Closed Opened 10 months ago by  **Devin Sylva**

RCA: Consul SSL Issue

Incident: production#1037 (closed)

Summary

A brief summary of what happened. Try to make it as executive-friendly as possible.

We discovered expired, self signed certificates on our consul servers. These certificates could not be renewed in the usual way because the signing key for the Cerficate Authority was no longer available. Existing TLS connections to the service were still up and passing traffic, but any change or network interruption would cause them to disconnect and not be able to reconnect. This is a problem because our database high availability setup uses those connections for service location. If any of the existing connections from web nodes or api nodes were interrupted, the wouldn't be able to find the database. If any of the connections from the database nodes were interrupted, the database would fail over and not be able to decide which is primary. Each of these situations would be very bad, and both of them together would render the entire site unusable until it was fixed. The problem in this case was that going to each machine and addressing the problem one at a time would not work. Even rolling out or pushing a change via Chef would leave us with each individual node non-functional for 1 to 30 minutes. All changes needed to be made (exactly) simultaneously, without allowing the database to fail over.

This left a lot of individual risks, and a lot of unknowns to test and validate. There were several possible solutions to work through, and after walking through them we decided that turning off validation of the certificates would both remove all of the risk, and allow time to come up with a proper solution for certificate management. All of the other options required a similar amount of effort and more importantly the same risk and process for simultaneously restarting the service everywhere. Other solutions explored were:

- Replacing the certificates and CA with another self signed cert
- Switching from a single custom CA to the system CA store and using sslmate
- Switching to a letsencrypt cert

Metadata

- Service(s) affected : Consul, Database, PGBouncer, Patroni, Web/API
- Team attribution : SRE
- Minutes downtime or degradation : 0 (10 seconds for consul, 1 minute for Patroni)

Impact & Metrics

Start with the following:

- What was the impact of the incident? (i.e. service outage, sub-service brown-out, exposure of sensitive data, ...)

The impact was the elevated risk in that any interruption in any established TCP connection would cause either a partial or total outage of GitLab.com, depending on the node(s) involved.

- Who was impacted by this incident? (i.e. external customers, internal customers, specific teams, ...)

Everyone using GitLab.com could have been impacted. In the end, nobody was impacted and nobody noticed who was not involved in the activity.

- How did the incident impact customers? (i.e. preventing them from doing X, incorrect display of Y, ...)

There was no impact to customers

Detection & Response

Start with the following:

- How was the incident detected?

This was detected by a restart of one of the database servers in staging. It could not re-connect to consul.

- Did alarming work as expected?

No. We had no alert for the expiration of this certificate, since it was never intended to go into production.

- How long did it take from the start of the incident to its detection?

3 days from certificate expiration to noticing it

- How long did it take from detection to remediation?

About 2 days of troubleshooting, planning, and a maintenance window to remediate

- Were there any issues with the response to the incident? (i.e. bastion host used to access the service was not available, relevant team memeber wasn't page-able, ...)
    - sshguard on consul servers was locking out the bastion hosts.
    - Behavior when testing in staging did not match behavior when testing in DR

Root Cause Analysis

- The SSL certificates were expired on the consul hosts
- Self signed certificates were in use and the CA key no longer existed
- No production readiness review was done
- These servers were originally a proof of concept and were later promoted to production
- Moving too fast due to the rush to switch the database high availability technology to Patroni

What went well

Start with the following:

- The process of restarting consul on all servers without causing an outage went exactly as planned.
- The team did an amazing job of covering all of the possible risks and planning around them.
- The handover between time zones was extremely helpful.

What can be improved

Start with the following:

- Using the root cause analysis, explain what can be improved to prevent this from happening again.

Our method of managing certificates is not optimal. Certificates should automatically renew in all cases.

- Is there anything that could have been done to improve the detection or time to detection?

All certificates should be monitored, especially in cases where they do not auto-renew - but even when they do A production readiness review should have caught this usage of a self signed certificate and its associated CA

- Is there anything that could have been done to improve the response or time to response?

We could have handed over the planning and response from the APAC shift to the Europe shift after the troubleshooting was finished. We decided instead to set up an emergency procedure and have the people who did the troubleshooting and testing be the ones to plan and execute the response. In retrospect this was the right decision - but if the situation had been more urgent, we could have reduced the time.

- Is there an existing issue that would have either prevented this incident or reduced the impact?

Yes: #1574

- Did we have any indication or beforehand knowledge that this incident might take place?

Since there were no alerts, we had no indication.

Corrective actions

- List issues that have been created as corrective actions from this incident.
    - production#1042 (closed)
    - #7484 (closed) - sshguard blocking consul servers

Guidelines

- Blameless RCA Guideline
- 5 whys

Edited 10 months ago by Devin Sylva

Linked issues 6

Relates to

- sshguard Blocking access to consul servers
  #7484 1
- consul: SSL certs expired on Aug 3
  production#1037
- Restart Consul agents to apply TLS configrations
  production#1042
- Upgrade Consul to Allow for Live TLS reloads
  #7472 Core Infra Te...
- Certificate monitoring
  #7487 5
- pgbouncer-leader-check.service can stop indefinately in times of failure
  production#1045

Related merge requests 1

Blog: The Consul Outage that Never Happened - Nov 8
gitlab-com/www-gitlab-... !28411 November '1...

Devin Sylva @devin added IncidentReview label 10 months ago

Devin Sylva @devin changed weight to **2** 10 months ago

Devin Sylva @devin marked this issue as related to #7484 (closed) 10 months ago

Devin Sylva @devin changed the description 10 months ago

Devin Sylva @devin marked this issue as related to production#1037 (closed) 10 months ago

Devin Sylva @devin marked this issue as related to production#1042 (closed) 10 months ago

Devin Sylva @devin marked this issue as related to #7472 (closed) 10 months ago

Devin Sylva @devin mentioned in issue production#1037 (closed) 10 months ago

Anthony Sandoval @AnthonySandoval mentioned in issue #7468 (closed) 10 months ago

Anthony Sandoval @AnthonySandoval marked this issue as related to #7487 (closed) 10 months ago

Anthony Sandoval @AnthonySandoval added workflow-infra Ready scoped label 10 months ago

Devin Sylva @devin marked this issue as related to production#1045 (moved) 10 months ago

Devin Sylva @devin mentioned in issue gitlab-com/www-gitlab-com#5096 (closed) 10 months ago

Devin Sylva @devin mentioned in merge request gitlab-com/www-gitlab-com!28411 (merged) 10 months ago

Anthony Sandoval @AnthonySandoval closed 8 months ago

ops-gitlab-net 💬 @ops-gitlab-net mentioned in issue #8119 (closed) 8 months ago

David Smith 🌴 @dawsmith added workflow-infra Done scoped label and automatically removed workflow-infra Ready label 7 months ago

Please register or sign in to reply