It Is Internet (Pty) Ltd T/A isoho.st
Registration: 2008/021004/07

Incident report: 2017-03-01

Description

A storage cluster issue caused blocked I/O for disk writes for all customer VMs. The fault was first reported by a customer at 11:29 on Wednesday 1 March. The root cause was rapidly diagnosed and corrective measures were taken within 30 minutes of the incident being reported. The nature of the problem meant that it took up to an hour from first report for most VM I/Os to return to minimally acceptable levels of performance. The storage cluster as a whole, however, experienced degraded performance for several hours thereafter due to rebalancing.

Analysis

The Ceph cluster utilized for the storage backing isoho.st VMs is a multi-terrabyte array operating at 63% of storage capacity used. The space is subdivided into several 4TB Object Storage Deamons (OSDs), each of which receives a portion of the data distributed according to a random hash value. This random distribution of data to OSDs is inherent in the design of Ceph.

One of the multitude of OSDs had been allocated too many objects by the random hash funciton, to the extent that it reached 95% capacity. At this point, the Ceph software is designed to pause I/O operations to the cluster as a safety measure in order to protect the integrity of the cluster.

Impact

All disk writes for customer VMs were blocked for between 30 minutes to an hour.

Actions

The immediate actions taken were to remove the offending OSD from rotation and to reweight the data distribution performed by the hash function. This reweighting precipitated the rebalancing of data amongst the available OSDs (leading to the aforementioned several hour long performance degradation). Such reweighting is not undertaken lightly due to the fact that it too is non-deterministic (it might simply cause a different OSD to become overloaded) and due to the fact that it impacts cluster performance (the rebalancing).

In the medium term, the only solution is to add more storage to the cluster, although at 63% utilization we should have had sufficient safety margin. Thus, we have immediately ordered more disks to be provisioned into the cluster. We are also evaluating further configuration changes to the number of placement groups that may result in more even distribution of objects between the available OSDs.

Further, we are investigating strategies to reduce our reliance on Ceph as a component of our technology stack. In the long term, we believe an architectural change may be the best way to prevent this kind of problem from recurring in the future.

Please direct any questions or comments to support@isoho.st