Data Centre Incidents – Summary report

Shared Service:	3C ICT
Date of Reporting:	12/01/2020
Completed by:	Sagar ROY, Interim Deputy Head of ICT

1.0 Introduction

1.1 This Report provides an overview of the resilience of the data centre services and information regarding the service interruptions that were experienced during Q3 20/21 that affected the councils.

1.2 The data centre service is managed by 3C ICT on behalf of the 3 partner councils and was delivered as part of a major workstream following the formation of 3C ICT in 2017. The design was developed in conjunction with a 3rd party supplier with scope of the work including procurement, commissioning, configuration of hardware as well as migration of systems and services from legacy environments to the new. The data centre project was completed in Q3 19/20 when the Shared Service Board agreed to the closure of the server room consolidation project, although services had actually been running live since Aug 2018 when the migration process was started. The 3C Shared Services Board had overall responsibility for the project with director level representatives from each council sitting on the project board supported by intelligent clients and 3C ICT project management and technical resources.

2.0 Background

2.1 The server room consolidation project was commissioned by the 3 partner councils to bring all their separate server rooms and data centres in to 2 purpose designed facilities. The aims and benefits included but was not limited to the following :- standardising of design, improve resilience and availability, up to date / modern equipment and hardware, up to date software and operating systems, enterprise class server management, etc.
Each council came from a different starting point with regards to their current server room and data centre facilities, but the end point would be running services from across 2 different sites and the decommissioning of the legacy equipment. From an SCDC point of view this involved redesigning and migrating specific services from the single facility that had been based at Cambourne to the new data centre that was distributed over 2 different sites.

2.2 The legacy data centre/server room set up at SCDC comprised of a single location with little or no built in resilience. Concerns had been raised that it was under provisioned with no standby equipment arrangement to cater for major failures. The RDS (Remote Desktop Service) environment had frequent failures, there was a reliance on dated backup technology and the data storage was of particular concern at the time as they were deemed already at the end of their life cycle. Some of the key services were running on equipment purchased in 2009 and demand had considerably outstripped capacity across a number of areas of the infrastructure. In addition, key parts of the infrastructure were not virtualised – which appears to be contrary to the agreed design and strategy.

2.3 Many of these issues and the impact they had on the council were brought out in the independent report SCDC commissioned from the EELGA (Paul Tonkins report) during 2018. It noted that the 'decision to sweat the assets had created an environment not fit for purpose'. Major service outages and Priority one incidents seem to be occurring several times a week and 'couldn't go a day without ICT breaking'[1]. One of the comments covered the future planned schedule of work from 3C ICT - that even if 'all that SCDC did was roll out CA – There will be a dramatic improvement'. The report also

[1] Appendix B shows the number of major outages per quarter for SCDC as a comparison to legacy SCDC setup

noted that consideration should be given to further investment. Fortunately this was already underway in the form of the infrastructure projects that 3C ICT were delivering across the 3 councils. The server room consolidation project, Council Anywhere and Wide Area Network to name but a few.

2.4 The speed, pace and scale of change for this entire infrastructure programme of improvement set this work apart from all the other projects that 3C ICT had undertaken (or likely to undertake in the next 3 to 5 years) and whilst we would have liked to have delivered all the improvements and benefits in a shorter period of time 3C ICT didn't always have enough capacity and resources for specific service areas and their projects. However, priorities were discussed and agreed with the intelligent clients across the 3 councils and some in areas additional funding was provided by the councils to support faster delivery or enhanced capability.

3.0 Current Data Centre Services – overview

3.1 The design of the current data centre service, which was delivered with direct support from an approved industry partner uses an 'Active-Active' arrangement across two data centres. This was used to provide much improved resilience, availability, capacity and fault tolerance of services. This was a significant improvement over what the 3 councils had available prior to the data centre being delivered. For example – SCDC would have to rely on tape back up in the event of hardware failure or data recovery once new or replacement hardware was delivered, whereas now, data is immediately available to restore to virtual servers. Almost on demand.

3.2 When operating normally, both halves of the data centre are actively used with services that we host spread out across both sites. There are elements of automated management that spreads the load equally amongst the available resources which helps mitigate most spikes and peaks in demands from service areas.

3.3 In agreement with the project board and the intelligent clients at the start of project, the service was spec'd so that either half of the data centre can run known live services out of a single site if needed. E.g. in a disaster recovery scenario if an entire site was lost or suffered a catastrophic failure. All the necessary data is also replicated in both halves, so data integrity is also maintained in the event of one site being lost.

3.4 There have been several instances over the past 12 months where the resilience and capacity improvements delivered by the new data centre have been proven. Key examples include Q4 19/20 where there was failure of a critical component and all services were able to be recovered within 30 mins. In Q2 20/21 all live services were migrated and moved to one half of the data centre for 2 days whilst major electrical power work was carried out at the other half of the data centre. As a result of the massive increase in working from home due to Covid-19 the infrastructure environment used to support remote access has been able to be scaled up to accommodate 4 to 5 times the original number it was configured to support, and most recently, Q4 20/21, all services we running live out of the 2nd half of the data centre whilst the equipment from the 1st half of the data centre was moved from Cambridge to Peterborough.

3.5 The move of the data centre was the culmination of an 11-month long project and is considered a once in a decade piece of work. Without this kind of data centre design and technology, interruption to data centre services would have lasted several hours in the case of the server failure and several days in the case of the power shut down and data centre move.

4.0 Recent Data Centre outages

4.1 During October and November, the Councils suffered from a series of data centre outages that severely impacted council operations. Staff working remotely from the office were unable to use or access services due to a widespread / cascading failure of network services within the data centre infrastructure. A list of these outages is provided in Appendix A which includes specific information about each incident and highlights those that are linked. On each occasion all efforts and resources were allocated to resolving the issues, supporting officers and staff, minimising the impact and investigating the root cause.

4.2 The main series of outages that were experienced, which had the biggest impact on the councils has since been traced back to the installation of the latest version of the underlying virtual server software. This upgrade took place during the Q2 20/21 in line with recommended good practice and as part of our routine processes to apply security and maintenance patches. Unfortunately this software introduced a previously unknown (by the software vendor) bug that a very specific combination of server hardware, software and network card triggered. It was noted by the supplier and the 3rd party who was brought in to assist 3C ICT with investigation after the first incident that 'this was a one in a million chance' that we'd uncovered the bug and that this had not been reported by any of their other customers globally. To address the issue an accelerated programme of replacing the network cards in all the servers took place over a series of nights and out of hours in order to minimise the impact on the councils – Many staff were working extended hours / working flexibly due to Covid. This work was completed in early Nov.

Note - 3C ICT had recently been subject to an audit review relating to server maintenance planning which reported high levels of assurance with our approach to server patching.

5.0 Findings and follow up activity related to the Data Centre stability issues.

5.1 The main cause of the data centre issues that have been experienced during Oct and Nov was a software bug in the network card drivers that are used within the server infrastructure. It wasn't until the 2nd major outage on the 20/10/20 that this was able to be fully diagnosed with assistance from the software vendor, 3rd party support company and hardware vendor. In order to provide assurance to the councils regarding the data centre design and the pending migration from Cambridge to Peterborough (which has since been successfully completed), 3C ICT engaged a couple of independent suppliers to quote on carrying out a health check and review of the data centre services.

5.2 A supplier has since been selected and work has started on completing this. As part of the discovery work carried out they reviewed our decision to upgrade the virtual server software earlier in the year that introduced the bug. Also a check of the change process 3C ICT followed to approve the work and allow it to proceed. They have said that all the necessary and expected checks were carried out to inform the decision to proceed with the upgrade – which they would have recommended customers adopt anyway to keep up to date with security patches. This included a hardware compatibility check which showed our combination of hardware and software was originally supported. They also agreed that the immediate action/remediation steps of replacing the network cards was the right approach to restoring stability.

5.3 The larger piece of work to carry out the deeper health check and design review is underway following the necessary agreements being put in place. This will take place in 2 stages – firstly to ensure that the necessary mitigations are in place to allow the data centre migration to progress at the end of Jan 2021, and secondly the wider review of the data centre design, capacity, resilience and future options, risks and consideration of moving services to externally hosted / cloud service providers as opposed to maintaining services 'on premise. The discovery and field work is expected to complete at the end of Feb with reports and findings being issued during March.

5.4 One of the major elements of the report will be detailed advice and guidance on a recommended approach to upgrading a key data centre component that controls and manages the networking within the infrastructure. This is also key to any future cloud adoption/migration strategy the council may wish to adopt. One of the benefits of this would be to open up more options to provide access to services hosted away from council run/operated data centres which could change the risk profile for service availability and service continuity. These options will be explored further within the final report.

5.4 As of the last week in Jan 2021 we've had almost 14 weeks of Data Centre running where the network card driver bug has not been encountered again. Whilst this hasn't completely addressed all the improvements that are needed as part of the ongoing data centre programme of work the replacement of the network cards and related changes has introduced much needed stability back into the infrastructure. This stability also allowed 3C ICT to continue with the data centre migration work safely – without risking data, service availability and service performance.

Appendix A
List of Data Centre service Incidents in Chronological order with a brief description, duration, services impacted, cause and any other related/relevant information.

Q3 – Tue 06/10/2020

Description Remote access to internally hosted systems and services affected
Duration - Started at 18:55. Resolved at 20:00. Outside normal working hours, but duration of interruption 1hr 5mins.
Services impacted - Remote/Working from home access to systems and services. Teams and email access via council smart phones not affected
Cause – planned maintenance work by network service providers revealed an underlying hardware fault which affected resilience of network services into the council.
Additional information – Agreement reached with all EastNet partners that planned maintenance work will now not take place before 8pm

Q3 – Wed 07/10/2020*

Description Remote access to internally hosted systems and services for all Council Anywhere users affected
Duration - Started at 18:00 on the 7th Oct. Resolved 12:30pm on the 8th Oct. Overlapped outside normal working hours and working hours. Elapsed from for interruption was 18.5 hrs. 4.5 hrs normal working hours.
Services impacted - Remote/Working from home access to systems and services. Teams and email access via council smart phones not affected
Cause – Initial root cause was thought to be a bug with Virtual Server software which caused Data Centre networking and routing failures. NOTE - This incident is linked to those on the 20/10/20 and 02/11/20 as well as the maintenance work on the 7th/ 8th Nov and 10th/11th Dec.

Q3 – Tue 20/10/2020*

Description Remote access to internally hosted systems and services for all Council Anywhere users affected
Duration - Started at 10:00. Resolved at 14:30. During normal working hours. Duration of interruption was 4.5 hrs
Services impacted - Remote/Working from home access to systems and services. Teams and email access via council smart phones not affected. Same symptoms and impact as incident on the 07/10/2020.
Cause – In depth diagnostics with the vendor now concludes that there is a Software driver bug with the type of network cards that are installed in the servers within the Data Centre and the version of the Virtual Server infrastructure we are running. NOTE. We updated to that version of the software in Mid Aug to keep up with security and stability patches. Additional software changes carried out in line with vendor recommendations.

Q3 – Sat 24/10/2020

Description Remote access to internally hosted systems and services for all Council Anywhere users affected
Duration - Started 12noon. Resolved 13:30 Outside normal working hours during planned maintenance work. Duration of interruption was 1.5 hrs
Services impacted - Remote/Working from home access to systems and services affected when Global Protect connectivity was interrupted. Teams and email access via council smart phones not affected.
Cause – Server security and stability patches took longer than normal to apply which caused a longer than expected interruption to some services. Downtime window is based on previous experience. NOTE. The subsequent monthly security patching and stability maintenance windows have been completed within the expected time frames.

Q3 – Mon 02/11/2020*

Description Remote access to internally hosted systems and services for all Council Anywhere users affected
Duration - Started at 16:15. Resolved at 22:00. Overlapped outside normal working hours and working hours. Elapsed for interruption was 5.75 hrs. 0.25hrs normal working hours.

Services impacted - Remote/Working from home access to systems and services. Teams and email access via council smart phones not affected. Same symptoms and impact as incident on the 07/10/2020 and 20/10/2020.
Cause – Software driver bug with the type of network cards that are installed in the servers within the Data Centre. NOTE. It was agreed that the software changes carried out wasn't enough of a mitigation and as there was no ETA on a permanent fix from the Vendor a decision was taken to replace all the network cards of this make/type/model across the environment. Accelerated purchase and procurement process for the network cards and 3rd party support carried out.

Q3 – Weekend of 7th and 8th Nov* and weekdays 10th and 11th Dec* – Planned maintenance work to replace all the relevant network cards across both data centres. There were planned outages during both sets of maintenance windows, but the all services remained up and operational.

Q4 – Fri 08/01/2021

Description Remote access to internally hosted systems and services for all Council Anywhere users affected
Duration - Started at 04:30. Resolved 09:20. Overlapped outside normal working hours and working hours. Elapsed time for interruption was 4.8hrs. 1.4hrs normal working hours.
Services impacted - Remote/Working from home access to systems and services. Teams and email access via council smart phones not affected.
Cause – Planned overnight maintenance to support the migration of the data centre from Cambridge to Peterborough required the Cambridge half of the data centre to be shut down and testing carried out. After the tests were successfully completed services were rebalanced across both halves. This process artificially created a situation where the network systems thought it was in a failed state when in fact it was live and resulted in VPN / security software failing to connect. It's been concluded that this set of circumstances can only occur during or after this kind of maintenance and isn't possible during normal operation. As a result procedures have been updated to prevent a repeat in future if a data centre shutdown and restart was required. NOTE – ECDC also experienced the same issue with their Global Protect service that morning which was resolved at 10:30, but there is nothing to suggest that the two services are linked in any way.

*linked incidents and activity – Data Centre Stability issues.

Appendix B
Other background data regarding history on incidents and major outages impacting SCDC:-
Q1 19/20 – 4 P1’s incl one specifically affecting members access to O365 (external fault)
Q2 19/20 – 4 P1’s incl comms room A/C failure – over heating at SCDC comms room caused network equipment to shutdown
Q3 19/20 – 2 P1’s incl one caused by BT affecting telephony services
Q4 19/20 – 2 P1’s incl one major failure of the server infra – resolved within 30mins because of fault tolerance in the data centre

Q1 20/21 – 2 P1’s – incl one 3rd party slow network performance issue
Q2 20/21 – 3 P1’s – incl one European wide Microsoft issue + 1 extended service outage (cert renewal issue)
Q3 20/21 – 6 P1’s – incl DC network card bug, one MS global issue,
*Q4 20/21 - 1 P1’s – Planned maintenance follow up