bit
internet technology

INCIDENT REPORT ON CEPH DISRUPTION 14/04/2023

Summary

On Friday, April 14th at 11.55 am, regular operations on the Ceph shared storage cluster of BIT encountered an unknown bug in the Ceph code. The cluster became unavailable, and recovery operations ensured that the cluster was available again on April 15th at 00.39 am. No data was corrupted or lost as a result of this incident.

Background Information

Ceph is a software-defined storage solution that enables data to be stored redundantly and made available via the network. It is known for its robustness, reliability, scalability, and flexibility. A simplified description of how Ceph works and the configuration choices made by BIT is provided below.

In the Ceph cluster at BIT, data is stored three times. The cluster consists of hundreds of hard drives (SSD, NVMe) distributed over dozens of storage servers, which in turn are spread over three physically and geographically separated locations. Each of these locations is a separate failure domain, which means that infrastructure at one location may fail without disrupting the service.

Each hard drive within the Ceph cluster is an Object Storage Device (OSD). Each piece of data is stored in three different Placement Groups (PG). The PGs are distributed across the OSDs in such a way that there are never two PGs in the same physical location. Data for different services is stored in different pools, with a pool consisting of one or more PGs.

There are five (monitoring) servers that maintain the status of the cluster in a Cluster Map and hand this over to clients. Clients determine where the data is available that they want to read or write using the Cluster Map. They use the primary PG, which is one of the three PGs, to do this. In the event of a hard drive failure, a storage node failure, or even an entire data center (failure domain), a new Cluster Map is created, and the data from the remaining two locations is read with no impact on service.

Data integrity was a key consideration in the design of the BIT Ceph cluster and is considered more important than availability. The idea behind this is that it is always possible to make intact data available correctly, and that it is not certain that corrupted data will be available correctly again. This means that once it is detected that there is only one copy of the data available, this remaining copy is automatically taken offline to prevent possible data loss.

Several weeks ago, an advanced optimisation was made to the cluster after this setting was tested on test clusters. This optimisation ensures that the primary PGs are better distributed across the different OSDs in the cluster and results in better read performance of the cluster. This setting proved to be conditional to the incident.

This storage platform is the basis for all cloud and shared services at BIT, including the “hard drives” of virtual servers, storage for virtual data centers, email, and web spaces.

Details

- 11.54 A new storage node is being added to the cluster, as has been done in the past. The monitoring servers create a new Cluster Map, which is published.
- 11.55 A large number of OSDs, in all three failure domains, segfault (crash) and abort, causing most services that rely on Ceph to become unavailable. The administrators lose access to the cluster.
- 11.56 The administrators regain access to the cluster through other channels.
- 12.04 An incident report is posted on www.bit.org.
- 12.10 The OSDs of the newly added storage node are shut down.

12.15 Several crashed OSDs are restarted but none come online.
12.25 Various logs from the cluster are analysed and checked for known bugs.
13.19 A conference call is set up with experts from Croit. Croit provides Ceph consultancy and employs several (core) developers of Ceph.
13.34 Kernel and Ceph logs from one storage node are shared with Croit.
14.43 Several OSDs are restarted; one additional OSD comes online.
14.55 Ceph debug packages are installed and started on a storage node.
15.37 A framework for analysing threading and memory usage by software is installed and started on a storage node.
16.31 Specific log rules from all storage nodes are shared with Croit.
17.40 The monitoring daemons on the monitoring servers are restarted.
17.45 All OSDs of a single storage node are restarted, no extra OSDs come online.
18.02 All OSDs of all storage nodes are restarted, some extra OSDs come online.
18.10 All servers and nodes in the cluster are restarted, no extra OSDs come online.
18.20 Based on the logs, it is suspected that the OSDs have crashed and continue to crash because there are "defective" PGs on the OSDs.
18.45 A "defective" PG is found on an OSD of which two copies are still online on other OSDs. This "defective" PG is removed from the specific OSD, and the OSD comes online.
19.05 Additional capacity is added to the cluster as it has become apparent that some OSDs may be brought back online.
19.43 Another "defective PG is dumped and removed, and the OSD starts without that PG and comes online.
19.51 The "defective" PG is imported onto the same OSD, causing the OSD to crash.
20.18 A mount is set on all storage nodes to another type of storage (redundant).
20.34 For clients who have requested it, their virtual machines are prevented from becoming available once storage is functional again.
21.04 The Ceph experts at Croit report that they have not seen a similar incident before. Therefore, one OSD will not be restored so that it can be used for forensics later.
21.16 The "defective" PG that was dumped at 19.43 is successfully imported onto an empty OSD, which comes and stays online.
21.31 The "defective" PGs are identified on all OSDs.
21.38 A start is made with dumping and removing the "defective" PGs on all OSDs to the mount with the other type of storage. Once an OSD is free of "defective" PGs, it is successfully brought back online.
23.13 Several services begin to become available again because at least two OSDs that hold their data are online.
23.30 A start is made to repair services whose storage is available again.
00.06 The cluster is recovering. PGs that are available on only one or two failure domains are replicated so that they are available on all three failure domains.
00.39 All pools and data are available again.
02.48 All services that BIT is monitoring have been restored.
03.02 Recovery is complete, and all PGs are available on all three failure domains. Balancing the cluster will continue for hours.
03.03 The incident is resolved.

Conclusion

The cluster is optimised with settings that distribute primary PGs more effectively across the OSDs. When additional capacity is added to the cluster, a new Cluster Map is created and made available to clients. The use of this new Cluster Map in combination with the optimisation settings triggers a bug, causing PGs in the cluster to become "defective". These "defective" PGs cause OSDs to crash, resulting in unavailability of the pools.

The crashed OSDs can be repaired by removing the “defective” PGs from them. These “defective” PGs can be successfully imported onto empty OSDs. BIT was able to reproduce the bug and recovery operations in a test cluster. The Ceph developers still need to analyse which part of the code introduced the bug.

Points for Improvement

To reduce the likelihood and impact of a similar disruption, several measures will be taken:

- Tests on test clusters will be evaluated and, where necessary, expanded or modified to include a wider range of operations that are also expected in production.
- The optimisation setting will no longer be activated on the BIT Ceph cluster as long as the bug in the Ceph code exists.
- A sample of the corrupt data and a description of the conditions under which the bug occurred have been shared with Ceph developers. They have investigated the bug report, confirmed the bug, and will submit a request for modification of the Ceph code.

Contact

If you have any questions regarding this report, please contact our Customer Care Department on 0318 648 688 or support@bit.nl.