCG2)
Reported by: Jos van Wezel
Date: June 05, 2011

Description

The content of a filesystem was lost after a third disk failed during rebuild of a degraded RAID6 disk-array. Analysis showed the failure one hour after rebuild automatically started. The array is part of a set of arrays that are logically grouped into a single GPFS filesystem. GPFS is configured to keep redundant metadata copies. With the inode information and the fact that content outside the broken array is not affected we were able to list all affected files and copy the non affected files to spare storage.

Impact

Loss of 4356 files of ALICE VO. List of file names was reported to ALICE. All data has been recovered by copying files from other sites.

Timeline

06-04-2011 05:58 broken HDD reported by monitoring software
06-04-2011 06:02 second broken HDD reported by monitoring software
06-04-2011 06:04 rebuild is not started because of faulty ADP unit
06-04-2011 09:25 operator contacts vendor and receives instructions to recover
06-04-2011 10:30 Vendor diagnoses remotely. Array is in degraded state
06-04-2011 10:30 Operator is instructed to replace HDD
06-04-2011 11:00 Nagios reports logical disk being down via GPFS trigger
06-04-2011 11:15 Vendor is contacted. Remote diagnoses shows problems with
on-board circuitry (ADP-unit)
06-04-2011 16:00 Problem escalation to manufacturer
07-04-2011 09:00 Notified ALICE contacts
11-04-2011 11:37 Report from vendor that content of LUN is lost
11-04-2011 12:30 KIT asks vendor for detailed analysis.
19-04-2011 14:00 Replacement of ADP unit, recovery of content started
20-04-2011 12:00 Storage gradually placed online
31-05-2011 16:31 After several mails between vendor and manufacturer
It is finally concluded that a third disk failed.

Analysis

The cause of the problem was several broken disks in the same disk array. For still unknown reasons 2 disks failed within minutes. The controller started using a third parity disk but it was flagged faulty with read errors. Surface/media errors on disks may go undetected until the content is actually read during rebuild of a degraded RAID array. The problem was amplified by the fact that circuitry in one of the disk enclosures, an APD unit, was also reporting errors. The ADP unit had to be replaced before further analysis and recovery could be performed. Spare parts were difficult to obtain for various reasons and hence repair and subsequent recovery were delayed. The vendor repeatedly contacted the manufacturer in Tokio but was unable get further information about the specific software actions taken by the controller unit on failure detection. Especially the effect of the broken ADP unit during the automatic rebuild remains unclear an may have had an effect.

KIT would like to that the contacts of the ALICE collaboration,Kilian Schwarz, Latchezar Betev and Christopher Jung for their help during analysis and recovery.