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/20191016-network eqsin

< Incident documentation

document status: final

Contents [hide]

Summary

An Equinix Singapore IXP peer flapped heavily, which overwhelmed the routing daemon on cr1-eqsin and caused all its BGP and OSPF sessions to flap or go down.

In addition to the external connectivity issues, as the primary transport link to codfw is on cr1, it caused the local caches to not be able to reach their peers in the main datacenters and serve 500 errors instead.

Impact

Estimated ~170k errors surfaced to users of eqsin PoP.

https://grafana.wikimedia.org/d/000000479/frontend-traffic?orgId=1&from=1571245200000&to=1571248800000&var-site=eqsin&var-cache_type=text&var-cache_type=upload&var-status_type=5

Detection

The following automated alerts got triggered:

- Varnish traffic drop between 30min ago and now at eqsin
- HTTP availability for Nginx -SSL terminators- at eqsin
- HTTP availability for Varnish at eqsin
- BFD status on cr1-codfw
- LVS HTTPS text-lb.eqsin.wikimedia.org - PAGE

This quickly pointed to an network issue in eqsin.

Was the alert volume manageable? yes

Did they point to the problem with as much accuracy as possible? yes

Timeline

All times in UTC.

- 17:15 SSL terminator alerts in eqsin fire, non-paging -- **OUTAGE BEGINS**
- 17:28 First page fires -- LVS HTTPS text-lb.eqsin.wikimedia.org
- 17:29 eqsin depooled
- 17:29 Recovery on its own -- **OUTAGE ENDS**

Conclusions

What went well?

- The issue was quickly identified
- The issue recovered on its own

What went poorly?

- A router's routing daemon should not behave that way but that router's model is known to be weak
- The logs didn't have any information on why OSPF and BGP were behaving that way.

Where did we get lucky?

- Several SREs were around when the issue started

How many people were involved in the remediation?

- 6 SREs

Links to relevant documentation

Depooling the site: DNS#Change GeoDNS

Actionables

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

- T236878 Improve resiliency of the eqsin transport link by either:
  - Terminating it on cr2-eqsin
  - Adding a 2nd link
  - Configuring link damping
- Replace cr1-eqsin with a better router (next FY)

Category: Incident documentation