Incident Report - Sunday, July 26, 2015

Incident #2015-88

Summary

Services located in the main Queen's Datacentre (Email/Calendar, PeopleSoft, Moodle, Wiki, QShare, internet connectivity, hosting services, web services, storage, authentication services and application services) were unavailable on Sunday July 26, 2015 at 11:30 am when the fire detection sensors in the main Queen's Datacentre, located in Dupuis Hall, activated the fire suppression system. As designed, this process cuts all power to the Datacentre. Kingston Fire Department, Queen's emergency services and It Services (ITS) staff reported on-site immediately.

The Kingston Fire Department inspected the room and concluded there was no active fire but ventilation was required to restore safe oxygen levels in the room. As prescribed, ventilation began and at 2:50 pm. Environmental Health and Safety confirmed the oxygen levels in the room were at a safe level for staff to enter. Upon entry, physical inspection and cleanup began and power was restored to the room at approximately 3:30 pm. The situation was assessed and it was determined that restoring services in the Dupuis Datacentre would take less time than moving the failing systems over to our disaster recovery site at HPCVL.

Following the service restoration plans in place for datacentre services, the main Queen's website was restored by approximately 3:45 pm while other systems were being restored. A hard shutdown of this nature often results in physical hardware failures (power supplies, disk drives, etc.) as well as corruption of systems, particularly for aging systems. Staff began troubleshooting and restoring equipment with services starting to come on-line at approximately 5:00 pm. All systems (production and development) were restored by Monday July 29, 2015 at 10:00 am. Facilities work also continued throughout the evening to provide cooling and power to the datacentre. At the time of entry, only one of the three air conditioning units in the datacentre was operable. The two UPS systems also required attention.

In most systems outages situations, the ITS Notification Tool is used to disseminate information however, during the first response process this tool was unavailable. Communications were made available to the campus users through ITS' Twitter account, @ITQueensU until the ITS Notification Tool was restored.

Although the incident itself was unfortunate, both the shutdown and recovery procedures were performed as planned. This incident does highlight the need for infrastructure renewal in the datacentre by replacing aging equipment.

Impact

All of the Dupuis Hall Datacentre services were unavailable from 11:30 am until 1:30 pm. At approximately 1:30 pm Internet connectivity was restored to campus by re-routing web traffic. The main Queen's web was available at approximately 3:30 pm and the majority of services came on-line between 5:00 pm and midnight on July 26, 2015, following the ITS restoration process. The two priorities identified for campus were SOLUS, as student registration was opening for all students at 12:00 am on July 27, 2015 and Moodle, required for the end of summer term on July 27, 2015. In addition to restoring SOLUS, extra systems capacity was required to handle the open registration volume. The capacity was provisioned and available by 12:00 am. Although the service was slow for the first 30 minutes, it stabilized by 12:30 am. Moodle was available to users at 6:30 am on July 27, in time for scheduled exams. On-premise email was available at 8:30 am July 27 along with the remainder of the application services, most in the development environment, by Monday afternoon.

The estimated incident cost for the University is approximately $65,000, in addition to over $5,000 in time costs for ITS staff to close the incident the following week.

Root Cause

A pinhole leak in a factory weld in the second compressor of an air conditioning unit caused mineral oil to leak under the datacentre floors. The subsequent reaction of the oil created condensation that was interpreted by the fire detectors as smoke in the area. Standard operating procedure when the fire system alarms go off is to automatically shut power down in that room in order to prevent further potential damage.

Resolution

The ITS team worked diligently late into the night in order to restore the facilities, servers and services to the campus community for Monday morning. Course registration was available for students by midnight and by 6:30 am on Monday, all systems were running except for on-premise email. Delays were experienced in restoring services due to some hardware failures and file corruption which is normal for a hard shutdown of systems. All services were up and running by Monday afternoon.

The incident review process creates a list of work to be done in the areas of Communications, Systems and Facilities, all as a result of the incident. The intent is to ensure processes and procedures are effective and efficient. More strategically, this incident highlights the need for an investment in infrastructure renewal. Replacing aging equipment is the proactive way to limit this type of incident and, should an incident occur, speed up the recovery time.

Communications (Internal)

- Twitter was used (@ITQueensU) through the cell network
- Direct telephone contact to various stakeholders
- ITS Escalation
- ITS Notification once system was restored

Communications (External)

- Twitter was used (@ITQueensU) through the cell network
- Infrastructure Operations posted the outages to the notification section of the ITS webpage which also sent out an email to ITSNOTICE-L

Lessons Learned

- There is a need to increase investment in facility equipment lifecycle in the datacentre. The ITS operating budget is stretched at the expense of proper maintenance and renewal of University's infrastructure systems. The air conditioning (AC) unit responsible for the outage is 14 years old and has begun to show signs of age. The lifecycle for these types of units ranges from 15 to 20 years. At the time of the incident the AC unit only had one of its two compressors functioning. This did not contribute to the incident but it does highlight the need to replace aging equipment. Note: there are two other AC units in the datacentre; one is twelve years old and the other is nine years old.
- Incident management processes and notification/communication for this type of event should be reviewed and enhanced.
- Although Queen's email resides in the Microsoft cloud (O365), authentication services are done through the Dupuis Hall Datacentre. This design requires review.

Action Items

Aspect		Priority	Prime	Due Date	Update
Incident response	**Reviewed: Fire Department, ERC <-> ITS, ITS Call-in, Business units**				
Follow up:	Create Incident Response sections in a Datacentre SOP (DC SOP) book and develop a distribution and update process.	M	Judy	Dec 2015	Paper copy and electronic
	Document responsibilities and contacts for Crisis Manager and Crisis Communications role	M	Terry	Dec 2015	Compile information into DC SOP Incident Response section.
	Document and communicate PPS Duty Manager contact process	M	Hugh	Dec 2015	Compile information into DC SOP Incident Response section. Include with FM200 procedure
	Document FM200 procedure to clear & test room	M	Hugh	Dec 2015	Compile information into DC SOP Incident Response section.
	Document process for engaging a Security resource to guard external doors when open for ventilation	L	Judy	Dec 2015	Compile information into DC SOP Incident Response section.
	Enhancements to external monitoring for early detection	H	Andy	Complete	NodePing need to add more contacts
	Add services monitoring for services without monitoring	H	Andy/ Dan B	Sep 2015	Events Calendar, SCOM server, terminal server RDP port
	Review and update (if necessary) contact and procedures documented at Campus Security	M	Ray	Dec 2015	
	Review with PPS the need for ventilation for smoke-producing work and other cutting and dust	M	Hugh	Dec 2015	July 30: building fire alarm activated from soldering/welding even though B36 alarm disabled. Verify PPS procedure for this work.

Aspect		Priority	Prime	Due Date	Update
	Engage Business units to develop their “fallback" plan if systems become unavailable	M	Gail	May 2016	Initiate with Enterprise Solutions team for Finance, Human Resources and Student Services.
	Communicate and update On-Call docs making carrying the On-call phone as well as pager mandatory	H	Judy	Complete	
Communications	**Methods: Phone, Twitter, Web, ITSC phone message** **Stakeholders: Escalation list, ITS Staff, hosting customers, Queen's community**				
Follow-up:	Hosting customers - list, contact numbers	H	Infra Managers	Sep 2015	Compile information into DC SOP Incident Response section.
	External client notification (school boards, ORION) - list, contact numbers	H	Ray	Sep 2015	Compile information into DC SOP Incident Response section.
	Add social media (Twitter) to standard notification channels	H	Commun-ications Team	Sep 2015	Compile information into DC SOP Incident Response section.
	Define the role of communications team	M	Commun-ications Team	Dec 2015	Compile information into DC SOP Incident Response section.
	Add EITAC to customer contact list	H	Brad/Terry	Sep 2015	Compile information into DC SOP Incident Response section.
	Define ITS Notification procedure for incidents (send email to ITS and EITAC with more detail - Notification on ITS page)	M	Brad	Dec 2015	Compile information into DC SOP Incident Response section. Verify ITSC phone message still in communication plan.
Facilities PPS/ITS	**Physical room access, power, AC, cabling, maintenance**				
Follow up:	Develop budget strategy for overhead cable trays (infrastructure renewal)	H	Gail	Sep 2015	~$40k – tray structure ~$10k – wire re-routing

Aspect		Priority	Prime	Due Date	Update
	Develop budget strategy for AC replacement cooling (infrastructure renewal)	H	Gail	Sep 2015	~$100k/unit with 15-20 year lifecycle
	Develop budget strategy for replacement for small UPS (infrastructure renewal)	M	Gail	Sep 2015	~$100k with 20 year lifecycle
	Budget and schedule yearly DC room cleaning	M	Hugh/Gail	Sep 2015	~$10k per datacentre (~$30k)
	Define AC maintenance schedule and contract (PPS or external). Develop operational process for monitoring.	M	Hugh	Nov 2015	Unknown cost
	Undertake an emergency power off design review	M	Andy	May 2016	
	Review AC run design as it is at maximum length	M	Hugh	May 2016	
	Investigate options for a DC ventilation plan with PPS	M	Hugh	May 2016	
	Document UPS procedures; post contact number on UPS	H	Hugh	Oct 2015	Compile information into DC SOP Incident Response section.
	Document physical layout of fire detector on floor plan	L		May 2016	Compile information into DC SOP.
	Provide staff with FM200 training (emergency shut off)	H	Ray	Oct 2015	Compile information into DC SOP.
	Submit PPS work order to inspect datacentre door latch to Chemical Engineer area as it blew open when fire suppression released.	H	Hugh	Sep 2015	
	Schedule regular fire sensor replacement with PPS	M	Hugh	Dec 2015	
	PPS to install guard by fire sensors to protect from AC belt dust	H	Hugh	Sep 2015	

Aspect		Priority	Prime	Due Date	Update
	Install ammeters on power bars	H	Ray	Complete	To be done as they are replaced
	Purchase fans to ventilate and/or cool room	H	Ray	Complete	
	Purchase cart to move equipment to storage	H	Ray	Complete	
	Contract initial room cleaning as a result of this incident	H	Ray	Complete	$10k
Systems					
Follow up:	Investigate: AD authentication servers - wireless client couldn't authenticate; AD in HPCVL - test with others off	H	Systems	Sep 2015	
	Investigate: AD service in Fleming did not allow authentication	H	Systems	Sep 2015	
	Review authentication design for O365 email.	H	Middle-ware	Dec 2015	
	Routing of fibre channel network all through Dupuis	M	Networks	May 2016	
	Review "auto boot" feature on servers	M	Systems	Dec 2015	
	Passwords - Keypass at alternate location	H	Brad	Sep 2015	
	DNS resolver redundancy needs investigation	M	Brad	Sep 2015	
	Review recovery plan for main queens WWW server	M	Brad	Dec 2015	Compile information into DC SOP Incident Response section.
	Regular CMDB copy at HPCVL	H	Brad	Complete	Primary to HPCVL, copy nightly to Dupuis

Aspect		Priority	Prime	Due Date	Update
	Decommission old systems to assist in restoration process.	M	Brad	Ongoing	Older systems and applications require more manual intervention and often hard shutdowns result in corrupt data or failed equipment.
	Dual power planning procedure for new installations	M	Hugh	Dec 2015	
	Monitoring - mailbox monitoring, bring up SCOM, external monitor ccs-server	H	Andy	Complete	▪ NodePing checking ccs-server sermon status page and www.queensu.ca texting Andy ▪ Sermon doing qwa login ▪ Sermon checking SCOM IIS active
	Update SQL servers	H	Brad	Sep 2015	Apply blade firmware bug update, change all auto power startup, system startup order
	Document in DC SOP all systems that require intervention. Include how to check and who to follow-up with	M	Brad, Terry	Dec 2015	Topaz, Oscar, EDRMS, doc should be in CMDB, list of systems needing special startup attention