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/20200501-vc-link-failure

< Incident documentation

document status: final

Summary

The virtual chassis link between asw2-d1-eqiad and asw2-d8-eqiad failed in two steps.

First on Friday the 1st where it was causing packet loss for hosts on D1 without any other signs of failures.

This packet loss caused connectivity issues between MediaWiki appservers (and API servers at a lower scale) and memcache servers. Resulting in a significant increase of MediaWiki exceptions being served to the users.

This got worked around for the weekend by depooling D1 servers. At this point the cause of the packet loss was unknown.

The day after, on Saturday, hosts in D8 started seeing the same issues as in D1. This time the switches were logging errors about the D1-D8 link. Disabling the link solved the issues.

Impact: This had little to no effect on traffic (Varnish_HTTP_Total), error rates (ATS availability) and latencies (Navtiming requests) for anonymous users. A increase in error rates (~1% of requests had errors, with a short spike to ~7.5%, Appserver errors) and an increase in tail latency (around plus 100%-150%, Appserver p95) has been observed for logged in users, though.

Contents [hide]

Timeline

All times in UTC. Friday 1st:

- 05:21 MW exceptions starts being reported on #wikimedia-operations <+icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops **OUTAGE BEGINS**
- 05:32 <+icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
- 05:33 < marostegui> wow those fatals really increased
- 05:33 <+icinga-wm> PROBLEM - MediaWiki memcached error rate on icinga1001 is CRITICAL: 5009 gt 5000 https://wikitech.wikimedia.org/wiki/Memcached https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=1&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
- Joe sends the following patches: https://gerrit.wikimedia.org/r/#/c/operations/puppet/+/593728/ https://gerrit.wikimedia.org/r/#/c/operations/puppet/+/593727/
- 7:40 Arzhel checks switches (especially D/D1) nothing out of ordinary
- Large but steady increase of TCP retransmits on mc1021/1029 - https://grafana.wikimedia.org/d/000000365/network-performances?panelId=15&fullscreen&orgId=1&from=now-7d&to=now&var-server=mc1029&var-datasource=eqiad%20prometheus%2Fops

- 8:29 <elukey> _joe_ there are some things that are off, namely a lot of traffic patterns showing spikes every 10m, I am wondering if in some twisted way the 10m TTL of the gutter is somehow exacerbating this problem
- 8:37 <elukey> so mw1331 for example doesn't show tkos, I think it is only the servers in D1
- 8:43 <_joe_> I will bring back mw1409 and mw1407 in the pool, and we can depool those servers in D1
- 8:54 <_joe_> !log depooled all servers in the app pool in rack D1 **OUTAGE WORKED AROUND**
- 8:55 +icinga-wm> IRC echo bot RECOVERY - MediaWiki exceptions and fatals per minute on icinga1001 is OK: All metrics within thresholds. https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
- 19:57 rzl depools api servers in D1 (mw1356-1362) at _joe_'s suggestion, in response to flapping alerts like "PROBLEM - PHP7 rendering on mw1361 is CRITICAL: CRITICAL - Socket timeout after 10 seconds"

Saturday 2nd:

- (Overnight) Wall of flapping PROBLEM - PHP7 rendering on mwXXXX is CRITICAL: CRITICAL - Socket timeout after 10 seconds **OUTAGE RESURFACE**
- 06:42 Giuseppe and Luca start investigating
- 06:52 Arzhel starts investigating
- 07:08 <XioNoX> asw2-d-eqiad> request virtual-chassis vc-port delete pic-slot 1 port 0 member 1 **OUTAGE ENDS**
- 07:49 <oblivian@cumin1001> conftool action : set/pooled=yes; selector: name=mw13(49|5[0-9]|6[0-2])\.eqiad\.wmnet

Detection

- <+icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
- <+icinga-wm> PROBLEM - PHP7 rendering on mwXXXX is CRITICAL: CRITICAL - Socket timeout after 10 seconds
- Did the appropriate alert(s) fire? Yes
- Was the alert volume manageable? Yes
- Did they point to the problem with as much accuracy as possible? No

The root cause didn't generate any logs at first, and when it did, those logs didn't trigger alerts.

Conclusions

- Packet loss through Virtual Chassis Fabric are difficult to pinpoint
- Higher layers monitoring worked as expected
- From history, this failure scenario has a low probability of happening, and is now documented

What went well?

- We had enough capacity to depool impacted mediawiki hosts
- Once the failure generated logs, the root cause and fix were quick to identify and apply
- SREs quickly identified D1 then D8 as common factor

What went poorly?

- The first VC link failure didn't generate any switch side errors
- The issue started happening on a Friday and re-appeared on a Saturday
- The issue would not have been noticed if SREs didn't look at alerts on a weekend

Where did we get lucky?

- SREs looked at alerts on a weekend

How many people were involved in the remediation?

- 4 SREs

Links to relevant documentation

Actionables

- Either re-cable or cleanup disabled cable - https://phabricator.wikimedia.org/T251663
- Add log alerting for VC link failure - https://phabricator.wikimedia.org/T251663

Category: Incident documentation