Information & Educational Technology Major Incident Report

Date and Time Major Incident Occurred	
Date: Jan 10-11, 2014	Time: 1200 Jan 10 to 1230 Jan 11, 2014

Executive Summary

Summary:

The Data Center suffered a severe service outage that began Friday, Jan 10, 2014, at noon and persisted intermittently until Saturday, Jan 11, at approximately 1300. Most of the service outage was characterized by the Private Cloud Service (PCS) hosted at the primary Quest Data Center being unavailable. Clients experienced unavailability of all associated servers and services. Almost all IET central services were affected by this outage. Campus and Medical Center clients were also unable to access departmental systems and services that are reliant on campus central authentications services.

Overlapping with this outage was a failure of the uConnect firewalls which prevented clients from accessing uConnect Active Directory, authentication, DNS and e-mail services from 1500 until 2000 on Fri, Jan 10.

Root cause:

The Netapp Storage Area Network (SAN) which supports all Private Cloud Service guests hosted at the Quest Data Center encountered an operating system bug which prevented it from processing incoming data requests.

A root cause for the uConnect firewall outage has not yet been determined.

Action steps taken during incident:

IET staff isolated the incident to the PCS and rebooted servers to reinitiate contact with SAN and restarted applications and services hosted within the PCS. Staff contacted VMWare and Netapp to assist in identifying the problem which led to the root cause.

Remediation effort to avoid future similar incidents:

Netapp had incorporated a bugfix for the issue in a later operating system release. That new operating system has been installed on SANs at Quest and the Campus Data Center.

Remedy and Pinnacle Ticket #(s)

The IET Help Desk opened 63 service now tickets created related to the outages. In addition, 346 calls were received by the help desk, but the service tickets are the only way to specifically correlate those to the outage. A recorded message describing the outage was placed on the help desk line. 116 calls were answered. 230 calls were dropped, possibly after the callers listened to the recorded message. However, we assume the majority of the 346 calls were related to the outage.

Services Impacted and Service Managers

List all services impacted and the service manager who is involved:

1. **Admissions**
2. **Banner**
3. **Central Authentication Services (CAS) (~100 minute downtime)**
4. **Computing Accounts**
5. **Electronic Death Registry System**
6. **Data Center File Services**
7. **DaFIS**
8. **DavisMail**
9. **Data Center Virtualization**
10. **Final Grade Submission**
11. **Geckomail**
12. **Kuali Financial Services**
13. **Identity and Access Management**
14. **IET Web Sites**
15. **MyInfoVault**
16. **MyUCDavis**
17. **ServiceNow and SSC Case Management**
18. **Shibboleth**
19. **Smartsite**
20. **Time Reporting System**
21. **Web Content Management System**
22. **uConnect Services (Active Directory, Exchange and Office 365)**
23. **UC Davis Directory Listings**
24. **UC Davis Home Site**
25. **Numerous department systems that rely on central authentication services**

Summary Description of Major Incident

The first outage of the PCS was Friday from noon until approximately 1500 hrs. Once the PCS was restored from this outage it took several more hours to restore all hosted campus services. Possibly related to the PCS outage, the uConnect firewalls stopped working properly. This caused uConnect, and all associated services such as Active Directory authentication, DNS services, Exchange, and Office 365 to be unavailable. While the PCS service and most dependent services were restored by 1730 hrs, uConnect continued to be unavailable until 2000 hrs due to the firewall issues.

At approximately 0000 hrs on Saturday, the PCS service started exhibiting similar symptoms on a smaller scale. While the earlier outage eventually affected all hosts, this new outage was only affecting two VMWare hosts. IET staff began troubleshooting and collecting data, but around 0300 hrs more hosts became affected by this second outage causing critical services to become unavailable. Administrators continued working on the problem, and PCS was restored at approximately 0700 hrs. At that point the process of restoring dependent services began. All services were restored by approximately 1300 hrs.

During this second outage uConnect authentication continued working. Some uConnect mailboxes, those hosted at Quest, were unavailable. Departmental system and services reliant on Active Directory authentication were not affected during the second outage.

Summary of Major Incident Impact

Please see list of services impacted. The majority of campus services were either down or degraded by the PCS outage. uConnect users on campus and at the Medical Center were unable to access services while the uConnect firewalls were down.

Summary of Major Incident Outcome

The PCS outage was temporarily resolved by rebooting the VMWare hosts. System administrators then restored individual services. The second PCS outage was also resolved by rebooting the VMWare hosts and restoring individual services. However, during the second outage the root cause was isolated and remediated. A permanent fix in the form of an operating system upgrade was scheduled and completed.

The uConnect firewall outage was difficult to diagnose. The firewall appeared to be working. No unusual cpu utilization or session counts were observed. uConnect is protected by two firewalls in a high availability configuration. IET staff failed from one firewall to the other and rebooted each of them, but these measures did not resolve the problems. The final resolution was to completely remove AC power from both firewalls for a short period, then restore power and reboot.

Detailed Major Incident Impact

The service outage affected the IET Private Cloud Service hosted at the Quest data center. IET has put much effort into deploying systems with redundant hardware and making this facility the primary location for most computing with over 500 virtual servers hosted within the PCS at Quest. Nearly all of these servers were unavailable during this event. A smaller virtualization footprint is also in place at the Campus Data Center and was not affected during the outages, but most critical services exist at Quest and were down.

The PCS service at Quest consists of Dell B620 servers running VMWare ESX 5.5. Storage is provided by a Netapp 6240 SAN. Identical equipment is deployed at the Campus Data Center, though the VMWare service has less capacity. All PCS data is replicated between the two SANs. In total, approximately 400TB of disk space and 6TB of RAM is in service.

The environment is designed to be highly available. All components in the SAN are redundant, and the two SANs and replicate data between them for disaster recovery. The VMWare software that is deployed, VCloud Suite Advanced, allows for any computer hardware to fail and for services to restart immediately on other hardware. Unfortunately the nature of this recent outage rendered all these contingencies ineffective.

The immediate fix was to limit the number of background jobs that can run on the SAN as the nature of the SAN OS software bug allowed background processes to take precedence over client access to storage and mounted drives. The permanent fix was to implement a SAN OS upgrade that incorporates a bug fix that addresses our issue. The former was completed Saturday evening. The OS upgrade was implemented the week of 1/13/2014; Tuesday for the Campus Data Center SAN and Wednesday for the Quest SAN.

Security Implications of Major Incident

There was no security implication to this incident. Security was not compromised, and we had no elevated security risk to the systems.

Detailed Major Incident Resolution

During the outages IET staff worked with both VMWare and Netapp to isolate a resolution and root cause. VMWare logs clearly identify the storage becoming unavailable. Netapp SAN logs verify this, and much in-depth analysis of logs and performance data with the IET SAN team led the Netapp engineers to identify a bug in the SAN OS code. This bug allows background jobs to take precedence over serving client data creating a race condition. While the VMWare hardware infrastructure provides much redundancy, it requires SAN connectivity for storage needs. When the SAN OS encountered this bug, the SAN became unavailable. When the SAN became unavailable, all VMWare hosts dependent on that SAN ceased working.

It was determined that two background deduplication jobs were running when the first failure happened. IET staff observed a SCSI queue exhausted error on one controller. That is an error that can occur with high IO, however we did not have excessive IO utilization which could have caused the error. At the time of the incident IO was higher than normal, but had been higher in the past without incident and was well below 50% utilization.

IET is still investigating whether the uConnect firewall problems are related to the PCS failure. The timing of the failure is certainly suspect, but there is very little interaction between the two services except that some uConnect guest and external clients that rely on uConnect are hosted in the PCS. There are two firewalls supporting the uConnect service, and it first appeared that each firewall might be alternating between active and failover causing sessions to drop. The uConnect team tried various troubleshooting steps to resolve the issue and ultimately needed to completely power off (physically remove power cords) both firewalls and bring them back up sequentially to restore proper service.

IET staff collected detailed performance data which was provided to Netapp engineering for post incident analysis. Netapp verified that there were no hardware issues and added further credibility to the assertion that the identified OS bug caused the problem.

During troubleshooting we re-implemented IO threshold limits to prevent spikes from affecting the SAN. While we do not believe high IO caused this issue, the bug asserts that high IO could contribute to a recurrence of the problem. We therefore felt it prudent to keep the IO threshold limits in place.

Next Steps and Recommendations

Our next steps will be to work with campus leadership and Data Center clients to identify priorities and solutions for business continuity.

During our analysis we identified the following vulnerabilities
- uConnect authentication servers live on a single VLAN.
- WebSSO is dependent on a single VLAN and single location.

- many servers on single virtual environment backended by a single SAN
- monitoring dependent on the PCS environment
- status.ucdavis.edu dependent on local infrastructure
- communications are tech focused and dependent on local infrastructure
- IET doesn't know all dependencies of our services, such as campus environmental/security monitoring and Medical Center services dependent on CAS authentication

Summary of Event Communications to Customer

Given the widespread nature of this outage it was difficult for IET to communicate with clients about the status of services. IET does have a twitter account @UCDavisStatus to which we post outage notifications and updates in addition to the standard status web site http://status.ucdavis.edu. While the status web site was unavailable during this outage, the twitter feed was updated. IET has also been investigating a cloud based notification system that is subscription based to replace the status page. We will accelerate efforts to deploy the service so that notifications are not dependent on the down services we are trying to notify on.

Once services were restored we sent detailed explanations to dc-clients@ucdavis.edu and tsp-info@ucdavis.edu

Major Incident Log

12/10
- ESX 5.5 Upgrade
- I/O Threshold Settings Introduced Latency.

12/18
- Systems running normally with higher latency
- Sympa performance issues. Troubleshooting began
- Same performance problems noted on Utilities VM also

1/7
- Isolated i/o threshold setting latencies

1/8
- Remove i/o threshold on sympa and utilities with improved performance

1/9
- Removed i/o threshold settings on all ports 2030 hrs

1/10
- 1200 Vmware failure. All hosts in dell chassis tank became unavailable
- fiber paths become unavailable
- esx hosts timeout become unresponsive

- 1220 chasis #1 unavailable
- 1330 - 2pm Hosts individually failing on second chassis bucket
- 1500 uConnect firewalls fail and uConnect becomes unavailable
- 2000 uConnect firewalls restored and uConnect available

1200 - 8pm
- 1342 CAS is restored on physical hardware
- 1400 Mail Routing restored on physical hardware but still in a degraded state
- 1300-1330 VM hosts rebooted
  - vcenter server down
  - AD Auth Issues
  - manual login to vm hosts (esx) reboots completed by 3:00pm
- 1500 VM environment restored
  - manual confirmation that all guests that should be powered up were powered up that didn't auto power up by 1700
- 1730 VM guests all powered on.
  - most restored by 5:30pm except uConnect
- 2000 uConnect firewall reboot resolved uConnect connectivity issues.
  - vCenter restored upon uConnect restoration
- a few lingering issues (directory services)

1/11
- 0000 VMware hosts (2ea - chassis) fail
- no additional failures
- troubleshooting with vmware
- 0330 mail routing started failing
  - uConnect unaffected except for mailboxes at quest (about half of the total users)
- 0430 7 hosts down, 3 up
- 0545 - 0702 started rebooting all hosts using remote access console on the blade chassis
- 05:55 started booting the guests on the first host that was rebooted.
  - AD domain controllers rebooted first.
- 1030 completed reboot of VM guests
- 1100 most major services restored
- 1230 remaining services restored

Symptoms
- VM host quest SAS storage
  - DC SAS ok, Quest SATA ok
- Cascading failure of Quest vmware guests
- Traffic spikes (figure channel, network, etc) but not unusual
- i/o thresholds on - ok, off - bad
  - turning off for sympa - ok, but on SATA
- VMware tech support
  - sys behaving as designed but that the i/o latency not normal
  - suggesting hba firmware upgrade - i/o threshold latency
- Netapp tech support

- performance issue at the Netapp OS level, didn't think it was an issue at the fiber channel level but rather on the OS level - requested we collect more data before and after an incident
Additional Comments
Additional comments may be added by any party involved or impacted by the major incident, including but not limited to service managers and IET executives.
Approved By
List all service managers and Directors who have approved the final Major Incident Report. 1. Mark Redican, 1/21/2014 2. Dave Zavatson, 1/21/2014 3. […]