FUJ00096450

Post Office Red Alert – Independent Technical Review
James Stinchcombe, Principal Solution Architect
27th April 2010

Executive Summary

There have been a number of problems with the Post Office service over the last few weeks, with the most severe being complete loss of the HNG-X Service on 26th and 27th March, with Horizon Banking being severely impacted on 1st April (with 80% of transactions failing). This has resulted in the account being placed on Red Alert.

As a result, an independent technical review has been requested to look at the specific problems and also take a wider perspective. This is short assignment to provide feed back to both the account and the Public Sector business unit.

This report is split down into the different areas considered and provides recommendations on how things can be improved. Given this is a short assignment it has not been possible to do a comprehensive review of all areas of the solution and where further investigation by the account would be desirable this is included.

The problems that are causing the Red Alert are not a single issue and there is no "magic bullet" to solve them. The teams need to continue to resolve the known issues while continuing to investigate the problems. It is likely that additional issues will emerge as the solution continues to be investigated.

Based on the observations in this report, the table below summarises the risk of significant business disruption for Horizon and HNG-X that the current solution represents. It also details which recommendations that help resolve or make progress on these issues

These recommendations should be considered as system improvements; not all will be necessary for re-starting HNG-X deployment. Completion of some of the activities is likely to take some time and needs to be planned accordingly.

Introduction

The HNG-X project for Royal Mail is providing a new, centralised service for the branches of the UK Post Office. It is currently part way through rollout, with approximately 5% of the 11,700 branches having been migrated from Horizon (the existing system provided by Fujitsu) to the new HNG-X solution.

To minimise running costs, HNG-X is a centralised, online only architecture. This means that failures in the data centres are likely to cause widespread outage to the service with knock-on business impact.

There have been a number of problems with the service over the last few weeks, with the most severe being complete loss of the HNG-X Service for periods of the branch trading day on 26th and 27th March, with Horizon Banking being severely impacted on 1st April (with 80% of transactions failing). This has resulted in the account being placed on Red Alert.

FUJ00096450

As a result, an independent technical review has been requested to look at the specific problems and also take a wider perspective. This is a short (2 day) assignment to provide feedback to both the account and the Public Sector business unit.

This report is split down into the different areas considered and provides recommendations on how things can be improved (marked as "Rx:"). Given this is a short assignment it has not been possible to do a comprehensive review of all areas of the solution and where further investigation by the account would be desirable this is included.

The following areas have been considered:

- Capacity - can it support the volumes expected and has this been proven in the live estate?
- Stability – how stable is the solution and what is causing the problems?
- Recoverability - how well does the system recovers from failures and the business impact of this?
- Supportability – can the solution be supported, including appropriate monitoring?

It is clear from the engagement that the account is working very hard to resolve the problems and even in the 2 days of the assignment several issues have been addressed. It is possible therefore that some information in this report may already be out of date.

Capacity

This section looks at the capacity of the solution in live and whether the volume testing is sufficiently representative.

Capacity – Live System

The live system has good capacity monitoring, with performance data being available in near real time. This has allowed a comprehensive look at the capacity of the key components of the system.

The following table shows the assessment of the capacity of the key components of the solution from a processor, memory disk and network traffic perspective:

Key component	Processor	Memory	Disk	Network
Branch Database Servers (BRDB) (holds data for HNG-X Branches)	Green	Green	Green	Green
Branch Access Layer (BAL) (access servers for Horzion Banking/debit card and all HNG-X)	Green	Green	Green	Green
Network Banking Persistent Store (NPS) Database (holds data for banking transactions – including reversals)	Green	Amber (1)	Green	Green
Branch Support Database (BRS) (holds historic records for	Green	Green	Green	Green

FUJ00096450

The evidence from the capacity monitoring shows that the problems with HNG-X are being caused by stability issues rather than capacity. It is the instability of the service (as highlighted in the next section) that is perceived by the end users as "timeouts".

The network banking NPS is showing signs of memory stress (1). Response times for banking shows that the system response times show large variations unpredictably at different times of the day. -This is consistent with memory stress. (NOTE for FJ only – instability implies that the system is not working, which is not true).

The root cause of this were changes to the NPS software, including upgrade of Oracle from 9.2 Horizon to 10.2 HNG-X and the hosting of APOP database on the NPS. NPS1 (which supports banking and APOP) is showing worse performance. NPS2 (which supports banking only) is showing less stress, but seems to have got slightly worse since the Oracle patchset has been applied.

It is worth noting that APOP is expected to grow and there is a change to improve APOP resilience by it to NPS2 as well.

The quickest and simplest solution to the problem is to replace the current server blades used to run NPS with server blades with more memory. By contrast, reducing memory usage by the application is likely to be difficult and time consuming, requiring both volume and resilience testing to be completed before it could be safely applied to live.

Upgrading the hardware can be achieved with low risk (e.g. by replacing one of the two NPS nodes, allowing this to run for a few days before replacing the other node). Options for this replacement are already under review. The Capacity Management Team has some suggestions on how this could be achieved.

R1: Urgently review the implementation of a hardware upgrade for the servers for NPS. This will be the most effective solution to the problem.

R2: Consider in the very short term, whether simple measures could be made to reduce memory usage on the NPS (e.g. by reducing the Oracle cache size or turning off some services) to improve stability.

The Branch Support database (BRS) is used to hold historic data from the HNG-X branches. It is used by the support community to investigate problems, so avoiding their workload from impacting counter performance.

Performance and other issues (for example switching off the replication during the data to remove possible causes of instability in the branch database; some outages and weekend upgrades) have resulted in the BRS being over 300 hours behind the branch data (it should be only a few minutes). This makes it unusable for support. As a result support have to use the Branch database itself for their work, making it more likely that there will be capacity issues with the service as it rolls out.

In addition, the capacity to allow the service to fall this far behind will be reduced as HNG-X rolls out. Given that the BRS is part of the solution to retain branch data for audit and archive, exhausting this capacity would be a significant issue.

FUJ00096450

One of the reasons for slow performance is the high Horizon workload. This has to be processed in large batches from the overnight processing, which is less efficient than when those branches are migrated to HNG-X. However unless the system is keeping up over a day (i.e. 24 hours behind) this could mean that it can never catch-up (particularly given the need for maintenance periods etc).

R3: Do not allow further rollout of HNG-X until the BRS is stable and keeping up (maximum of 1 day behind).

It is unclear from the performance data available why BRS is not keeping up – it appears to have sufficient headroom in processor, disk, memory and network resources to process the data significantly faster (although some disks are busy for short periods of time). There are a number of performance fixes in the pipeline and in addition Oracle has made some recommendations which are being reviewed (CAN WE CHECK THIS IS STILL VALID – haven't a number of fixes been implemented into live already? Have the Oracle recommendations been implemented?).

R4: Engage an expert in Oracle performance to review the performance of BRS to understand the existing performance and how to improve it.

Capacity – Volume Testing

Given the problems with the live service, questions have been asked as to how representative is the Volume Testing environment (VOL). The following is a short summary:

- The hardware for Volume testing is representative with some minor differences that are not material and in general mean that VOL has slightly less capacity than the live solution. (A spot check by the VOL team has shown no significant differences in layout).
- The number of BAL (branch access layer) servers is fewer to simulate running under a failure condition.
- Banking is run without standby agents.
- VPN has fewer servers as it only needs to show how one cluster scales.
- There are fewer instances of the support systems than would be found in live (including VPN, ACD, BMX, DEA, DWS).
- The large databases (e.g. BRS) are sized at 10% of live – as is normal practice for this type of testing.
- VOL does not have POLFS or POLSAP.

The differences above do not invalidate the volume testing and testing on this environment should be very similar to those expected to be seen on the live service. However given the problems with NPS on live, it is recommended again to look at the agents and BAL to ensure their memory usage on the database servers is representative.

R5: Review Volume Testing with regards to memory usage for agents / BAL to ensure that memory footprint on servers is accurate.

FUJ00096450

In general volume testing has progressed well - it has found a number of problems with the solution and largely proved that the solution should be capable of supporting the full volumes.

It should be noted however that volume testing is only as good as the data profiles used to generate the test data. Given that HNG-X is so different to Horizon, it would be worth checking experience from live to date to ensure the data profiles (particularly reports) are accurate.

R6: Consider information from those HNG-X Branches already live to see whether data being generated is representative.

In addition, the volume test cycle 3 has not yet completed. The main remaining planned tests to be run are those associated with stressing the system beyond the expected volumes. However there are also some limitations of the tests already run in cycles 1 and 2 that need to be addressed.

The following types of testing need to be run before it can be considered safe to rollout HNG-X to the full estate:

- Testing with representative data for report volumes (current testing was limited in this area).
- Stress testing – running the system at higher than contracted volumes. This is needed to ensure the system is stable (this is planned as part of cycle 3).
- Resilience Testing under volume – while running at full workload, fail key components (e.g. Branch database node; Branch access layer node, NPS node etc) to prove the system will recover correctly from such failures. Without this there can be no confidence the live system will recover from such failures (testing to date in this area has only be done with very small, manually generated data).
- Some testing was done without the BRS due to rig problems – this needs to be resolved. Full details of which tests were impacted is available from the VOL team.

R7: Complete performance & stress testing before rolling out HNG-X to full volumes.

FUJ00096450

Stability

As covered in the previous section, many problems in the current solution seem to be being caused by "stability" issues (i.e. things don't work correctly) as opposed to capacity (i.e. there is insufficient resource to do the work).

Due to the operational immaturity of HNG-X, it is to be expected that there will be some problems, however many of those already identified have not yet had fixes applied. This makes diagnosing the remaining problems difficult.

More importantly, during the recent period of instability there have been more low level failures than would be expected. This in turn has resulted in more recovery actions. Since the recovery itself has a number of problems (see "Recoverability") this in turn causes significant business problems.

There are a number of these problems that have been identified by the account that fall into the area of "stability" including:

- Lots of stale network connections resulting in a firewall running out of resources – causing failure of banking on 1st April (fixed by OCP on 6th April – CONFIRM this is correct).
- Old network drivers for many servers (the "PV driver" issue) which result in network connections and packets being dropped. This results in more failures than would otherwise be the case in both banking and HNG-X.
- Problems with Oracle "hanging" on the Branch Database servers. There have been no reported incidents since the LCK patch was implemented on 22/04/10 – NEED TO CONFIRM THIS IS CORRECT.
- The network being configured to present more traffic than needed to the BladeFrame server chassis (rather than just presenting traffic for those servers within the BladeFrame, it was also presenting data for other servers, which the BladeFrame then has to discard). This seems to resulting in the chassis itself being overloaded and unable to process all the traffic – this has been reproduced in VOL and looks likely as the root cause for the problems with Horizon on 6th April (Mark Jarosz has the details for this). It is understood that an OCP has been applied to resolve this issue.
- Servers generating significant broadcast traffic – when little or none is expected (understood some have now been fixed by OCP on 8th and 9th April).
- Memory issue on the NPS (see previous section) causing connections to be dropped.
- The ACE Blades (that handle load balancing to the BAL) have a known issue that will cause HNG-X counters to fail when transaction rates in the system grow (exact volume unclear – but is assessed to be greater 50% volumes). An update on any planned fix would be useful here.

R8: Apply all known fixes to live that will resolve stability issues as quickly as possible. This is needed to reduce both the likelihood of failures and also allow the remaining problems to be diagnosed.

Recoverability

FUJ00096450

There is clear evidence from the solution that both Horizon (as a result of the PCI changes) and HNG-X are not able to recover correctly from failures.

The most worrying are reconciliation BIMS exceptions – these are failures that require manual intervention by both Fujitsu and Royal Mail. Many of these failures result in end customers of the Post Office not being paid money (the exception shows that a bank believes it has paid out the money, whereas the Horizon/HNG-X system knows it did not pay out in reality).

The trend for these types of failure is clear:

- Before PCI/HNG-X change – 7 to 8 exceptions per month
- March 2010 – 263 BIMS exceptions
- April 2010 – forecast over 500 BIMS exceptions

In March, this excess volume was a combination of higher than expected days plus some high volumes associated with the major outages. However April's data is showing 30 or more problems per day and does not seem to be associated with major issues (it is this that the forecast is based on). Joanne Ball has detailed breakdown of the data if required.

These volumes are assuming current HNG-X branches. As HNG-X rolls out this is likely to grow unless the problems are fixed and identified. It is worth noting that these volumes are those that require manual intervention. Other reconciliation errors are frequently exceeding the report threshold of 30,000 per day.

This exception workload is causing significant (and unsustainable) workload on both Fujitsu and Royal Mail including:

- Investigation of each BIMS to ensure it really is a failure (Fujitsu Support)
- Handling of the problem with the banks etc (Royal Mail)
- Additional call volume from end customers asking why they haven't been allowed to take out their money (Royal Mail and other banks).

R9: Urgently Review causes of failures causing BIMS exceptions for both Horizon and HNG-X. Ensure these are fixed before further rollout of the HNG-X solution.

There are also other areas in HNG-X where the system does not recover well from failures. For example on 25th March an overnight job on the BRDB failed to run and this resulted in none of the HNG-X counters being able to login the next morning (a change has been applied to make the solution more defensive). There are undoubtedly others of a similar nature that have not been found in testing and may only be found in live operation.

Supportability

This section looks at how supportable is the solution - monitoring, and diagnosing problems.

One of the problems with the live system at the moment is the ability to monitor the solution. There are two broad areas that need to be monitored:

FUJ00096450

- Failures in the solution – technical faults that may or may not result in loss of service.
- "Wellness" – is the system behaving as expected at a business level?

The second is particularly important as it makes up for any shortfalls in the monitoring solution. It is also needed to ensure it known that the system has recovered from any failures.

There are a number of known problems with the current monitoring solution and as a result often the first thing operations know there is a problem is when Post Masters phone up to say there is a problem. Working out what is the root cause of the incident is then difficult as many problems end up with similar behaviour as far as the Post Master is concerned (e.g. timeouts).

R10: Review current issues with monitoring solution to make sure they are given appropriate priority. Ideally, these should be fixed before HNG-X is rolled out further to ensure that incidents can be responded to in a timely way (this is understood is being considered as part of the second work stream in the Red Alert).

R11: Review approach to monitoring of live based on lessons learnt with the current problems to see if the solution provided is sufficient – particularly with the need to monitor "wellness" in business terms that are understood by Post Office.

There are also a number of issues with the ability of the SSC to efficiently and effectively diagnose problems or to provide evidence to the offshore teams. These include:

- The diagnostic logs for the branch access layer are spread over 40 files (10 servers, two instances per server, two logs per instance) with each interaction from the branch randomly using a particular instance. This means that tracing a problem that a branch has is very time consuming since there is no tooling support for this.
- All evidence for offshore has to be sanitised before it can be sent to remove financial transaction information (in order to comply with the Data Protection Act). The tools to do this sanitising aren't trusted, so the SSC have to manually inspect each file before shipping.
- Some of the requirements in the diagnostic area have not been met, in generally making support less efficient than was budgeted for.

As a result of this, the stability and the recoverability issues the SSC are significantly overloaded. This in turn makes it slower to diagnose the root cause of problems and increases the risk that things are missed.

R12: Review how problems are diagnosed on HNG-X in the light of recent experience to understand the gaps and their business impact.

Within the design area, it is clear that each area (e.g. Database, Networks, Platforms, Applications etc) is working well and making good progress on their issues. However there was no evidence of a senior architect ensuring that these different strands are brought together effectively and being communicated (for example there is no single list of technical issues that need to be resolved). This doesn't mean that the teams

FUJ00096450

weren't working together – but rather that there is a lack of technical leadership. As a result things are likely to be taken longer to resolve than necessary, agreeing priorities can become difficult and things are likely to be missed. This should be resolved once the new CTO joins the account.

R13 Until the CTO joins, consider nominating one of the senior architects to provide technical leadership across the design teams and provide a single point of contact.

FUJ00096450

Conclusion

The problems that are causing the Red Alert are not a single issue and there is no "magic bullet" to solve them. The teams need to continue to resolve the known issues while continuing to investigate the problems. It is likely that additional issues will emerge as the solution continues to be investigated.

Based on the observations in this report, the table below summarises the risk of significant business disruption for Horizon and HNG-X that the current solution represents. It also details which recommendations that help resolve or make progress on these issues:

Consideration	Horizon Recommendations	HNG-X Recommendations
Capacity (ability of solution to support volumes)	R1, R2	R3, R4, R5, R6, R7
Stability (is the solution stable)	R1, R2	R8
Recoverability (Can the solution correctly recover from failures)	R9	R9
Supportability (ability to support and monitor the solution)	n/a	R10, R11, R12, R13

These recommendations should be considered as system improvements; not all will be necessary for re-starting HNG-X deployment. Completion of some of the activities is likely to take some time and needs to be planned accordingly.