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/20151217-cxserver

< Incident documentation

Contents [hide]

Summary

service-runner migration of cxserver (https://phabricator.wikimedia.org/T101272) along with its Puppet config was planned and cxserver service was intrupped during this deployment. See Timeline for full details and conclusion on lesson learned.

Timeline

- 08:00Z Language team daily meeting. Service-runner migration is expected to happen today, but time is not known as it was not on Deployments. Runa was assigned to check with Kartik.
- 12:22Z Kartik announced in Language team chat that deployment is in progress and asks Santhosh to be around. These messages are not delivered on-time, given Santhosh does not respond and that Niklas asks at 12:40 whether deployment is in progress after seeing notices on #mediawiki-i18n.
- Kartik, Alex and Marko set service-runner migration of cxserver on Thursday 5.30 PM IST (https://wikitech.wikimedia.org/wiki/Deployments#Thursday.2C.C2.A0December.C2.A017)
- Alex depooled sca1001 to make sure cxserver is unintrupted. sca1002 continued to serve traffic.
- Alex disabled puppet and salt on sca1002 so that no changes happen to it at all and it continues to serve traffic.
- Alex merged https://gerrit.wikimedia.org/r/#q,250910,n,z and deployed it on sca1001 (service-runner migration for cxserver).
- Kartik deployed, https://gerrit.wikimedia.org/r/#q,258435,n,z (Update mediawiki-contenttranslation to 1eed8b4).
- Kartik tested that cxserver is OK at API end-point.
- At this point Alex and Kartik thought cxserver on sca1001 was OK. Monitoring reported an OK state. Unfortunately testing/monitoring was not complete enough and hence several problems detailed below surfaced.
- Alex pooled sca1001. However it is not used as it fails. That is not immediately obvious however. Monitoring keeps on saying OK, however pybal uses a different check and that fails. Effectively sca1001 is unpooled.
- Alex depooled sca1002, ran puppet on it and starts salt. However pybal does not depool it as it is the only backend left.
- Kartik deployed cxserver on sca1002.
- cxserver on sca1002 no longer works.
- Pages for LVS cxserver service are sent to ops.
- Language pairs (Registry) were empty ({}) on cxserver.
- 12:36Z #mediawiki-i18n starts sending notices that Apertium in WMF is down.
- Announces the service disruption at twitter: https://twitter.com/WhatToTranslate/status/677486473055105024
- 12:44Z-12:45Z Alex notices that / returns 404 and fixes the pybal configuration to use /_info https://gerrit.wikimedia.org/r/#/c/259669/ for monitoring.
- 12:47Z Alex restarts pybal to load the new config on lvs1003 first and a couples of minutes later on lvs1006 as is the process.
- 12:50Z Kartik asks in Language team chat whether everything is OK.

- 12:53Z Alex also fixes the paging config: https://gerrit.wikimedia.org/r/#/c/259672/ 🔗
- 12:54Z Niklas replies that MT/dictionaries are not working.
- 12:56Z _joe_ notes that there are lots of stacktraces: https://phabricator.wikimedia.org/P2434 🔗
- 13:00Z We believe based on the stacktraces that there is something wrong with the config, Niklas is trying to check what it could be.
- 13:20Z It was determined that defaults-merging no longer happens and that we need to set everything in puppet.
- 13.31Z OK pages are sent to ops since /_info endpoint works fine. But other endpoints don't.
- Kartik with help of Alex and Niklas updated registry in hieradata since it was not picking from cxserver's config.yaml.
- 14:00Z Santhosh joins team chat to help debugging, Niklas goes back to non-WMF things.
- Alex deployed multipe hieradata patches: https://gerrit.wikimedia.org/r/#/c/259680/ 🔗, https://gerrit.wikimedia.org/r/#/c/259695/ 🔗, ..
- Language pairs (Registry) were OK.
- Niklas, Marko and Santhosh found that probably proxy was causing an issue for cxserver to connect to apertium and restbase.
- 14:57Z Alex deployed fix for Proxy: https://gerrit.wikimedia.org/r/#/c/259700/ 🔗
- 16:00Z Niklas returns, status is that MT and page loading apis are not working and that we are having proxy issues
- 16:10Z-16:20Z Niklas points out that Yandex is using global proxy, not Yandex specific proxy as intended and that we could fix it.
- 16:20Z-16:26Z Santhosh makes two fixes in cxserver: https://gerrit.wikimedia.org/r/#/c/259731 🔗 and https://gerrit.wikimedia.org/r/#/c/259733 🔗
- 16:26Z Niklas notifies Alex about the above fixes, Alex asks to wait as he has found some clues related to why no_proxy_list doesn't work.
- 16:36Z Alex says he got it working and starts preparing patches.
- 16:50Z-17:05Z https://gerrit.wikimedia.org/r/259741 🔗 and https://gerrit.wikimedia.org/r/259743 🔗
- The above bug took quite a while to fix unfortunately. There were no logs and only an unhelpful 403 HTTP error reported by cxserver.
- 17:02Z-17:05Z Status check: Yandex was still down, page loading and Apertium were back.
- 17:07Z Santhosh has identified a bug in service-runner migration that causes MT routes to fail
- 17:10Z-17:16Z Patch in and merged https://gerrit.wikimedia.org/r/#/c/259746/ 🔗
- We are reverting the two cxserver patches made earlier not to break current proxy config.
- 17:35Z Kartik deployed cxserver update and askes to test: https://gerrit.wikimedia.org/r/259752 🔗
- 17:36Z Niklas founds that it is CERT_UNTRUSTED error again
- 17:40Z Niklas suggests using previous strictSSL = false fix until we figure this out. It was clarified that we did not actually use ca-fix before service-runner.
- 17:46Z Santhosh provides a testing script to debug the issue. Five minutes later Niklas figures out how to actually run it.
- 18:02Z Debugging with the script goes on without providing clear clues. We confirm that ca file is read correctly and rule that out.
- 18:13Z eureka: Niklas figures out that we pass ca: file_path instead of agentOptions { ca: file_path } then it works.
- 18:18Z-18:24Z Patch by Santhosh submitted to gerrit and merged.
- 18:38Z Kartik deployed cxserver: https://gerrit.wikimedia.org/r/259772 🔗
- 18:00Z Announces service back status https://twitter.com/WhatToTranslate/status/677563660055748609 🔗
- 18:40Z MT confirmed working.

Conclusions

1. Better monitoring for all endpoints in cxserver.
2. Test and check config in Beta. Currently, there is no way until we merge 'cxserver/deploy' patch. So, this should have done earlier rather than on deployment day.
3. Schedule early and make sure relevant people are around to debug issues and help with testing.
4. cxserver need more error path testing. Also defensive coding for bad configurations and environments.
5. Migration to different endpoints needs changes in LVS/pybal monitoring/configuration.
6. (Language team) hangout chats are not reliable. We should we use other communication channel during outage investigations where communication and cooperation is important.

7. Some of the problems could have been avoided with more recent nodejs. Language team had already planned to expedite that and filed a request.
8. In absence of monitoring, have test checklist to confirm that important things work before claiming things are working.

Actionables

- Status: █ **Done** Add monitoring for cxserver ([bug T121776](#))
- Status: █ **Done** Switch cxserver to use Node.js 4.2 ([bug T121072](#))

Category: Incident documentation