Service Incident Report

By: Onno Zweers, systems programmer, SURFsara.

Description

This is a report about service instabilities in the SURFsara grid storage cluster, that started with a power outage on October 24, 2013. The power outage had left several storage controllers and disks broken and filesystems damaged. Some of the replacement controllers also broke. We have 14 such controllers; 8 have been replaced as a result of this power outage.

Situation

The instabilities occurred in a group of seven racks. Each rack has the following layout: 4 servers, attached to 2 DDN (Data Direct Networks) storage controllers of type S2A9900, and 5 disk enclosures attached to the controllers. Each server had an XFS file system of 102TB in a RAID6 compatible redundancy. Each server, controller and enclosure is attached with a redundant (double) power supply. The controllers were named hive7 until hive20. The servers were named bee34 until bee61.

Lessons learnt

- xfs_repair has improved over the years; in all cases it was able to make unmountable filesystems mountable; in only one case the root inode was damaged, but this could be fixed by moving the main directories back from “lost+found”. We could not prevent some file level data loss though.
- Procedure for fixing XFS errors:
    - Try to mount the filesystem; if successful, it will replay the logs.
    - If that fails: xfs_repair -n (dryrun, does not fix anything yet). This will instruct what to do next. That could be a xfs_repair -L, which throws away the log entries without replaying them. This could mean data loss.
    - If the filesystem was fixed: move all data away to a fresh filesystem. This will show errors in individual files. If dCache is used to migrate data away, be sure it does adler32 checksum checks.
- xfs_check is worthless on really big filesystems, because its RAM requirements cause system hangups.
- Copying files into a glusterfs filesystem based on 10 servers and replica factor 1, turned out to be slow. Glusterfs only reached up to 500 MB/s, while dCache could migrate files at almost 1100 MB/s on a 10Gb/s interface, checking the checksums on the fly. In both cases there were 4 to 8 parallel transfers.
- dCache refuses to use a filesystem that is read-only mounted. This can be painful, because dCache is the preferred way to migrate data away from unstable filesystems. You can set pools to read-only in dCache; but still dCache wants to be able to write in the storage filesystem.
- When power comes back on, the peak usage can be higher than expected; enough to blow a fuse. The power provisioning in these seven racks was perhaps not sufficient; in two of the seven racks, the remaining feed turned out to be insufficient to keep all systems running. We have started a review of the power provisioning in all our systems.
- DDN S2A9900 controllers seem to be sensitive to power issues. Even a normal restart on old controllers can be dangerous; it can reveal a hardware failure that is not noticeable during normal operation. A DDN SFA10000 controller that is part of our HPC cloud cluster had many issues, thus keeping our cloud admins

Postbus 94613
1090 GP Amsterdam

Science Park 140
1098 XG Amsterdam

T +31 (0)20 800 1300
F +31 (0)20 800 1401
E info@surfsara.nl
www.surfsara.nl

BANK 66.41.95.067
IBAN NL29 IN GB 0664 1950 67
BIC INGBNL2A
BTW NL8226.55.287.B01
KvK 50277375

busy for many days. The DDN support was very good though. Hardware of other brands in our datacenter had no serious issues.
- Some engineers are great. Jan Schwaratzky of DDN is one of them.

Timeline of events

Times are in 24h notation CET.

- 2013-10-24, 6:03: A power outage occurred in feed1.
- 2013-10-24, 7:27: We arrived and started investigating.
- 2013-10-24, 7:45: Power was restored.
- 2013-10-24, 8:48: We discovered that in two racks (with hive7/8 and hive11/12), the disk enclosures were off. In these racks, the fuses of feed2 were blown. We assume that power fluctuations in feed1 caused a surge on feed2. We switched the fuses back on. Then we switched on the enclosures and power cycled the controllers. Three controllers showed a “bootup failure”. The other, hive11, remained dead.
- 2013-10-24, 11:44: DDN shipped 4 controllers.
- 2013-10-25, 14:25: DDN engineers had replaced the 4 controllers hive7, 8, 11, and 12. Mounting the filesystems proved difficult. Some filesystems could not be mounted because of “can’t read superblock” errors. Other systems required an “xfs_repair -L”, which clears the metadata log, which can cause some data loss. Our slightly dated experience with xfs_repair was that it could wipe filesystems clean, so we proceeded with caution, one filesystem by one. We ran some “xfs_check” commands first, but they caused server hangups, probably because it tries to load filesystem info into RAM, and 96GB RAM proved to little for 100TB filesystems. Some systems returned “input/output error” or “sense key: medium error”. We reported these to DDN at 16:24.
- 2013-10-25, 17:14: We noticed a critical LUN in hive7. Rebuilding of the two failed disks was unsuccessful. We started copying data from the attached server to another, to have a backup. We had started adler32 checksum checks but they slowed down the copying of data so we cancelled the checksum checks on this server.
- 2013-10-28, 9:47: 5 out of 8 affected servers showed no issues with adler32 checksums. On two other servers, there were IO errors on a few dozen files. One node, bee44, still could not mount its filesystem; for this, we requested vendor support.
- 2013-10-28, 13:58: Bee37, a server attached to hive8 could not mount its filesystem anymore. We rebooted it. After that, the filesystem was mounted.
- 2013-10-28, 21:07: We created/enlarged pools on other storage systems to migrate data away from the affected unreliable filesystems. Since we have configured dCache to perform an adler32 check on every operation, this meant that migrated files were not corrupt. DDN was able to have us mount the bee44 filesystem, but strongly recommended to mount it read-only. Because of this we could not start dCache on this node. Also, to make this possible, DDN had to disable one disk enclosure (two disks per RAID6 configured LUN). This meant that data had lost redundancy. We started copying the files at risk to a glusterfs filesystem.
- 2013-10-29, 10:12: Bee37 again had IO errors. A restart fixed it.
- 2013-10-29, 13:02: User activity considerably slowed down the migration process. We temporarily switched off the SRM door, to give all priority to the migration.
- 2013-10-29, 21:04: Bee37 again had IO errors. We reported this to DDN.
- 2013-10-30 (night): A DDN engineer accidently cut of the storage from three servers. A reboot the next morning fixed it.
- 2013-10-30 14:38: DDN could not access hive9 remotely. They asked us to powercycle it. Foolishly enough we did that, and then it got a bootup failure. It had to be replaced before the two connected servers could mount their filesystems.
- 2013-10-30 22:04: Hive9 was replaced and the two attached servers were back in production. Bee37 had crashed again. Another reboot fixed it.
- 2013-10-31: Lots of data migrations to get the data to safe filesystems, and have dCache do checksum tests. After some days, a few dozen files had to be declared lost.
- 2013-11-01: The copying of data from bee44 without using dCache proved to be slow and impractical. We decided to ignore DDN’s advice and mounted the filesystem RW; then we started dCache and migrated the files to other pools. This succeeded without problems.

- 2013-11-01, 16:21: Hive9 collapsed again with “Disk slot GH, Invalid long response CRC”.
- 2013-11-01, 23:35: DDN started a rebuild of 4 disks attached to hive9. The two attached servers could not mount the storage: “structure needs cleaning”. We decided not to do an “xfs_repair -L” to delete the log yet, but to wait until after the weekend, because we were uncertain how risky it would be.
- 2013-11-04, 9:40: DDN got hive9 operational again. We ran an xfs_repair -L on the two filesystems and that got them in a usable state.
- 2013-11-04, 16:24: In our haste to migrate data we had forgotten to create new Atlas tape pools for the ones we had deleted. After Atlas reported this, we quickly created new tape pools.
- 2013-11-05: Official end of our support contract for this hardware. DDN never even mentioned this fact and kept on providing valuable support.
- 2013-11-15: File migration and sorting out file issues continued. We found that bee37 was still crashing with IO errors. In the afternoon, hive8 (attached to bee37) turned out to be broken. DDN had notably more difficulty to find replacement controllers. No more were available in Europe, so this one had to be shipped from the USA. One would almost think we had exhausted DDN’s European spare supply.
- 2013-11-21: The new hive8 was operational.
- 2013-11-22: The new hive8 was broken.
- 2013-11-25: DDN shipped two new controllers from the USA. The ETA was a few days later. We still had a few servers that were empty after migrations, and we offered DDN to use one of these controllers (hive11/12) to speed up the process. DDN gladly accepted.
- 2013-11-26: DDN replaced hive8 with hive12. One server was back up; the other had media errors.
- 2013-11-28: DDN was able to repair the media errors by verifying LUNs. The other server was back up.
- 2013-12-02: DDN Installed a new controller in the location of hive12. They also replaced hive7 because it had a broken fan. The old one became our new spare on site.
- 2013-12-09: Still sorting out exactly which files were lost by running database queries. Meanwhile: preparing the new hardware so that we can decommission the DDN hardware.
- 2013-12-23: Started migration of files with dCache from the old hardware to the new hardware. We reached bandwidths of up to 18GB/s and were able to move 1PB of data within a day (see graph below). We moved more than 2PB of data.
- 2014-01-08: Migration of all data and services completed.