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/20171116-s5-dewiki-wikidata

< Incident documentation

Contents [hide]

Summary

s5 (de.wikipedia.org/wikidata.org) was read-only for periods of time and s5 suffered from reduced capacity due to a primary master server hardware crash (https://phabricator.wikimedia.org/T180724)

Once the master was correctly replaced by another host replication broke in numerous slaves due to an ongoing schema change (T174569) (which is something habitual), some slaves had the schema change applied and some others didn't.

As the new master was running ROW based replication, the schema change (adding a column) was incompatible with those slaves that already had the new column, and hence, replication got broken. Rolling back the schema change was safest option chosen, despite it would take 10 hours.

s5 was left with 1 master and 2 slaves serving all traffic for around 10 hours until all the servers got the schema change reverted and were able to go back into the pool.

Timeline

This is a step by step outline of what happened to cause the incident and how it was remedied.

- 2017-11-16 16:48:00 UTC: db1071 is restarted for regular maintenance (while this is not terribly relevant, it made it unavailable).
- 2017-11-16 17:20:05 UTC: Icinga complains about SSH service not working on db1063
- 2017-11-16 17:21:28 UTC: jynus notices increasing lag for s5 on all slaves
- 2017-11-16 17:25:42 UTC: Pages to all of ops are being sent, more people show up to help.
- 2017-11-16 17:26:31 UTC: First efforts to login into the out-of-band console after realizing SSH is unresponsive
- 2017-11-16 17:27:15 UTC: Realization out-of-band isn't working in any way
- 2017-11-16 17:28:24 UTC: godog notes kernel stack traces in the central syslog for db1063. Box is presumed dead, problems with hardware RAID controller
- 2017-11-16 17:28:48 UTC: jynus notes that the best possible candidate for promotion (db1071) is not a good choice as it was being restarted and under maintenance (replication was not up to date)
- 2017-11-16 17:32:01 UTC: On site personnel is being called to the rescue, however ETA is larger than desirable
- 2017-11-16 17:36:43 UTC: All slaves are found to be equally caught up to the master, decision to failover to one of them is taken. db1070 is picked (this turns out to be important) https://gerrit.wikimedia.org/r/#/c/391870/ https://gerrit.wikimedia.org/r/#/c/391869/
- 2017-11-16 17:39:19 UTC: bblack shuts off db1063's switchport to make sure we don't end up with a split brain if and when db1063 comes back online (note: while this is good, the first thing jynus did was to stop db1070 replication once we decided to failover, plus we do not start mysql automatically in the first place precisely for that)
- 2017-11-16 17:40:32 UTC: An "All Halt" on all merges and deployments is called for the first time
- 2017-11-16 17:41:36 UTC: First slave reports having db1070 as master and having caught up with it. Rest follow
- 2017-11-16 17:51:19 UTC: mw2251 reported as being down causing some minor confusion. Ends up being

an unrelated hardware problem tracked in https://phabricator.wikimedia.org/T180724
- 2017-11-16 18:01:35 UTC: jynus reports db1070 is now the master and is set as R/W
- 2017-11-16 18:03:23 UTC: replication reported as broken. ROW replication method plus in-progress schema changes applied to only some of the slaves is the cause. Only 2 replicas out of 10 in total are capable of working with db1070 as master.
- 2017-11-16 18:04:00 UTC: discussion begins on what to do
- 2017-11-16 18:09:02 UTC: cmjohnson1 reports that fixing the hardware RAID error on db1063 via changing the battery didn't work. Will proceed with swapping the entire card
- 2017-11-16 18:19:29 UTC: wikidata editing is reported as giving timeouts. dewikipedia is assumed to be the same. Discussion about what to do goes on. Options include: rolling back last 30 mins and using the 7 servers, continuing with current 4 servers and try to fix the other 7, going read-only for a few hours while trying to fix things, reinserting edits (counted at 555) at some servers.
- 2017-11-16 18:27:15 UTC: In order the split brain becoming worse while debating options, read only is enforced again for s5.
- 2017-11-16 18:44:03 UTC: Decision is taken to proceed with the 4 servers and fix the other 7. The 7 servers start being depooled from configuration. The process of reverting schema changes on 1 of the 7 servers starts.
- 2017-11-16 18:48:02 UTC: s5 goes back to r/w since the decision was taken.
- 2017-11-16 18:51:03 UTC: First stage of above process worked fine, starting it throughout all servers
- 2017-11-16 19:00:00 UTC: Reevaluating deployments. Decision holds to pause them while fixing the situation.
- 2017-11-16 19:12:35 UTC: db1063 is back but network separated on purpose and with mysql stopped. A while later network is plugged in but box remains unused.
- 2017-11-16 20:15:00 UTC: After fully depooling all 'bad' servers, things have cooled off. Proceeding with degraded performance for the next few hours for s5 wikis, but otherwise working.
- 2017-11-17 04:48:00 UTC: reverts for logging and recentchanges tables are done for most hosts. https://phabricator.wikimedia.org/T180714#3768985
- 2017-11-17 04:51:00 UTC: archive table done https://phabricator.wikimedia.org/T180714#3768990
- 2017-11-17 04:56:00 UTC: ipblocks, filearchive, oldimage, protected_titles done. Servers are now catching up https://phabricator.wikimedia.org/T180714#3768994
- 2017-11-17 05:09:00 UTC: All hosts are ready to be pooled. https://gerrit.wikimedia.org/r/#/c/391995/
- 2017-11-17 07:15:00 UTC: All hosts pooled again.

Conclusions

What weakness did we learn about and how can we address them?

- IDRAC/iLo not working was an issue as the source of the problem took a bit longer to be identified.
  - Maybe trying to monitor the idrac in a way that we can check if the log-in can actually happen
- In order to avoid replication issues with on-going schema changes, the new master should've been running STATEMENT based replication.

Actionables

Explicit next steps to prevent this from happening again as much as possible, with Phabricator tasks linked for every step.

- s5 primary master db1063 crashed (phab:T180714)

Category: Incident documentation