Incident documentation/20110623-Performance

< Incident documentation

On 23rd June 2011 at about 15:35 UTC, WMF Engineering reported access and performance problem with some of its web properties, which was subsequently resolved by 16:30 UTC.

A summary of the sequence of events, observed symptoms, impact, root cause analysis and actions/lessons learned are noted below.

Contents [hide]

Sequence of Events

UTC time

- 15.35 Roan reported intermittent access and performance problems with the site. TechOps was paged
- 15.50 Roan rolled back an earlier commit - 03:18 logmsgbot: tstarling synchronized php-1.17/includes/db/DatabaseMysql.php 'Database::getLag() fix from r90644'
- 15.58 Ryan restarted pybal on lvs3 (load balancer)
- 16.17 Ariel noticed the wrong file was reverted and Ryan corrected it (reverted the 3:18 commit)
- 16.20 DB error logs still inundated by errors
- 16.24 Upped the max threads configuration in db.php. No impact observed
- 16.24 Escalated issue to Tim
- 16.28 Tim made a change to disable max threads altogether
- 16.32 Site seems to be stabilized and the earlier db error messages are gone

Observed Symptoms

- DB Error Log was showing lots of "LB failure with no last connection" errors which started to appear suddenly at 10:53 UTC
- They appeared to be caused by queries from LoadMonitor::postConnectionBackoff cycling thru each slave for a wiki, then bailing
- Most of the errors seem to be pointing to db16 though there were a few to db36
- While the site started to experience access and performance issue, Nagios was reporting problems with AMS squids but TPA squids look fine.
- A couple of US Apache servers started to report connection timeout problems
- php exception.log started flooding at 15:14 format is "date time server db: " with nothing else on the line; Those error messages stopped at 16:30
- reference to the earlier (3:18 UTC) commit - http://svn.wikimedia.org/viewvc/mediawiki/branches/wmf/1.17wmf1/includes/db/DatabaseMysql.php?r1=81457&r2=90644

Impact

- Problems were reported on all wikis. An s7 outage would hit logged-in users on all wikis because centralauth lives on s7.
- Users experienced performance degradation, page timed out issue and/or trying to authenticate into the various WMF sites.
- The duration of this problem lasted about 1 hour.

Root Cause Analysis

At this time, the root cause has not been determined yet. However we have identified (as well as eliminated) some possibilities:

- It is unlikely the R90644 caused the problem because reverting it did not solve the problem.
- It is possible that LoadMonitor was preventing most queries to S7 / CentralAuth due to the number of connected threads on these db's. However, much or all of this load seems to have been generated by LoadMonitor itself. From analyzing the slow query logs for the duration of the problem, all the entries came from the load-balancer.
- Threads_connected and load got very high on the S7 db's but as soon as LoadMonitor was disabled, the load went away. This was the only place $conn was accessed in getReaderIndex(). Is it possible that $conn was occasionally (and wrongly) point to S7 / CentralAuth even on non-authenticated requests? This is still to be investigated.
- With most authenticated requests failing and potentially some percentage of non-authenticated requests (if the above theory pans out), the squids started failing because large percentage of requests were retrying failed back-end requests.

Recommendations/Lessons Learned

- Tim suspects there was a load spike on S7 and load balancer never recovered from it (possibly a bug with load balancer). Given that, we should continue to disable max threads. Domas and Tim have discussed about measuring the load on a DB without connecting to it. Maybe it is time to revisit this option.
- Asher recommends investigating LoadMonitor to determine whether it is the root cause. Meantime, he has modified the code to capture and display more meaningful error messages should this happen again.
- When Ryan wanted to escalate to Tim, he could not get to the contact list on office.wikimedia.org (because CentralAuth was impacted). Suggest we keep a copy of the contact list on fenari.
- Domas:
- Mark:
- Ryan:
- Roan: The particular failure mode that probably occurred here (centralauth DB down, affects logged-in users on all wikis) is not one that I really thought of, or that is documented to my knowledge. I should've investigated whether uncached anon requests went through on non-s7 wikis; they probably did (given the DB issues and the Apache CPU graphs) and that would've made it more obvious where to look. Documenting this failure mode and making people at least vaguely aware of it will help diagnosing it in the future. It would probably have been useful if I had already known that the centralauth DB was causing the cluster-wide issues, either by observing that anon backend requests worked but logged-in requests didn't, or by noticing most DB names/IPs in the DB error log (the LB errors that spammed the log didn't contain any; we fixed that, right?) were s7 DBs.

Category: Incident documentation