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/20200206-mediawiki

< Incident documentation

document status: in-review

Contents [hide]

Summary

Widespread service timeouts starting immediately after moving group2 wikis to 1.35.0-wmf.18. Reverting to 1.35.0-wmf.16 immediately brought most services back to life.

A single root cause was responsible for two incidents, this one on Thursday the 6th and again on Friday the 7th. See Incident_documentation/20200207-wikidata for more relevant details on this incident.

Impact

All sites were unresponsive or at least partially offline for ~8 minutes. (Except for cache hits, which would be 'popular' pages for non-logged-in users.)

Detection

~100% of icinga alerts fired simultaneously.

Timeline

All times in UTC.

- 20:22 or so: start of deploy of wmf.18 to group 2 wikis
- 20:24 Icinga starts alerting for High average GET latency for mw requests on appservers in eqiad and various other things
- 20:25 scap completes: <twentyafterfour@deploy1001> rebuilt and synchronized wikiversions files: all wikis to 1.35.0-wmf.18 refs T233866
- 20:25 Icinga starts alerting for restbase endpoint health on all restbase servers
- 20:28 Icinga starts alerting for Varnish HTTP text-frontend health on various cp servers and Apache HTTP on mw servers
- 20:29 Grafana is not available to European users (and SREs)
- 20:30 (twentyafterfour@deploy1001) Scap failed!: 9/11 canaries failed their endpoint checks(http://en.wikipedia.org)
- 20:30 --force is used to revert the Mediawiki deploy: <twentyafterfour> sync-wikiversions --force
- 20:31 Icinga recoveries start coming in for Varnish HTTP text-frontend, PHP7 rendering on appservers, etc, but Restbase alerts remain
- 20:45 Restbase is only reporting errors from wikifeeds (in k8s), restbase is restarted on restbase1016 and

restbase1027 but that did not help recoveries

- 20:47 akosiaris kill all pods for wikifeeds in eqiad. They were in a throttled CPU downward spiral. https://grafana.wikimedia.org/d/35vIuGpZk/wikifeeds?orgId=1&var-dc=eqiad%20prometheus%2Fk8s&var-service=wikifeeds&fullscreen&panelId=28&from=1581020337234&to=1581022780963
- 20:52 Restbase alerts recovered

Conclusions

The root cause was a typo in a config setting for Wikibase.

On Jan 16 [1] this config change `565074` was deployed. This would have caused all Wikibase client wikis to read items from the new wb terms store. It did not take effect because of a typo in the name elsewhere in the config files. This means that the default setting in the Wikibase extension was used, with all clients reading from the old store.

On Jan 22 the default setting in the Wikibase extension was changed to have clients all read from the new store. This did not make it into a branch until wmf.18, because of All-Hands. It went live to groups 0 on Feb 4th, and group 1 on Feb 5th. This would have impacted Commons and Wikidata but they do very few Wikibase client reads compared to the bulk of the wikis. The change went live to group 2 on Feb 6th, when we saw the outage.

A similar but smaller scale incident occurred the next day when trying to fix that config variable typo. Friday's incident report: Incident documentation/20200207-wikidata.

The similarity of the two incidents (onset, type of impact, graphs [2], [3]) led us to believe that the root cause of Friday's incident is also the root cause of Thursday's incident. The root cause of Friday's incident was narrowed down to the specific config change because only that single change was deployed at the time of the incident on Friday, which is unlike Thursday when we were routinely deploying a larger batch of changes during the MediaWiki train.

The Monday rollout of wmf.18 to group2 without incident confirms this understanding.

What went well?

- deployer immediately noticed the deploy was bad and started reverting
- monitoring detected the incident, many additional people got paged and were online quickly

What went poorly?

- canary checks did not block the bad deploy
- scap revert could have been quicker, canary checks blocked the revert at first and revert had to be forced
- wikifeeds pods had to be killed to fix restbase
- after this incident, the underlying cause was not at all clear because (as is usual) many changes were rolled out together in the branch deploy

Where did we get lucky?

- reverting fixed most of the issues and besides wikifeeds pods we did not need additional service restarts

How many people were involved in the remediation?

- about 8 SRE, 2 DBA, 1 Releng, 2 Performance engineers

Links to relevant documentation

- TODO: specific command on our hosts? https://kubernetes.io/docs/concepts/workloads/pods/pod/
- TODO: add 'revert' section on https://wikitech.wikimedia.org/wiki/Scap ?

Actionables

- T244535 - wikifeeds: Fix the CPU limits so that it doesn't get starved
- When text-esams was down, Grafana was not available to European SREs. Workarounds below were mentioned:
  - echo $(dig +short text-lb.eqsin.wikimedia.org) grafana.wikimedia.org | sudo tee -a /etc/hosts
  - ssh grafana1002.eqiad.wmnet -L3000:localhost:3000
- conversations about moving monitoring interfaces outside the normal traffic path (Herron). continue them and turn into a ticket
- T243009 - Make scap skip restarting php-fpm when using --force

- T217924 - Make canary wait time configurable
- T244544 - add a force-revert command to scap to shorten the time it takes to revert
- T244533 - Slow query hitting commonswiki
- T183999 - scap canary has a shifting baseline (scap, why did canaries not catch the bad deploy (tangentially related))
- Consider moving to more of a continuous deployment model with only groups of related changes being deployed together (User:20after4 is thinking about this)
- As an underling problem. Typos are really easy to happen on mediawiki-config.
  - T183999 - Define variant Wikimedia production config in compiled, static files
  - T220775 - Consider creating a puppet-compiler equivalent for mediawiki-config.git

Categories: Incident documentation in-reviews | Incident documentation