Lost files at TRIUMF due to hardware/firmware issue

Incident description and summary

On December 18 2016, at 6:30am, one IBM dcs3700 storage unit (180*3TB disks, 12 arrays) had an issue. Following a disk failure in one of the arrays and while a spare drive was taking over, the controller issued a power cycle on the failed disk, thereby triggering a firmware bug. This caused a parity problem on the array and the LUN became inaccessible from the host. The affected LUN has a 36TB capacity and contained 30TB of ATLAS data in 185,933 files. After extensive efforts in trying to recover and repair the LUN while working with the vendor, we managed to recover most files, but unfortunately, in the end, 2921 files were not recoverable and had to be declared lost and with no other replicas found on the grid. The data volume lost is close to 500 GB.

The current firmware was updated in the summer of 2016 to 8.20.12.00 which has a bug in it as confirmed by the vendor. The latest firmware versions from 8.20.20 have the bug fixed.

Vendor explanation

To reiterate the sequence of events as explained by our Product Field Engineering team, on December 18th at 6:30 AM, drive 1/4/2 hit a Predictive Failure Analysis (PFA) threshold and was transitioned to an impending failure state. Before the drive was marked failed, the controllers started a copy-on-fail operation and at this same time, before the drive was able to copy the data off to a hot spare, the controllers triggered a drive power cycle due to the drive timeouts. These two simultaneous operations resulted in a write failure on the portion of the stripe intended for the piece that drive corresponded to. This resulted in redundancy becoming inconsistent and when detected by pre-read redundancy, the entire stripe was marked unreadable.

Repair and recovery actions performed

On December 25, IBM repaired the problematic array, and verified no parity errors found at the hardware array level. However, the xfs filesystem could not be mounted and its Structure needed cleaning; and we have made further recovery requests to vendor.

In parallel, while doing the array parity verification process, we started a file system duplication, and we recovered 149,167 files (out of 185,933) from the duplicated filesystem image.

On Jan. 3rd, IBM confirmed this incident was caused by a firmware bug, and no rollback action can be performed. The problematic array needs reformatting before putting it back into production.

On Jan. 4Th, we declared an initial batch of about 17k files bad to ATLAS since those files had another replica on the grid and they were recovered successfully.

After another several days of further work, we managed to recover another 14,232 files from the problematic LUN after a second round of filesystem image duplication. Files were unique to TRIUMF.

In the end, we had to declare 2921 files lost from this incident (~500 GB of ATLAS data).