2021-06-17 sto4 ceph cluster

Owner of the incident process

Name: Anders Bruvik

Incident commanders: Anders Vaage, Anders Bruvik, Janne Johansson

Overview

On the morning of Thursday 17. June, one of the users of the cluster reported that they got a high number of Gateway timeout errors. An extreme high load from another customer was observed, but given the design of the cluster, it was not unexpected that high load from one customer would lead to temporary issues of this kind. No further investigations were done until the night, when the cluster stopped serving requests. The following night and weekend were spent doing various investigations and attempts to restart the cluster, and at first it was believed that stopping the services would allow the cluster to catch up again, but the queues of stalled operations and the massive backlogs combined with the errors it caused, showed that normal recovery would not occur and on Monday morning June 21. The approach was changed from cluster repair to data recovery and restores.

Work to automate the restore process of customer data went on from Monday 21. of June until the middle of August. It was a two phased process, building tools for restoring online- and offline data and meta data and reverse engineer the CEPH algorithms where data is stored so customer data objects could be restored. During the last weeks in July, the different tools were put together in order to be able to automate the restore process. On june 17, a setback was encountered costing at least two weeks of work, see separate incident report. In August, all data that was unique to this cluster had been restored and handed over to the customer.

Impact

On June 18, 2021 at 2:15 the cluster became unavailable due to high load. After several attempts at making the cluster available again during the weekend, the high number of objects made it impossible for the cluster to get back into a stable state – and instead it was decided to start a process to restore the data from the cluster.

From June 21. to early august, work was done on writing code for the restore process, and during the first three weeks of august all data classified as critical was restored. We are still in the process of verifying that all data was restored correctly, but assuming it is, we have successfully restored all data that was unique to this cluster.

One customer was using the cluster for off-site backups for their backup solution. That customer was offered to use the client based Backup-as-a-Service meanwhile. For the other affected customers, a temporary solution have been made available at other sites.

At the time of writing (10. September) - the cluster is still not operational, but it’s decided to do a redesign to make it more suited for the intended use case. The cluster will be reinstalled when the design is ready.

Timeline

- June 17. at 10.13: Users were complaining about high error rate from the cluster
- June 18, at 2:15 the cluster became unavailable due to high load

- June 18 an incident response team were formed with participants from Safespring and with ceph consultans
- June 18: Metadata databases were compacted, investigations were made
- June 19-20: Further attempts to get the cluster into a working state
- June 20: Decision to start restoring data
- July 15: Progress was made in restoring metadata and data
- July 18: A setback due to an incident in the cluster. The restoring process had to start again
- August 17: The first buckets were restored. Work to verify the restored data was started
- August 24: After a few modifications and corrections, the first full bucket was restored
- September 6: All buckets that had data that were unique to the cluster was restored. Verification is ongoing
- September 9: incident retrospect meeting
- September 10: Incident closed

What happened

On June 17. A high load on the cluster was observed, caused by high traffic from a customer. Because the design of the cluster means limited number of parallel write operations, a large number of write operations from one user can lead to timeout errors for other users. It was thus decided that the errors were within expected behaviour for a cluster under load, and no further investigations were done at this time.

On June 18, the cluster became unavailable. Further investigations were done, and the following conclusion were made:

The high load customer data was highly imbalanced in terms of object sizes versus number of objects which made the cluster work more than anticipated to keep lists of objects and object metadata in sync among all parts. At some point the sync could not keep up with the incoming data rate and the errors started showing for all customers.

On June 18, the response team was expanded with experts from a subcontractor specializing in CEPH.

The cluster monitor databases were growing quickly, under normal operations they should be at around 1GB size, but they were growing by several GB per hour. There was a lot of slow operations in the cluster, approximately 50000 queued operations were observed. The metadata databases for the OSDs were stored on hard drives, and the cluster had a total of 31TB of metadata, which were more than the cluster could handle.

On the evening of Friday June 18. We started compacting metadata databases, and we also started to investigate why peering between OSDs was blocked, as this prevented repairs.

On Saturday, June 19, work was focused on getting the cluster to a status where it would repair itself. Metadata was reduced to 14TB. 14 placement groups were not active. Different strategies were tried, among them exporting and removing a blocked shard to an external disk to trigger repair - but unfortunately the results were not positive. One theory was that metadata was wrong, but we don’t know which, and there was a large amount of metadata. The conclusion after the day was that the cluster was in a state where it did not move forward, and we did not know why

Different experiments to try to get the cluster into a repairing state were performed on Sunday and Monday: On Sunday one of the blocking OSDs was stopped and restarted with highest level of debugging. We analysed several gigabytes of log messages from this OSD, and found no error messages. We also tried to stop OSDs, force replacement groups to OSDs etc., but on Monday afternoon, we had exhausted our options for restoring the cluster, and a decision was made to start to look at strategies for data recovery.

Now it turned out that some commands did not work: It was possible to get data from Placement Groups that were online, but for the 14 inactive placement groups, we had to start a project to write customized code to get the data. At this point we also knew that it would take days or maybe weeks before we could get any data at all, so we also made a plan for communication and set up recurring meetings with customers etc.

Restoring data started in the middle of July, but on July 18, an issue was encountered where a number of active OSDs became inactive, which again lead to restore activity in the cluster, and data was moved around. This again meant previously restored metadata was of no use, so the recovery process had to be restarted. The second round of restoring was concluded in august, and the data is currently being verified with the customer.

Root Cause Analysis

We have listed some factors that we believe contributed to the incident:

- the large number of objects would be very hard for any CEPH cluster. The cluster design is good, but not suited for this data. The average object size is only 180kb - while design of the cluster is optimized for 4mb objects. Before the incident the cluster contained about 9 million objects. When the incident occurred, the offending customer had uploaded 680 million objects
- Metadata is hosted on too slow hard drives to handle above mentioned number of objects,
- Erasure coding was chosen to drive down cost, but is inefficient when handling large amounts of objects, because the need of a lot of calculations, which leads to unreasonably long repair times for corner cases

- Focus for the design was archive storage- low cost
- High load from one customer filled up the cluster quickly.
- Few placement groups with many objects - imbalanced cluster

There were 360 placement groups in the cluster, shared among 8 pools, but not evenly. So in the end, data was only using 64 placement groups.
The high IO lead to an imbalanced cluster, and while the number of placement groups normally is tuned, the very fast increase in the number of objects meant that tuning was not possible before it was to late.
There was no rate limiting in place. Rate limiting would have meant that the cluster would fill slower and would probably have prevented some of the errors other user were seeing. But even with rate limiting in place, there were too many objects in the cluster to be handled by the spin drive based metadata databases, and we would still have a situation were the cluster could become unstable and unable to repair, it would only have taken longer time.
When error messages were observed on June 17, no further investigations were done. It was discussed whether this could have made a possible difference, but we have not concluded. A cluster with a limited number of spin drives, will reach a very high latency for a percentage of operations when the load is high, so it is not unexpected to see errors. The incident was triggered by a high number of small objects, not the high load, so if the very fast increase in number of objects had been discovered, maybe a Ceph expert could have understood that this could lead to an issue, but this is outside what we would expect a “normal” Ceph operator to know.

Resolution

After exhausting the options to save the cluster, it was decided instead to restore data. This was concluded by the end of august. After that, the cluster will need to be reinstalled before it comes back into service.

Outcomes

What did not go so well?

Restoring took longer time than expectedl, because of an incident during the restore process, this is documented in a separate incident report.

The incident happened before vacation time, and having people on vacation during the incident was unavoidable, and contributed to a longer restore time.

Our public status page turned out to be less than ideal for long running incidents, so we need to look at options for better communication during long running events and for private cloud customers.

What went well?

This was a long running incident, so there is probably a lot to say here... A few things:

The cooperation between the different teams at Safespring, the external consultants and the customers were good.

People from SUNET contributed with metadata from the Nextcloud installation, help with debugging and verification of restored files, we managed to get quick and useful communication between the involved parties.

Develop tooling that made it possible to restore all data.

Action items

- It has been decided together with SUNET to go through a new design phase before we reinstall the cluster. It needs to be verified that the design is suited for the planned usage, and probably at least add solid state drives for metadata, and look for options to increase the number of parallel operations the cluster can handle
- Other solutions are being investigated (based on NVME or block storage) for the backup data solution
- Rate limiting for S3 clusters has been implemented, both to prevent out of control growth, but also to ensure that all users get a minimum service quality.
- Limitations for number of objects per bucket and bucket per cluster has been implemented, for better control
- Further improvements to monitoring is planned - number of objects, latency times, error rate etc
- Work to improve the rate limiting is planned
- A better solution to communicate with users, especially for private cloud will be implemented in the future
- Solutions to allow customers to subscribe to events from the cluster will be investigated and put on a future roadmap.