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/20190920-d2 switch failure

< Incident documentation

document status: in-review

Summary

At 23:56 UTC on Friday 20th, the top of rack switch asw2-d2-eqiad went down causing all servers in rack D2 to go offline. As this switch was also row D virtual-chassis master and spine to the upstream routers all row D servers suffered a brief connectivity loss. Some services were still degraded after the situation stabilized but were eventually recovered.

Impact

1. 14 Hosts in D2 lost connectivity. (an-presto1001, an-worker1092, an-worker1093, analytics1076, backup1001, cloudelastic1004, cloudstore1008, cp1087, cp1088, flerovium, kafka-main1004, labstore1007, ms-be1043, ms-be1048)
2. Row D hosts had their connectivity impacted (eg. packet loss) for a few seconds the time the virtual-chassis fails mastership to fpc7.
3. Some services (e.g. AQS, mobileapps) continued to be impacted long after the initial blip, and even after recovery of the failed switch. The general theme here for at least some of the nodejs services seems to be related to application-layer caching of transient DNS failures (from the very brief blip of their network reachability towards recursive DNS (and everything)), and they were fixed with service restarts. Other related APIs and services (e.g. restbase and such, and MW APIs that interact with these things) also had alerts due to this indirectly.
4. Logstash was overwhelmed by all the carnage and fell behind and threw a lot of alerts itself. Since some services' alerting relies on logstash, this also confusingly caused logstash-derived alerts to persist long after the underlying issues were resolved.
5. dbproxy1016 and dbproxy1017 (which live in D1 and D3, thus were indirectly affected by the spine failure) failed healthchecks on their primaries and failed over to their secondary choices (both of which were different ports on db1117). We (eventually) reloaded haproxy on these hosts to recover towards the primaries.
6. Cloud NFS clients lost access to dumps and scratch until those services were manually failed over. Toolforge Kubernetes clients saw steadily increasing load until they could mount the volumes again (switch recovery) even after failover because they do that until rebooted with a changed fstab.
7. The maps Cloud VPS project lost NFS access to home directories and project dirs until manual failover.

Detection

The first alert came from Icinga showing commons was down, but **the switch failure wasn't clear until the alert storm had settled and correlation could be drawn from down hosts in Icinga cross-referenced with**

Netbox.

```
   23:57 UTC - <+icinga-wm> PROBLEM - Host commons.wikimedia.org is DOWN:
/bin/ping -n -U -w 15 -c 5 commons.wikimedia.org
   < Alert storm here >
```

Timeline

This is a step by step outline of what happened to cause the incident and how it was remedied. Include the lead-up to the incident, as well as any epilogue, and clearly indicate when the user-visible outage began and ended.

All times in UTC.

- 23:57 **OUTAGE BEGINS** <+icinga-wm> PROBLEM - Host commons.wikimedia.org is DOWN: /bin/ping -n -U -w 15 -c 5 commons.wikimedia.org
- Cascade of alerts
- 23:58 <+icinga-wm> PROBLEM - Host asw2-d-eqiad is DOWN: PING CRITICAL - Packet loss = 100%
- 23:59 <+icinga-wm> RECOVERY - Host asw2-d-eqiad is UP: PING OK - Packet loss = 0%, RTA = 0.80 ms
- 00:07 Identified as limited to Eqiad D2
- 00:17 < librenms-wmf> Critical Alert for device asw2-d-eqiad.mgmt.eqiad.wmnet - Juniper alarm active
- 00:21 Chris and John called, John responds.
- 00:27 Logstash identified as thrashing on non-UTF-8 messages
- 00:47 AQS recovery
- 01:06 John arrives onsite.
- 01:15 Switch power cycled.
- 01:35 Mobileapps recovery
- 02:14 dbproxy failover recovery
- 02:15 **OUTAGE ENDS**
- 05:58 logstash caught up

Conclusions

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

What went well?

- Automated monitoring detected the incident
- Outage was root-caused quickly
- Multiple SREs were available to investigate the issue

What went poorly?

- Alerting noise was overwhelming IRC
- Root cause for the switch failure could not be determined
- Mediawiki logging rate went through the roof during outage, ~130k messages/s at peak

Where did we get lucky?

- Rack D2 didn't have any critical servers
- John was available to drive to the datacenter

How many people were involved in the remediation?

- 6 SRE, 3 WMCS, 1 Dev

Links to relevant documentation

Where is the documentation that someone responding to this alert should have (runbook, plus supporting docs). If that documentation does not exist, there should be an action item to create it.

Actionables

- Investigate the switch failure - https://phabricator.wikimedia.org/T233645
- Make it easier to distinguish per-host icinga spam vs real whole-service-level icinga spam (cf. alerting infrastructure roadmap)

- Investigate solutions to non-UTF8 logging crashing logstash mutate plugin (and ultimately the pipeline) https://phabricator.wikimedia.org/T233662
- Emergency response to logstash being backlogged https://phabricator.wikimedia.org/T233735
- Fix MediaWiki spammy logs during such outages https://phabricator.wikimedia.org/T233739

Categories: Incident documentation in-reviews | Incident documentation

This page was last edited on 27 September 2019, at 21:21.

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