Incident Report - August 26th, 2015

Incident #2015-92

Summary

On August 26, 2015 at 9:00 am, a large number of centrally managed ITS services began paging down. These services were all hosted from Dupuis Hall on our legacy Unix servers, as well as our shared virtual machine hosting environment.

ITS staff were quickly dispatched to determine the cause of the service failures. A large delay in storage response time was found. Diagnostic data was gathered on the storage arrays and then the primary storage array was restarted. The restart occurred around 10:50 am. Upon restart the systems quickly returned to their normal operating state.

The diagnostic data has been sent to the vendor for further analysis.

Impact

Users of almost all of the ITS production enterprise services were unable to get to the service - or at a minimum were experiencing slow performance. This was also the beginning of add/drop which would have inconvenienced many students. These services included, but were not limited to: SOLUS, PeopleSoft Finance and Human Resources, MyQueensU, the Queen's University web site (www.queensu.ca), Moodle and Single Sign-On.

Root Cause

Diagnostic data was gathered before restarting the system and was then submitted to the vendor for analysis. The vendor was able to attribute the failure to a bug in the storage software which is fixed in the next revision which has recently been released.

More simply, a pool of disks got "too full". When they got too full they became very inefficient and the software on the storage starting locking up which caused the widespread performance degradation.

The vendor described the problem as:

Cause of bad failure - Background operations and possible race condition associated with the evacuation of slices while running low on space in a pool. The pool has trouble dealing with low capacity especially when provisioning LUNs, starting compression (which they saw occurring) and deletion of snapshots.

This is a new behaviour we have not encountered before, and is due to the recent increase in storage usage.

Resolution

The short term solution was to reboot our primary storage processor. New disks are now added to the storage pool to prevent this from reoccurring immediately. Long term, we will need to enhance our monitoring of the storage pools to ensure at minimum, 10-15% of them are free. We will also no longer be using the storage compression feature.

Communications (Internal)

ITS infrastructure staff became aware of the service issues around 9:00 am. At the same time the Support Centre began to receive pages and followed notification processes. Around 9:30 am the Manager of Infrastructure Operations was designated Crisis Coordinator. The Manager of Systems and Storage was designated Technical Lead. All Systems members were contributing in their area of expertise.

ITSPP Communications (External)

A notice was intended to be posted as soon as possible to the ITS webpage and Notification Tool. Unfortunately, it was not accessible due to the storage issues.

The Crisis Manager posted to Twitter:

9:04 am –https://twitter.com/ITQueensU/status/636177384975278082 "ITS is experiencing issues in Dupuis Data Centre. Multiple services are affected and may appear unresponsive. ITS is working on the issues"

9:36 am - https://twitter.com/ITQueensU/status/636185431671996416 "ITS is estimating that services will start being available at 11am. More updates will come as we know more."

10:07 am - https://twitter.com/ITQueensU/status/636193300429570048 "ITS has fixed an issue with the storage system and services are now coming back online. ITS will tweet when all services are up."

10:49 am - https://twitter.com/ITQueensU/status/636203785807921152 "ITS is reporting all services are restored. If you are having issues please contact the Support Centre at 36666."

The Crisis Manager also sent emails to the specific hosting customers.

10:20 am - Reported to our hosting customers, ITS-L and ITADMIN-L that we were having widespread outages (via email)

11:45 am - Notification sent out to ITS-L and ITADMIN-L via Notification Tool indicating systems were back up and running

3:18 pm - A second email alerting users to the issues being fixed at 11:00 am.

Lessons Learned

- We will need to more carefully monitor our storage pool allocations. We are also aware of a new software patch that we should install as soon as reasonably possible that apparently addresses many of these issues. The patch is very recent, and preceding that we were up to the latest patch cycle.

Action Items

- Schedule software update with storage vendor for both HPCVL and Dupuis Hall Datacenters.
- Improve storage pool monitoring to avoid pools becoming “too full”.