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/20170719-ldap

< Incident documentation

Contents [hide]

Summary

On July 19th, 2017 at 20:06 (approximately), some LDAP queries to the "Labs" set of LDAP servers started failing. Affected access included various production services behind password authentication (including monitoring tools such as Icinga, Logstash and Tendril), as well as Cloud VPS (SSH logins and sudo commands) and logins to the ToolForge bastions. The ultimate cause was the expiry of the "WMF CA 2014-2017" that was the issuer of a handful of certificates, the most impactful ones being the LDAP server certificates.

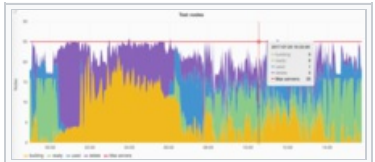
Timeline

- 20:08 checker.tools.wmflabs.org alerts, pages opsens
- 20:09 Daniel, Chase and Madhu start investigating
- 20:12 Chase restarts LDAP on serpens
- 20:12 Chase restarts LDAP on seaborgium
- 20:16 Bryan and AndrewB join the investigation
- 20:18 Filippo joins the investigation
- 20:19 AndrewB restarts slapd on seaborgium
- 20:29 Faidon joins the investigation
- 20:29 Filippo points out the certificate has expired
- 20:31 Faidon points out that it's the CA that has expired
- 20:34 Faidon finds all affected CAs to determine impact; ldap-labs, ldap-corp and labvirt-star
- 20:35 Faidon confirms inbound wikimedia.org email (dependent on ldap-corp) still works
- 20:36 Chase points out that labvirt-star affect admin-things only
- 20:37 WMCS team nd opsens determine impact on Cloud Services
- 20:51 Faidon creates a new WMF 2017-2020 CA, pushes it to puppet-private and operations/puppet
- 20:53 Faidon pushes new certificates for ldap-labs.{eqiad,codfw}.wikimedia.org, force-runs puppet on seaborgium
- 20:56 Daniel runs puppet on einsteinium (Icinga server), which is still broken
- 21:00 Faidon realizes puppet didn't trigger slapd restarts, restarts them manually
- 21:01 Daniel reports that Icinga is back
- 21:05 Faidon pushes new certificates for ldap-corp.{eqiad,codfw}.wikimedia.org
- 21:06 Daniel restarts Apache on grafana-admin
- 21:12 Jaime fixes Tendri
- 21:16 Daniel restarts Apache on piwik
- 21:23 Madhu restarts Apache on thorium (runs hue, yarn, and pivot)
- 21:25 Chad restarts Apache on logstash100[1..3]
- 21:29 Daniel restarts Apache on tungsten (xhgui)
- 21:31 Chad restarts Gerrit/Apache on cobalt/gerrit2001
- 21:34 Chase restarts nslcd on labstore1004/1005

- 21:34 Jaime points out the m5 database shard has issues and went down, probably due to connection overload; asks WMCS team whether it's OK to pool back db1009
- 21:35 Daniel restarts Apache on graphite1001
- 21:39 Jaime reloads haproxy on dbproxy1005 to repool db1009 (m5)
- 21:44 Daniel fixes netmon1001/librenms (re-enable puppet once to get new CA, restart Apache, disable puppet again)
- 21:52 Daniel restarts Apache on netmon1003/servermon
- 21:56 Daniel restarts Apache on graphite200*
- 21:59 Daniel restarts Apache on dbmonitor2001
- 21:59 Daniel restarts Apache on hafnium/labmon1001
- 03:32 Andrew service uwsgi-labspuppetbackend restart on labcontrol1001
- 03:34 Andrew service nova-network restart on labnet1001
- 04:05 restarting rabbitmq-server on labcontrol1001
- 06:24 Daniel restarts Apache on netmon1002
- 07:48 Filippo restarts diamond on serpens/seaborgium to restore LDAP server data on dashboards

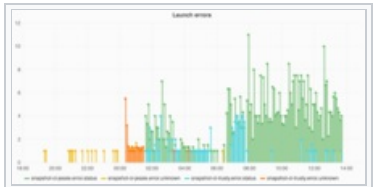
CI/beta

Nodepool/openstack


State of the Nodepool pool

Nodepool maintain a pool of instances on Cloud service. If OpenStack is in trouble, there is quickly no instances left and Jenkins jobs can no more execute. The Nodepool service went down around 00:10, partially recovered at 06:00 and was back operational at 14:00.

Main task: T171158


OpenStack launch errors

- 00:10 Nodepool can no more delete instances from OpenStack.
- 00:30 No more instances available CI is on halt
- 05:30 TimStarling on contint1001 restarted jenkins
- 05:46 TimStarling on contint1001 restarted zuul and zuul-merger
- 06:00 Nodepool finally manage to have some instances spawned. Probably due to openstack being fixed meanwhile
- 07:40 Antoine joins IRC
- 09:17 Antoine drops integration-slave-docker-1000 (irrecoverable) replaces it with integration-slave-docker-1003
- 10:30 The Nodepool pool shows suspicious rate of instances in delete/building mode (Graph)
- 11:00 T171158 filled. Some instances refuse to spawn: //Failed to allocate the network(s) with error Maximum number of fixed ips exceeded, not rescheduling.//
- 14:00 Andrew forces refresh the fixed-ip quota for contintcloud tenant. Nodepool pool is fully back

castor

Castor is a central cache, for changes being merged, some jenkins job attempt to rsync to a single instance. If it is not available, the job blocks and instance are kept idling until killed with the job failing due to a timeout.

Antoine has not investigated what happened before 07:09 when the instance got rebooted. After (and maybe before) Jenkins was no more able to SSH into it and thus the castor job could not run.

The instance was not reachable via salt or ssh. Since it is fully puppetized a new one got created and service got restored at 09:03 (with an empty cache).

Main task: T171148

- 07:09 Giuseppe reboots castor (ssh unreacheable?)
- 07:40 Antoine joins IRC
- 07:55 Antoine disables Castor for CI, therefore restoring CI services albeit in a degraded mode (no caches) - T171148
- 09:03 Antoine replaced castor instance with castor02 with an empty cache. Instance was deadlocked (no salt no ssh)
- 09:08 Antoine refilled operations/puppet cache in castor

beta

Main task: T171174

Lot of instances had ssh access broken and puppet broken on most.

- 11:43 Antoine starts investigating failures on beta
- 12:30 Puppet master completely broken
- salt -v '*' cmd.run 'systemctl restart nslcd'
- 14:31 Antoine has finished cleaning out the beta cluster puppet master. Instances recover

T171174 is still ongoing. Any instance that had puppet failing can not be reached by ssh, they need puppet to be fixed the nscld to restart.

Conclusions

Annexes

- State of the Nodepool pool

Incident tasks

- T171148 - CI jobs are blocked because castor is unreachable
- T171158 - contintcloud instance refuses to launch due to "Maximum number of fixed ips exceeded
- T171174 - a lot of beta cluster instances are not reachable over SSH

Actionables

- Update certs:
  - labvirt-star.codfw.wmnet.crt (done in https://gerrit.wikimedia.org/r/366514)
  - labvirt-star.eqiad.wmnet.crt (done in https://gerrit.wikimedia.org/r/366505)
  - labtestservices2001.wikimedia.org.crt (done in https://gerrit.wikimedia.org/r/366515)
  - ldap-corp.codfw.wikimedia.org.crt (done in https://gerrit.wikimedia.org/r/#/c/366462/)
  - ldap-corp.eqiad.wikimedia.org.crt (done in https://gerrit.wikimedia.org/r/#/c/366462/)
  - ldap-labs.codfw.wikimedia.org.crt (done in https://gerrit.wikimedia.org/r/#/c/366428/)
  - ldap-labs.eqiad.wikimedia.org.crt (done in https://gerrit.wikimedia.org/r/#/c/366428/)
- Add an icinga check for that CA so we notice before it expires next time
- Update documentation at WMF CA

Category: Incident documentation