THE UNIVERSITY OF BRITISH COLUMBIA

Information Technology

Critical Incident Review:

EduCloud Outage – May 15, 2018

Released: June 26, 2018

EXECUTIVE SUMMARY

On May 15th, 2018, UBC experienced an outage that affected a large number of enterprise services. The cause was a portion of the storage subsystem for UBC's EduCloud server and storage infrastructure which hosts services to the University, such as Campus-Wide Login (CWL), Faculty and Staff email (FASmail), Human Resources Management System (HRMS), Financial Management System (FMS) and the Student Information System (SIS).

The cause was determined to be the failure of the storage system and the resulting lack of failover, which should have occurred. The storage hardware has been replaced by the vendor and the cause of the lack of failover is under investigation by the vendor.

The storage system in question is designed for fault tolerance by dividing the service load across four control units, but the failure of one of these units caused a 19 second delay between 2,150 servers and their storage. The service impact of one node failing should have been mitigated by the remaining three nodes; ultimately this was not the case on this occasion and as a result the impact of the outage was higher than it should have been.

When the faulty unit reset itself (as intended), it attempted to rejoin the remaining functioning cluster of control units. For reasons that have yet to be determined by the manufacturer, it consumed all resources available to the cluster resulting in a 19 second delay. A few seconds of delay is within acceptable tolerance levels; however, the extended period of time adversely affected the servers connected to the storage, effectively taking them offline. This resulted in each server requiring a manual reset, along with the central virtualization controller. The manual reset was required to stabilize vCentre, a centralized management application that manages virtual machines and ESXi hosts centrally.

With CWL impaired, many other services that were not part of this group of servers were also adversely impacted, both locally such as Wi-Fi authentication, as well as cloud-hosted services such as Canvas.

The issue was first noted by UBC IT staff in the early hours of Tuesday morning and was escalated at 03:00 to a "Code 3" level (the highest level of escalation within UBC IT) after initial investigation and analysis. The on-call Code 3 team was mobilized at that time to conduct further investigation and attempt remediation in anticipation of the potential impact to university operations at the start of the business day. The relevant technology manufacturers were engaged throughout the investigation and remediation.

The nature and impact of the outage created particular challenges relative to communicating with the UBC community as most regular channels were unavailable. Web bulletins, social media posts, telephone calls, text, and in-person meetings were used heavily to provide as much information to the community as possible, on a regular cadence throughout the incident.

In its entirety, the outage lasted from approximately 01:00 until 13:00 on the day, although core services such as FASmail were reinstated by 11:20. It should be noted that local email service at the Okanagan campus was operational throughout the incident given there is a separate instance in place for that campus.

EduCloud Server Service:
A UBC hosted and managed cloud-based platform that provides secure, multi-tenant private clouds to the BC higher education community, pooling infrastructure resources into a virtual data centre managed by UBC IT. EduCloud is built on VMWare's virtualization platform.

Node:
A virtual machine or a physical server with one or more hard-drive disk drive or solid-state drive

Virtualization Controller:
The virtualization controller (VMware vCenter) is the software which centrally manages the entire virtual infrastructure. This virtual infrastructure is comprised of servers, storage, networks, firewalls, high-availability rules, CPU and memory allocation, performance logs and other configuration items.

Code 3:
A UBC IT incident response protocol invoked when an enterprise service is experiencing an unexpected disruption or outage, resulting in a significant impact to the operations of the university and/or a large number of people. It is the highest escalation notification alert within UBC IT. An incident can be deemed as a Code 3 any time of the day. An on-call emergency response team is alerted to review, communicate and resolve the incident.

Prior to the event, a mitigation strategy had already been designed and initiated based on a similar event that occurred at the end of 2017 which resulted in service disruption across UBC. The deployment and testing of replacement storage infrastructure had been completed, and its implementation into the production environment was scheduled for May 18th to avoid any potential disruption to the exam schedule. This was unfortunately 3 days after this critical incident occurred. Subsequent to the critical incident, the work to migrate services to the new infrastructure went ahead as planned.

Through the course of our Critical Incident Review (CIR) process – which included feedback from service owners, UBC IT leadership, key stakeholders and in this case the addition of Risk Management Services - a number of key issues have been identified related to technical architecture decisions, communication protocols and critical incident management processes.

Each of these issues presents an opportunity for improvement and have provisional resources and timelines identified to support the specific action items that have been developed as part of the CIR process. Progress on these action items through to completion will be tracked and be reported back to the IT Advisory Committee (ITAC), UBC Executive and the IT @ UBC Community.

While the cause is determined to be faulty hardware, there were a number of lessons learned including: how we design systems and services to ensure greater redundancy across our infrastructure, how we maintain documentation on our systems, the expectation of uptime vs cost, and how we respond and communicate during a critical incident where regular communication channels are not available.

Recommendations resulting from the critical incident review include:

1. Determine the future state technology architecture for high availability services based on cost/benefit analysis of the risk the University is willing to assume.
2. As an immediate interim step, reconfigure FASmail to operate across a number of storage clusters with independent virtualization controllers on the existing infrastructure.
3. Develop and implement a sustainable approach to the creation and on-going updating of architectural documentation that is readily available in the event of a critical incident.
4. In concert with Risk Management Services, develop and implement a sustainable approach to the creation and on-going updating of procedural documentation that is readily available in the event of a critical incident.[1]
5. Review current configuration and capacity of Wi-Fi authentication services and conduct cost/benefit analysis to determine if the cost of architecting for high availability is of sufficient value.
6. Consult with community application owners to review service agreement and ensure that the applications they host on EduCloud are configured with an appropriate level of resiliency to meet their operating needs.
7. In concert with Risk Management Services, review critical incident management and communication protocols with a view to harmonize where possible.
8. Examine options for consistent communications channels that can be used in the event of a system-wide outage by UBC IT and the IT @UBC Community to facilitate critical incident communications and coordinate remediation efforts.

[1] Recommendations 4, 7, and 9 are underway as of June 2018. Adoption of consistent routine and non-routine incident classification, framework, and communications is being reviewed and revised with a core group of representatives from UBC IT Vancouver and Okanagan, Risk Management Services, and University Counsel.

9. Refine Critical Incident Response communications protocol including checklists, contact information, and updated UBC channels including representatives from the AVP, Communications portfolio and with key campus communicators who own stakeholder channels that crucial for mass communications.
10. Develop and implement a sustainable approach to the creation and on-going updating of directory information that is readily available in the event of a critical incident.
11. Review protocol with manufacturer and establish communication channel (with redundancy) to be used in the event that enterprise services are unavailable.

BACKGROUND

The EduCloud storage outage experienced on May 15th 2018 was similar in nature to an earlier issue experienced in December 2017 which was addressed by UBC IT. At that time the storage controller software had been updated at the recommendation of the manufacturer; however, unbeknown to both parties, the updated code had a bug that caused the system to become unstable resulting in a service outage. UBC IT followed the recommended steps provided by the manufacturer to resolve the issue, by returning the version of software on the storage controller to a “known good” version i.e. a version without the bug.

Given the severity of the outage, UBC IT worked with the manufacturer to develop a comprehensive plan to remove any chance of a reoccurrence of the issue, which included at the manufacturer’s cost, replacing the entire storage infrastructure. The replacement storage infrastructure (Nimble Storage) was provided to UBC in April 2018.

Deployment and testing took place between April and May, with implementation to the production environment scheduled to begin on May 18th, timed as such to avoid the exam “blackout” period during which no major planned system or service changes are made. The critical incident happened 3 days before the scheduled implementation of the new infrastructure.

ROOT CAUSE

The critical incident of May 15th and ensuing outage was the result of the failure of Node 1 in one of UBC’s EduCloud enterprise storage clusters. This storage cluster consists of four nodes and is designed to continue operating in the unlikely event that one of the nodes were to fail. In this particular instance, this did not happen resulting in a full outage of the storage cluster.

As part of the root cause analysis, the manufacturer determined the cause of the storage failure was due to Node 1 running out of memory. The node “crashed” then reset, during which time the other 3 nodes handled the workload, but when the restarted node attempted to rejoin the remaining functioning cluster, it consumed all resources available to the cluster and caused a 19 second delay between 2,150 servers and their storage.

A few seconds of disruption is within acceptable tolerances, but an extended disruption of 19 seconds was enough to put the VMWare virtualization controller and the servers into an unresponsive state.

At the time of writing UBC IT continues to work with the manufacturer’s highest level of engineering support) to determine exactly why Node 1 ran out of memory and why rejoining the cluster caused the delay which in turn impacted the servers originally attached to the storage.

SEQUENCE OF EVENTS

At 00:45, Tuesday May 15th, UBC IT’s Systems Team was notified that several services were experiencing issues. Through a series of predefined triage activities and escalation points within UBC IT and with the manufacturer (see full timeline in the Event Log), it was determined by 03:00 that the issue warranted “Code 3” status given the impact on multiple services and systems. Code 3 is the highest level of escalation within UBC IT and is initiated typically when there is impact, or threat of impact, to campus-wide services. It is a protocol reserved for incidents that would classified as non-routine and/or major.

The affected systems and services included:

- Authentication services
- UBC Faculty and Staff email (FASmail)
- Financial Management System/ Human Resource Management System
- UBC Wireless Network
- Student Information System (SIS)
- SharePoint
- Canvas
- Connect
- Skype for Business
- Bronze web hosting & blogs
- Kaltura
- Workspace
- Anything using CWL for authentication would likely be impacted

By 03:30, the Code 3 team was fully mobilized. The UBC IT Code 3 Team is comprised of IT staff representatives from the Operational Service Teams, Client Engagement, Communications, Security, and IT Service Centre, each with specific roles and responsibilities in the event of escalated incidents. Additional members from UBC IT's senior leadership team are included as needed.

At 04:00, the manufacturer provided confirmation that the storage was ready to return to service. Although the storage was deemed "healthy" by the manufacturer, the servers using it were still in an unusable state as they could not reconnect to the storage automatically, meaning the services they hosted were effectively offline. In order to restore the virtualized servers, UBC IT staff initiated a reboot of the central management application (VMWare's vCentre) that governs the entire virtual server infrastructure used to run EduCloud.

In doing so, it was determined that vCentre was unable to process the volume of activity in its attempts to reconnect services, prompting further investigation with the manufacturer. It was determined that due to the way in which the virtualized servers lost connection was creating issues with vCentre. It was recommended that UBC IT "remove" all the servers from the configuration, restart vCentre and then "add" them back to the storage cluster. It was understood that this would take a series of hours to complete, but presented a low risk remediation with a higher likelihood of success. A Code 3 Team conference call was initiated at 07:00 including UBC IT senior leadership team members.

A recovery plan was presented and approved by UBC IT senior leadership at 07:00 and executed between 08:00 and 11:00. All virtual servers were reconnected to VCentre, allowing service restoration to begin. Priority services were online by 10:00 with the remainder restored by 13:00. All services were fully validated by 15:00.

ENGAGEMENT AND COMMUNICATION STRATEGY

Prior to initial campus-wide communications members of the UBC Executive were contacted via text to advise of the outage. In addition, the members of the Centre for Teaching Learning and Technology (CTLT)wereadvised of the outage and impact to services, on and off campus (of particular note Canvas, given CWL was unavailable and was required for authentication) as well as the Associate Vice President Enrollment Services and Registrar due to the impact to SIS ecosystem and online voting.

Initial campus-wide communications began at 08:00, with the first UBC IT Bulletin posted, a voice-mail message updated at the IT Service Centre (both campuses), and tweets released via @UBCITNews. Campus communicators were contacted where possible to request assistance in redistributing messages (retweeting) through their various Twitter channels (which have a greater audience reach than @UBCITnews alone). In addition, Client Services Managers began calling departmental stakeholders directly, and together with Desktop Support Zone Leads began visiting departments in-person to provide information and gather concerns.

Social media monitoring and responding to inquiries began by 08:00 and continued throughout the morning and afternoon to engage with and respond to users until email services were restored. Overall sentiment across all channels was mostly neutral.

The Code 3 team had an hourly team cadence call throughout the incident to provide status updates and, where appropriate, determine next steps. Each Code 3 call was immediately followed by a communications and engagement briefing to confirm key messages and outreach to the campus community. As a number of communication channels were unavailable due to the outage, it was determined that communications would be through phone (landline and mobile), texts, social media channels, in-person contact (walk-ups at information kiosks and departmental visits), and the use of external email providers (such as Gmail).

COMMUNICATION TOUCHPOINTS

SOCIAL/DIGITAL	IN-PERSON	IT SERVICE CENTRE
Twitter 8,722 impressions 111 original tweets 88 retweets 75 likes 13 conversations **UBC IT Bulletin page** 80,000 views **UBC IT website** 10,000 page views 50% unique views	44 departmental visits by Client Service Managers and Zone Leads, as well as multiple phone calls throughout the morning to key stakeholders including the Executive. WiFi volunteer kiosks at 7 Vancouver campus locations provided with outage script.	338 calls handled, with peak calls occurring between 08:30 am – 09:30 am.

Full details of the communication and engagement timeline are described in the Event Log.

MITIGATION STRATEGY

The mitigation strategy was developed well in advance of the May 15th outage and was primed to go live according to a predefined schedule built around exam "black-out" times. All activities relating to the deployment and testing of new EduCloud storage infrastructure had been carried out during the months of April and May and were completed in readiness for the replacement of the old infrastructure. Data hosted on the former storage infrastructure (HPE-3Par) was scheduled to be migrated to the new predictive flash storage, called HPE Nimble Storage on May 18th (three days after the unplanned outage).

At time of writing, the storage migration has been completed ahead of schedule successfully. All critical services are now hosted on the Nimble Storage. The 3Par storage is still available for non-critical workloads such as development, verification and testing.

The replacements costs of the Nimble Storage were absorbed by the manufacturer.

IDENTIFIED ISSUES & RESULTING ACTIONS

The events surrounding the May 15th incident have highlighted several areas for improvement. Feedback has been provided via the UBC Executive, Deans, departmental contacts, Risk Management Services and from the general campus community following the incident. In addition the Code 3 team has conducted several debriefs, with support from Risk Management Services.

The following are the significant issues that require action:

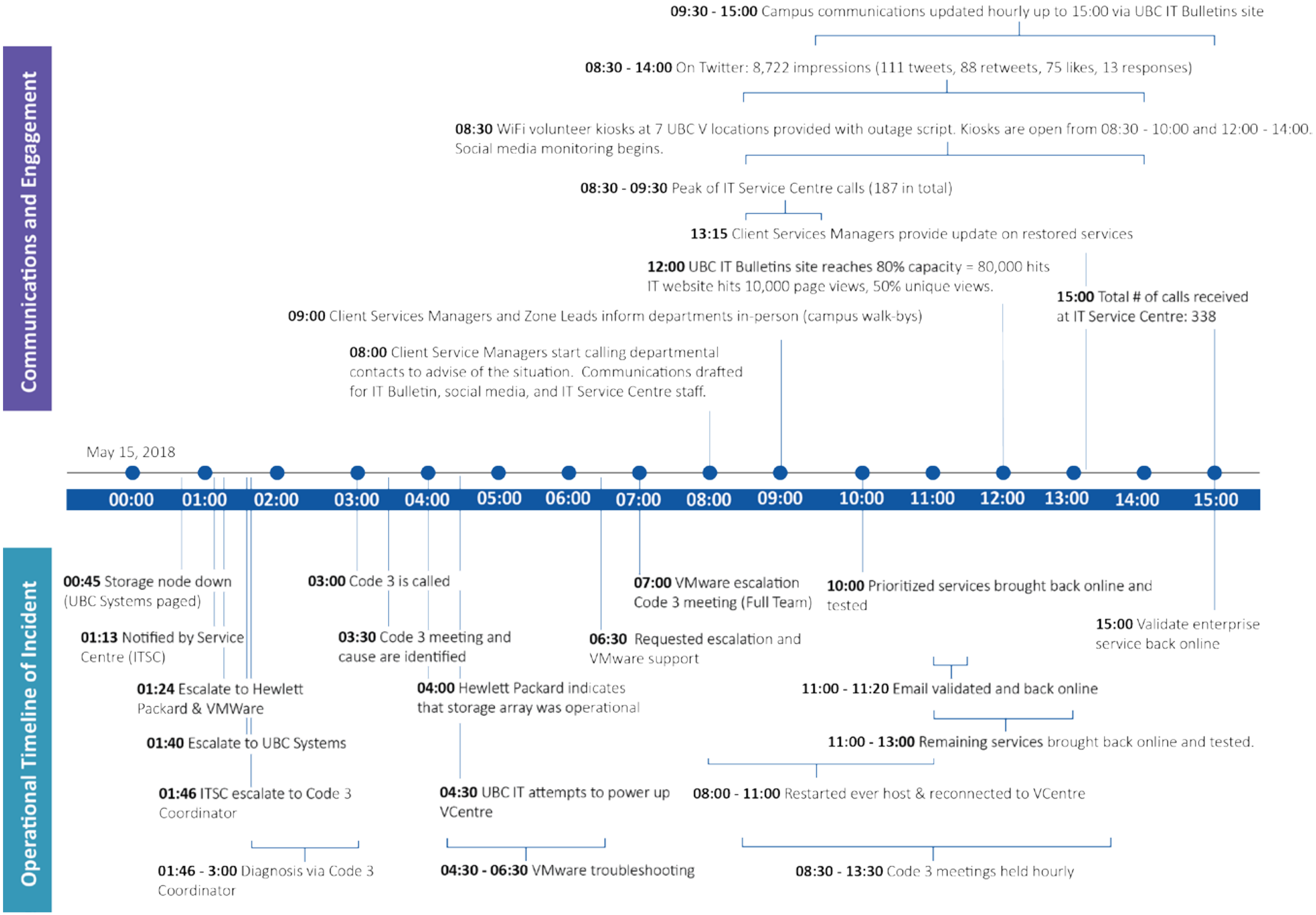
TECHNICAL

ISSUE	ACTION
Critical services are currently not hosted in a high availability, geographically redundant configuration. With all services hosted within the UBC Vancouver infrastructure, the impact of a single outage event such as the incident of May 15th has the potential to be more disruptive to university operations than if they are hosted independently from one another in different locations. Of particular issue is the current configuration of FASmail and its reliance on one storage cluster and one virtualization controller located at UBC Vancouver.	1. Determine the future state technology architecture for high availability services based on cost/benefit analysis of the risk the University is willing to assume. o Briefing note to be prepared for UBC Executive and ITAC as appropriate. o Accountable: Chief Operating Officer, UBC IT o Responsible: Senior Manager Systems; Director of Architecture, Strategy and Planning 2. As an immediate interim step, reconfigure FASmail to operate across a number of storage clusters with independent virtualization controllers on the existing infrastructure. o Plan, configure, test and deploy for September 2018 o Accountable: Chief Operating Officer, UBC IT o Responsible: Senior Manager Systems
Lack of up-to-date and accessible architectural and procedural documentation.	3. Develop and implement a sustainable approach to the creation and on-going updating of architectural documentation that is readily available in the event of a critical incident. 4. In concert with Risk Management Services, develop and implement a sustainable approach to the creation and on-going updating of procedural documentation that is readily available in the event of a critical incident. o Documentation updated and available in soft and hardcopy for August 2018 o Accountable: Senior Manager, Systems o Responsible: System Team
Wi-Fi authentication failed over to UBCO but there was insufficient capacity to handle the load, resulting in degraded service for some Wi-Fi users.	5. Review current configuration and capacity of Wi-Fi authentication services and conduct cost/benefit analysis to determine if the cost of architecting for high availability is of sufficient value. o Analysis conducted and briefing note prepared for UBC Executive and ITAC as appropriate. o Accountable: Chief Operating Officer, UBC IT o Responsible: Senior Manager Systems; Senior Manager UBCNETwork and Infrastructure Facilities
Lack of understanding as to the criticality of applications hosted on EduCloud by community application owners outside of UBC IT.	6. Consult with community application owners to review service agreement and ensure that the applications they host on EduCloud are configured with an appropriate level of resiliency to meet their operating needs. o Consultation with IT @UBC stakeholders completed by Fall 2018 o Accountable: Senior Manager, Systems o Responsible: System Team

COMMUNICATIONS, ENGAGEMENT & PROCESS

ISSUE	ACTION
UBC IT does not operate an enterprise communication system that can be used in the event of a system-wide outage. During this incident, all enterprise UBC communication systems were disabled, resulting in the use of ad hoc "out of band" communication channels such as social media, cellular phones, text messaging and private emails. Insufficient level of technical information provided to the IT @ UBC community.	7. Examine options for consistent "out of band" communications channels that can be used in the event of a system-wide outage by UBC IT and the IT @UBC Community to facilitate critical incident communications and coordinate remediation efforts. o Include involvement of Risk Management Systems, Disaster Recovery Program, and other relevant communication stakeholders o Recommendation to be presented to CIO for October 2018 o Accountable: Deputy CIO o Responsible: Director of Communications, UBC IT
Communication protocols not consistent with other service providers at UBC (specifically Building Operations and Risk Management Services). Reliance on cascading communications when contacting faculty and departments, resulting in information not always reaching key stakeholders such as IT @ UBC system administrators.	8. In concert with Risk Management Services, review critical incident management and communication protocols with a view to harmonize where possible. o Defined, documented and adopted process and protocols for October 2018 o Accountable: Chief Operating Officer, UBC IT o Responsible: Senior Manager, ITSC Desktop
Unclear escalation protocol, as well as roles and responsibilities as they relate to the updating of critical incident information through the UBC.ca site and social media channels.	9. Refine Critical Incident Review communications protocol including checklists, contact information, and updated UBC channels. o Updated protocols in place for September 2018 o Accountable: Director of Communications, UBC IT o Responsible: Communications Team, UBC IT
Lack of up-to-date directory information in hard copy to be used to contact key stakeholders. Staff resorted to using their own personal contact lists.	10. Develop and implement a sustainable approach to the creation and on-going updating of directory information that is readily available in the event of a critical incident. o Process in place and updated and available in soft and hardcopy for August 2018 o Accountable: Senior Manager, Desktop ITSC o Responsible: Manager Service Centre
Manufacturer alert protocol initial failure. Specifically system alerts were emailed by the manufacturer to UBC when our email services were already offline.	11. Review protocol with manufacturer and establish out-of-band communication channel (with redundancy) to be used in the event that enterprise services are unavailable. o Protocol established, tested and in place July 2018 o Accountable: Chief Operating Officer, UBC IT o Responsible: Senior Manager Systems

EVENT LOG



Communications and Engagement

09:30 - 15:00 Campus communications updated hourly up to 15:00 via UBC IT Bulletins site

08:30 - 14:00 On Twitter: 8,722 impressions (111 tweets, 88 retweets, 75 likes, 13 responses)

08:30 WiFi volunteer kiosks at 7 UBC V locations provided with outage script. Kiosks are open from 08:30 - 10:00 and 12:00 - 14:00. Social media monitoring begins.

08:30 - 09:30 Peak of IT Service Centre calls (187 in total)

13:15 Client Services Managers provide update on restored services

12:00 UBC IT Bulletins site reaches 80% capacity = 80,000 hits
IT website hits 10,000 page views, 50% unique views.

15:00 Total # of calls received at IT Service Centre: 338

09:00 Client Services Managers and Zone Leads inform departments in-person (campus walk-bys)

08:00 Client Service Managers start calling departmental contacts to advise of the situation. Communications drafted for IT Bulletin, social media, and IT Service Centre staff.

Operational Timeline of Incident

00:45 Storage node down (UBC Systems paged)

01:13 Notified by Service Centre (ITSC)

01:24 Escalate to Hewlett Packard & VMWare

01:40 Escalate to UBC Systems

01:46 ITSC escalate to Code 3 Coordinator

01:46 - 3:00 Diagnosis via Code 3 Coordinator

03:00 Code 3 is called

03:30 Code 3 meeting and cause are identified

04:00 Hewlett Packard indicates that storage array was operational

04:30 UBC IT attempts to power up VCentre

04:30 - 06:30 VMware troubleshooting

06:30 Requested escalation and VMware support

07:00 VMware escalation Code 3 meeting (Full Team)

10:00 Prioritized services brought back online and tested

15:00 Validate enterprise service back online

11:00 - 11:20 Email validated and back online

11:00 - 13:00 Remaining services brought back online and tested.

08:00 - 11:00 Restarted ever host & reconnected to VCentre

08:30 - 13:30 Code 3 meetings held hourly

GLOSSARY

3Par - A type of storage offered by Hewlett Packard Enterprise.

Client Services Managers (CSMs) - UBC IT staff who help coordinate the delivery of IT services and project support for faculties and departments across UBC.

Code 3 – A UBC IT incident response protocol invoked when an enterprise service is experiencing an unexpected disruption or outage, resulting in a significant impact to the operations of the university and/or a large number of people. It is the highest escalation notification alert within UBC IT. An incident can be deemed as a Code 3 anytime of the day. An on-call emergency response team is alerted to review, communicate and resolve the incident.

EduCloud Server Service - A UBC cloud-based platform that provides secure, multi-tenant private clouds to the BC higher education community, pooling infrastructure resources into a virtual data centre managed by UBC IT. Educloud is built on VMWare's virtualization platform.

Hewlett Packard Enterprise (HPE) - Hewlett Packard Enterprise (HPE) is one of UBC IT's storage manufacturers.

IT Service Centre (ITSC) - IT Service Centre (ITSC) is the UBC Information Technology service desk that provides support for the majority of UBC's enterprise-level online and telecommunications services.

Node - A virtual machine or a physical server with one or more hard-drive disk drive or solid-state drive.

Storage Array - A dedicated data storage system also known as a disk array that contains a group of hard disk drives.

vCentre - A centralized management application that manages virtual machines and ESXi hosts centrally.

Virtualization Controller - The virtualization controller (VMware vCenter) is the software which centrally manages the entire virtual infrastructure. This virtual infrastructure is comprised of servers, storage, networks, firewalls, high-availability rules, CPU and memory allocation, performance logs and other configuration items.

VMware - A virtualization and cloud computing software provider.