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/20160608-gallium-disk-failure

< Incident documentation

This is an ongoing incident, page opened as a placeholder 09:27, 8 June 2016 (UTC)

Summary

- On Wednesday June 8th shortly after midnight, gallium.wikimedia.org encountered hard disk / raid failure that is unrecoverable. That caused the whole CI infrastructure to be entirely unavailable since that servers hosts Jenkins and Zuul.

Tracked in Phabricator
Task T137265

Timeline

All times are in UTC.

- 23:56 Icinga ** PROBLEM alert - gallium/MD RAID is CRITICAL **
- At this point Jenkins has lost most of its executor / Zuul must be misbehaving and the Zuul status page would show changes pilling up. Jobs are no more triggering.
- 02:xx YuviPanda looks at the RAID alarm
- 02:51 legoktm: / partition on gallium is currently read-only for some reason
- 02:56 Legoktm files https://phabricator.wikimedia.org/T137265
- 03:07-03:50 yuvipanda runs an fsck -n . Reports back on https://phabricator.wikimedia.org/T137265#2363081 . Output shows corruption
- 04:10 yuvipanda and Kunal agree to not page since that has been going on for a while and European ops are about to roll in anyway. gallium is intentionally NOT rebooted for fear the unpuppetize parts get lost or the issue goes worth.
- 04:58 Moritz diagnose the RAID and find /dev/sda2 as failed and the array need rebuild.
- 07:30-07:58 Giuseppe further investigate raid / acknowledge the alarms in Icinga
- 08:15 hashar shows up and catch up with ops. Stops Zuul/Jenkins that are useless at this point.

Most of this initial delay is due to the incident happening at an odd time (SF Evening, Europe night) and only impacting CI / developers which tends to be low traffic at that point.

During the European morning:

- Jaime took backups to db1085 and dealt with the disk failure + RAID with confirmations/support from Faidon/Mark.
- Giuseppe allocated a server and installed Jessie pairing with Antoine to polish up the puppet scripts
- Antoine rebuild a Zuul package for Jessie and tested it, provided info about the CI context / indicate which data are important and which one can be dropped
- 15:00 contint1001.eqiad.wmnet been working on by Giuseppe and passing puppet with all proper roles, new partitions layout and Zuul masked in systemd. Needs Jenkins and docroots to be restored.
- 15:00 faulty disk is being replaced on gallium by Chris
- 17:00-18:00 Jenkins data are pushed to contint1001. The RAID array is rebuilding. We agreed to keep CI down until the array is complete to prevent additional I/O from Jenkins
- 18:55 Gallium rebooted. Mark confirms RAID is all good. Jenkins and Zuul spawned just fine and the service

is resumed.

Conclusions

What weakness did we learn about and how can we address them?

- hosts lacked a backup despite it has been identified and set up a year or so ago by operations (bug T80385)
- gallium was 5 years old and still on Precise. Should have migrated it early 2016 to a newer host / Jessie
- multiple Jenkins instances could be used to get redundancy there
- Zuul is AFAIK not architected to run as a cluster so it is the SPOF

Actionables

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

- Status: 🟥 **Unresolved** Add contint to backup (bug T80385)
- Status: 🟥 **Unresolved** Validate and publish Zuul Debian package for Jessie bug T137279
- Status: 🟥 **Unresolved** Migrate to contint1001 asap bug T137358 and others
- Status: 🟥 **Unresolved** Decide on best option to replace gallium bug T133300

Ideas:

- Have more than one Jenkins master for CI (co masters)
- Setup a dedicated Jenkins for daily jobs / long jobs not triggered by Zuul
- Monitor the smartctl status. Wouldn't have helped as no attributes went over the threshold.
- Run the online "offline" SMART self-test regularly to update some of the values. Wouldn't have helped as these disks only have Seek_Time_Performance and Offline_Uncorrectable as offline updateable attributes, also might cause degraded performance when it is running. that would not lead to an error anyway.

References

- http://bots.wmflabs.org/~wm-bot/logs/%23wikimedia-operations/20160608.txt
- http://bots.wmflabs.org/~wm-bot/logs/%23wikimedia-releng/20160608.txt
- Lot of activity and synchronization happened on a non public channel.

Category: Incident documentation