Post Mortem Tier-1 Service Incident ZFS Corruption in a DDN dCache Pool 20110209

From WikiPIC

Incident Start: 21st January 2010 17:00 CET

Incident End: 8th February 2011 around 17:00 CET

Contents

Description

File system corruption was detected in two ATLAS dCache pool servers, this disabled both dCache pools (250 TB of data). Software (SUN/ORACLE) support was contacted and started working on this issue while in parallel at PIC we started to work out the evaluation of the affection either at dCache pool level and file system level. ZFS filesystem could be partially recovered hence giving the chance to serve about 60% of the affected dataonly few days later, while the migration to a safer place was in action. At the end few hundred files were lost due to corruption, but during these two week recovery process about 40% of the data was not online.

N.B. The affected dcache pools were implemented on HP blades running Solaris and mounting the disks via a Fiberchannel SAN. This SAN is based on DDN S2A9900 solution (http://www.datadirectnet.com/9900)

Impact

868.472 files (about 250 TB) were offline for ATLAS during 17 days, affecting data processing and data consolidation. Data was slowly taken online but only had the final picture (unique files/lost files) on the 8th of February. Cloud was brokeroff during some days and data processing workload was manually steered by ATLAS once they knew the affected files. The net cost was the loss of 1136 files that could not be recovered from cached/tape local copies.

Time line of the incident

Incident started cooking on 18th January 2011:

- 18/01/11 16:00 - After an scheduled downtime one controller from each DDN controller couplet (cabddn1(a|b), cabddn2(a|b)) failed to boot. Hardware support (Omega/DDN) contacted and confirmed a HW failure (cabddn1a and cabddn2a). System is online but not redundant. Waiting for two new controllers to replace the failed HW.
- 19/01/11 12:00 - Only one DDN controller is on local stock, cabddn1a is successfully replaced by the new one with the same firmware version. DDN controller couplet cabddn1 is fully redundant again (transparent intervention).
- 21/01/11 11:00 - Two RMA DDN controllers arrive, one is setup as the faulted cabddn2a replacement but this RMA comes broken. The second RMA DDN controller is installed but can not boot due to newest firmware in the RMA controller. The new (RMA) cabddn2a is powered but software faulted because of firmware mismatch (the controller is not active). Since it is Friday, the Omega/DDN technician modifies the FibreChannel (FC) switch to work with port based mapping instead of WWN based mapping (as has been in all 4 FC switches for over a year). With port based mapping it should be possible to (manually) switch to the faulted controller if the active fails. All this operations are transparent and system is working OK.
- 22/01/11 10:00 - First IO errors detected by ATLAS shifters (GGUS 66409)
- 23/01/11 15:30 - Massive file system corruption detected by the system in an ATLAS dCache pool server: dc005_1. dCache pool goes offline (disabled) due to [too many] IO errors. ZFS disk pool is unusable (not possible to read/write data).
- 24/01/11 12:30 - Massive file system corruption detected by the system in an ATLAS dCache pool server: dc004_1. dCache pool goes offline (disabled) due to [too many] IO errors. ZFS disk pool is unusable (not possible to read/write data).
- 25/01/11 whole day and 26th early morning - Standard SUN/Oracle procedures performed (Solaris upgrade, ZFS debugging commands, etc) but not succeeded.
- 26/01/11 13:00 - SUN/Oracle engineer (Victor Latushkin) started looking at the issue via the SunSharedShell (remote shell).
- 26/01/11 15:22 - It is clear to the SUN/ORACLE engineer that the corruption is due to the disk array controller, HOST:LUN crosstalk is proven (information from HOST_A:LUN_X written in HOST_B:LUN_Y).
- 26/01/11 21:00 - Both dc004 and dc005 dCache pool ZFS disk corruption is partially recovered and now it is possible to mount the FileSystems using a Solaris 11 LiveCD (from this moment, 60% of the affected files become online and accessible for ATLAS). Started copy of the recoverable data to NFS servers, building the "failed to copy" list in order to know which files are lost. Copy is working at ~150MB/s per server, so it is expected to copy all the files (and get the list of lost PNFSIDs) in 9.6 days.
- 27/01/11 16:00 - NFS copy for both dCache pools data continues working OK, list of unrecoverable PNSIDs grows slowly (meaning that most data is being recovered). Failed attempts to serve the data via dCache, the system doesn't want to work with a ReadOnly FileSystem.
- 28/01/11 15:30 - The dCache pool metadata is not corrupt (BerkeleyDB). dCache metadata is copied to a production dCache server (in a dedicated path) and data directory from the broken dCache pools is mounted via NFS in ReadOnly mode. Finally the pool refuses to start because the I/O test fails. dCache.org is contacted (RT#6086) to see if there is any workarround.

- 28/01/11 17:21 - A list of affected dCache files is provided in GGUS#66409. The list was not available before because the conversion from PNFSID to path+file without affecting the production PNFS was costy (list was finally generated in a distributed fashion, using the local PIC computing farm).
- 28/01/11 18:00 - dCache.org provides a patched binary which does not perform the I/O test. Special thanks to Paul Millar and Patrick Fuhrman.
- 28/01/11 20:40 - dCache pools dc005_1 and dc004_1 are ReadOnly and online in dCache but less than 30% of the data contained in dc005_1 is seen by dCache because the data directory is corrupted (so the PNFSID list is not complete when the system asks for it). On dc004_1 NFS export is paused and a dCache migration process started. NFS data export from dc005_1 continues working.
- 28/01/11 23.50 - dCache pools automatically disable every now and then due to I/O errors while trying to read data files. A workarround look which sets the dCache pools as readonly (not disabled) every minute is deployed.
- 30/01/11 22:50 - Further details on the extensive unavailable file list provided in the GGUS#66409
- 02/02/11 09:36 - Provided file list of recovered files to the moment in GGUS#66409. File migration via dCache migration/NFS copy is still ongoing.
- 03/02/11 15:42 - Status report of the file migration provided by PIC in GGUS#66409. Expected 80 hours to finish the file migration and get the list of files lost by corruption.
- 04/02/11 09:28 - Priority file list to be recovered provided by ATLAS in GGUS#66409
- 06/02/11 15:43 - NFS file copy process from dCache pool dc005 finished successfully. List of lost files in this pool is available. Migration of data from dc004 using dCache migration module is still ongoing; list of lost files from this pool will be available only when migration ends.
- 08/02/11 08:44 - Provided partial list of lost files in GGUS#66409
- 08/02/11 15:02 - File migration of dc004 ends, additional time was required due to orphan files migration error (dCache tries to migrate files which no longer exist in the PNFS, this triggers errors and wastes time).
- 08/02/11 16:46 - Complementary and finel list of lost files provided in GGUS#66409
- 08/02/11 17:03 - Besides the lost files we still had some offline data (from the broken dCache pool dc005_1). After talking to the dCache.org developers in the dCache Tuesday's weekly meeting (this afternoon) a way to bring the data online was provided. At the moment all data is back online. The dCache servers have been a bit loaded for the last hour, but everything looks OK now

Details

- Investigation from SUN/ORACLE engineer determined that file system corruption in both servers happened between 17h and 19h CET on January the 21st due to "LUN crosstalk"; data which should have been in a LUN from dc004 was in dc005 and vice versa.
- https://gus.fzk.de/ws/ticket_info.php?ticket=66409
- http://rt.pic.es/Ticket/Display.html?id=839 (PIC internal)
- http://rt.pic.es/Ticket/Display.html?id=914 (PIC internal)
- http://rt.pic.es/Ticket/Display.html?id=915 (PIC internal)

Analysis

The degradation of services due to the controllers failures after the Scheduled Downtime lead to a hardware

replacement that triggered the mixing of data in different LUNs, zpools detected this crosstalk, ZFS got corrupted and the pool was marked as faulty resulting in a I/O failure for these 250TB. Once discovered this evidence, efforts were concentrated in the recovery of the zpools and minimize data loss. That was patched by SUN/ORACLE engineer and after this we started the migration of data out of the affected pools to a safer place, while in parallel PIC engineers were working out list of affected files and list of real lost files.

Now the priority is to understand the two issues that remain as a mystery: a) blown up of chip controllers at poweroff time and b) the cause to trigger the crosstalk between LUNs.

Follow-up

There are two issues that need clarification and are being investigated.

The first one is the hardware failures during the scheduled downtime poweroff, two controllers broke (PCI chipsets blew up). Affected hardware was shipped to US and is currently under forensics services in DDN headquarters.

DDN engineers believe that controllers are breaking at PIC because of high temperatures inside the controllers after shutdown. In order to solve this issue the controllers that now are directly attached to the disk enclosures will be moved so that at least one U of free rack space is left between the controllers and the disks. Alsso a firmware upgrade to version 6.10 has been done, v6.10 has some cooling improvements.

On the other side the source of the ZFS corruption is not understood by DDN. SUN/ORACLE engineer pointed out a disk array failure, now all parties are collecting information as something strange happened at some point when the controllers were replaced. Both issues are treated with high priority as knowing the source of the corruption is vital to take the proper actions to prevent another ugly disruption like this.

DDN engineers revealed that similar problems were observed in other S2A9900 deployments. This issue can show up when Fiber Channel paths are modified without notifying the S2A9900 controllers, in PIC's case this happened because Fiber Channel switch zoning configuration was modified from WWN based to port based without service disruption, after the change the FC attached servers have to reloguin but the controller doesn't rediscover the paths again. This issue is solved in S2A9900 fimrware version 6.10, now initiators attempting to issue commands without a login will receive LOGO.

A different but correlated issue is the difficulty to extract the list of files present at a certain pool, this requires a dump of the PNFS DDBB and it's translation to Storage File Names and in a different dimension the extraction of real lost files, that can only be achieved once the data is attempted to be read. Chimera would help in diminishing the latencies, but we also want to investigate the feasibility of having a background process to have a local DDBB with file/pool mapping. Informing the VO is critical in this cases so they can react either on data management and processing workflows.