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

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/20150401-LabsNFS-Overload

< Incident documentation

Contents [hide]

1 Summary
2 Timeline
3 Conclusions
4 Actionables

Summary

In a matter almost identical to the previous day's incident, labstore1001 suffered two consecutive bouts of super high load / kernel stalls, to a point where NFS service was unavailable (or nearly unavailable). The root cause was identified, this time, by careful monitoring of disk activity at which point it was realized that one of the storage shelves housing 12 of the 72 disks was extremely unresponsive, causing long I/O queues to pile up with only a minuscule fraction of read and writes actually making it to the disks.

After a server reboot, the same phenomenon could be observed where even detection of the errant shelf was erratic. While preparing the backup server (labstore1002) to take NFS service over, one last attempt was made on the primary server by doing a power cycle and cold boot. At that point, the shelf returned to complete responsiveness and was again available for use. The filesystems and underlying raid arrays were checked, remounted, and NFS service resumed. Labs file services return to full function minutes later as instances recovered.

All five shelves are since operating normally. An inspection was done on-site to ascertain that no cabling issues were to blame but, since normal operation was observed, it was decided against manipulating the shelves or server at this time. A further, more in-depth inspection will take place at the next scheduled opportunity.

Timeline

```
02:15 User reports of intermittent slowness of NFS start coming in
02:21 NFS service mostly stalls, shinken-wm reports outages
02:45 Investigation on labstore1001 in progress, with signs of clear
overloading.  I/O is still possible, but very slow
03:06 Andrew restarts nova-network for an unrelated reason
03:10 Filesystem recovers, load and disk activity on labstore1001 returns to
normal
03:29 load and disk activity on labstore1001 breaks again
03:32 Andrew restarts nova-network again on the off-chance that the issue is
network-related - no effect
03:37 link between the issues and and a specific group of disks noted (the
difference between the active and broken period was key in discovery)
03:47 No hardware issue located, rebooting the server
03:58 server fails to recognize shelf, communication timed out during detection
phase
04:00 preparing labstore1002 for takeover as a backup plan
04:06 attempting power cycle to cold boot server and controller (labstore1001)
04:13 system boots with no issues with the shelf
04:18 raid array and filesystem journal recovery begins
05:10 recovery complete, all filesystems and raid arrays are in good state
05:13 NFS restarted, file service recovering
05:24 Labs back to normal with no issues - raid array of the errant shelf is in
rebuild mode
06:55 Labs operating normally for 90 minutes, Marc going to bed
```

Conclusions

While the ultimate cause of the communication failure between the server and the disk shelf cannot be known for certain, that a cold restart fixed the issue seems to point to the RAID controller having been wedged in an improper state requiring a full reset. That the issue resolved itself briefly before failing again seems to indicate that the issue was intermittent, and is likely to have been the cause of the previous day's outage given that the symptoms were essentially identical.

Actionables

1. Inspection of the affected hardware to eliminate obvious mechanical issues [1]
2. Use the next scheduled maintenance window to make a more throughout investigation to determine whether equipment replacement/service call is warranted. [2]

Category: Incident documentation