UIS report following critical incident on Monday 29/1/2018

On 29 Monday at 15.15 a power outage in the machine room (MER) in the Roger Needham Building (RNB) caused the loss of two critical services, wireless for Eduroam users and email for Hermes users, for approximately 90 mins. Other services were also impacted. This report focuses on three key elements to ensure lessons are learnt and help improve services:

1. What caused the initial power loss to MER in the RNB?
   *Led by Dr Rachel Hooper, Deputy Director, Departmental Operations*

2. How can communications before, during and after the incident be improved?
   *Led by Nigel Thornton, Interim Head of UIS Communications*

3. Why did a power loss in the MER cause service failures?
   *Led by Steve Riley, Deputy Director, Service Operations*

**Incident timeline** - the key events on 29/1/2018 were as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.15</td>
<td>Following a routine test of the fire Suppression System in the MER the fire alarm was triggered and the RNB evacuated. The fire alarm process causes the power to be shut down to the server room, impacting critical and non-critical services</td>
</tr>
<tr>
<td>15.45</td>
<td>Power restored to MER cabinets and boot-up procedure commences</td>
</tr>
<tr>
<td>16.32</td>
<td>Eduroam service restored</td>
</tr>
<tr>
<td>16.45</td>
<td>Hermes service restored</td>
</tr>
<tr>
<td>17.30</td>
<td>Critical service status set to green. All services fully restored by 10.30am on Tuesday.</td>
</tr>
</tbody>
</table>

1. **What caused the initial power loss to the machine room (MER) in the RNB?**

   **Issue: How did a routine fire safety test cause the power loss?**
   To enable normal working during safety tests it is necessary to isolate the alarm system in the MER. The incident report from the University’s fire safety systems maintenance provider (BBC) shows that an experienced tester inadvertently reinstated the isolated output causing a power shutdown when the system was tested. The report highlights the lack of a consistent process for testers isolating outputs for this system.

   **ACTION:** BBC has created a standard operating procedure (SOP) for the MER to reduce the risk of human error. BBC will also produce similar SOPs for other critical installations, including the Soulsby Data Centre and CNH and ensure implementation.

   **Issue: Were the lessons from the 2014 incident implemented?**
   In 2014 a fire safety test in the Uninterruptable Power Supply (UPS) in the RNB also resulted in a power loss in the MER, due to the activation of a power shut off. All the recommendations from that incident report were implemented. The system has been tested on numerous occasions since without an issue.

   The incident on 29/1/2018 involved a second power shut off in the MER itself.
ACTION: It has been agreed with the University’s fire safety officer that the automatic power shut off activation within the MER can be removed. Such a power shut off is not required at any UIS data centre and is a legacy of the RNB’s original construction.

2. How can communications before, during and after the incident be improved?

**Issue: Service line status information inconsistent**
The recently launched ‘service status line’ is a key channel for incident communication. Throughout the incident the subscriber text alerts were issued and the telephone messaging updated. However, the web display remained falsely showing all green status. This was because the MER hosts the system which provides the UIS website with updated data.

ACTION: An existing plan to create a resilient web update will be actioned.

**Issue: Email updates too slow**
Once restored, email updates were issued to service desk ticket holders but not to the wider CO community in a timely manner. An unclear and untested incident communication and escalation processes caused a delay in CO comms.

ACTION: UIS Comms to draw-up, and drive adoption of, an incident comms checklist.
ACTION: UIS Comms to run incident comms training sessions.
ACTION: UIS Comms and UIS Service Operations to review the incident escalation process to ensure the right comms resource is in the right place during an incident.

**Issue: Twitter updates did not reach the CO community**
The loss of emails and the newsfeed system meant that the @UniCamITStatus twitter feed was one of the few operational channels available during the incident. However, because the UIS website doesn’t display the @UniCamITStatus feed and it has few direct followers, it didn’t have the reach to communicate effectively.

ACTION: Ensure tweets are visible on the website, actively promote Twitter as a source of incident information.

3. Why did a power loss in the MER cause service failures?

**Issue: Eduroam users lost connection**
Wireless controllers and infrastructure were sustained during the outage. However, because two radius servers (which are part of the user authentication) were co-located in the MER, 20,000 Eduroam users were unable to connect to the wireless network. Note: 10,000 users connected to the UniOfCam wireless were not affected.

ACTION: One radius server has now been relocated to the West Cambridge Data Centre (WCDC), with one in the MER. This is a temporary fix until we establish a process to ensure that VMs are not co-located in the same location in the future.

**Issue: Email for Hermes was unavailable**
The primary Hermes server is usually in the WCDC, with a secondary server in MER. However, at the time of the incident the WCDC server was being patched, with the MER server providing user-facing
Loss of power to this user facing server resulted in Hermes being unavailable between 3:15pm and 4:45pm

ACTION: UIS needs to improve change control to ensure that Service Managers are aware of any work being done in any of the Data Centres, in which their services reside. UIS will review how to use the Change Control Board more effectively, to ensure that people understand what work is being carried out when and how their services may be impacted.

**Non-Critical Systems**
The outage of non-Critical and non-Important services (i.e. the “Standard” services) is within the 99% availability target. No further actions required.