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/20170119-Labstore

< Incident documentation

Contents [hide]

Summary

Labstore1004 experienced a sudden spike in load starting at 21:40, reporting a load of 24, where normal load is <5. We investigated for usual signs of overzealous clients, but even with some clients with high use, it looked fairly normal. There were some NFSd threads that were in D-wait state but no other correlating indications of disk contention or abusive clients, and clients seemed to be operating normally. DRBD status was reported as normal and in sync. The load kept climbing upto 60, at which point we decided to fail-over to the secondary DRBD server, Labstore1005. There was some trouble during the failover due to labstore1004 not cleanly releasing devices on umount even after stopping nfs-kernel-server, and delay in forcing labstore1005 to become primary because of previous primary status of labstore1004. Due to this delay in bringing labstore1005 up as the NFS server, some clients reported stale filehandles - which was automatically resolved once service was restored.

Labstore1004 was rebooted as part of the failover, and post reboot - even though it was up, icinga started to periodically report that it was down. There was also some packet loss seen in controlled ping tests to Labstore1004.

Timeline

This is a step by step outline of what happened to cause the incident and how it was remedied.

- [20:41] Icinga reports high load on labstore1004
  - [20:41:40] <icinga-wm> PROBLEM - High load average on labstore1004 is CRITICAL: CRITICAL: 66.67% of data above the critical threshold [24.0]
  - [20:43:40] <icinga-wm> PROBLEM - High load average on labstore1004 is CRITICAL: CRITICAL: 88.89% of data above the critical threshold [24.0]
- [20:47] Chase and Madhu check for usual suspects - overzealous NFS clients, and even though some tools had fairly high activity, it was not above normal.
- [21:08] Chase enlists help of ema and volans, who look at drbd status, and bandwidth - and things look normal. Volans looks at dmesg and finds logs from NFSd like - [Jan19 20:13] INFO: task nfsd:95840 blocked for more than 120 seconds
- [21:24] Load keeps climbing upto 60, ema notices that a number of nfsd threads are in D state (ps xafwulawk '$8 == "D"'), and the number is piling up, but no correlation with disk contention or abusive clients.
- [21:29] Chase restarts nfs-kernel-server, but it has no impact on climbing load
- [21:34] Chase decides to failover to labstore1005, and brings down labstore1004 with nfs-manage down. He notices that it wasn't cleanly releasing devices through umount, nfsd threads were still persisiting, and the both Tools and Misc resources were still in primary status. To make sure labstore1005 would see it as unavailable, and come up as primary, he reboots labstore1004.
- [21:40] - Chase reports he did the failover to labstore1005, but there was additional delay because labstore1005 was refusing to come up as primary due to previous primary state of labstore1004, and he had to use --force in order to bring it up as primary.

- There is temporary NFS unavailability across labs/tools due to the delay in bringing labstore1005 up, and Yuvi emails the labs list that there is an nfs outage ongoing

- [21:41] Yuvi notices stale filehandle reports in bastion-03, which was also reported by a few other clients - the

filehandle issues resolve once nfs service from labstore1005 is fully restored

- [21:50] Yuvi verifies that all tools instances are okay, emails labs-l again saying service has been restored
- [22:15] Icinga alerts of labstore1004 unavailability, even though uptime says it has been up for 39 minutes (since reboot during failover)
  - [22:15:57] <icinga-wm> PROBLEM - Host labstore1004 is DOWN: PING CRITICAL - Packet loss = 100% [22:16:27] <icinga-wm> RECOVERY - Host labstore1004 is UP: PING OK - Packet loss = 0%, RTA = 0.28 ms
- [01:05] The alerts continue intermittently, and volans reports that he sees 2% packet loss in a controlled ping test from wtp1008
- [01:07] Madhu silences icinga and decides to investigate after gym, given that labstore1005 was primary, load was normal, and service was stable.
- [06:20] Madhu is back and confirms variable packet loss from other prod nodes to labstore1004
- [06:41] The icinga downtime runs out and it alerts again, *joe* comes in and starts looking. He suspects problems with the cable, or network card, or the switch.
- [07:03] Faidon investigates and diagnoses that labstore1004 and 5 are misconfigured, because both eth0 and eth1 were on the same network, which causes non-deterministic behavior, which may be the cause of periodic unavailability to icinga.
- [16:18] Chase, working with cmjohnson swaps out the eth0 cable for labstore1004, but icinga reports of downtime continue
- [17:22] Chase shuts down eth1 on labstore1004 to test - and notices that the icinga alarms stop. DRBD replication continues to go smoothly even without eth1 (though it's set up over eth1)
- [01:04] Switch asw-c2-eqiad, which is the switch for the row C2 where labstore1004 is hosted - reboots. Faidon notices that during the period of labstore1004 reboot post load spike on 19th, the switch logged the following errors

```
Jan 19 21:35:29  asw-c-eqiad fpc1 MRVL-
L2:mrvl_fdb_mac_entry_mc_set(),1089:Sanity Checks Failed(Invalid Params:-2)
Jan 19 21:35:29  asw-c-eqiad fpc3 MRVL-
L2:mrvl_fdb_mac_entry_mc_set(),1089:Sanity Checks Failed(Invalid Params:-2)
Jan 19 21:35:29  asw-c-eqiad fpc3 MRVL-
L2:mrvl_fdb_mac_entry_rebake(),482:fdb_mac_entry_mc_set() failed(-1)
Jan 19 21:35:29  asw-c-eqiad fpc3 RT-HAL,rt_entry_topo_handler,4121:
l2_halp_vectors->l2_entry_rebake failed
```

- [01:22] The switch investigation continues, and there seems to be a link between labstore1004's link state and weird errors logged in the switch. More details - https://phabricator.wikimedia.org/T155875

Conclusions

The high number of nfsd threads in D state(we are not sure why this happened) seems to have caused the load increase on labstore1004. Since failing over to labstore1005, the set up has been stable.

Actionables

- Recable eth1 to be a direct crossover (to avoid eth0 and eth1 being on same network) (Task T155832)
- Investigate whether nfs-manage up should always have the --force option to bring up the failover node as primary (Task T157478)

Category: Incident documentation