Incident Report - April 20, 2015

Incident #2015-74

Summary

On the morning of April 20th, 2015 the Systems and Storage Team was alerted, via automated email messages, that there were considerable delays in our shared storage system. These delays could cause slower than expected performance of many enterprise applications, potentially rendering them unresponsive.

This issue is typically a result of very large amounts of data being copied or generated over long periods of time. Upon investigation, there was no indication that this was the case. The Systems and Storage Team then proceeded to evaluate related systems for signs of issues or faults. A server was found to be non-responsive and although it was uncertain if the server was the cause of the issue, a reboot was required.

During the reboot procedure for the non-responsive server, the entire server cluster (14 servers) rebooted simultaneously. The widespread reboot caused an outage to the majority of ITS core services as well as to some of our hosting customers. Services affected included PeopleSoft, Single Sign-on and Moodle (including customers).

The reboots began at 10:05 am. After a 20-minute reboot period, nearly all services had returned to normal operating states. An urgent support ticket was filed with our server vendor and they have provided a replacement part that should address the issue going forward. ITS is continuing to work with the vendor to verify the root cause by looking for known issues regarding large scale reboots as well as the potential of operator error. Reboot procedures have been changed to ensure that servers that require rebooting will be more directly targeted. This will avoid the possibility of servers being affected by a bug or by operator error. Plans have been implemented for improved redundancy that will lessen the impact should this hardware component fail again.

Issue

High disk latency on our production enterprise shared storage resulted in all VMWare-based services becoming either slow or unresponsive for short periods of time. While working towards resolution of solving the latency issues, a system hard locked. Due to either an operator error or a bug, the entire production VMWare cluster was then rebooted. This reboot brought the systems back up to full health.

Impact

During the period of the slow disk performance, users would have experienced slow page loads on services such as PeopleSoft, Single Sign-on logins and a variety of customer-hosted applications. The Queen's website appeared to be largely unaffected by the issue. During the widespread server reboot all of these services were unavailable.

Login service (STS/ADFS) for Office 365 was also affected. Users who were logged into Office 365 before the outage would not have been affected, but all new logins would have failed.

Root Cause

It has been determined that a failed fibre converter in our server infrastructure was causing an excessive amount of port disconnections which resulted in degraded storage performance and the eventual locking up of a server.

Resolution

The failed component was manually taken offline, which immediately restored performance. We have since received a replacement component. The component will be installed immediately but will not be active until the exam period has ended.

Communications (Internal)

The manager of Infrastructure Operations coordinated communications for ITS application owners to keep them apprised of the situation. Once it was confirmed by application owners that their applications were healthy, alerts to campus were communicated that indicated the systems were fully operational.

ITSPP Communications (External)

Notifications (April 20th 2015)

- 8:29 am - Indicating possible degradation
- 9:47 am - Indicating some service issues (server 4 unresponsive)
- 10:07 am - Wider outage including Office 365
- 10:10 am - Services returning to normal, but not all up yet. ETA of 10:30 am
- 10:36 am - All services up except QShare
- 10:47 - QShare now up (x2; two different posters)

Lessons Learned

- Customers in our hosting environment were not notified appropriately.
- Technical knowledge acquired through troubleshooting.
- Increased awareness of this specific issue.
- Monitoring process was in place but on-call was not sent alert.

Action Items

- Create distribution list for all hosting customers. Hosting customers will receive more detailed and more frequent updates.
- Configure monitoring to page on-call on early indication of storage-related issue(s).
- Revise server reboot procedure to ensure single servers are targeted.
- Improve redundancy of server infrastructure to reduce impact of failed components.