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!

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/20190213-cloudvps

< Incident documentation

Contents [hide]

Tracked in Phabricator
Task T216218

Summary

Hardware failures have caused downtime in several CloudVPS projects.

Timeline

All times are UTC.

On 2019-02-10:

- 09:47 icinga reports an issue with the hardware RAID in cloudvirt1024. This event is ignored by the WMCS team.

On 2019-02-13:

- 09:42 Incident 20190213 - PAWS is solved.
- 12:10 icinga reports an issue with the hardware RAID in cloudvirt1018.
- 12:12 icinga alert for cloudvirt1018 is acknowledged by Arturo, who start looking into it.
- 12:20 Arturo confirms issues in storage of cloudvirt1018. Tries to depool/drain/evacuate the server without success.
- 12:30 Arturo decides to try rebooting the server to see if that fixes anything enough to allow a controlled drain of the server.
- 12:37 cloudvirt1018 doesn't boot again due to issues in the RAID. All virtual machines running in this server are unreachable (64 instances). Severe data loss could be produced.
- 13:00 Arturo is notified by Moritz and Chris via IRC that cloudvirt1024 has HW issues as well. To this point, nobody at the WMCS team was aware.
- 13:22 Giovanni tries refreshing disk states (pool/depool) in cloudvirt1018.
- 13:37 cloudvirt1018 is back online. No data loss apparently. VM instances aren't back online yet. We suspect the issues were related to the RAID controller being overly defensive.
- 14:00 Andrew, Giovanni and Arturo start thinking on next movements, if reallocating all the workloads to other servers or try rescuing cloudvirt1018.
- 14:05 Arturo estimates that draining cloudvirt1018 and cloudvirt1024 could cost 3 Full-Time-Engineer (FTE) a full day worth of work.
- 14:10 a disk replacement for cloudvirt1024 is coming. Since the server is working (with a damaged drive) we decide to continue using and don't drain it yet.
- 14:15 Giovanni and Arturo start VM instances in cloudvirt1018 to try using the server until we see further evidences of issues.
- 15:00 Some VM instances on cloudvirt1018 aren't coming back online healthy. Andrew starts looking into it.
- 15:25 another reboot for cloudvirt1018 due to extended filesystem checks by Giovanni. All VM instances are unreachable again.
- 15:49 Andrew confirms VM local storage may be corrupted for some VM instances.
- 17:00 still dealing with local VM storage corruptions.
- 18:47 after fixing by hand almost all VM storage corruptions, we still see issues in /var/lib/nova/instances in cloudvirt1018. We decide to drain it.

- 18:50 Arturo left the team for rest.
- 19:38 icinga alert for cloudvirt1024, storage has more issues.
- 19:52 we decide to drain all workload from cloudvirt1024.
- 22:01 Giovanni left the team for rest.

On 2019-02-14:

- 05:49 all workloads from cloudvirt1018 and cloudvirt1024 have been drained.

Conclusions

The initial issues with cloudvirt1024 were ignored for 3 days by the WMCS team, even though a phabricator task existed, automatically created by icinga. This resulted on a sudden increase of human workload when hardware issues piled up.

When we were unable to access VM storage in cloudvirt1018 after the first reboot, there was very strong risk of massive data loss. We currently don't have backups or other redundancy solutions for VM storage other than RAID setups in cloudvirt servers themselves.
We delayed a lot our decision to drain the failing cloudvirt servers because that takes a lot of effort and tedious manual intervention.
All of this can be improved if we implement a shared storage mechanism for VM storage. Exploring this is currently one of our team goal.

Fixing VM disk storage is also a very manual and tedious work, which is not automated at all.

Hardware, specially disks, can fail. We need to be ready to handle that and to minimize impact to our services. In our current situation, the 24h replacement deal with hardware vendors may not be enough, and we should consider having spare disk on-site to be able to quickly address disks issues.

Also, independent incidents can randomly happen at the same day, which requires lots of extra efforts, focusing and extended working hours by the team.
In the concrete incidents of this day (also Incident with PAWS), we took benefit of our team being spread into different timezones, and we were able to work non-stop on incident response for 24h in a row.

Links to relevant documentation

- To fix VM disk corruption, we follow this documentation: How to fix VM disk corruption.
- To drain a hypervisor, we use this script: wmcs-cold-migrate.
- To fix/deal with failing RAIDs, we have some draft documents: Hardware troubleshooting runbook. HDD & SSD failures.

Actionables

- CloudVPS: create wmcs-vm-fsck script phab:T216132
- Order spare cloudvirt SSDs for eqiad phab:T216041
- Test Ceph for instance storage phab:T90364
- Increase visibility of auto-generated tasks for RAID errors phab:T216133
- Update raid controller firmware phab:T216733
- Add a single puppet setting to disable VM creation phab:T216767
- Write some docs re: capacity planning and how to handle being under capacity phab:T216768
- Planned failovers for nova services

NOTE: Please add the #wikimedia-incident Phabricator project to these follow-up tasks and move them to the "follow-up/actionable" column.

Category: Incident documentation

This page was last edited on 21 February 2019, at 21:20.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. See Terms of Use for details.

Privacy policy About Wikitech Disclaimers Code of Conduct Developers Statistics Cookie statement Mobile view