


Closed

Opened 8 months ago by  David Smith 🌴

# Incident review for 1222- Gitaly down on file-13

Incident: [production#1222 \(closed\)](#).

## Summary

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

- Service(s) affected : Gitaly
- Team attribution : Gitaly / GCP - cloud provider
- Minutes downtime or degradation : 77 minutes degraded

For calculating duration of event, use the [Platform Metrics Dashboard](#) to look at appdex and SLO violations.

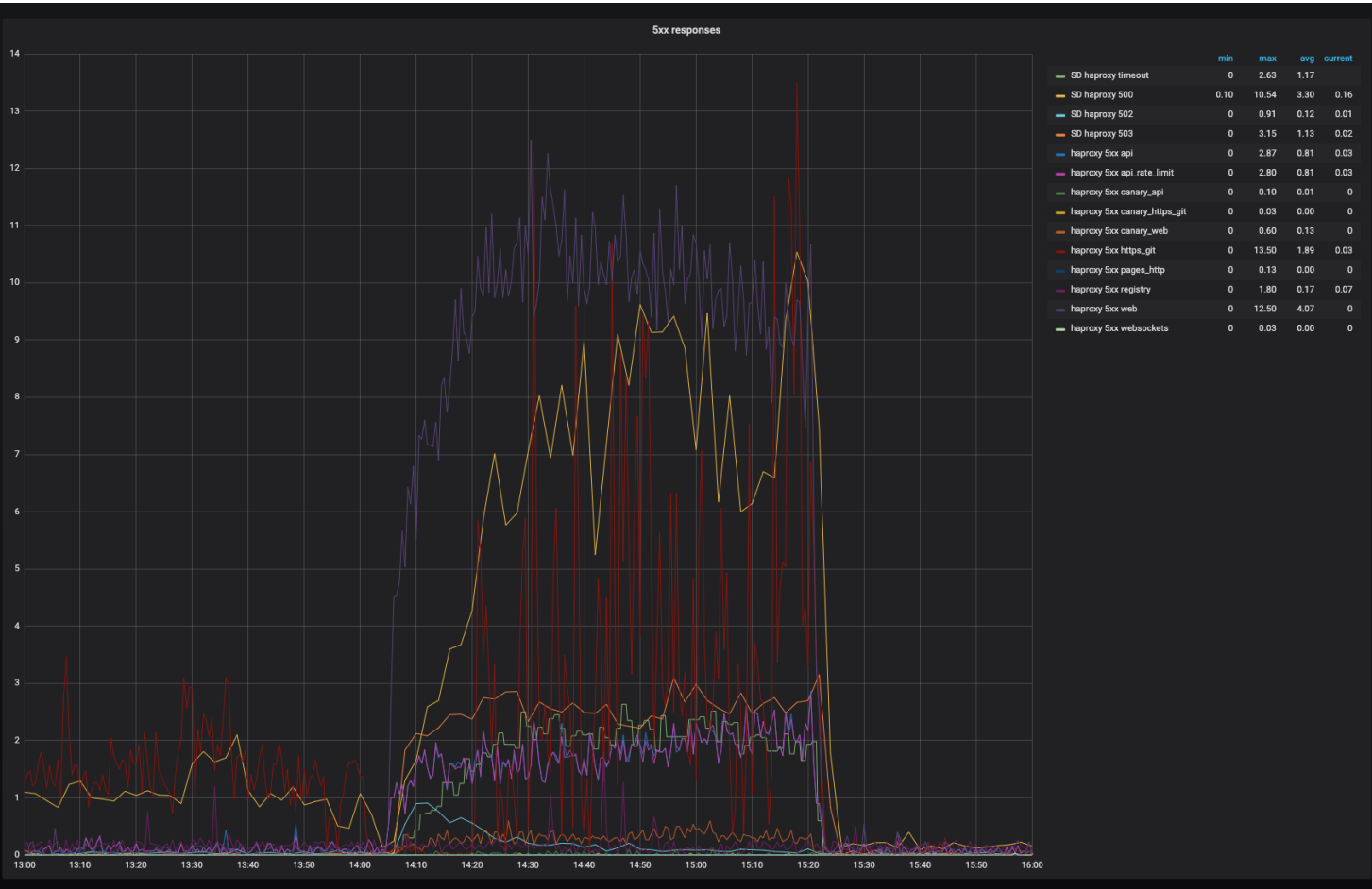
## Impact & Metrics

Start with the following:

- What was the impact of the incident? (i.e. service outage, sub-service brown-out, exposure of sensitive data, ...) Users who had data on File-13 could not interact with their git repos. In some cases, users who had data on those nodes would see 5xx errors on <https://gitlab.com>
- Who was impacted by this incident? (i.e. external customers, internal customers, specific teams, ...) Noted above - customers who had data on file-13
- How did the incident impact customers? (i.e. preventing them from doing X, incorrect display of Y, ...) unable to use GitLab.com
- How many attempts were made to access the impacted service/feature? unknown
- How many customers were affected? unknown
- How many customers tried to access the impacted service/feature? Include any additional metrics that are of relevance. As a guess - there were 99941 5xx requests during the degraded state. The prior friday about that time, there were 23183 5xx errors

Provide any relevant graphs that could help understand the impact of the incident and its dynamics.

sum(increase(haproxy\_backend\_http\_responses\_total{environment="gprd",code="5xx"}[80m])) at 15:21 UTC on Oct 4.



## Detection & Response

Start with the following:

- How was the incident detected? Paged via pagerduty for Gitaly Down on file-13

- Did alarming work as expected? Yes
- How long did it take from the start of the incident to its detection? 2 minutes
- How long did it take from detection to remediation? 14:06UTC - 15:21 UTC - 75 minutes.
- 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, ...) Response from Google / Rackspace - no visibility into what was happening to the node in the GCP console.

## Root Cause Analysis

The purpose of this document is to understand the reasons that caused an incident, and to create mechanisms to prevent it from recurring in the future. A root cause can **never be a person**, the way of writing has to refer to the system and the context rather than the specific actors.

Follow the "**5 whys**" in a **blameless** manner as the core of the root-cause analysis.

For this it is necessary to start with the incident, and question why it happened. Keep iterating asking "why?" 5 times. While it's not a hard rule that it has to be 5 times, it helps to keep questions get deeper in finding the actual root cause.

Keep in min that from one "why?" there may come more than one answer, consider following the different branches.

### Example of the usage of "5 whys"

The vehicle will not start. (the problem)

1. Gitaly was down on File-13
2. From later correspondence - the underlying host for file-13 was having maintenance done. There were problems that made file-13 unavailable.

Further questions:

1. Could we have been notified of the maintenance?
2. Why didn't we see anything in the GCP console about the status of the node?
3. What can we do to be able to recover a failed file-nn/Gitaly node sooner?

## What went well

Start with the following:

- Response time of engineer on call
- Everyone joined the incident zoom quickly

## What can be improved

Start with the following:

- Using the root cause analysis, explain what can be improved to prevent this from happening again.
- Is there anything that could have been done to improve the detection or time to detection?
- Is there anything that could have been done to improve the response or time to response?
- Is there an existing issue that would have either prevented this incident or reduced the impact?
- Did we have any indication or beforehand knowledge that this incident might take place?

## Corrective actions

- List issues that have been created as corrective actions from this incident.
- For each issue, include the following:
  - - Issue labeled as [corrective action](#) .
  - Include an estimated date of completion of the corrective action.
  - Incldue the named individual who owns the delivery of the corrective action.

## Guidelines

- [Blameless RCA Guideline](#)
- [5 whys](#)

Linked issues ⓘ

1


Relates to

 [Gitaly down on file-13](#)

production#1222




**David Smith** 🌴 @dawsmith changed due date to October 10, 2019 8 months ago




**David Smith** 🌴 @dawsmith · 8 months ago Owner

[@mwasilewski-gitlab](#) - even though this was more of a S3/S4, putting together an [IncidentReview](#) - I filled in some things, but thought you might want to add. I'm going to follow up with Rackspace/Google on some questions from here.


cc [@glopezfernandez](#) @ansdval



**David Smith** 🌴 @dawsmith marked this issue as related to [production#1222 \(closed\)](#). 8 months ago



**David Smith** 🌴 @dawsmith added [IncidentReview](#) label 8 months ago




**Andrew Newdigate** @andrewn · 8 months ago Maintainer

Corrective action: [gitlab-org/gitlab#33616 \(closed\)](#): "RootController, DashboardController and other "root-level" controllers should tolerate single Gitaly node failure"



**Anthony Sandoval** @AnthonySandoval closed 7 months ago



**ops-gitlab-net** @ops-gitlab-net mentioned in issue [#8465 \(closed\)](#). 7 months ago

Please [register](#) or [sign in](#) to reply