the Availability Digest

www.availabilitydigest.com

Help! My Data Center is Down!

Part 1: Power Outages

October 2011

Nothing strikes fear in the hearts of IT management (and corporate management for that matter) like that of losing the entire corporate IT infrastructure. To make sure that this never happens, companies invest heavily in their data centers using technologies ranging from fault-tolerant NonStop systems to redundant data centers.

The Uptime Institute defines a four-tier certification for data centers.[1] The most reliable, a Tier IV data center, has redundant everything and can achieve an availability of 0.99995. This represents an average downtime of ten minutes per year.

Many companies consider this unacceptable since a data center may virtually never fail; but should such a failure occur, it might be days before IT services can resume. The company might well be out of business by then. Therefore, it is common to provide one or more backup data centers so that operations can continue within a few hours of a data-center failure.

Nevertheless, there is a disturbing incidence of significant data-center outages. These incidents show that it is not enough to try to protect against any event – fire, flood, power outage, network failure, and so on – that might take down a data center. Some event that was not even envisioned is going to happen sometime, and it will take down a data center somewhere. If this is your data center, your Business Continuity Plan had better specify how the enterprise is going to continue in the absence of IT services for hours or even days.

In this series, we review from the archives of the *Availability Digest* some Never Again horror stories that show some of the unlikely events that have taken down entire data centers, including in some cases the disaster recovery site as well. In this first part, we look at some unusual power outage stories.[2]

Google's App Engine

The Google App Engine is a compute cloud service for users to develop and host their web applications in Google's data centers. The App Engine virtualizes applications across not only multiple servers but also across multiple data centers to ensure that no fault will take down the applications. Should a server fail due to equipment problems or even due to an entire data-center outage, the applications on that server will be rapidly migrated to a surviving server, where they will continue to run. Or so Google thought.

[1] Data Center Tiers, *Webopedia*.
[2] Our thanks to The Connection for giving us permission to reprint this series of articles.

On Wednesday, February 24, 2010, a Google data center hosting the Google App Engine suffered a major power outage.[3] The power was out for about thirty minutes, but Google's backup power kicked in and continued to power the data center.

The only problem was that for some reason (never disclosed), about 25% of the servers did not receive backup power and subsequently went down. This caused their applications to fail over to the surviving servers in the data center, which were not configured to handle that much additional load. The server failures cascaded until the entire data center went down.

The simultanous failure of all the servers was an event that had never been considered. There was a great deal of confusion about how to handle the multiple server failures. The first decision was to fail over the data center. The documented procedures were followed but led to an unsuccessful failover.

It then looked like the primary data center had returned to operation, so processing was returned to it. This turned out to be an erroneous observation since the primary data center became operational only in the presence of reduced load. It could still not handle the full load, and it once again failed. Finally, knowledgeable technical personnel were reached; and the backup data center was brought successfully online. Two-and-a-half hours had passed since the initial failure. If things had gone properly, it would have taken only about ten minutes to fail over to the backup data center.

A post-mortem showed that the failover procedures had just been improved to make them more automatic. Unfortunately, the failover documentation was still in the process of being updated. Parts of the documentation for the new failover procedures incorrectly referred to the old data-center configuration rather than to the upgraded configuration. Clearly, the newly documented failover procedures had not been tested; nor had the staff been trained. Otherwise, the errors would have been found.

The Coffee Pot Fiasco

It was time to upgrade the nodes in another company's active/active system. This was a major upgrade involving new hardware and a new operating system. The company had successfully applied rolling upgrades to its nodes in the past by taking down one node at a time, upgrading it, and reintroducing it into the system.[4]

As best practices dictate, the system at each node was powered by a separate circuit protected by an uninterruptible power supply (UPS). When the new system was rolled in at one of the nodes, the installers found that all of the UPS power connectors were being used. There was not one available for the new system. As a consequence, the new system could not be powered up.

So as not to delay the upgrade, the new node was temporarily connected to the facility's unprotected power. Though this power source was not protected by a UPS, the plan was to correct this problem in short order by adding an additional connector to the UPS output.

Evidently, no record was made of this issue on any task list. The required power connector change was forgotten; and the upgraded system continued to run successfully for quite a while on the unprotected power source.

As time went on, the load on the unprotected circuits gradually increased as the company grew. More and more people meant more lighting, more heating, more air conditioning, and more workstations.

One fateful day, an employee performed a normal, everyday task. He or she plugged in the coffee pot to make fresh coffee. This was the straw that broke the camel's back. The coffee pot blew the circuit

[3] Poor Documentation Snags Google, *Availability Digest*; April 2010.
http://www.availabilitydigest.com/public_articles/0504/google_power_out.pdf
[4] Active/Active Save #1: Coffee Pot Takes Down Node, *Availability Digest*; November 2006.
http://www.availabilitydigest.com/private/0102/a-a_save_1_coffee_pot_takes_node_down.pdf

breaker, taking down everything that was on that circuit. This included dropping primary power to the upgraded system, which had never been moved to the UPS circuit.

The node, however, kept on running for a while. It was supported by an internal UPS that kept it operating long enough to save its state and to shut it down gracefully following a primary power failure.

Fast action on the part of the staff at the site restored the primary power in just 35 seconds – an admirable feat. Unfortunately, the system's internal UPS only lasted for 30 seconds. The node shut down and suffered a 30-minute outage until it was brought back online.

If the system had not been an active/active system, its users would have been denied service for a half hour. However, as it turned out, the users assigned to the failed node were quickly switched over to surviving nodes and suffered no apparent outage.

The Planet Blows Up

The Planet provides dedicated servers for a variety of companies, many of which are web-hosting companies. Operating six data centers in Texas, The Planet is the largest, privately-held dedicated-server hosting company in the world.

On Saturday, May 31, 2008, at 5:55 PM, an explosion took down The Planet's Houston data center.[5] A short circuit in a high-volume wire conduit set a transformer on fire, which then caused an explosion of battery-acid fumes from the UPS battery-backup system.

The explosion was strong enough to blow down three walls surrounding the electrical equipment room on the first floor of the data center. It blew apart the power-transfer switch that transferred the data center from utility power to backup diesel generator power, thus knocking out power to the entire data center. Fortunately, no one was injured.

For safety reasons, the fire department evacuated the building and directed that the backup generators could not be turned on. It wasn't until after 10 pm that staff was allowed back into the building to assess the damage.

The staff was able to move some customers to new servers in other data centers, but limited cooling capacity in the data centers limited this to only a few customers. Shortly after the explosion, The Planet had to deny further requests for reprovisioning.

In total, 9,000 servers went down. Three thousand of the affected servers were on the first floor of the data center, and 6,000 servers were on the second floor. Since The Planet provided dedicated servers for thousands of web-hosting companies, the web sites that were taken down measured in the millions.

The Planet's staff was able to restore power to the second-floor servers; and around 5 pm Monday evening - two days after the explosion, the second-floor servers were once again operational.

Restoring power to the first-floor servers was a much more difficult challenge due to the extensive damage. Each of these servers was brought online as soon as possible; but four days after the explosion, full service to all customers had yet to be restored.

[5] The Planet Blows Up, *Availability Digest*; September 2008.
http://www.availabilitydigest.com/public_articles/0309/planet_explosion.pdf

A Truck Downs Rackspace

Rackspace, headquartered in San Antonio, Texas, provides web-hosting services for thousands of web sites around the world. It operates eight data centers – four in the U.S. and four in the U.K. Founded in 1998, it has grown at a rate in excess of 50% per year.

Rackspace is very conservative in its power management. It uses triplexed power sources –two independent power sources and a diesel generator to provide power during a switchover.

In the early evening of Monday, November 12, 2007, at 6:30 PM, Rackspace suddenly lost power to its Dallas data center. Unbeknownst to Rackspace, a trucker had passed out and rammed a transformer that fed the data center. The transformer exploded, and the data center went black.[6]

As planned, Rackspace's emergency diesel backup generator kicked in; and the data center came back to life and continued in operation with but a brief interruption. This allowed Rackspace operators to switch to their secondary power source – a completely separate utility line feeding the building. At this point, the emergency generator had done its job and was shut down. Triple modular power redundancy had paid off.

However, fifteen minutes later, the secondary power source shut down. This time, the blackout was requested by the emergency personnel trying to free the trapped truck driver so as to avoid electrocution of not only the truck driver but also the emergency workers. Things were happening so fast at the scene of the accident that Rackspace was not notified by the electric utility of the intent to shut off the data-center power.

Again, the emergency generator started up and continued to power the data center. The diesel generator was designed to power the data center indefinitely (so long as fuel was available), and the data center was once again operational with little impact on the hosted web sites.

But a serious and unanticipated problem became apparent. With each interruption in power, the air-conditioning chillers had to recycle. It would take them about a half hour to recycle before they were effectively cooling the data center again. The chillers were down for about fifteen minutes as a result of the first power outage, and they would have been back on line in another fifteen minutes, a delay accounted for in the data-center design. However, with the second interruption in power, the chillers had to once again recycle.

With thousands of powered servers pumping out heat, the temperature in the data center was rapidly climbing to a dangerous level. Management realized that this extended time without air conditioning would cause the servers to overheat and could cause significant damage to the hardware. Therefore, management reluctantly decided to shut down all of the servers in the data center to protect them. The Dallas data center was now completely nonoperational.

Once power and cooling were restored, all of the thousands of servers had to be restored to service. Most of the web sites were up by the following day, Tuesday. However, they had been down for hours.

The Rackspace failure could have been avoided by proper disaster planning. Nowhere, evidently, in Rackspace's business-continuity planning was the concept of data-center redundancy. The N+1 redundancy that Rackspace had built into its emergency UPS (uninterruptible power supply system) and even into its HVAC (heating, ventilation, and air conditioning) system and whatever redundancy that it had in its server farms came to naught when the entire data center was taken down.

[6] Rackspace – Another Hosting Service Bites the Dust, *Availability Digest*; December 2007. http://www.availabilitydigest.com/public_articles/0212/rackspace.pdf

Triple Redundancy Failure on the Space Station

In June, 2007, a triply-redundant attitude and environmental-control computer provided by Russia failed on the International Space Station (ISS).[7] Had this been a mission to Mars, it would have been fatal. Only the space station's proximity to Earth, which put it in range of support and resupply missions, prevented a tragedy. Though the problem was circumvented in a few days by the space-station crew, it took weeks for the station crew and ground engineers to determine the source of the problem.

The crew quickly determined that the failure was caused by the simultanous loss of power to all three computers. Power had been shut off by a surge-protection unit designed to protect the computers from power surges beyond the capabilities of their own power filters. A NASA internal technical report describing this failure said, "On 13 June, a complete shutdown of secondary power to all (three) central computer and terminal computer channels occurred, resulting in the loss of capability to control ISS Russian segment systems."

Russian officials were quick to blame NASA for "zapping their computers" with "dirty" 28-volt power from a newly-installed solar array. This was the first of many bad guesses by top Russian program managers and would distract engineers trying to get to the real source of the problem.

In the meantime, the computers had to be fixed – and fast. The station crew assumed that some external interference such as noise in the 28-volt power supply was responsible for generating false commands inside the computers' power-monitoring system and caused it to send shut-down commands to all three computers. Based on this reasoning, the crew bypassed the power monitoring system to two of the computers by using jumper cables. These two computers were now subject to damage by power surges, but by now the power system had settled into a steady state and was generating clean power.

The astronauts spent their time disassembling the power control boxes and the associated cabling in order to look for clues that might lead to the cause of failure. Though multiple scopes and probes failed to find the problem, their eyes and fingers did.

What they discovered was that the connection pins from the power-monitoring unit were wet and corroded. Continuity checks showed that the command lines in the cable coming off the unit had failed. Even worse, one of the command lines had shorted. It was the power-off command line that went to all three computers. The shorted condition created the disastrous power-off command. The jumper cables had bypassed the false power-off command and had allowed the computers to function properly once again.

But what had caused the corrosion? Water condensation. The problem was traced to a malfunctioning dehumidifier dripping condensate water. The situation was aggravated by a stream of cold air from another location on the dehumidifier that at times cooled the cables below the dew point at which moisture could condense. As temporary further protection, the crew rigged a thermal barrier between the computers and the dehumidifier. The thermal barrier was built using a surplus reference manual and ordinary gray tape.

Once the problem was understood, it became clear that the system suffered from a fatal design flaw. The supposedly triply-redundant design included a single point of failure – the external power monitoring unit that, by itself, could turn off all of the computers. Should it fail (as it did due to condensation), the triply-redundant system was down.

[7] Triple Redundancy Failure on the Space Station, *Availability Digest*; November 2007. http://www.availabilitydigest.com/public_articles/0211/iss_tmr_failure.pdf

And Then There Was the Great Northeast Blackout

On August 14, 2003, much of the Northeast United States and neighboring Canada lost power. It would be several days before power was fully restored.[8]

How did such a disaster happen? Through a chain of events that started with a tree.

Power lines sag in hot weather. They also sag due to the heat generated by the electrical current that they are carrying. A high-voltage transmission line can blast a tree to its roots. This takes a tremendous amount of power and instantly overloads the transmission line. Therefore, it is imperative that trees under these transmission lines be kept trimmed so that they will not come in contact with the transmission lines.

The policy of FirstEnergy in Ohio was to trim trees every five years. However, they did not always stick to this schedule; and the result was that some trees under its transmission lines had grown too tall. August 14th was a hot day, pushing 90 degrees Fahrenheit in the Ohio area. Air conditioners and fans were imposing heavy demand on generating capacity. Between the heat of the day and the heat generated by the electrical current, transmission lines were seriously sagging. One apparently zapped a tree that was too tall. This caused the transmission line to shut down, putting more load on the remaining transmission lines.

These transmission lines sagged even further due to the increased load. Over the next two hours, two more transmission lines were taken out of service by tree contact. Problems rapidly increased until a giant power surge took down the entire Northeast electric grid. 508 generating stations at 256 power plants, including 22 nuclear power plants in the U.S. and Canada, went offline. Over 40 million people in the U.S. and Canada were without power. The financial impact ran into billions. It wasn't until the next evening that partial power was restored.

Where were the operators during this time? It turns out that the GE monitoring system in the Ohio control room had failed due to a software bug, and the operators were blissfully unaware of the problem. GE Energy subsequently sent a patch correcting this bug to all of its customers around the world.

Summary

None of these events was foreseeable. This speaks to the needs of a good Business Continuity Plan. During the Risk Assessment phase, it is perhaps not so useful from an IT perspective to try to identify all of the events for which strategies must be planned.[9] Rather, the Risk Assessment should focus on the results of the events. It doesn't make any difference whether a fire or a flood or an explosion takes down a data center. What is important is that the data center is down. How will business continue if it loses its data center, no matter how that happened?

In our next part of this series, we will look at network problems that impacted IT operations. Again, we will find that many of these could not have been imagined.

[8] The Great 2003 Northeast Blackout and the $6 Billion Software Bug, *Availability Digest*; March 2007. http://www.availabilitydigest.com/private/0203/northeast_blackout.pdf

[9] Of course, other aspects of the BCP, such as personnel management, communication, and so on may depend upon the nature of the event.