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/20200108-mw-api

< Incident documentation

document status: final

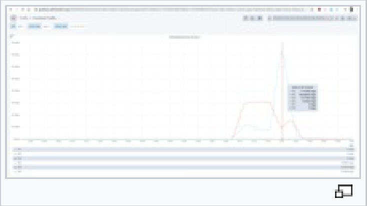
Contents [hide]

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 Actionables

Summary

A MediaWiki configuration change making EventBus use TLS for eventgate-analytics was merged. POST requests from MediaWiki application servers to eventgate-analytics started timing out. This resulted in HTTP server errors being served to users in all data centers. The very high rate of application server requests timing out caused troubles elsewhere in the stack, with 50% of the Varnish frontend caches in esams crashing due to memory allocation failures. Exacerbating the problems, although the Varnish frontend processes were automatically restarted, ATS-tls kept seeing them as down.

Impact



Between 15:06 and 15:32, ATS backend requests to the application servers resulted in multiple errors, between 3K and 7K 5xx per second. See Grafana.

User facing impact has been two-fold: between 15:21 and 15:30 several requests received 502 error responses, with a peak of 2759 errors per second at 15:26. Between 15:06 and 15:45, many requests received no response at all. See Grafana.

At the same time of this incident, one day before, the 2xx response rate was between 133k and 136k responses per second. During this incident it was ~100k rps between 15:11 and 15:21, and as low as 70k rps around 15:37.

All analytics events in the mediawiki.api-request and mediawiki.cirrussearch-request streams produced between 15:05 and 15:30 were lost.

Detection

The SRE team was first notified about the issue by various Icinga PHP7 rendering alerts on IRC, shortly followed by multiple pages regarding api.svc.eqiad.wmnet socket timeouts, as well as ATS TLS and Varnish reduced availability.

Timeline

All times in UTC.

- 15:04: Scap sync started for wmf-config/ProductionServices.php: Make EventBus use TLS for eventgate-analytics - https://phabricator.wikimedia.org/T242224

- 15:06: varnish-fe crashes on both cp3050 and cp3054 (Cannot allocate memory)
- 15:08: First of many similar alerts on irc: PROBLEM - PHP7 rendering on mw1280 is CRITICAL: CRITICAL - Socket timeout after 10 seconds
- 15:09: varnish-fe crashes on cp3058 (Cannot allocate memory)
- 15:10: PROBLEM alert - api.svc.eqiad.wmnet/LVS HTTP IPv4 #page is CRITICAL
- 15:10: varnish-fe crashes on cp3062 (Cannot allocate memory)
- 15:10: <otto@deploy1001> Synchronized wmf-config/ProductionServices.php: Make EventBus use TLS for eventgate-analytics - https://phabricator.wikimedia.org/T242224 (duration: 06m 10s)
- 15:11: _joe_ points out on IRC that there are known issues with php-fpm+TLS, and asks ottomata to revert
- 15:11: ottomata tries reverting but the deployment fails
- 15:12: PROBLEM alert - icinga1001/ATS TLS has reduced HTTP availability #page is CRITICAL
- 15:17: PROBLEM alert - text-lb.esams.wikimedia.org_ipv6/LVS HTTPS IPv6 #page is CRITICAL
- 15:17: Cause of failure in reverting the change identified: php-fpm cache checks
- 15:19: RECOVERY alert - text-lb.esams.wikimedia.org_ipv6/LVS HTTPS IPv6 #page is OK
- 15:19: otto@deploy1001 sync-file aborted: REVERT Make EventBus use TLS for eventgate-analytics - T242224 (duration: 06m 33s)
- 15:19: ema notices that several varnish frontends in esams has crashed by looking at the icinga web ui (the events are reported as warnings)
- 15:20: varnish-fe crashes again on cp3058 (Cannot allocate memory)
- 15:20: ottomata tries syncing again
- 15:21: deployment still hanging on checking php-fpm cache
- 15:22: varnish-fe crashes again on cp3050 (Cannot allocate memory)
- 15:23: varnish-fe crashes again on cp3062 (Cannot allocate memory)
- 15:23: thcipriani mentions on irc that --no-php-restart needs to be passed to scap
- 15:24: ottomata tries running scap once again but gets an error: extra arguments found: --no-php-restart
- 15:25: _joe_ comments out temporarily the php-conditional restarts from scap, which were preventing the deployments from finishing
- 15:26: PROBLEM alert - icinga1001/Varnish has reduced HTTP availability #page is CRITICAL
- 15:26: Another deployment attempted, this time successfully: <otto@deploy1001> Synchronized wmf-config/ProductionServices.php: REVERT Make EventBus use TLS for eventgate-analytics - T242224 (duration: 00m 34s)
- 15:26: RECOVERY alert - api.svc.eqiad.wmnet/LVS HTTP IPv4 #page is OK
- 15:29: RECOVERY alert - icinga1001/Varnish has reduced HTTP availability #page is OK
- 15:30: PROBLEM alert - text-lb.esams.wikimedia.org_ipv6/LVS HTTPS IPv6 #page is CRITICAL
- 15:30: RECOVERY alert - icinga1001/ATS TLS has reduced HTTP availability #page is OK
- 15:30: vgutierrez notices very high ats-tls CPU usage on various nodes
- 15:35: PROBLEM alert - text-lb.esams.wikimedia.org/LVS HTTPS IPv4 #page is CRITICAL
- 15:36: ats-tls restart on cp3050. Process up at 15:37:43
- 15:36: ats-tls restart on cp3056. Process up at 15:37:00
- 15:37: rolling restart of ATS backends in text esams to reclaim some memory by applying https://gerrit.wikimedia.org/r/#/c/operations/puppet/+/562849/
- 15:37: esams text traffic depooled in DNS
- 15:41: ats-tls restart on cp3052. Process up at 15:41:29
- 15:43: ats-tls restart on cp3054. Process up at 15:43:14
- 15:43: RECOVERY alert - text-lb.esams.wikimedia.org_ipv6/LVS HTTPS IPv6 #page is OK
- 15:43: RECOVERY alert - text-lb.esams.wikimedia.org/LVS HTTPS IPv4 #page is OK
- 15:45: ats-tls restart on cp3058. Process up at 15:46:45.
- 15:46: ats-tls restart on cp3062. Process up at 15:48:02.
- 16:00: esams text traffic re-pooled in DNS

Conclusions

What went well?

- Incident detection worked properly, the SRE team was quickly notified about the issue both on IRC and by SMS.
- Within 5 minutes from incident detection, a member of the SRE team identified the root cause and suggested

a course of action.
- Communication was effective.

What went poorly?

- The configuration change triggering this incident was known to be problematic by the Service Operations team. The change was merged without code review.
- Scap does support canary deployments, yet the problems were not identified on the canary and the problematic change got rolled out to all application servers.
- Although root cause detection was fairly quick, reverting the change took 15+ minutes. The deployment procedure was unclear and confusing. Carrying out the revert successfully required a manual change to scap.
- Varnish frontend crashes should not happen regardless of what is going on at the application layer. Such crashes are important events and must show up on IRC as critical.
- The interaction between ATS-tls and Varnish frontends when Varnish crashes under high load is buggy and not well understood. It took way too long to find out that ats-tls was in trouble once the varnish crashes were noticed (10+ minutes).
- Depooling esams resulted in eqiad issues almost as soon as traffic moved. Eqiad is not able to sustain esams traffic.

Where did we get lucky?

- _joe_ was around and paying attention when the issue started.
- Varnish frontend crashes affected only one DC (esams), albeit a very important one.

How many people were involved in the remediation?

- 6 SREs.

Actionables

- Status: 🟥 **TODO** Make scap skip restarting php-fpm when using --force https://phabricator.wikimedia.org/T243009
- Status: 🟥 **TODO** Scap should be able to wait longer on the canaries https://phabricator.wikimedia.org/T217924
- Status: 🟩 **Done** varnish-fe should not crash trying to allocate memory https://phabricator.wikimedia.org/T242417
- Status: 🟩 **Done** varnish parent should be able to send signals to its child https://phabricator.wikimedia.org/T242411
- Status: 🟩 **Done** Raise severity of varnish child restart to critical: https://gerrit.wikimedia.org/r/#/c/operations/puppet/+/563174/
- Status: 🟥 **TODO** ats-tls should immediately mark varnish as up after the latter restarts https://phabricator.wikimedia.org/T242620
- Status: 🟥 **TODO** Per-cgroup CPU graphs would allow to quickly find out misbehaving processes: https://phabricator.wikimedia.org/T183146

Category: Incident documentation