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/20191007-wmcs-network

< Incident documentation

document status: final

Summary

During the upgrade of the WMCS OpenStack control plane, all cloud VMs presented with the wrong originating IP for outbound traffic. That broke several services on many VMs, most importantly DNS, LDAP, and NFS.

Impact

The largest impact during this outage was NFS access from Toolforge. That prevented many grid and k8s jobs from running properly, and also produced a torrent of alert emails.

CI tests produced incorrect failure messages for part of outage due to DNS failures.

Ssh access to most VMs was broken for about an hour.

Detection

The problem was immediately evident, as many of the issues produced shinken and icinga alerts. The team working on the upgrade didn't respond immediately because some level of background alerting was already expected as part of the upgrade process.

Timeline

All times in UTC, and approximate

- 14:00 Andrew begins upgrading OpenStack services on cloudcontrol1003, cloudcontrol1004, cloudnet1003, and cloudnet1004. Those four hosts (as well as all cloudvirt hosts) are marked for two hours of downtime in icinga. Horizon is put into maintenance mode.
- 14:00 - 15:30 various unexpected issues arise during upgrade (most importantly relating to the scripted Neutron schema upgrades failing) which extends the expected Horizon outage window. Nothing user-facing (other than Horizon) is broken up to this point.
- 15:40 At this point, cloudcontrol1003 and cloudnet1003 are fully upgraded. On Andrew's request, Arturo disables the currently-active neutron server on cloudnet1004, failing all network traffic over to cloudnet1003. At this point, all VMs present to the outside internet (including WMF production) as originating from Neutron on cloudnet1003.
- 15:45 **OUTAGE BEGINS**
- 15:50 Lots of things are starting to break and throw alerts. NFS, and the Cloud DNS recursors all use ACLs based on the origination IP of incoming traffic. Because VMs are presenting with the wrong IP, they are unable to access either DNS or NFS. Toolforge VMs cannot access NFS, ssh to VMs fails, and CI jobs fail due to

DNS resolution failures.

- 16:00 The WMCS team identifies the issue as relating to the origination IP (thanks to additional logging channels on cloud-recursor0). Andrew hacks the DNS recursors to allow traffic with any origination IP; this resolves some of the outage (DNS and LDAP) but NFS is still inaccessible.
- 16:20 (approximately) After a good deal of digging into logs and iptables rules, Jason determines that we should just restart all the Neutron serves.
- 16:30 **OUTAGE for most VMs ENDS** Once all the services are restarted, origination IPs return to normal and all communication works properly. NFS clients nevertheless fail to reconnect to NFS servers, so the Toolforge outage continues. Brooke and Hieu start restarting jobs and/or rebooting hosts to force NFS reconnects.
- 19:06 Krenair detects more VMs (most inside of Toolforge, some outside) with failing NFS connections and restarts them as well. Among his reboots are is the Toolforge mail server which results in a huge backlog of pending emails (from the outage an hour previously) getting sent all at once.
- 19:15 Most toolforge VMs are rebooted, normal toolforge behavior is restored. **USER-FACING TOOLFORGE OUTAGE ENDS**
- 19:20 Final toolforge VM rebooted - k8s master
- 20:15 Bryan clears out the tail end of the pending mail from the queue, ending the deluge of alert emails

Conclusions

What went well?

- Mostly all-hands-on-deck situation and all WMCS people were involved in the operation. Team response went well.

What went poorly?

- Database encoding issue added delay to the operation.
- Neutron network misbehaved and caused several issues inside CloudVPS.
  - the network setup can be complex to debug.

Where did we get lucky?

- DBA team were around to advice with the DB enconding issue.

How many people were involved in the remediation?

At various times: Andrew, Jason, Arturo, Brooke, Hieu, Bryan, Krenair

Links to relevant documentation

- https://tools.wmflabs.org/sal/tools
- phab:T234834

Actionables

Some Phabricator tickets opened as a result of this incident:

- Various user visible errors in Cloud VPS projects following OpenStack upgrade on 2019-10-07 https://phabricator.wikimedia.org/T234834
- CloudVPS: m5-master databases for openstack may require re-enconding https://phabricator.wikimedia.org/T234830
- nova-conductor running out of mysql connections https://phabricator.wikimedia.org/T234876
- CloudVPS: update DNS record for eqiad1 routing_source_ip https://phabricator.wikimedia.org/T234836

Category: Incident documentation