Search Wikitech

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!

Incident documentation/20190823-network codfw

< Incident documentation

document status: final

Summary

A provider outage on our primary transport link between eqiad and codfw caused it to be in a constant flapping (going down and up) state.

This flapping caused routing re-convergence churn and packet loss between the two sites.

On the application level, this translated to elevated 5xx/s from Varnish from ulsfo, eqsin, and codfw from 21:20 to 21:55 UTC. Varnish reported "No backend" for many of the requests. Host checks in Icinga were flapping "TTL exceeded" and service checks flapping "No route to host."

Impact

Surfaced a bit more than 52,000 5xx responses.

https://grafana.wikimedia.org/d/000000479/frontend-traffic?panelId=2&fullscreen&from=1566594998796&to=1566597283616&var-status_type=5

Detection

Monitoring caught and reported the issue via SmokePing and Icinga.

Timeline

All times in UTC.

- 2019-08-23 21:20 **OUTAGE BEGINS**
- 21:25 Investigation begins
- 21:33 Zayo (the link's provider) reports issue with service (email unnoticed)
- Lots of errors and recoveries - flapping
- 21:41 Arzhel starts investigating
- 21:46 Brandon called
- 21:47 Decided to depool codfw (ended up not needing it)
- 21:48 Arzhel promotes backup link to primary
- 21:55 **OUTAGE ENDS**
- 2019-08-25 01:37 Link stops flapping

Conclusions

What went well?

- The root cause was quickly worked-around once the cause (network link) was identified.

What went poorly?

- Due to the frequency of the flapping Icinga checks for link status, OSPF and BFD didn't trigger, causing SREs to think of an application layer issue
- The work-around (failing over to the backup link) is not documented and requires Netops to be done.
- Nothing paged even though it had user facing impact

Where did we get lucky?

- Giuseppe, and Filippo responded outside of their office hours.

How many people were involved in the remediation?

- No Incident coordinator appointed - 5 SREs

Links to relevant documentation

- Global traffic routing#Disabling a Site 2
- Network monitoring#OSPF status

Actionables

NOTE: Please add the #wikimedia-incident Phabricator project to these follow-up tasks and move them to the "follow-up/actionable" column.

- Those two will help mitigate the consequences of an overly flapping link:
  - Configure interface damping on primary links - T196432
  - ospf link-protection - T167306
- This one will make it easier (down the road) to a non-netops to failover a link if the need arises:
  - Configuration management for network operations - T228388
- This one is about having better monitoring and alerting by replacing Smokeping by something Prometheus based
  - Investigate/setup prometheus blackbox_exporter - T169860

Category: Incident documentation

Privacy policy   About   Disclaimers   Code of Conduct   Developers   Statistics   Cookie statement   Mobile view
Wikitech

WIKIMEDIA
a project

Powered By
MediaWiki