Create account Log in

Page Discussion Read View source View history Search Wikitech

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

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/20190916-kubernetes-dns

< Incident documentation

document status: final

Contents [hide]

1 Summary
1.1 Impact
1.2 Detection
2 Timeline
3 Conclusions
3.1 What went well?
3.2 What went poorly?
3.3 Where did we get lucky?
3.4 How many people were involved in the remediation?
4 Links to relevant documentation
5 Actionables

Summary

2 services (citoid, cxserver) in eqiad had partial failures. In the case of citoid, results were not augmented enough, in the case of cxserver, translations were not returned.

Impact

- Citoid was not augmenting results for citation with zotero data for 46 minutes.
- cxserver was not returning translations for usage in ContentTranslation in eqiad for 35 minutes.

Detection

Detection was by both by human as well as Icinga.

Timeline

CoreDNS was enabled in codfw+eqiad today over the course of many hours (a process slowed down on purpose). Unfortunately Calico network policy rules were not correctly applied in eqiad (since the mid August migration to helmfile). The reason was the calico-policy-controller which does syncing between the Kubernetes NetworkPolicy objects and the Calico store, had configuration referencing the codfw datacenter.

The CoreDNS change required all pods to be restarted for the change to be applied, which was done in a slow rolling restart for loop. When it was understood that there is an issue, it was stopped, saving the other services.

- 13:43 Citoid partial outage begins. No alerts, it is evident however in Grafana dashboards
- Citoid was the service that suffered the most timewise, from 13:43 - 14:29

https://grafana.wikimedia.org/d/NJkCVermz/citoid?refresh=5m&panelId=15&fullscreen&orgId=1&from=1568641410756&to=1568644177940&var-dc=eqiad%20prometheus%2Fk8s&var-service=citoid

- Cxserver was the other service that was hit, shortly after. Smaller time window, it did alert, although no SMS sent, as it was partial

https://grafana.wikimedia.org/d/F7rttgqmz/cxserver?refresh=1m&panelId=15&fullscreen&orgId=1&from=1568642277671&to=1568643894326&var-dc=eqiad%20prometheus%2Fk8s&var-service=cxserver

- Helmfile deploys were also problematic in eqiad due to tiller failing DNS requests to the Kubernetes API, touched one deploy, namely the first deploy of Wikifeeds in eqiad, no alert (and none seems to be required).
- 14:29 full recovery

The fix was merging the change mentioned below and applied using helmfile. Luckily enough, kube-system namespace was not touched yet by the breaking change, so there was no need to work around it.

All times in UTC.

Conclusions

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

- Humans are the weakest link. The process went perfect in codfw and hence not everything was checked before applying it in eqiad.

What went well?

- Automated monitoring detected the incident when it reached cxserver
- It was root caused relatively quickly
- It only partially impacted editors, mildly inconveniencing them.

What went poorly?

- Citoid never logged anything, making the root cause finding a tad slower
- The CoreDNS internal kubernetes service wasn't adequately tested.

Where did we get lucky?

- The operator was looking into graphs from the beginning, spotted the issue and was able to react.

How many people were involved in the remediation?

- 1 SRE

Links to relevant documentation

New service so nothing yet.

Actionables

- Fix the calico configuration https://gerrit.wikimedia.org/r/537121. ✓ **Done**
- Monitor coreDNS. https://phabricator.wikimedia.org/T234545
- Monitor zotero indirectly via citoid https://phabricator.wikimedia.org/T234544

Category: Incident documentation

This page was last edited on 8 October 2019, at 06:25.