TRIUMF Tier-1 Incident Report

Lost files on September 16 2013

Description

We experienced an unexpected behavior on a DCS3700 storage system when using the IBM Storage Manager tool to reorganize LUNs and prepare the system for a new storage layout. Several LUNs (10) disappeared from the production host group. The LUNs were manually restored back into the production host group right away. However, the corresponding XFS filesystems were damaged, causing files to be lost.

Impact

In total 93,832 files are lost (16 TB) out of 1,561,849 files stored on the affected storage unit. Four LUNs were verified to be good and were remounted on the hosts; but 6 LUNs needed a repair (39 TB each, 234 TB in total). These problematic 6 LUNs were hosting 828,681 files on them that were not accessible for some time; the LUNs belong to our disk group, providing dcache pool service for the space token areas: ATLASDATADISK, ATLASGROUPDISK ATLASLOCALGROUPDISK, and ATLASSCRATCHDISK. Both ATLAS production and analysis jobs were affected by the missing files.

Time line of the incident

All timestamps below are in the Vancouver time zone.

The incident happened at 13:38:45 September 16 2013. LUNs remapping was done at 13:40:15, followed by a few actions: dcache pool service stop, filesystem umount, multipath stop and SAN devices rescan etc. We noticed that 6 LUNs could not be mounted because of unclean XFS filesystems.

At 13:57:49, on 4 out of 6 LUNs an XFS repair was initiated at different time within the next few hours, the result showed that XFS put all the recovered files into lost+found directory.

At about 16:00, adler32 script started to check all the recovered files. Also, an initial assessment about data potentially lost was in the range between 25k and 164k for best and worst case scenarios.

At about 18:00, we started to prepare another tool for second recovery method, which goes through the inode tree to recover files.

At different time around midnight on the same day, the recovered files were being rsynced to other storage systems.

At 07:20 on September 17, an assessment regarding the worst case scenario was that 164k files would be lost.

At different time on the night of September 17, 4 LUNs files started registering to dcache on other storage nodes. The secondary file recovery started on those 4 LUNS, but was not successful. We started the primary recovery procedure on the remaining unclean 2 LUNs.

In the morning of September 19, all recovered files were back online. The final list of lost files was then known (see table below). We examined the final list and issued a savannah ticket to ATLAS DDM operations. By looking closely at the affected datasets it revealed that about 19k files have replicas on the Grid and could be recovered by DDM from other sites. Another ~18k files appear to be “dark data” without catalog information but this needs further checks from ATLAS. We have also alerted affected users individually and provided them with their list of lost files and jobset id's.

Analysis

The problem was caused by a SAN interruption, causing the XFS superblock on several LUNs to be damaged.

The first XFS recovery procedure has been used at the site for several years. Unfortunately this time around it did not fully work. To do further repairs, a secondary recovery tool was attempted but was not successful.

We are revisiting our procedures to avoid loss of data in the future. In particular, we are exploring to perform regular backups of XFS file system meta data. We will continue also to work on testing an XFS recovery tool for future considerations.

The summary of lost files

Space token	Number of files on damaged file systems	Files lost after repairs
atlasdatadisk	756,401	73300
atlasgroupdisk	10,584	77
atlaslocalgroupdisk	14,263	705
atlasscratchdisk	47,370	20287
Others	63	2
Total	828,681	94371