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/20200325-codfw-network

< Incident documentation

document status: in-review

Summary

Unexpected loss of internal connectivity to codfw hosts and services for 5 minutes, creating user-visible failed queries for users whose traffic hits our eqsin and ulsfo edges, when they were using services that are active/active (Swift, Maps, Restbase API, ...)

Cause was maintenance that required a linecard restart on cr1-codfw, which exposed a flaw in codfw's network design.

Loss of some external connectivity to codfw was expected, and the site was CDN-depooled before the maintenance began. However, what was unanticipated was that cr1-codfw would hold VRRP mastership for the duration of the linecard restart, so it tried to act as the default gateway for all hosts in the cluster, while effectively being a black hole for routing anywhere outside the cluster (as the linecard being rebooted has both all the cross-cluster links and the router-to-router interconnect).

There was a second OSPF flap/convergence event around 12:22, however it doesn't seem to have been impactful.

Impact

~28k queries lost for queries terminated in ulsfo and eqsin against active/active services
https://logstash.wikimedia.org/goto/bcab629e395fc8a71ef9ac5d525c1ec7

Although this was <1% of global HTTP traffic at the time, upload-lb requests in ulsfo and eqsin were very much affected -- so users of Wikimedia Commons images, or of map tiles whose traffic terminates in those datacenters. Impact on upload-lb in ulsfo was ~10% of requests failed for the interval; in eqsin, about 1.5%.

This also created Kafka mirrormaker delays, the impact of which is TODO

Detection

Automated: Icinga pages for service IPs in codfw, in addition to alerts for socket timeouts against many hosts (especially appservers).

Since all codfw appservers could not be reached, there *would* have been lots of alert spam in #wikimedia-operations (one per appserver) -- except that icinga-wm got `Excess Flooded` off of IRC.

Timeline

see also:

- all OSPF-related logs across all routers
- all 'neighbor up/down' logs across all routers

All times in UTC.

- 11:35 <cdanis> depool codfw for router maintenance T248394 https://tools.wmflabs.org/sal/log/AXEReW8N0fjmsHBaqHGh
- 11:50:08 re0.cr1-codfw mgd[96118]: UI_CMDLINE_READ_LINE: User 'cdanis', command 'request chassis fpc restart slot 5 ' **OUTAGE BEGINS**
- 11:50:17 first socket timeout reported from icinga1001. icinga2001 sees no socket timeouts or unreachable hosts at any point, as most of its traffic flows via the switches in codfw -- aside from frack hosts (as those also traverse the core router)
- 11:50:20 first user-visible HTTP 502 response
- 11:53:07 TTL exceeded event seen for cross-cluster traffic from icinga1001: icinga2001 icinga: HOST ALERT: pfw3-eqiad;DOWN;SOFT;1;CRITICAL - Time to live exceeded (208.80.154.219)
- 11:53:27 Icinga meta-monitoring fails against icinga2001 -- expected; external connectivity via some routes was likely to be impacted
- 11:54:48 first page sent: `search.svc.codfw.wmnet;LVS HTTPS IPv4 #page;CRITICAL;HARD;3;CRITICAL - Socket timeout after 10 seconds`
- 11:55:21 last socket timeout & first recovery reported from icinga1001
- 11:55:41 final OSPF state change: `cr1-codfw.wikimedia RPD RPD_OSPF_NBRUP: OSPF neighbor fe80::aad0:e5ff:fee3:87c5 (realm ipv6-unicast ae0.0 area 0.0.0.0) state changed from Loading to Full due to LoadDone (event reason: OSPF loading completed)`
- 11:55:44 icinga1001 reports TTL exceeded events for eqsin hosts: `Mar 25 11:55:44 icinga1001 icinga: HOST ALERT: upload-lb.eqsin.wikimedia.org;DOWN;SOFT;1;CRITICAL - Time to live exceeded (103.102.166.240)` and also reports PING CRITICAL for text-lb.ulsfo.wikimedia.org_ipv6 and upload-lb.eqsin.wikimedia.org_ipv6
- 11:56:02 end of the penultimate spike of user-visible HTTP 502 responses
- 11:56:43 start of the final spike of user-visible HTTP 502 responses
- 11:56:54 end of the final spike of user-visible HTTP 502 responses **OUTAGE ENDS**
- 11:57:51 first recovery for the unreachable-from-icinga1001 ulsfo/eqsin hosts

Conclusions

What went well?

- automated monitoring detected the incident very well

What went poorly?

- a lot of Icinga spam due to many host-level service checks
- root-causing the issue took a while and required deep network expertise

Where did we get lucky?

- Linecard took only 5 minutes to finish rebooting
- Exposed a design flaw in the codfw network during scheduled maintenance, rather than unexpectedly -- an actual hardware failure in the same spot would be ugly
- codfw wasn't the primary DC

How many people were involved in the remediation?

- 1 SRE during incident; 2 SRE + 1 SRE director investigating afterwards

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

Explicit next steps to prevent this from happening again as much as possible, with Phabricator tasks linked for every step.

NOTE: Please add the #wikimedia-incident Phabricator project to these follow-up tasks and move them to the

"follow-up/actionable" column.

- Add linecard diversity to the router-to-router interconnect in codfw. phab:T248506
- Consider plumbing a backup router cross-connect via a new VLAN on the access switches.

Categories: Incident documentation in-reviews | Incident documentation