the Availability Digest

www.availabilitydigest.com

Help! My Data Center is Down!

Part 5 - Upgrades

February 2012

Data centers are extraordinarily complex. They include hundreds or thousands of servers and storage subsystems with their applications and operating systems, all interconnected by vast internal networks. A failure in any one of these components can bring some if not all data-center functions to their knees.

However, major failures are not always caused by hardware or software. A disturbing number are caused by upgrades that go wrong. If a fallback plan has been put in place, then such a failure is typically not a problem. However, in too many cases, data centers have undertaken an upgrade with no backout procedure in place. If the upgrade fails, major applications will be down – for hours and sometimes for days.

In our previous articles on data-center failures, we focused on failures due to power, storage subsystems, and network faults. In this article, we look at some major data-center outages due to faulty upgrades. The stories are all true and are taken from the Never Again archives of the *Availability Digest*.

IRS Goof Costs U.S. Taxpayers $300M

The IRS (the Internal Revenue Service) is responsible for collecting taxes in the U.S. As part of this responsibility, the IRS uses a complex fraud-detection system. However, tax laws change; and fraud perpetrators get smarter. After several years of successful operation of its fraud-detection system, the IRS realized that it had to upgrade the system if it were to continue to be effective at catching fraud. To implement this upgrade, the IRS asked for competitive quotes for a new web-based system. The winner began work on the new system in 2001 with an expected completion date in late 2005, just in time for processing the 2006 tax returns.

However, as often happens, the project experienced delays. As a result, the system cutover was rescheduled to January, 2006, and then to February. Though there were numerous warning flags, the contractor repeatedly voiced confidence that a February, 2006, cutover would be achieved. Based on these assurances, the IRS decided to shut down its current system in late 2005 in anticipation of the new system becoming available shortly in February. So confident were they in the new system that a contingency plan to recover from a failed cutover was never even created.

The ax fell when a March, 2006, test showed that the new system could not even process a day's worth of data in a day. In effect, it could not keep up with the workload and would never work! And there was no fallback plan – the old system was gone. The result was that the IRS paid an estimated $300,000,000 or more in fraudulent or improper income tax refunds for tax returns filed in 2005.

PayPal Upgrades with No Fallback Plan

PayPal provides payment processing services for online merchants, auction sites, and others. Now owned by auction giant eBay, PayPal processes over $50 billion USD per year and services 200 million accounts. It is used in 190 countries and supports 19 currencies.

Clearly, a great deal of today's ecommerce flows through PayPal. Its services must be extremely reliable, as billions of dollars of revenue for millions of small online merchants depend upon it. An extended outage could put many small merchants out of business. That is what happened when critical PayPal services went down – not for hours but for weeks.

The problem occurred when PayPal attempted to upgrade its Instant Payment Notification system with no rollback plans in the event of an upgrade problem. Suddenly, many ecommerce customers could not process orders. PayPal accepted orders from buyers with no problem and extracted its fees. Upon completion of each transaction, sellers expected from PayPal an Instant Payment Notification message, which would allow them to process the order. Instead, they received an "invalid order" message. Consequently, merchants could not process their orders.

This left a situation in which buyers were told that their order was accepted and that their money was removed from their account. However, they never received the goods they ordered. Not only were merchants denied the revenues upon which they depended, but they also were swamped with negative complaints by buyers, complaints that sank their ratings on eBay and other sites.

Clearly, the upgrade had gone wrong; but PayPal had not prepared a fallback plan. It took them two weeks to get merchant services once again operational.

BlackBerry – OMG, It's Déjà Vu!

From April, 2007, to December, 2009, the Blackberry service (provided by Research in Motion, RIM) suffered four multi-hour and in some cases multiday outages, all caused by failed upgrades.

April, 2007 – Blackberry service suffered a two-day outage, and it took another day to clear up the backlog of messages before service returned to normal. Blackberry reported that the outage was caused by the "introduction of a new noncritical system routine" designed to optimize cache performance. Since when is messing around with cache noncritical?

February, 2008 – Half of all North American subscribers suddenly found their email screens empty. This outage was caused by an upgrade to RIM's routing system. For redundancy purposes, RIM provides two IP networks for its North American service. RIM clients are split between these paths. The upgrade took down one path, taking out half of the North American subscribers. It seems that there was no way to switch these subscribers to the "redundant" path.

December 2009 – RIM issued an upgrade to its BlackBerry Messenger instant messaging service and encouraged all subscribers to download it. A few days later, the upgrade caused BlackBerry to suffer a major outage that took down email, Internet browsing, and instant messaging across North and South America. It was hours before service was restored. RIM released a new upgrade a few days later and directed its subscribers to download this upgrade.

December, 2009 – Just a week later, it became apparent that the new upgrade was also faulty. Users in the Americas and Asia Pacific reported problems. RIM immediately issued a third upgrade, which fortunately corrected the problem. However, it took from Tuesday to late Thursday (Christmas Eve) before email was freely flowing again. RIM explained that the problem stemmed from the two updates having a flaw that caused an unanticipated database issue.

Even Mighty Google is Not Immune

Similar to the Blackberry service, Google has suffered many major outages due to upgrade failures.

October, 2008 – Following a major upgrade to iGoogle, Google decided unilaterally and without prior warning to update its Google Apps portal pages to look more like its iGoogle personalized home pages. Suddenly, links were broken, buttons were misconfigured, and strange "gadgets" caused confusion, preventing access to many Google Apps services. It took days for Google to correct the problems.

February, 2009 – Gmail was down around the world for two-and-a-half hours, earning Gmail the infamous nickname of "Gfail." The cause was a new feature that it had installed to keep email geographically close to its owners. In preparation for the update at one of its European data centers, Google routed users to another nearby data center. This inadvertently overloaded that data center, which caused a cascading effect from one data center to another, ultimately taking down all of Gmail services.

September, 2009 – To perform a router upgrade, Google staff took down several Gmail routers. What the staff had underestimated was the additional load that this would put on the remaining routers. The overloaded routers rejected traffic that they could not carry, and this traffic was rerouted through other routers that then became overloaded. Within minutes, all of the routers in the Gmail network were overloaded; and Gmail crashed.

Summary

An upgrade to any data-center component is a complex operation. It should be properly planned, and all cognizant personnel should be available during the upgrade. There is too much of a chance that regardless of the effort put into the upgrade plan, something will go wrong.

If something does go wrong, there had better be a fallback plan of some sort that will allow operations to continue while the upgrade is corrected. Fallback plans come for free in active/active networks. The other nodes in the application network are currently operational and are handling the transaction load. No action has to be taken to ensure continued operations. In an active/backup configuration, the alternate system perhaps can be put into service. If there is no redundancy, the fallback plan typically involves backing out the upgrade and returning to the original system.

So far in this series, we have focused on technical failures. But a disturbingly large number of failures are caused by human actions, whether accidental or malicious. In fact, human error played a role in many of the failures that we have described in these articles. In our next article, we will look at some spectacular data-center failures that were caused by people.