the Availability Digest

www.availabilitydigest.com

More Never Agains VII

September 2012

Since our Never Agains VI summary that was published in the April 2012 issue, catastrophes continue to plague the IT industry. We already have reported on several – the DNS-Changer and Flame viruses, Royal Bank of Scotland's two-week outage due to a software bug, Knight Capital's disastrous high-frequency-trading algorithmic bug that literally destroyed the company, and, in this issue, the Go Daddy outage that took down millions of web sites worldwide. Below are some others that have made headlines during this period.

Nasdaq's Facebook Glitch

Facebook's IPO, scheduled to open at 11 AM on Friday, May 18, 2012, was expected to be one of the biggest IPOs in Nasdaq's history. As it normally does, Nasdaq sets the opening price via an auction held shortly before the IPO's opening. This is done with its IPO Cross system.

As planned, Nasdaq opened bidding on Facebook shares with IPO Cross before the IPO time. It terminated bidding and set the IPO price based on bids made by brokers during this pre-IPO auction. IPO Cross then checked to see if any orders had come in after the close and found that orders were still flooding in. It recalculated the opening price and checked again. Orders kept coming in and the price kept being recalculated.

The recalculation cycle went on well past the scheduled opening time of 11 AM, thus delaying the Facebook IPO. To correct the problem, Nasdaq switched to a backup copy of IPO Cross and opened the IPO a half-hour late, at 11:30. Unfortunately, the backup system ignored orders that came in after 11:11 AM.

As trading began, the system was so far behind that brokers did not get confirmations of their order executions for over two hours after the executions had occurred. They had no idea what had become of their orders. If their orders had come in during the dead interval from 11:11 to 11:30, they did not get executed at the opening price. They got executed at the rising price following the opening.

As a result, many brokers lost a lot of money. Nasdaq has set aside $13 million in a compensation fund for these brokers and has modified IPO Cross so that it does not recalculate opening prices.

Software Bugs Make Amazon Outage Worse

Over the weekend of July 2, 2012, severe thunderstorms moving in from the Midwest left millions without power in the Northeastern United States. Several major web sites were taken down – Netflix, Instagram, Pinterest, and Amazon. Most web sites came back up quickly, but Amazon's EC2 cloud service recovery was plagued by unforeseen software bugs.

Amazon structures its EC2 cloud with independent Availability Zones within regions. Users who configure backup sites in alternate Availability Zones will automatically fail over to their backup site if their primary site fails. At least, that is the intent. It didn't happen in this instance.

The problems started when power failed at Amazon's U.S.-East Region data center and the backup generators failed to start, taking down an Availability Zone. The control plane that connects Availability Zones then experienced problems and overloaded, making it difficult for users to fail over to other Availability Zones within the region. Then, a bottleneck appeared in its server reboot process, which meant that it took longer to bring the EC2 servers back online. Due to effects on several pieces of hardware, it took several hours to validate that the databases were correct.

Finally, a critical bug reared its head in Amazon's Elastic Load Balancer (ELB), which routes traffic to servers with capacity. The bug caused Amazon to erroneously rapidly scale the ELB's to larger servers, thus compounding the control plane overload.

The result was that backup applications in other Availability Zones were substantially unreachable and therefore ineffective.

Salesforce.com Downed by Power Upgrade

Salesforce.com is the leading provider of CRM (Customer Relationship Management) cloud-based services. It leases space for its data center in Silicon Valley from Equinox. Equinox is a builder of data centers in which customers lease space to install their own systems.

On Tuesday, July 10, 2012, Equinix was in the process of upgrading some electrical equipment that supplied power to the data center. As Equinix attempted to transfer the power load seamlessly from the old equipment to the new equipment, the transfer met an unexpected failure. This caused all of Salesforce.com's systems to lose power abruptly and unexpectedly.

Though power was quickly restored, it took Salesforce.com seven hours to bring its systems up and to restore service.

Calgary City Services Disrupted by Data Center Fire

On Wednesday, July 11, 2012, a transformer in the IBM data center of the city of Calgary, Alberta, Canada, exploded. The resulting fire knocked out all city services for three days. Though the city had a backup system, it was in the same facility as its production system. The water sprinklers that were triggered by the fire also destroyed the backup system.

The fire knocked out key public services for the city and for medical institutions. It took out the city's 311 emergency service and Alberta's property and vehicle information databases. It also disabled the medical computer network for Alberta Health Services, forcing the postponement of hundreds of elective surgical procedures.

In order to recover, IBM Canada flew backup tapes to a backup facility in Ontario.

A similar incident occurred in Dallas County, Texas, the previous week when a water-main break flooded the basement of the Dallas County Records Building, which houses the electrical supply equipment for the data center located on an upper floor. The result was a three-day outage for the county, which did not have a backup system.

These incidents show the necessity of having a backup site remotely located so that no single incident can take down both the production and backup systems.

Twitter Downed By Failover Fault

On Thursday, July 26, 2012, just before the start of the Olympic games in London, Twitter messages stopped flowing. Was it a message overload caused by the Olympics? It turned out that this was not the case. Rather, its primary servers had failed; and as Twitter tried to switch over to its backup data center, that data center failed also. A coincidental failure or a failover fault? Sounds like the later.

Twitter users in the Americas, Europe, Africa, Asia, and Australasia were all affected. It was two hours before service was restored. Visitors to the site were greeted by a half-formed message that said "Twitter is currently down." However, the message fields that were supposed to give the cause and the estimated recovery time were filled with computer code.

In its early days, Twitter was known for its unreliable service. However, more recently, its availability has significantly improved. Twitter claims 99.96% to 99.99% availability. However, just last June, Twitter went down again for about two hours due to a cascading software bug that caused its servers to crash.

Twitter has 140 million active users and sends about 340 million tweets per day. The record for a sporting event is 15,000 tweets per second. When Twitter goes down, people notice.

Hosting.com Taken Down by Human Error

Hosting.com provides cloud-based and managed hosting services for customer web sites, applications, and data. Early one Friday morning, shortly after midnight on July 28, 2012, as the company was conducting preventative maintenance on a UPS system at its data center in Newark, Delaware, a maintenance engineer executed an incorrect circuit breaker operation sequence that resulted in a total power loss to one of the data center suites within the facility.

Power was restored within 11 minutes, but over 1,100 customer web sites were offline for up to five hours as servers were rebooted and databases were recovered.

Hosting.com offers a backup option that adds about 30% to the cost of hosting, but only several dozen of its customers at the affected location had elected to purchase it.

This incident underscores the need for CIOs to carefully evaluate the cost of downtime versus the cost of high availability. It seems that many are willing to take the risk, and in this case may have lost. As one customer said, "They're asking for extra money based on incompetence." This is someone who should learn something about availability.

Hundreds of Millions Lose Power in India

India is the third largest nation in Asia and is the world's second most populous nation. It has been enjoying strong economic growth; and as its economy has grown, so has its need for electric power. However, its power infrastructure has been unable to meet its growing needs. India has missed every annual target to increase electrical power capacity since 1951.

The lack of attention to India's power generating capacity came to a head on Monday, July 31, 2012, when its northern electrical grid failed, leaving 300 million people in the dark. Power was restored by early evening.

However, this was just a prelude to the major disaster. Around noontime on Tuesday, the next day, three of five India grids collapsed. These were the northern grid, the eastern grid, and the northeastern grid. This left more than half of the nation's 1.2 billion people without power. Transportation stopped. Businesses closed. The lack of traffic lights caused massive traffic jams.

Economists have estimated that India's power problems could reduce its GDP by 1% to 2%.

Cut Cables Disable Wikipedia

Wikipedia is the crowd-sourced online encyclopedia that has grown into the fifth busiest site on the web. It is strictly funded by donations. In its early years, it used to joke that downtime was its most profitable product – whenever it was down, users would get a web page asking for donations. It has since improved its availability to the point that this technique no longer works.

Wikipedia operates two data centers, one in Tampa, Florida, and one in Ashburn, Virginia. While Ashburn services most of the Wikipedia traffic, many database services are provided by its site in Tampa. To provide this connectivity, Wikipedia uses two 10-gigabit fiber channels connecting the two sites.

However, on Monday afternoon, August 6, 2012, Wikipedia's latest availability record was broken when a backhoe severed the two cables between its Florida and Virginia data centers. Users received partially complete pages that had much of their content missing.

It took a little over an hour for the cables to be patched, and Wikipedia services were restored an hour after that.

Summary

Half of the outages that we have described above were caused by power failures – Amazon, Salesforce.com, Hosting.com, and half of India. Others were plagued with failover faults, either because the failover was compromised (Amazon, Twitter) or because the backup system was damaged (Calgary).

The experience of the city of Calgary shows the foolishness of depending upon a backup system that is not remote enough from its production system so as not to likely be affected by an event that takes down the production system.