Toolforge webservices are in the final stages of migrating to the toolforge.org domain .
Please help us clean up older documentation referring to tools.wmflabs.org!

Incident documentation/20160829-ToolLabs-Outage

< Incident documentation

Contents [hide]

Summary

On labstore1001, a snapshot device of the scratch device, made for testing - was mounted accidentally on to the directory that was the acting as the scratch NFS share. This caused NFS server to back off and stop serving traffic, and all the file handles held by NFS were rendered stale. It affected all of tools - puppet runs failed across hosts, the proxy didn't serve traffic, and instances with NFS mounted couldn't be ssh-ed into. It was fixed by stopping NFS server, fixing the mounts, and starting it again. The outage lasted ~30 minutes.

Timeline

- 22:49 - 23:04 - Puppet runs on all tools hosts fail, tool labs home page unreachable
- 23:05 - Chase and Yuvi look, Chase attempts to restart nginx because tools-proxy is serving 499s everywhere, and Yuvi bumps up worker connections and restarts again - which doesn't help.
- 23.15 - bd808, Chase and Yuvi spot check a bunch of nodes by ssh-ing into them, find that they cannot ssh into nodes with nfs, and can ssh into nodes without - Chase suspects the issue lies with storage and not networking.
- 23.21 - Chase pokes Madhuvishy to see if she was doing anything with nfs - she says she isn't.
- 23.28 - Chase notices super high load on spot checked exec nodes and tools bastion, and outbound for NFS on labstore1001 was doing nothing, so restarts nfs-kernel-server on labstore1001 and 1003(this was not necessary, he did it for fun anyway).
- 23.29 - Puppet runs succeed, things start to recover
- 23.33 - Madhuvishy realizes she had actually messed up the scratch nfs share on labstore1001, by mounting a snapshot device created for testing on the directory that serves scratch(/srv/scratch). She reports it, and Chase figures out that when the shared device was mounted over, nfs backed out and stopped serving traffic, after nfs server restart that was fixed but all the old file handles to scratch had been rendered stale. It isn't possible to unmount the incorrectly mounted backup device since it's being used, so he stops nfs server, switches the mount over, also fixes a binding mount at /exp/scratch. and start the server again.
- 23.34 - Things are fixed at this point, Madhuvishy checks all the tools hosts and makes sure there aren't any stale scratch mounts.

Conclusions

- It is still not clear why the Proxy served 499 - the proxies don't have NFS enabled. Could it be because they were trying to hit the enough hung tools hosts for data and failing - if so why not 5xxs and why 499?
- Labstore1001 is a huge single point of failure for all of tools. We couldn't failover to another box and debug this one without breaking everything, nor could we be testing in some place that isn't part of the live system, and such mistakes wouldn't cause all of tools to fail. (This is being worked on)

Actionables

Explicit next steps to prevent this from happening again as much as possible, with Phabricator tasks linked for every step.

- Madhuvishy wears kid gloves on labstore1001 for eternity

- The labstore overhaul - HA setup cannot happen fast enough (bug T126083)

Category: Incident documentation