Committee: General Purposes

Date: 27th September 2012

Agenda item: 8

Wards:

Subject: IT Systems Failure

Lead officer: Mark Humphries, Assistant Director Infrastructure & Transactions

Lead member: Councillor Mark Allison

Forward Plan reference number: Not Applicable

Contact officer: Richard Warren, Head of IT Service Delivery

RECOMMENDATIONS

That Members note the information and details relating to the failure of the Councils IT systems on 5th March 2012.

---

1. PURPOSE OF REPORT AND EXECUTIVE SUMMARY

Following a failure of the Councils IT systems and infrastructure on Monday 5th March 2012, a comprehensive investigation was completed and a detailed report prepared for CMT.

The report identified the main reasons for the failure and made a number of recommendations in respect to upgrading some of the Councils existing IT infrastructure, and changing some of its current operating procedures in order to minimise the risk of a similar failure at some point in the future.

Following the meeting of the General Purposes Committee on the 14th March, Members requested that a report on the incident, lessons learned and future options be bought back for consideration and discussion at the next meeting of the committee.

2. BACKGROUND INFORMATION

What is a SAN?

A SAN is a core part of the IT infrastructure and holds the majority of the Council's data. It is a highly complex device using a large number of high and low speed disk drives in a highly resilient format. Organisations deploy SANs as an effective and reliable method to store and retrieve data.

Merton's SAN system has dual power supplies, dual controllers and "hot swappable" disk drives making it very robust and under normal circumstances an extremely reliable piece of equipment.

There are two different types of disk drive within a SAN device, Tier 1(15K) and Tier 3 (7.2K) drives, with K being an indicator of the speed at which the disk drive spins and therefore the relative speed at which data can be retrieved from the system.

In normally circumstance the data is stored on the Tier 1 drive although the system is intelligent and looks to see what data is being used and how quickly it needs to be accessed, before moving it around. This ensures that any data which has not been accessed for a longer period is stored on the slower speed drives (longer term storage) allowing the system to continually adjust and rebalance to deliver optimal performance.

The Councils IT infrastructure and systems have been designed and installed to be resilient and cope with unexpected equipment failures by incorporating two SAN devices. One of these devices is in the main production unit located within the Councils main data centre on the 6th floor. A second device which serves as the Disaster Recovery (DR) unit is located in a computer room on the 1st floor of the Civic centre. Future disaster recovery and business continuity facilities will be covered by a reciprocal agreement with L.B Wandsworth whereby each council hosts the others DR equipment.

System servers are connected to the SAN using connector cards, we typically install two cards in each server so they go through different fibre switches to ensure resilience of connection. Each fibre switch is connected to a different production controller, fundamentally this is to ensure that data is always accessible to the servers in case of any failure on the SAN device.

3. EVENTS LEADING TO LOSS OF SERVICE

The SAN is the central storage system for all of the Council's critical applications and servers and is used to deliver services across the whole authority. At 9am on Monday 5th March the Councils production SAN equipment suffered a critical failure causing a significant impact on the operation of the IT systems. The failure resulted in almost all of the Councils critical applications and systems being either unusable or operating very slowly. The failure was initially caused by the loss of two faulty disks in the main production system. This was then further compounded by the subsequent failure of a third disk whilst the system was going through the automatic data recovery process from the two faulty disks.

The SAN device is configured and designed in such a way that in the event of a failure or loss of one of the disks, the data it contains is automatically restored onto another spare disk using a RAID (Redundant Array of Inexpensive Disks) data recovery process. However, because the system suffered the loss of two disks almost simultaneously - an extremely rare occurrence - this effectively meant the normal automatic recovery process could not be completed.

The system continued to try and recover the data and prevent any loss which is how it is designed to operate i.e. on a fail safe basis. However, the failure of two disks at the same time resulted in the system conflicting with its inbuilt programming causing it to go into a looping process and lock up. The system was unable to resolve the problems during the normal rebuild process.

At the same time as the recovery process was being implemented, system users were also trying to access data that was stored on the two failed disks. When taken together these two events effectively consumed most of the

processor and memory resources available on the system resulting in extremely poor performance across the Councils systems.

Despite the best efforts of the manufacture to restore the system remotely through external network connections, all attempts to recover the system failed. At around 15:00 it was decided to invoke a failover to the Councils Disaster Recovery system as it was considered very unlikely that the production system would be recovered in time to be able to operate normal services ready for the following day.

4. RECOVERY OF SYSTEMS

Members of the IT services team and representatives from Dell/Compellent who manufacture the equipment worked through the night of 5th March to reconfigure all the servers to use the 1st floor Disaster Recovery SAN. By the morning of 6th March the failover to the DR storage system was completed and most of the Councils critical applications were available with the exception of the Carefirst system. The production SAN was not able to be fully recovered for another 36 hours after the initial failure and due to problems with the specialist server that operates the Carefirst system, this was not fully operational for a further three days.

After failing over to the DR system users started to complain about the poor performance of many of the systems and applications. This was due to the decision to size the capacity of the DR SAN at roughly 50% of the performance of the main production system. This decision was made on the basis that approximately only 50% of the Councils systems and applications were considered to be mission-critical in the event of a disaster. The DR system was not specified to run LBM's entire server estate, which itself has increased significantly since the SAN equipment was first installed in late 2009.

Over the following days a decision was taken to move a number of servers back to the 6th floor SAN, as the general speed and responsiveness of the some systems was so poor that they were practically inoperable. The transfer continued up to the point where the split of the systems operating across the two SAN's was 60% on the 1st floor and 40% on the 6th floor.

Whilst the systems were operational and relatively stable, further works still needed to be undertaken in order to fully restore them, and this would require some further programmed shut down of services. It was felt that no further works should be undertaken on the systems until such time as we had fully identified the root cause of the failure and fully understood the requirements for any upgrading or repair works. A fully co-ordinated approach was also adopted to potentially avoid the need for any unnecessary loss of service.

5. CAUSE OF THE SAN FAILURE

Failure of disk drives within SAN equipment is a normal occurrence. Systems are designed and configured to deal with routine failure and recovery without any significant loss of data or performance. However, it is clear that the loss of three disk drives in very quick succession was the primary reason for the

failure of the production SAN equipment. It is apparent from further analysis that the failure rate of the Tier 1 (15k) disk drives was far higher than would normally be expected on this type of system. Given the production system has 48 disk drives and the DR system has 24 this represented an extremely high failure rate of nearly 50% in just over 2 years. As a comparator there have been only 2 failures of similar drives in our other systems which total almost 90 drive units.

The subsequent investigation found that since the two systems were installed in late 2009 there had in fact been 23 failures of the Tier 1 (15k) disk drives in the production SAN and 13 failures of the same drive in the DR unit.

The IT services team had already contacted the system support provider to alert them that the failure rate on the Tier 1 (15K) disk drives had increased significantly in the first few months of 2012, and asking for this to be investigated

6. POST FAILURE ACTIONS

Dell/Compellent reported the excessive number of Tier 1 Cheetah K6 drive failures at Merton to Seagate who manufacture the disk drives. These drives are used extensively all over the world and are seen as being high quality equipment with a good reputation for reliability.

Some of the failed drives have been sent to Seagate for examination, initial findings seem to indicate that the cause of the drive failures is a mechanical problem associated with the failure of the drive bearings possibly due to a lack of lubrication. Dell/Compellent are still awaiting an official response from Seagate, and we are actively pursuing this matter through the equipment supplier.

In advance of the root cause analysis from Seagate and to minimise the risk of any further disk failures, Seagate and Dell/Compellent have replaced all of the Cheetah K6 drives in both of the SAN systems with newer K7 drives. The latest release of drive firmware was also applied to the new K7 drives after they were installed.

Further remedial work was not undertaken until the Council were confident that a robust plan had been developed and agreed, which would minimise the potential for any further unnecessary disruption or loss of IT services – See section 11.

7. LOSS AND DISRUPTION TO THE CAREFIRST SYSTEM

Following the failure of the SAN equipment there was also an associated problem with the specialist server that operated the Carefirst system. The Carefirst system is based on UNIX operating system and Oracle database, it is very complex in its design and requires a high level of specialist knowledge to maintain and support. Unfortunately due to a combination of reductions in staffing levels and people leaving to take up employment opportunities elsewhere, the council's ability to maintain and support the Carefirst system is currently below what is deemed to be the normal acceptable level. This then

resulted in delays in completing repairs and routine maintenance activities which was a contributory factor in the length of time it took to restore the system after the initial failure.

Prior to the incident the Council had also been using a company called OLM who are the supplier of the Carefirst system to perform a number of tasks to upgrade the system, this support was provided on a day rate basis and outside of a formal service contract. Following the incident OLM were contacted about providing assistance in respect to repairing the system, and initially they were very reluctant to help but following some discussions at a fairly senior level they agreed to send an engineer to site to see if they could repair the server. Unfortunately, the task was beyond the skills of the engineer sent to complete the repair and it was eventually completed by members of the IT services team.

As the Carefirst server was also coming to end of life the decision was made to replace it earlier than scheduled and also organise a new support contract with a different company who would install and configure the system, this work has now been completed and the Carefirst system is stable and operating reliably. The old original server is currently being completely rebuilt, which was far more cost effective than purchasing another new server. When the works have been completed it will be used to provide a proper backup server for the system and will be relocated together with the Disaster Recovery SAN at Wandsworth.

8. LESSONS LEARNT

Following the failure of the SAN equipment and associated IT services, the Assistant Director of Infrastructure and Transactions completed a detailed review of the processes and procedures for dealing with this kind of incident.

The purpose of the review was to identify any specific lessons learnt and make recommendations for improvement in terms of the Councils internal processes and procedures for managing this type of incident should something similar occur again at some point in the future.

On the basis of this review it was recommended that the main areas of focus and improvement should be:

- Because of the life and limb nature of the services that the Carefirst system supports, it is essential that the Council develops and maintains robust Disaster Recovery arrangements and associated infrastructure to minimise the potential for a loss of the service.
- Ensure that the Councils IT Disaster Recovery and Business Continuity processes and procedures are properly defined and are physically tested on a regular basis in order to ensure that they are suitable, sufficient and remain fit for purpose.
- Support directorates to have detailed and robust emergency and contingency plans in place for the delivery of critical services in the unlikely event the Councils IT systems should completely fail, as it wasn’t clear from within departments what their arrangements were.

- Consider and develop a detailed plan that deals with effectively communicating information to citizens and members of staff in the event that the Councils IT systems should fail and that the normal channels for disseminating information are no longer available (i.e.) webpage, intranet, email and landline telephones.
- Undertake a review of the current support and maintenance arrangements for systems in light of the recommendations that come out of the BDO project on rationalisation of IT systems.
- Undertake a review of the current arrangements and process for how CMT respond to major incidents and emergencies.

9. RECOMMENDATIONS FOR REMEDIAL WORKS, UPGRADING OF THE EXISTING SAN EQUIPMENT AND OTHER LESSONS LEARNT

Following the failure of the SAN unit and an analysis of the Councils requirements in terms of improving its current Disaster Recovery arrangements, a range of recommendations were made which cover:

a) The need to minimise the risk of any further reoccurrence of the problems caused by the failure of the Tier 1 (15k) disk drives

b) A requirement to create a fully documented IT Infrastructure Disaster Recovery plan and fully automate the Disaster Recovery failover and failback processes between the two SAN devices.

c) Upgrade the capacity of the existing Disaster Recovery SAN to more closely match the performance of the production unit therefore enabling the Council to continue to deliver business critical services in the event of a failure of the main production unit. However it should be noted that some additional work is also required to update the identification of the business critical systems across the Council and then agree the order in which services would be restored in the unlikely event that we should suffer a similar complete system failure at some point in the future.

d) An opportunity to upgrade and future proof both the main production and Disaster Recovery SANs to add sufficient additional storage capacity to the system in order to accommodate the Councils data storage requirements through to 2015, based on the current levels of usage.

e) Relocate the Disaster Recovery SAN and various other business critical systems including the new Carefirst UNIX server to L.B Wandsworth in order to strengthen and improve the Councils current DR and BC arrangements.

f) Implement a programme of regular timetabled shutdown of systems in order to undertake upgrades, planned maintenance and repair of systems on dates agreed in advance.

10. CONSULTATION UNDERTAKEN OR PROPOSED

Discussions have taken place with departmental client representatives and colleagues from the IT Business Improvement division to fully understand the issues that resulted from the failure of the SAN and understand what

adjustments and improvements needed to be made to the internal communication processes in order to ensure that information about systems and hardware is shared across the two parts of the IT function.

11. PROPOSED TIMESCALES

Following agreement and approval of the recommendations by CMT, a programme of works was drafted identifying the timescales for completion of the various tasks which are shown below:-

Item	Description	Timescale
a	Minimise the risk of any further reoccurrence of the problems caused by the failure of the Tier 1 15k disk drives.	Works have been fully completed.
b	Create a fully documented IT Infrastructure Disaster Recovery (DR) plan and fully automate the DR failover and failback processes between the two SAN devices.	Works to be fully completed by 31stDecember 2012
c	Upgrade the capacity of the existing Disaster Recovery SAN to more closely match the performance of the production unit.	Works have been fully completed.
d	Upgrade and future proof both the main production and Disaster Recovery SANs.	Works have been fully completed.
e	Relocate the Disaster Recovery SAN and various other business critical systems including the new Carefirst UNIX server to L.B Wandsworth.	Works to be fully completed and equipment re-located by 31st December 2012
f	Implement a programme of regular timetabled shutdown of systems in order to undertake upgrades, planned maintenance and repair of systems on dates agreed in advance	Dates agreed for quarterly weekend shutdowns and first one successfully completed 25th & 26th August

12. FINANCIAL, RESOURCE AND PROPERTY IMPLICATIONS

All of the costs associated with the replacement of the disk drives and the transfer of existing data between the new storage devices have been met

directly by the manufacturer, and this was done as a gesture of goodwill in recognition of the premature failure of the disk drive equipment, at an estimated cost of £60,000.

The manufacturer has also agreed to supply and install the additional equipment that was required to complete an upgrade of the system as described within section 9 at a significantly reduced price. (Discounted by approximately 50% against the normal market price)

The costs for completing the works are:

- Create a fully documented IT Infrastructure Disaster Recovery plan and fully automate the Disaster Recovery failover and failback processes between the two SAN devices - £10,000.
- Upgrade the capacity of the Disaster Recovery San to match the production unit - £70,000.
- Upgrade and future proof both the main production and Disaster Recovery SANs to add sufficient additional storage capacity to the system in order to accommodate the Councils data storage requirements - £45,000.

Total cost of works £125,000.

As part of the agreed 2012/2013 capital programme there is already £100,000 allocated to a scheme to improve Disaster Recovery arrangements which is being used to fund the majority of the project, but leaving a requirement to identify an additional £25,000.00 of capital funding in order to fully complete the recommended upgrading works, which will be met from the IT Transformation budget.

There will also be a requirement for some growth in respect to revenue to cover the additional cost of support and maintenance of the upgraded systems which is estimated to be approximately £18,000.00 per year, which will be met from the disaster recovery revenue fund.

There are no other property or resource related implications for this report.

LEGAL AND STATUTORY IMPLICATIONS

None for the purposes of this report.

HUMAN RIGHTS, EQUALITIES AND COMMUNITY COHESION IMPLICATIONS

None for the purposes of this report.

CRIME AND DISORDER IMPLICATIONS

None for the purposes of this report.

RISK MANAGEMENT AND HEALTH AND SAFETY IMPLICATIONS

Upgrading the DR SAN and the subsequent plans to move the equipment to L.B Wandsworth during the summer period will significantly reduce the risk in respect to future loss of service.

No specific Health and safety implications have been identified.

BACKGROUND DOCUMENTS

None for the purposes of this report.

This page is intentionally blank