2021-06-17 Sunet Drive S3 Storage Incident

Name: Richard Freitag
Date: 2021-07-27

Summary

On the morning of Thursday, the 17th of June, Sunet Drive reported a lot of gateway timeouts (504) for the object storage in Sto4. The problem was reported at 10:13 CET to Safespring via their support portal. Subsequent investigations led to the conclusion that another customer was unintentionally causing a larger than expected load on the object storage, which as a result had to be taken offline. A detailed technical description of the incident can be found in the document "2021-06-17 sto4 ceph cluster". As a result, instances of Sunet Drive using Sto4 as their storage backend had to be taken offline and prioritized buckets/customers are at the time of writing waiting for the results of the restore initiative from Safespring.

A main reason for the severity of the incident was the decision to assign backup responsibilities to the customers and end-users of Sunet Drive. This led to the situation where the only copies of certain files resided solely in the affected object storage.

Customer Impact

The following customers, including an assessment of the criticality, are impacted by the outage:

Customer	Description	Criticality
Stockholm University	The primary S3 bucket, automatically created user-buckets, as well as manually assigned project-buckets were impacted. Detailed description below.	High
SciLifeLab	The pilot environment set up for SciLifeLab was affected for a limited number of S3 buckets.	Low
Pilot users	A low number of users of the pilot environment having been assigned storage in Sto4 were impacted.	Low

Impact on Stockholm University

Stockholm University (SU), as the only production user of Sunet Drive during the incident was impacted due to the architecture and design of Sunet Drive. Their instance is comprised of three components using S3 as its technical backend:

- The primary bucket of the Nextcloud instance

- User-buckets automatically created and assigned for each user of Sunet Drive
- Project-buckets requiring approval and manual assignment from SU

The Nextcloud instance was taken offline on the day of the incident and preserved for forensic analysis to support the restore initiative.

Forensic analysis of the SU Sunet Drive Nextcloud Node

The following information was extracted from the SU Sunet Drive Nextcloud node and provided to Safespring:

- A list of all configured userbuckets
- A list of all users from Stockholm University (411 in total)
- The latest storage report that has been sent to Stockholm University
- The list of all configured project buckets

The latest storage report that was created two days prior to the incident and had been sent to Stockholm University showed the following:

- 129 GB of data was stored in the primary bucket
- 1161 GB of data was stored in respective userbuckets
- 80 of 411 users had >14MB in their home folder, indicating potentially unique files
- 81 of 411 users had files in their userbucket, indicating files that require potential restore

In total, the Nextcloud instance contained 772399 files and the metainformation has been extracted from the Nextcloud database, containing the following information to support the restore efforts: path, size, mimetype, checksums. The restore initiative faces two major challenges:

- The primary bucket contains files as objects, handled via Nextcloud. A full restore of all objects is required to be able to reenable the Nextcloud instance.
- The Nextcloud sync-client introduced “online only” files in on of their recent versions and it is unknown which files do not have a local copy on the customer computers.

Resolution

At the time of writing, the restore efforts are still ongoing. All affected end users have been contacted and critical S3-buckets have been prioritized. Due to the timing of the incident before the summer holidays, it is expected that more data/buckets will be prioritized in August. The Sunet Drive pilot environment has been selected as the current replacement for Sunet Drive and is currently being hardened to avoid further incidents. The outcome of this is described in the next section.

Outcomes

The main outcome of the incident is the development and implementation of backup and mirroring of the customer data. This will eventually be done by answering the following question: *“How much data do we potentially lose if a disaster happens in one datacenter?”*. A pragmatic approach based on available skills and technical solutions will be implemented. This means, that certain technologies will not be taken into consideration within the scope of this incident (e.g.: use of CoW filesystems or RAID setups spanning multiple datacenters).

Currently there are two approaches to implement this, both with the help of the tool rclone:

- Approach 1: Replication and backup from S3 to S3
- Approach 2: Backup via TSM

Summary of approach 1: Replication and backup from S3 to S3

Replication from S3 to S3 has the goal to create an identical copy of an S3 bucket in another S3 bucket residing in another datacenter. A third bucket can be used for backup of changed files, essentially resulting in a simple implementation of "copy on write".

```
rclone sync sto3:bucket1 sto4:bucket1.clone --backup-dir
sto3:bucket1.backup/Y-m-d_H-M
```

This approach when regularly executed results in a mirrored copy of the data in bucket1 and bucket1.clone, while changed files are being saved in bucket1.backup in a timestamped folder. Essentially, all changed data will be stored and only deleted if actively implemented.

The frequency of the backup needs to be scaled with the amount of data, due to the eventual time to compute the compare operations between the buckets. Depending on the amount of data, the frequency should be minutes to hours for frequently changing data, and hours to days for infrequently changing data.

Summary of approach 2: Backup via TSM

TSM backup is currently available and provided in Sto2, providing another option for scheduled backups of changed data. A dedicated backup-worker VM can be configured to mount S3-buckets via fuse/rclone (filesystem in userspace). Recent improvements of fuse have led to the possibility of using TSM to backup this type of data. The backup-schedule needs to be aligned with the performance of TSM file operations.

Conclusion

The incident shows again that disasters can happen and that the risks need to be mitigated as early, but also as reasonable as possible. Eventually, this will result in a more resilient and performant storage solution.