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/20150825-Redis

< Incident documentation

Contents [hide]

Summary

A surge in the number of Cirrus Elastica Write jobs happened to send to the jobqueue more jobs than could be processed. This resulted in a progressively larger memory usage by the redis servers, which resulted in their inability to trim the AOF file we use as persistency. This resulted in Redis servers filling up their disks and finally becoming unable to process any job and in ocg malfunctioning.

Timeline

- Aug 24 17:58 ElasticSearch indexes get freezed. Chase froze the index to do a test restart of an Elasticsearch node with alternate allocation settings for T109104
- Aug 24 18:15 Chase tries to unfreeze the indexes as the test is over but issues the wrong command "mwscript maintenance/showJobs.php --wiki enwiki --group --thaw" instead of "mwscript extensions/CirrusSearch/maintenance/freezeWritesToCluster.php --thaw". The command 'succeeded' without error but did not unfreeze the cluster.
- Aug 24 18:00 - Aug 25 04:00 the network traffic and the number of submitted jobs continues to grow, as all ES jobs get resubmitted as they fail, and this causes a thundering herd effect.
- Aug 25 00:00 Around this time, all of the redis hosts reached a very high memory mark. This made in turn the background trimming of the AOF file impossible. Disk space is quickly exhausted
- Aug 25 03:57 First critical alert on rdb1003's disk
- Aug 25 04:10 (timing is approximate) Connections to redis become difficult given the huge in-memory dataset that probably makes the CPU be a bottleneck, the number of jobs processed plummets.
- Aug 25 04:19 Redis alerts on rdb1003 that has meanwhile filled its disk completely start to come out
- Aug 25 05:45 Giuseppe wakes up and sees the alerts, investigation starts
- Aug 25 06:00 After determining the disks are full on all redises due to a huge AOF, Giuseppe finds out there is no way to trim them at the moment. Tim, Ori and Aaron join the investigation
- Aug 25 06:40 An attempt to take a snapshot on rdb1002 (after failing to do so on other servers) is started. This will finally prove to end successfully. Ocg service starts flapping.
- Aug 25 07:00 Tim notices the network graph for the redises has a huge increase since last afternoon, a similar increase in the jobqueue is found by Giuseppe. Aaron offers to wipe some of the queues clean
- Aug 25 07:11 After attempting some recovery strategies without success, it's decided to wipe clean rdb1003/rdb1004 and they become available again. At this point, ocg and the jobqueue start working again
- Aug 25 07:35 After the snapshot (27 GB of compressed data!!) is finally done on rdb1002, the wipe/restart cycle is performed on the other servers, rdb1001/1002; This ends the practical outage, root cause search is ongoing.
- Aug 25 08:23 Aaron finds out that elastica jobs are indeed abnormally high. This from the write freeze still being in effect.
- Aug 25 08:37 David finds out that the indexes of ES are indeed frozen, unfreezes them.

Conclusions

Our redis installation has a large number of flaws that need to be corrected; our monitoring is not adequate and we should add paging on critical things like redises and other critical storage. Finally, a stricter policy of logging what people do in production to the SAL should be enforced - the outage would've been solved much earlier if Aaron, Tim and Giuseppe didn't have to spend about one hour trying to guess what was happening.

Actionables

- Status: 🟥 **Unresolved** Allow vm.overcommit on redis hosts with persistence turned on (bug T91498)
- Status: 🟥 **Unresolved** Monitor memory and disk usage of the redis process (bug T110169)
- Status: 🟥 **Unresolved** Monitor Elasticsearch index freezing (bug T110171)
- Status: 🟩 **Done** Update Elasticsearch for missing time period (bug T110179)
- Status: 🟨 **On hold** Maintenance scripts should fail on unknown parameters (bug T110209)

Long term goals

- Better metrics from the job queue: bug T62105

Category: Incident documentation