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/20150205-SiteOutage

< Incident documentation

Contents [hide]

Summary

Wikimedia sites suffered from a full site outage starting at 17:10 UTC. Most wiki pages would not load and result in HTTP 503 errors from our Varnish caches.

This was caused by a loss of power by a critical network switch during maintenance on different equipment in the same rack, which caused a full rack to go offline for about 10 minutes until the failure was detected. The impact of this single rack failure was unacceptably high, as all memcached servers in eqiad reside in this rack, representing a single point of failure. After rack connectivity was restored at 17:21 UTC, the sites remained down due to cascading failures in several backend layers, preventing the new HHVM application servers from serving pages due to multiple configuration issues. Due to our recent migration of MediaWiki from PHP servers to HHVM and the many changes across our application stack, the failures took longer to investigate and narrow down than usual. The issue of HHVM application servers being unable to serve pages was tracked down to an overload in our new logging system to Logstash along with suboptimal timeout values in the MediaWiki configuration.

By 17:47, it was once again possible to load pages, but users were not able to log in. This was caused by memcached proxies not correctly reestablishing connections when the Memcached servers came back online. Full functionality was restored at 18:06 when all nutcracker (memcached proxies) instances were restarted.

Timeline

All times are UTC.

- [17:10] asw2-a5-eqiad (network switch) lost power
- [17:12] Icinga reports multiple memcached and Varnish caches (in rack A5) down
- [17:12] Users report site downtime with 503 errors from Varnish; all sites are affected. Icinga reports HHVM application services as critical, which is expected as cascading failures.
- [17:12] Chris (on-site engineer) reports that server downtime may be due to power rearrangement work in a rack. Operations starts investigating and suspects a full rack power outage
- [17:12] Chris investigates the power issues on the related servers. The Operations team scrambles to gather exact information
- [17:17] Chris realizes that not the servers, but the switch in the rack is down. He power cycles it and reports this
- [17:21] The network switch finishes boot and the servers in rack A5 come back online. Icinga reports all rack hosts as up. The site is expected to recover shortly after
- [17:23] The site doesn't recover as expected, and Icinga keeps reporting unresponsive HHVM application servers. The Operations team investigates the HHVM servers, and tries to mitigate the blocked, backlogged servers by restarting in batches now memcached and Varnish are available again. This continues for 20 minutes with modest improvements but doesn't fully stabilize the site.
- [17:44-7] Giuseppe discovers HHVM instances blocked on logging by MediaWiki to Logstash due to a too high timeout, and Ori disables Logstash logging entirely
- [17:47] The web sites mostly recover and articles start loading again. Some issues remain due to memcached still being unavailable to the appservers, the most obvious being users unable to log in.
- [18:01] Ori correctly theorized that nutcracker (memcached proxy on all application servers) was not properly proxying memcache queries to memcached servers and restarts most nutcracker instances.

- [18:06] All remaining nutcracker instances are restarted, and all functionality recovers.

Conclusions

Multiple problems resulted in these cascading failures today:

- Despite all equipment in rack A5 having a redundant power configuration, power rearrangement work still caused a critical network switch and two servers to lose power. Power rearrangement work needs to be done more carefully in announced maintenance windows, and this would avoid investigation delays due to a lack of information.
- All memcached servers in eqiad reside in a single rack, which is a single point of failure that has been known for some time. We need to address this issue as soon as possible by spreading these servers across multiple racks and rack rows, just like everything else.
- Memcached being unavailable caused a storm of log messages, overloading our Logstash infrastructure. This, and the unnecessarily high timeout on logging failures caused MediaWiki/HHVM to get blocked, unable to serve further requests. The new MediaWiki logging needs to be decoupled from logging infrastructure downtime.
- Nutcracker (memcached proxy) needs to be tested with more failure modes and fixed to automatically recover from memcached failures without requiring manual intervention

Actionables

These tasks will be collected and further developed at https://phabricator.wikimedia.org/tag/incident-20150205-siteoutage in the following days.

- Split memcached service across multiple racks, and remove this single-point-of-failure
- Decouple logging infrastructure failures from MediaWiki logging
- Nutcracker needs to automatically recover from MC failure

Category: Incident documentation