Stockholm 2024-09-05

INCIDENT REPORT REGARDING VMWARE STORAGE OUTAGE IN STOCKHOLM

GleSYS has released the following Reason for Outage (RFO) report. This document outlines the background and remediation actions in response to the storage outage in Stockholm on September 4th, 2024.

BACKGROUND

Our storage system is a crucial part of our infrastructure, providing the foundation for data storage and management across our VMware platform. To maintain optimal performance, security and reliability, it is essential to keep the system's software up to date.

The procedure to upgrade the storage system is designed to be non-disruptive, which means no impact on data availability or performance during the upgrade. This is achieved by upgrading the storage controllers one at a time and performing a controlled failover to ensure the array remains online and available throughout.

Unfortunately, a fatal error occurred on the active storage controller during a routine upgrade causing the storage system to go offline. This error was not detected by the health checks run prior to and during the upgrade. The standby controller, which was still in the process of being upgraded, was unable to take over. Even when the upgrade had completed, it was not possible to perform a failover due to the nature of the error and the potential risk of data loss.

While this resulted in a prolonged outage for our VMware VPS customers with disks backed by the affected storage system, it is important to note that we have multiple storage systems in place. Not all systems were impacted, and the outage was limited to those relying on this specific storage array.

SEQUENCE OF EVENTS (CEST)

2024-09-04

05:00 The planned maintenance window begins.

05:25 Storage system software upgrade process initiated.

05:30 (Automated) Pre-upgrade health checks are completed successfully.

05:35 (Automated) Upgrade of Controller A (Standby) begins.

06:15 (Automated) Upgrade of Controller A (Standby) is completed.

06:19 (Automated) Upgrade of Controller B (Active) begins.

06:20 (Automated) Failover from Controller B to Controller A is successful.

06:20 Fatal error detected on Controller A (Active) causing the data service to crash. VMs lose connectivity to storage.

06:22 Alerts are triggered. GleSYS begins troubleshooting the issue.

06:45 Vendor support is engaged and begins troubleshooting the issue.

07:05 (Automated) Upgrade of Controller B (Standby) is completed.

07:25 Vendor support attempts to failover to Controller B to bring the system back online, but this fails. The storage system is preventing the failover from occurring due to a built-in failsafe to protect from data loss.

07:30 The incident is escalated to the vendor's engineering team.

08:14 The vendor's engineering team performs a failover to Controller B in a controlled manner to ensure no data loss.

08:14 VMs regain connectivity to storage.

08:20 Controller A is rebooted and rejoins the cluster as the standby controller in a healthy state.

CONCLUSIONS

While the upgrade was necessary, the unexpected outage that arose underscored several areas where we can improve.

Restoring backups while the storage was offline posed a challenge as some components (not the backup data itself, but other components) required to perform the restore resided on the offline storage. We need a clear plan to ensure we can quickly restore backups in these disaster scenarios.

There is room for improvement with regards to our communication. While we have incorporated some lessons learned from previous incidents, our planned maintenance on Statuspage where we were posting updates automatically closed at 08:00 CEST, which caused confusion as the incident was still ongoing. This issue has been identified in a recent incident, but implementation of our new routine is not yet in place.

The vendor has not yet provided their root cause analysis of the outage as they are still investigating the issue. We will take the necessary actions once it is provided to minimize the risk of a similar incident occurring in the future.

The incident demonstrated that even with careful preparation, unforeseen technical challenges can still occur. Ultimately, the situation was resolved without data loss, but the experience serves as a critical learning opportunity.

Sincerely,

Stefan de Vries
GleSYS