REASON FOR OUTAGE

RFO Date	10/08/2022
Author	Chris James, Director
Incident start time	04/08/2022 2:26am
Incident end time	04/08/2022 4:57am

What happened

At 2:26am on 04/08/2022 our monitoring detected that the clustered storage used within our Cloud Cluster 1 (hosting approx. 90% of active cloud server clients) was performing slowly.

While not initially affecting Cloud servers, we troubleshooted the slowness to try and resolve the issue without server downtime. Unfortunately, while this was taking place Cloud servers gradually started experiencing extreme slowness or downtime as the primary storage failed to keep up. By 4:00am it appeared that most Cloud servers within this cluster were experiencing downtime or extreme slowness.

At approx. 4:50am we resolved the issue within the storage cluster which caused most Cloud servers to come back online and start responding to requests at expected speeds. Unfortunately, approx. 2% of Cloud servers had entered a “hung” state due to having limited access to the primary drive so needed to be manually rebooted to get back online after the storage problem was resolved.

Summarised investigation and fix

We found some log entries within the cluster related to (non-service impacting) slow network checks of drives that started around the same time as a recent software upgrade and likely led to the outage.

Our software vendor had not seen behaviour like this before and suggested firmware updates and settings changes to fix. After continuing to see the same errors after this work we then downgraded the underlying kernel within the operating systems across the cluster which appears to have resolved the errors from occurring.

We are confident the problem will not happen again.

Technical details (skip this section to avoid ‘tech jargon’)

On a basic level, our storage cluster works by combining all hard drives within each node, through a high speed 40G network, with replicas of data stored across three separate drives (within three different nodes). The software keeps an active process in the cluster for each hard drive to manage heartbeat health checks, data reads/writes, management tasks, etc. The clustered storage is used for all primary drives within Cloud servers.

While working on this issue it appeared that 3 hard drives (in the cluster of 200+) were reporting extremely slow response times. When this was established, we restarted the processes for those hard drives which immediately improved the situation. A problem on a few drives like this should be automatically detected and "worked around" by the storage cluster without noticeably affecting performance so we needed to investigate why this caused cluster-wide issues.

When looking over the log files for the previous week we noticed some minor log entries relating to single drive network heartbeat checks. While these can happen from time to time in periods of high activity, from analysing the logs we could see them happening more regularly for a few seconds at a time. While the log entries do not indicate anything service impacting at that moment, we believe these were a sign of possible upcoming issues like we saw on 04/08/22 so considered a long-term fix to be when these log entries no longer happen.

Working closely with the software vendors they suggested updating to the latest firmware on our 40G network cards and making some changes to the clustered storage settings to break up blocks of data into smaller pieces. This was done on the night of 08/08/22 with the storage cluster rebalance taking 12-24 hours to complete following the settings change. Unfortunately, the log entries related to slow network health checks on some hard drives continued to happen after this work was complete although general cluster performance was improved from the already good performance due to the settings changes.

From further log analysis and communications with our software vendor, we could see that the log entries in question started to happen soon after our underlying software upgrade on 26/07/22 so we focused our investigation on the difference in the software versions.

The decision was made to roll back the kernel on all hardware nodes to an older version in case any driver updates were incorporated into newer kernels causing the problems we have been seeing. This work was done overnight on 09/08/22 with no service impact to clients.

Since the kernel rollback we have seen no further log entries related to hard drive heartbeat checks which we are confident will resolve the matter and not cause any reoccurrence of the issues experienced on 04/08/22. We are also confident that this is related to the problem seen in the underlying software upgrade on 26/07/22 (announcement link).

What we are doing to prevent reoccurrence

We are continuing to work closely with the software vendors to establish how this bug progressed into the enterprise version of their software. While the issue only presents itself in very specific circumstances (when using the exact network cards that we use) and has not occurred for many others, we will be staying on the older kernels until we have conclusive information from our software makers (or the underlying operating system vendor) that there is a bug and it has been fixed.

In addition, we are reviewing our infrastructure update procedures to assess whether this issue could have been avoided. While we always test all updates and work in a lab environment, due to not having the exact same hardware (on the same firmware versions) the bug did not present itself in any tests. Going forward, we are looking to incorporate detailed analysis of all kernel changelogs into our upgrade planning.