Main page
Recent changes
Server admin log (Prod)
Server admin log (RelEng)
Deployments
SRE/Operations Help
Incident status

Cloud VPS & Toolforge
Cloud VPS documentation
Toolforge documentation
Request Cloud VPS project
Server admin log (Cloud VPS)

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Incident documentation/20150617-LabsNFSOutage

< Incident documentation

Contents [hide]

Summary

At some point near the end of Jun 17, 2015 the filesystem backing the NFS storage used by LABS suffered a catastrophic failure, preventing most of Labs from working. Because efforts to recover the filesystem did not succeed, the decision was made to restore from a June 8th backup to a fresh set of volumes.

Service returned gradually as NFS files were restored to the different projects; tools was back in function early on the 19th with most other Labs project returned later that day. A prose summary with updates was maintained here.

Timeline

2015-06-18

- [00:30] Labs NFS switches readonly, Andrew notices and consults Gage for opinion. Only symptom in logs is ext4 on labstore 1001 hitting a single bad inode and switching to readonly for protection
- [00:36] Andrew calls Marc
- [00:37] Gage believes a fsck and remount is likely to fix the issue
- [00:40] Marc arrives online, diagnoses
- [00:45] Marc arrives at the same diagnosis; with the logs reporting a single (comparatively minor) issue, decides to halt NFS and fsck before remount
- [02:21] fsck progresses, but first signs of serious trouble show up as it begings reporting multiply-allocated blocks in the filesystem pointing to severe damage
- [05:13] fsck still working, but shows signs of distress. Giuseppe joins the effort.
- [05:20] Filesystem estimated to be damaged beyond putting back in function, Mark paged
- [06:37] Mark arrives.
- [06:50] Evaluating the backup options. Labstore1002 backup found to be good, hardware issues with labstore2001
- [07:30] Plan formulated to rebuild a new volume to restore the labstore1002 backup after moving some extents to the older raid6 drives
- [08:49] labstore1001 rebooted without assembled raid arrays to allow 1002 to take over
- [09:07] While attempting to reboot 1002 to start the recovery process, labstore1002 H800 controller fails to pass POST and server does not boot
- [09:09] Chris contacted. Attempts to repeatedly powercycle 1002 continue, hoping flea power is the issue.
- [09:39] 1002 boots; some Jessie/Precise diffences rear their heads and boot does not work
- [12:20] Issue found (Jessie ignores the mdadm.conf AUTO stanza); fixed and rebooted
- [12:41] Moving old data to make room for new volume group begins
- [13:35] New volume group (not thin, on raid 10) created
- [13:50] Attempt to restore from the backup using dump(1).
- [15:08] dump found to not scale right and would take too long to complete
- [15:15] Plan B: resize the backed up filesystem and do a block-level copy

- [18:27] resize found to be unusable as it would take many days to complete
- [18:30] Plan C: create new filesystems, selectively rsync data to them, and bring NFS service back gradually
- [18:52] Rsync of tools started, for all but four tools that are outliers in size
- [19:40] Rsync working well, Marc goes for a nap while it completes
- [23:52] Rsync done, Marc called for second Rsync

2015-06-19

- [02:54] Restoring NFS service for tools
- [03:40] Issue with the Jessie NFS configuration found (ports switched)
- [03:59] Issue fixed at the firewall level,
- [04:05] Beginning to restart tools project with the new NFS
- [04:18] gridengine returns, begin to restart all grid nodes
- [04:30] Rsync for originally excluded tools begins
- [04:48] Rsync for most other projects begins (excluded are maps, osmit, deployment-prep and mwoffliner)
- [06:30] Tools is back online
- [14:22] Issue found with NFS speed
- [14:54] Issue determined to be confusion about the scratch space mount, restarting NFS and redoing the exports fixed it.
- [15:40] fsck of the old filesystem started
- [17:23] maps project delayed until the weekend (too big), osmit rsync'ed and avaliable
- [17:31] deployment-prep restore started
- [18:08] deployment-prep available to NFS
- [18:30] rsync of maps started on a new volume

Conclusions

It's not clear what caused the corruption of the filesystem; the logs contain no error or indication of issues before the single (relatively minor) hit on a broken inode, at which point the filesystem automatically switched to read-only mode. There are a number of plausible hypoteses about the underlying cause[1] but the net result was extensive damage to the block allocation structures of the filesystem (mostly around the files being actively written at the time – log files being the hardest hit).

Recovery time was long because of the large amount of data to restore and the requirement to keep the previous filesystem for recovery (restricting the amount of space avaliable for manipulation of the filesystem and restoration) as well as the raw quantity of data, where a number of projects stored a large number of files that may have been safely discarded had they been properly noted.

Troubleshooting, recovery and team coordination was made more difficult and further lengthened the recovery time because of the comparative complexity of the system as a whole, as well as its poorly maintained state: inexistent or inconsistent configuration management, inconsistent environments (precise/jessie, configuration files) between primary and backup systems, on-going hardware issues in in both of the backup systems, multiple prolonged migrations in flight.

1. ↑ The more likely of which are (a) the secondary server having accidentally assembled and written to the RAID arrays despite the volumes not having been active; or (b) an issue or incompatibility with the then-ongoing pvmove over the thin volume holding the filesystem

Actionables

(All should be tracked in https://phabricator.wikimedia.org/tag/incident-20150617-labsnfsoutage/)

- Maintain labstore systems better, by employing standard operations team practices such as configuration management
- Reduce NFS server SPOFs (e.g. by employing sharding)
- Reduce the size of the filesystem(s) underlying NFS to speed backups and recovery
- Make certain that all hardware issues (labstore1002 & labstore2001 in particular) are fixed (Task T102626)
- Formulate a rigid, well-known backup plan across servers and locations and apply it (Task T103691)
- Simplify the NFS server setup: no added complexity unless absolutely needed (Task T102520, Task T103265, Task T94609, Task T95559)
- Reduce reliance on NFS for projects that do not strictly require a networked filesystem for their operation (Task T102240)

Category: Incident documentation