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/20170419-restbase

< Incident documentation

Contents [hide]

Summary

Two Cassandra nodes failed to return to service after OOM exceptions, resulting in an elevated error rate.

Timeline

> *Everything that transpired here occurred **after** a routine switchover of RESTBase from eqiad to codfw, but **before** the switchover of mediawiki.*

- 2017-04-18T23:01:00: restbase1018 disk failure
- 2017-04-19T01:21:00: urandom: T163292: Starting removal of Cassandra instance restbase1018-a.eqiad.wmnet
- 2017-04-19T03:53:00: urandom: T163292: Starting removal of Cassandra instance restbase1018-b.eqiad.wmnet
- 2017-04-19T11:28:27: 2005-c DOWN (cause: OutOfMemoryException)
- 2017-04-19T11:29:17: 2010-b DOWN (cause: OutOfMemoryException)
- 2017-04-19T11:29:00: Elevated 5xx rate
- 2017-04-19T11:36:17: 2010-b UP (service started, but node did not join cluster or answer requests)
- 2017-04-19T11:40:27: 2005-c UP (service started, but node did not join cluster or answer requests)
- 2017-04-19T12:38:00: urandom: T163292: Starting removal of Cassandra instance restbase1018-c.eqiad.wmnet
- 2017-04-19T13:05:00: `nodetool removenode` invoked to force completion of restbase1018-c
- 2017-04-19T13:11:17: 2010-b UP (joined; answering requests)
- 2017-04-19T13:13:17: 2005-c UP (joined; answering requests)
- 2017-04-19T13:10:00: Some requests failing w/ User restb has no XXX permission on... errors
- 2017-04-19T13:28:00: urandom: `cqlsh -f /etc/cassandra/adduser.cql`, recreating user/perms (as-needed)

The previous day (the 18th), a disk failure in restbase1018 resulted in outages of the 3 instances running there. This occurred *after* the planned switchover to codfw and so resulted in no client-facing impact. Since it was apparent that all 3 instances were unrecoverable, work began immediately to decommission the affected instances, and restore redundancy.

Sometime later, restbase2005-c.codfw.wmnet and restbase2010-b.codfw.wmnet went down due to OOM exceptions. The cause of these OOMs are likely the result of known issues, but it would be unusual for them to occur outside of the data-center where change-propagation updates are being processed (eqiad at the time), so more investigation is needed. Regardless of the cause though, the loss of two instances resulted in the elevated 5xx rate as it became more difficult to achieve quorum for some results.

Normally such outages would be short-lived, however, due to a bug in Cassandra, pending range movements (such as though generated by the on-going `removenode` operations) can block normal startup. Once this became apparent, the `removenode` for restbase1018-c was force-completed, 2005-c and 2010-b started up normally, and the elevated error rate subsided.

The replica corresponding to row 'd' of the `restb` user and/or roles apparently resided on 1018-c, once the `removenode` was forced, 1018-c was no longer a part of the cluster (it was no longer a downed node). Another of the row 'd' nodes became responsible for the replica, but without a record to serve, authentication failures occurred ("User restb has no XXX permission on..."). [1] This impacted a fraction of the queries [2] in eqiad until a recreation of the user and permissions forced re-replication. Since client requests were being served from codfw, only change-propagation was impacted, and all failed requests subsequently succeeded after retries.

Conclusions

- Any impact was minimal and relatively brief but resulted in a lot of uncertainty in the run-up to the data-center switch over
- While the two codfw nodes failed to fully start due to the restbase1018-b removal, the restbase1018-c removal should not have been started until this was resolved. Had this been postponed, the restbase1018-c removal would not have required force-completing, and the authentication failures could have been avoided.

Actionables

- Consider increasing `system_auth` keyspace replication (Task T158908)
- ~~Follow-up with upstream regarding startup failure during topology changes~~, CASSANDRA-12281
  - Consider upgrading to 2.2.9
- Resolve OOM shutdowns (Task T156199, Task T144431, and Task T163506)
- Eliminate RAID-0 blast radius with a JBOD configuration (Task T160570)

Footnotes

1. ↑ `system_auth` is queried at LOCAL_ONE consistency level
2. ↑ Driver load-balancing notwithstanding, ~1/3 of requests would have failed

Category: Incident documentation