IT INFRASTRUCTURE RESILIENCY REVIEW

23 February 2017

Introduction

Objective of the review:

This report presents the findings of the IT Infrastructure Resiliency review carried out by PA Consulting following the outage at the Strand Data Centre in October. The objective of the review was to understand what went wrong with the infrastructure technology and how it was managed. Whilst this is important, the review also looks forward: making sure that in the future the College manages IT in such a way that appropriate and informed risks concerning technical resilience, business impact of system interruptions and/or failures, reasonable user expectations and financial affordability are accounted for and understood by management.

Approach and methodology:

The review was conducted in two parallel work-streams:

1. Understanding the core technology issues that caused the outage
2. Getting a rounded view from the non-technical areas around culture, engagement with the users and the wider IT and College IT governance

The findings are based on approximately 30 interviews with stakeholders from IT and other areas of the College as well as reviewing the available documentation. Findings were mapped to PA's assessment frameworks and conclusions validated with the relevant interviewees.

Structure of the report:

- Executive summary

1. Review findings
    1. Technology management
    2. Data management
    3. Business relationship management
    4. IT Governance & decision making
    5. Team management
2. Recommendations
3. Appendix
    - List of interviews
    - Incident timeline

EXECUTIVE SUMMARY

A storage system hardware failure that should have been manageable without outage, created a chain of events with catastrophic impact to the College

On the 17th October 2016 one of the four controllers within the principal HP storage system located in the Strand Data Centre failed. There was no user impact. HP hardware engineers then arrived on site to replace the component that had failed. In theory the storage system should have returned to a normal state. However, the system went offline and simultaneously many of the storage disks within it started failing leading to a complete loss of data. At this point what had been a routine incident with no impact to users was escalated to senior management in the IT team. A pre-documented business continuity process was instigated to establish a cross functional response team to co-ordinate the incident resolution and subsequent service recovery. The response team produced regular progress updates to the affected user communities, though it was difficult to provide accurate recovery time estimates on account of the fact that the performance of backup systems had never been tested.

At the time of the incident there were multiple backup systems implemented and had they performed as intended the data could have been recovered and the incident would have been annoying but not damaging. Unfortunately the backup systems collectively failed to provide an adequate service and some of data was lost. Much effort has been expended by the College to recreate data (e.g. Admissions). However in some instances data may be lost forever.

The cause of the backup failure was due to the IT technical team not fully understanding the importance of the tape back ups within the overall backup system and not following the back up procedures completely. In addition some data has consciously never been backed up on tape due to capacity constraints and the potential impact of this was never communicated to the College.

It was later established, in an assessment of the incident by HP, that the inability of the storage system to return to service after the defective hardware was replaced was due to a flaw in the firmware responsible for keeping the hardware controllers functional. HP had issued an updated version of the firmware weeks before the incident which they claim would have allowed the replacement controller to be installed in the storage system without a service outage. The IT team had not had the opportunity to apply this routine firmware update before the incident.

The nature of the failure and inability to completely restore all data raises a number of immediate questions

Did the College buy the right technology and support from HP?	**Yes** – the technology supplied by HP was modern and fit for purpose. The College sensibly purchased additional "proactive support" when the system was installed four years ago. However, this support package does not provide the level of risk assessment and change management advice which is now available through an "enhanced support" option introduced by HP in 2015. "Enhanced support" would have been, and would still be, appropriate for this complex technology.
Are systems adequately backed up today ?	**Partial** - The IT team have now moved more systems over to the new distributed backup system to relieve the load on the legacy tape backup system. This is now able to correctly backup all of the remaining systems and file stores and the success reports created by the backup systems are reviewed and acted upon on a daily basis. However this falls short of a complete backup restoration test which is the only way of completely ensuring that the backup system works correctly.
Is there a strategic roadmap to ensure that the College gets the levels of resiliency required in the future?	**No** – there are a number of solutions currently being considered to eliminate the need for the Strand Data Centre which is no longer fit for purpose. However nothing has yet been presented to the College that explains the levels of resiliency that a proposed solution would deliver including guaranteed recovery times from a major failure.

What needs to change to prevent another catastrophic situation from occurring again ?

Over the past four years the IT leadership team have managed an ambitious transformation programme that has introduced impressive new technology and operating processes. The business (and some in the IT team) have struggled to keep up with this change so the IT team must now build a closer collaborative relationship with the business and within itself.

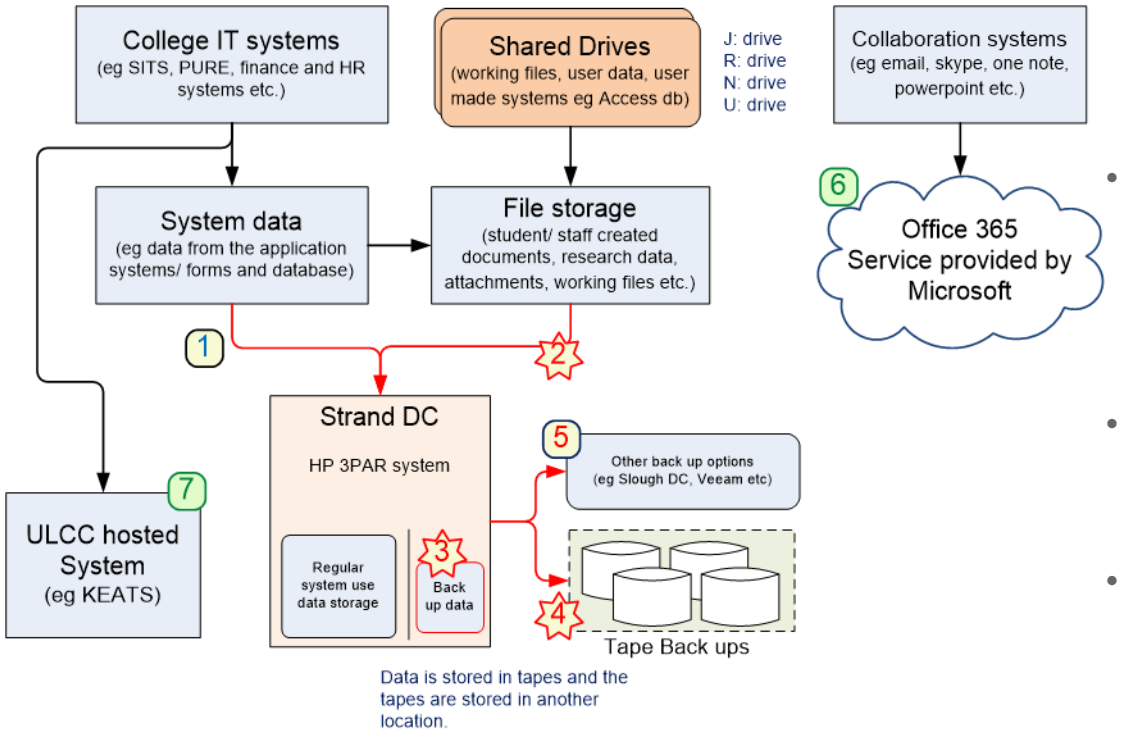
The IT team did not understand the criticality of the tape back ups and did not ensure that these were reliable	The consequence of the outage was severe, users were unaware they stored critical data inappropriately and the IT team were unaware of its importance	The IT team has an inadequate understanding of the needs of different user groups	IT team has not been able to engage the College in IT governance sufficiently to collectively understand the exposure to risk	The rapid growth in IT team size has meant that team members have been over relying on processes and focussing narrowly
• Migration from the Strand Data Centre took longer than expected. • A complex infrastructure in transition for a number of years. Modifications were made to the original design (e.g. Hardware redundancy configuration) without understanding of consequences • The importance of tape backups was underestimated, when these failed intermittently the root cause was not fixed • User Data criticality not understood, and important data was consciously not backed up on tape	• Users stored valuable academic research and College administrative data in shared drives • IT did not know how users used shared directories or their importance • Users perceive that no guidance or policy has been provided for appropriate data storage • IT data archive service is not known to users	• IT is perceived as being distant from users and engagement is reliant on process • IT do not understand the needs of all the different user groups and are perceived to lack empathy with some • The IT culture is not aligned with user expectations hence they disengage	• IT is doing many things at once, which has overwhelmed College stakeholders • Insufficient time has been given for the senior IT governance team to constructively challenge IT plans • IT were not able to negotiate windows for Disaster Recovery tests which is the only way to demonstrate the ability to recreate systems and data a complete system destruction incident.	• IT teams following process mechanically with a narrow focus on their own work • Too many initiatives competing for attention, and priorities are not clear • Hand-offs and decisions not properly understood by team

These recommendations should be considered in conjunction with any other remediation activities already planned and a focused programme initiated

Theme	Recommendation
Technology management	• Backup coverage must be reviewed and tested. If necessary tape library capacity increased, even if only required for an interim period. (Note: Backup system capacity has now been increased)
	• The Daily Service Review update coverage must be revised and communicated so that all team members understand the significance to the business of each item in the review.
Data management	• Develop and communicate clear policies and provide guidance on how best to use data services from IT. Ensure that the IT data services are linked to these policies and meet the user requirements. • Support this with a coordinated college-wide culture change programme so users do use these services in the right way.
IT Governance and decision-making	• The wider College IT governance structure needs to be reviewed and adjusted, decision making delegated based on complexity, scale and investments. • Prioritise and focus on completing the critical things first. • Review the financial model and funding approach for IT investments, aligned with the IT governance. • Improve communication with the senior stakeholders and encourage them to “own” the decisions.
Business relationship management	• IT must build a closer collaborative relationship with the business. • Adjust the internal IT culture and adopt a relationship based approach for user engagement, less reliant on process alone.
IT team management	• The culture within the IT teams should move towards a more collaborative approach that is not heavily reliant on process alone. • The IT team members should be educated about the overall technology strategy and roadmap, priorities made very clear so that they know where they should be focusing effort and how their work helps the combined IT and College goals.

REVIEW FINDINGS

The complex architecture at the time of the incident meant the complete failure of the data storage system required restoration from a variety of sources



Simplified summary of King's College storage infrastructure

- On the flawed assumption that the storage hardware was very resilient the core College IT systems and data **(1)** and file storage **(2)** were backed up on a different location of the same storage unit **(3)**
- Some of the systems and shared storage were also backed up to an independent tape unit that also allowed for offsite storage of backup tapes **(4).** This system was overloaded and not all data was successfully backed up
- Other systems were backed up to a newer independent storage solution **(5)** that was located in a Slough data centre. This performed well
- Some systems and directories were either inadvertently (e.g. Admissions portal attachments) or consciously not backed up onto an independent device
- The Office 365 systems are provided independently by Microsoft **(6)** and were not impacted
- The KEATS system was hosted on a separate ULCC infrastructure **(7)** and was not impacted.

The IT team did not understand the criticality of the tape back ups and did not ensure that these were reliable

Findings

1. This migration of the Data Centre located in the basement of the College's Strand building to a new purpose built facility at Slough has taken longer than the initial estimate of a year. Infrastructure has gained in complexity over time, with adjustments being made to the original design (e.g. Resiliency of the original HP hardware was reduced to deal with performance issues).
2. With the prolonged migration the backup architecture has been in transitional state for a number of years. Currently a number of backup mechanisms and approaches are deployed; multiple backup copies are stored on the SAN being backed up with the expectation that it is fault tolerant, newer systems are also backed up on a new independent backup system, whilst other systems went to an independent but near end of life tape library.
3. Owing to capacity constraints some shared drives were deliberately never independently backed up to tape. A conscious decision, without documented rationale, was made to rely on the volume level back ups within the same storage device (SAN), which was not communicated to data owners.
4. In some instances systems and data (e.g. Admissions Portal) were migrated onto the new HP hardware and owing to an incomplete understanding of how the systems worked the tape library was not configured to backup all data.
5. Tape backups failed regularly and some folders were not backed up properly for several months. Daily Service Review (DSR) updates on the tape back up status were not being reported correctly (e.g. backups were declared as a success when they contained repeated failures to back up some shared drives). This was compounded by the team not comprehending the business criticality of the data being backed up to tape so these issues were not escalated.
6. The HP technology is complex and at the time of implementation the College sensibly purchased additional proactive support from HP. This did not provide much advice on change management or risk assessment. HP subsequently (2015) introduced enhanced support which they claim would have provided this advice.

Recommendations

- Backup coverage must be reviewed and if necessary tape library capacity increased, even if only required for an interim period.
- The Daily Service Review update coverage must be revised and communicated so that all team members understand the significance to the business of each item in the review.
- All systems should be subject to an annual recovery test to establish it is possible to recreate the system and associated data in the event of a destructive failure. This is the only way of establishing that the backup systems are functional and that the IT team and the business understand what is required to recover from a severe failure.
- **The first two recommendations described above have now been implemented.**

The consequence of the IT outage was severe as users were unaware they stored critical data inappropriately and the IT team were unaware of its importance

Findings

1. Users unaware they were storing valuable data (e.g. research data, strategic plans, budgets etc) in inappropriate shared drives (file storage) including user-defined systems (eg Access databases)
   - Terabytes of static reference data was being backed up daily clogging up the available tape back up capacity
   - Faculty admin teams built their own systems with complex data connectivities (eg Access database with hard coded links to enterprise data stores) and stored them in the shared drives, which made it harder for the IT team to achieve service restoration
2. Users did not know how best to use the IT services to store different types of data. There is a user perception that no guidance is provided about when to use Sharepoint, OneDrive, shared drives and other cloud storage options or how to request storage or retrieval of archive data
3. The data governance strategy and policies are not linked to the supporting and enabling IT services. The IT team did not know what type of data was being stored by the users in shared drives. Nor did the IT team have a catalogue or user to folder mapping. All of which hampered the recovery effort
4. IT naturally focused on the big legacy systems for tape back ups and de-prioritised the file storage (shared drives). IT did not understand the risks being accepted as a result nor did they communicate it well to the user communities.

Recommendations

- Develop and communicate clear policies and provide guidance on how best to use data services from IT. Ensure that the IT data services are linked to these policies and meet the user requirements.
- Initiate a culture change programme within the College to raise awareness amongst the user communities and to help them look after their data in the right way and use the appropriate data services from the IT team;
- Make use of other teams in the College who could help IT in embedding this culture change, collaborate with them more closely (eg Library Services, Research ethics committee, Data governance and strategy committee, Faculty readiness leads etc.)
- Investigate how best to “catalogue” the data and maintain the information – what type of data is stored where, who needs it and what is the relative importance.

The IT team has an inadequate understanding of the needs of different user groups

Findings

1. Over the past four years the IT team have made great efforts to provide a good and consistent level of service across the College. They have now defined a comprehensive set of processes to support this. The user communities prefer a more relationship led approach, so unfortunately the focus and reliance on process focus has created a "process wall" that distanced the IT team from user engagement.
2. The user communities find the IT processes too rigorous and too rigid even for simple things (i.e. small ad hoc projects). This alienates them further from IT and encourages shadow IT (e.g. DIY Access databases) as the finance and budgeting is devolved to the faculties.
3. Different user groups across the College (faculty admin staff, professional services, academics and students) have different needs. IT teams lack this understanding of the different user groups and their unique requirements (e.g. IoPPN research data is accumulated over decades and some research data is generated by specialist devices such as MRI scanners).
4. Without a dialogue users can't provide the relevant feedback to IT on proposed solutions and help define the business requirements that would then support infrastructure design decisions. Consequently IT are having to infer requirements without acknowledgment or expectation setting with the business. (e.g. IoPPN research data has complex confidentiality requirements which apparently cannot be implemented on the new collaboration solutions).

Recommendations

- King's must build a closer collaborative relationship between its business and IT function, that supplement formal IT governance forums. For example, identifying influential power users across the College and informally discussing ideas for service improvements reinforces the partnership ethos that should exist between IT and the business.
- Adjust the internal culture and move towards an engagement with the users built on closer working relationships. Processes are important and they should not be thrown away move away from being overly reliant on process alone for user engagement.

IT team has not been able to engage the College in IT governance sufficiently to collectively understand the exposure to risk

Findings

1. The technology roadmap has a large number of initiatives within it. The volume of IT initiatives is overwhelming and the business stakeholders have found it difficult to help IT to prioritise appropriately.
2. The senior level IT governance - the IT Governance Subcommittee (ITGS) was set up two years ago. Meetings only occurred quarterly. This frequency combined with the large number of projects did not give sufficient time for the senior stakeholders to constructively challenge IT plans, particularly those that improved services that were hidden from user view (e.g. storage backup)
3. IT has not been able to convince users on the need for doing full Disaster Recovery tests and negotiate windows for these to occur. The infrastructure limitations mean that any such test will involve downtime which business has so far refused, without properly understanding the consequence. Had these occurred it would have demonstrated that the backup systems were not functioning correctly..
4. The information going into ITGS ("the booklet") is often very detailed but it is not presented at a level where senior stakeholders could fully understand the implications of all the options presented.
5. Faculties have relinquished their IT budgets to KCL IT for pan College IT infrastructure transformation projects. However, common solutions have failed to satisfy all the diverse needs so Faculties perceive IT investment decisions have forced on them (e.g. Inability to cope with complex IoPPN research data directory structures)

Recommendations

- The wider College IT governance structure needs to be reviewed and adjusted. Consider how the IT governance can be improved by delegating decisions at different levels based on scale, complexity and investments e.g. director level and escalate to ITGS by exception or large investments.
- Prioritise and focus on completing the critical things first, free up some capacity to tackle critical things with the right level of attention
- Review the financial model and funding approach for IT investments, aligned with the IT governance.
- Improve communication with the senior stakeholders and get them to "own" the decisions

The rapid growth in IT team size has meant that team members have been over relying on processes and focussing narrowly

Findings

1. The IT team has grown from c. 115 members in 2014 to approximately 350 members. This has created a reliance on a structured process-based approach to work together. Otherwise it would be impossible to keep on top of everything.
2. The teams are following processes mechanically, focusing on things that they have to do within their step of the process and not thinking about the wider implications of what services mean to the business (e.g. key members of the IT team were unaware of the fact that admissions portal data was missing even after several weeks of the outage occurring).
3. Large number of initiatives are competing for attention from the IT teams (specifically platforms team). The teams are not clear on the priorities and keeping a number of things running at the same time. As a result the design decisions and the implied risks did not cascade well enough (e.g. the decision taken to reduce the resiliency of the storage hardware to improve performance). The team are not able to fully mitigate all the risks and act on them appropriately.
4. The lack of awareness of business context meant that the IT teams misinterpreted business risk (e.g. the decision not to independently backup some directories to tape).
5. The daily reviews and other such procedures implemented to catch things falling through gaps did not work. The teams followed process without awareness of the wider context (e.g. Tape backups were reported as completing rather than as being successful giving a false sense of security).

Recommendations

- The culture within the IT teams should move towards a more collaborative approach that is not heavily reliant on process alone. If the teams understand the wider purpose of IT and how they help their colleagues across the College, then they are more likely to collaborate and engage with other teams in IT rather than mechanically following their own bit of the process.
- The overall technology strategy and road map for priorities should be made very clear so that all stakeholders and teams know where to focus effort and how their work helps deliver the College's goals.

RECOMMENDATIONS

Summary Recommendations [1 of 2]

Theme	Issue	Recommendation
Technology management	• Backup architecture has been in transition for a number of years and has complexity. • Migration to Slough has taken longer than initial estimate of a year consequently delayed adoption of a clean back-up strategy. • Tape back ups failed regularly and root cause not fixed. • Some data consciously not backed up due to capacity constraints in tape back-up	• Backup coverage must be reviewed and if necessary tape library capacity increased, even if only required for an interim period. (Note: This has now occurred) • Regular system recovery tests
	• Daily Service Reviews not effective as checking mechanism as updates on back up status were not being reported correctly.	• The Daily Service Review update coverage must be revised and communicated so that all team members understand the significance to the business of each item in the review.
Data management	• Users stored valuable research data in shared drives, that IT did not know • No guidance or policy for appropriate data storage • IT data archive service is not known to users	• Develop and communicate clear policies and provide guidance on how best to use data services from IT. Ensure that the IT data services are linked to these policies and meet the user requirements. • Support this with a coordinated college-wide culture change programme so users do use these services in the right way
IT Governance and decision-making	• IT is doing many things at once, this overwhelms stakeholders • Insufficient time for senior IT governance to constructively challenge IT plans • IT were not able to negotiate windows for DR tests	• The wider College IT governance structure needs to be reviewed and adjusted, decision making delegated based on complexity, scale and investments. • Prioritise and focus on completing the critical things first, free up some capacity to tackle critical things • Review the financial model and funding approach for IT investments, aligned with the IT governance. • Improve communication with the senior stakeholders and get them to “own” the decisions

Summary Recommendations [2 of 2]

Theme	Issue	Recommendation
Business relationship management	• IT is distant from the users and engagement is reliant on process • IT do not understand the needs of different user groups and are perceived as lacking empathy • The culture is not aligned with user expectations hence they disengage	• IT must build a closer collaborative relationship with the business • Adjust the internal IT culture and adopt a relationship based approach for user engagement, less reliant on process alone.
IT team management	• IT teams following process mechanically with a narrow focus on their own work • Too many initiatives competing for attention, priorities not clear • Hand-offs and decisions not properly understood by team	• The culture within the IT teams should move towards a more collaborative approach that is not heavily reliant on process alone. • The IT team members should be educated about the overall technology strategy and roadmap, priorities made very clear so that they know where they should be focusing effort and how their work helps the wider IT and College goals.

APPENDIX

I. List of interviews

II. Incident timeline

Appendix I – List of interviews

Role
IT Business Assurance Manager
Project Officer, King's Futures
Chief Information Officer
Director of Admissions & Registry Services
Director of Real Estate Management
Business Continuity Manager
Chief Operating Officer (Arts & Sciences)
Assistant Chief Operating Officer (Arts & Sciences)
Director of Management Accounting
Director of Administration (Arts & Humanities)
Director of Administration (IoPPN)
Director of Planning and Service Transformation
IT Delivery Manager
Head of Transition & QA (IT)
Director of Marketing
Head of End User Services (IT)
Director of IT Governance
Head of Platforms (IT)
Hed of Architecture (IT)
Director of Library Services
Associate Director of Library Services (Collections & Research)
Chief Accountant
President, King's College London Students' Union
Communications and Campaigns Director
Compute & Storage Manager
Director of IT Procurement
IT Risk & Continuity Manager
Director of IT Solutions
Compute & Storage Engineer
Director of Strategy & Operations (Fundraising)
Head of Operations (Fundraising)
Interim Executive Dean of the IoPPN

Appendix II – Incident timeline

Objective of this Appendix

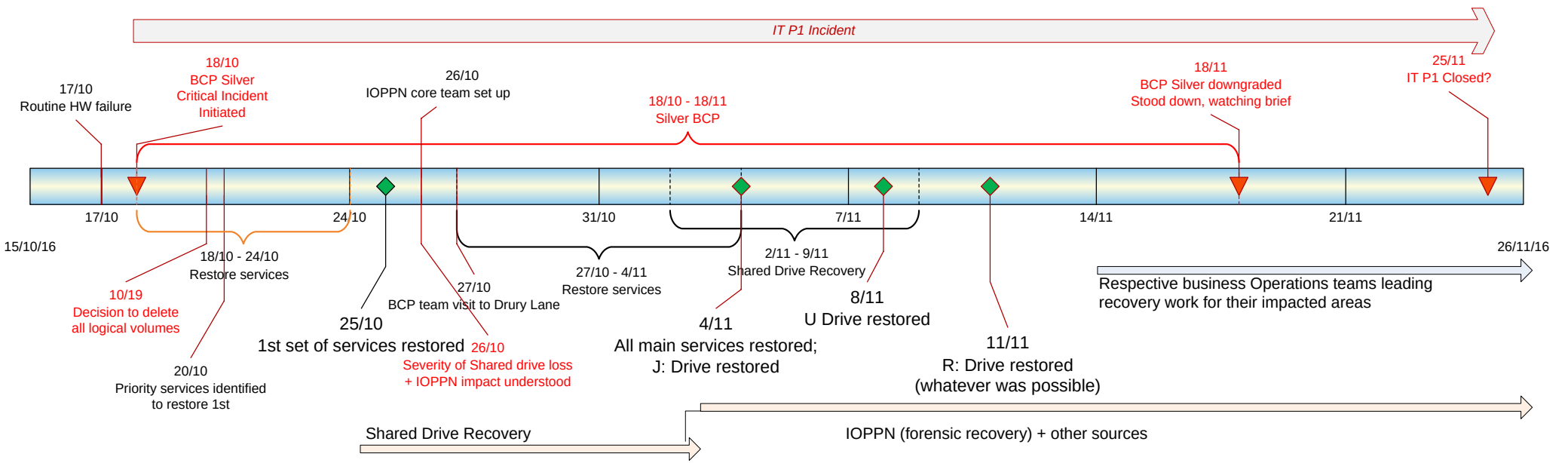
The incident timeline presented overleaf is a high level depiction of the incident from the start on Monday 17th Oct until the BCP Silver team handed over the recovery and restoration work to the respective operations teams in the week ending Friday 25th Nov.

The top half of the timeline shows the events and when the systems started to come back.

The bottom half of the timeline shows how the incident was being managed and led and the structure of the BCP group.

IT did a pretty good job of managing the actual incident. There was a structured approach and all the right people were present in making the key decisions with respect to technology actions.

The recovery work was hampered by wider issues which we have highlighted in the main report.



BCP Silver Lead

Ian T

Nick O'D

IOPPN Recovery: Richard B

Shared Drive & IOPPN
Recovery Group : Laura C

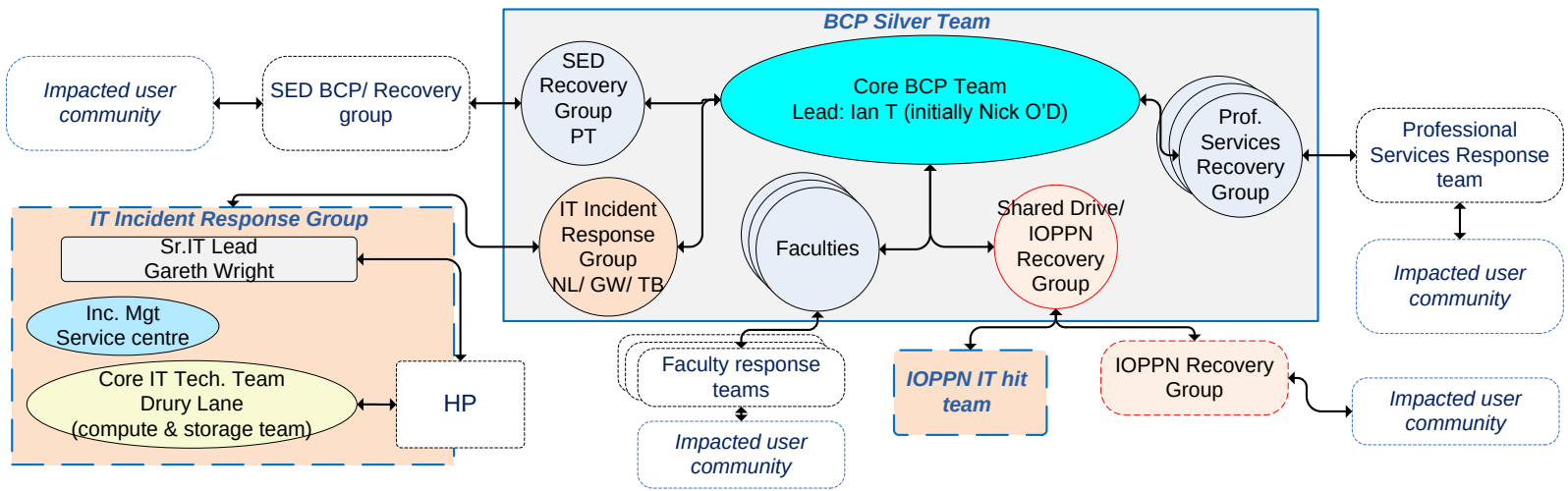
IT Lead

Trevor B

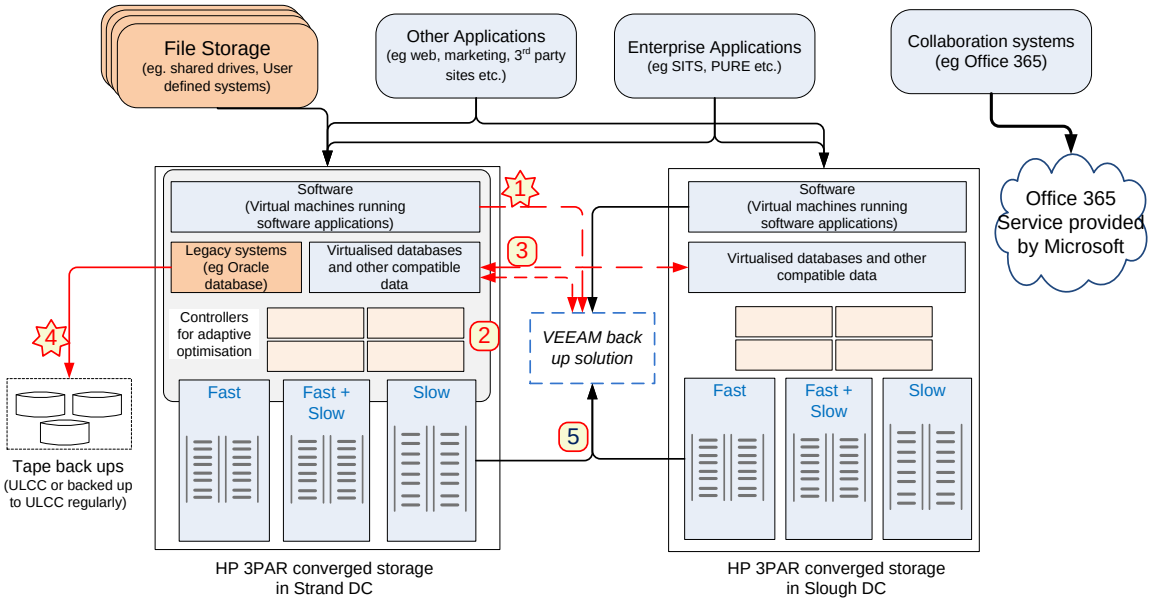
Gareth W

Mark K is the point person at Denmark Hill

IOPPN Recovery Group: IT hit team: Mark Keep (with TB)



The hardware failure in Strand HP 3PAR lost the primary back up data, creating a reliance on secondary tape back up and these were not reliable



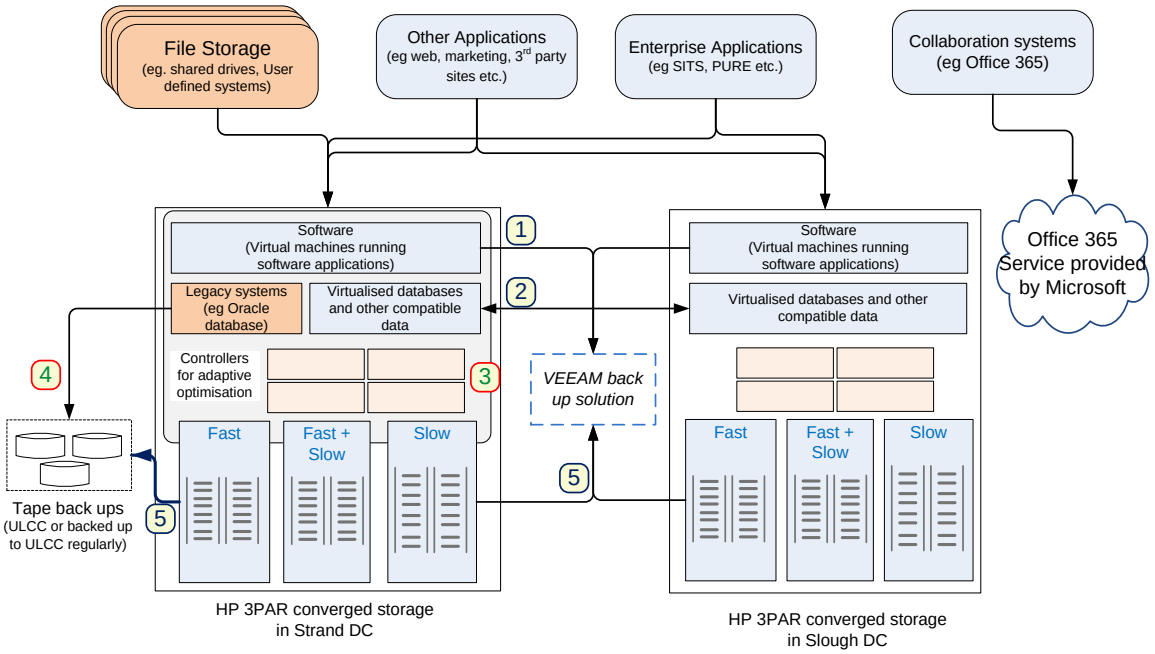
Back up situation in Oct 2016 (at the time of the incident

The hardware failure caused the Strand HP 3PAR storage device to lose all data that was stored in the machine. This meant that only available back ups were the tape drives. As the Netbackup solution did not work correctly, shared drive data was lost and had to be recreated from other sources.

1. Some (but not all) of the virtual machines were backed up using the long term strategic solution (Veeam).
2. The operating system level snapshots (1st level) and volume level snapshots (2nd level) being backed up in a separate partition within the same HP 3PAR machine.
3. The back up data (above) was not replicated. Solely reliant on the same HP 3PAR machine in Strand for 1st level and 2nd level back up options.
4. Legacy workload (eg., oracle database, Solaris hosts, physical SQL & File cluster nodes, physical rack servers, solaris and workloads currently incompatible with Veeam) was solely reliant on the Netbackup tape back up solution.
   - Some data was never backed up due to capacity constraints on the Netbackup tape solution.
   - Backup jobs failed regularly, some repeatedly
   - Shared drive data was using the Netbackup option
5. Some data was being backed up by both Veeam and Netbackup whilst some data was not backed up anywhere on tape

Some additional backups taken by a retired backup device of legacy systems had been retained and it was possible to restore some useful data. This was lucky. Arguably the data should have been destroyed as it was on a retired backup system that was known to contain confidential data.

The immediate back up resilience is improved and will help avoid a similar level of impact as back up options have been strengthened



1. All the virtual servers in both Strand and Slough are being backed up using Veeam.
2. The back up data is being stored in separate disk storage and being replicating to the opposite side.
3. No longer reliant on volume level snapshots as sole back up option for any data
4. Netbackup capacity freed up by removing previously untested virtual workloads that have now moved to Veeam.
   NetBackup is only being used for workloads that are currently incompatible with Veeam. Any remaining non-virtual servers and workloads, (eg. Solaris hosts, physical SQL & File cluster nodes, physical rack servers) are being backed up to tape at ULCC, either directly to ULCC, or to a local tape library in the Strand which then duplicates the backup data to ULCC.
5. All workloads that were previously not backed up to tape, are now being backed up either to tape, or to Veeam.

These steps ensure that data has been backed up and is available to be restored when needed. Without testing the recovery it will be difficult to predict how well or how soon will the systems come back from a similar outage.