

Help! My Data Center is Down!

Part 2: Storage Outages

November 2011

More and more, the data center is becoming part of the lifeblood of a company. If the data center goes down, so do much of the services that a company provides to its customers, vendors, and employees.

To avoid this disaster, most companies spend a lot of time and energy on hardening their data centers and, in some cases, even investing in backup data centers. As the saying goes, however, "The best laid plans of mice and men often go awry." Events that are so improbable that we cannot even think of them can and do happen.

In our previous article in this series, we discussed several unimaginable, power-related events that took out data centers, with outages lasting hours and even days. These ranged from a truck driver's heart attack and a battery-room explosion to the simple act of plugging in a coffee pot. The failure to keep a tree trimmed triggered the great Northeast Blackout of 2003.

In this article, we look at some spectacular storage-system failures. Corporate data is one of the most prized assets of a company. Companies do everything they can to protect the integrity of their data, from maintaining real-time remote backups to long-term offsite storage. Unfortunately, the press is replete with horror stories of companies that have lost their data for long periods of time or forever. Following are some Never Again stories showing some spectacular storage failures, pulled from the archives of the *Availability Digest*.¹

The State of Virginia Loses Twenty-Six State Agencies for a Week

On Wednesday, August 25, 2010, a large storage area network (SAN) in a data center providing most services for the State of Virginia began sending alert messages that something wasn't right.² However, a redundant, fault-tolerant EMC Symmetrix SAN with automatic failover was in service; and the SAN continued to operate properly.

The maintenance staff determined that one of the SAN controllers needed replacing. A few hours later, a technician replaced the board; and pandemonium erupted. Evidently, the technician pulled the good board of the fault-tolerant pair. At that point, the dual SAN crashed. 485 out of the state's 4,800 servers were knocked offline.

¹ Our thanks to The Connection for giving us permission to reprint this series of articles.

² The State of Virginia – Down for Days, *Availability Digest*, October 2010.
http://www.availabilitydigest.com/public_articles/0510/virginia.pdf.

System testing showed that the crash had caused severe database corruption. The only alternative was to rebuild the massive database from magnetic tape, a lengthy process. All data since the last tape backup was lost.

Database recovery from tape took almost a week to complete. Even then, a significant amount of data could not be restored. Thousands of photographs and signatures submitted by residents for drivers' licenses during the four days prior to the outage were lost. For an entire week, twenty-six of the state's agencies were down, including the Motor Vehicle Bureau, Social Services, and the Department of Emergency Management just as Hurricane Earl was approaching.

American Eagle Loses Its Online Store for Days

American Eagle Outfitters is a multibillion dollar retailer. Online apparel web sites account for 12% of its revenue.

In July, 2010, the loss of a major storage subsystem caused American Eagle's online services to come crashing down.³ As the staff attempted to fail over to the backup system, the secondary storage system failed as well. The joint failure of the primary and secondary disk drives has a probability of occurrence of much less than a million to one, but it happened.

American Eagle keeps magnetic-tape backups of its databases. Restoration of 400 gigabytes of data was started. However, for some reason, the staff was only able to get a restoration rate of one gigabyte per hour. At this rate, it would take over two weeks to restore the database.

American Eagle had built a remote disaster-recovery site, and staff initiated a failover to the remote site. However, they soon discovered that the remote site was not yet operational - a big surprise to American Eagle since IBM was supposed to have had the site operational months earlier.

American Eagle's online stores now were down. It took four days to restore purchasing capability to the web sites. Several important ancillary functions, including order tracking, wish lists, and order history, were inoperative for another four days.

A Maintenance Error Costs DBS Bank \$200 million

DBS Bank is the largest bank in Southeast Asia and a leading bank in Singapore and Hong Kong. In 2002, DBS contracted with IBM to run much of DBS' data-center services. IBM built new data-center facilities in Singapore and Hong Kong to house the DBS IT systems.

One morning in July, 2010, IBM operations staff began to receive alert messages.⁴ The messages indicated instability in a communications link within a major storage system used by most of the bank's mainframe applications. However, the bank's systems are highly redundant and are designed for high resiliency. Consequently, the storage system was still fully functional.

The support group deduced that it was a cable problem, and a cable replacement was scheduled for the wee hours of the morning. The new cable worked temporarily; but after a few hours, the alert messages resumed and the standby SAN took over.

The local service technician decided that he could probably fix the cable and started fiddling with it. The result – he took down the standby SAN as well. Gone were the bank's online services, ATM services and

³ [American Eagle's Eight-Day Outage](http://www.availabilitydigest.com/public_articles/0509/american_eagle.pdf), *Availability Digest*, September 2010.

⁴ [Singapore Bank Downed by IBM Error](http://www.availabilitydigest.com/public_articles/0508/singapore_bank_outage.pdf), *Availability Digest*, August 2010.

POS services. Fortunately, the database was still intact; and the staff was able to restore services in ten hours.

Nevertheless, the Singapore Monetary Authority directed the bank to set aside S\$230 million (about USD \$180 million) additional regulatory capital for operational risk.

JournalSpace Closes its Doors Following a Malicious Database Attack

Started in 2003, JournalSpace was a popular and growing blog-hosting service. In December, 2008, JournalSpace lost its entire database and was unable to recover. The database's demise was the malicious act of a disgruntled employee – even worse, the IT manager.⁵ JournalSpace claims that it had caught the IT manager stealing from the company. They summarily fired him; but he did a slash-and-burn on his way out, overwriting the entire database with garbage.

This should have been only a minor irritant because all that was needed to cure the problem was to restore the database from the backup copy. The problem? No backup copy! It was, of course, the IT manager's responsibility to ensure that a backup copy was periodically taken and preserved. His backup strategy was to use a RAID 2 mirrored disk. If one disk failed, the database was still available on the mirror. Unfortunately, upper management should have known that this was not a backup strategy at all. True, it protected against a hard-disk failure. But it did not protect against a site disaster – or a malicious act.

After valiant efforts, JournalSpace was unable to recover the database. Thousands of bloggers lost years of their work. One month later, JournalSpace closed its doors forever.

A Major South Pacific Bank Replicates Database Corruption to all its Systems

A major bank in the South Pacific ran three redundant nodes for its critical ATM, POS, and online banking services – a production node, a disaster-recovery (DR) node, and a development node. The development node could be pressed into service as the production node if need be.

In December, 2008, an operating-system patch was made to the production system to correct a processor problem.⁶ The patch had worked on earlier versions of the operating system to correct the problem but had not been tried on the current version being run by the bank. The bank installed the patch anyway, and it seemed to work fine.

However, the patch was actually causing write errors, which corrupted the production disks. The errors eventually brought down the production system. When the bank tried to fail over to the DR site and then to the development site, the same problem occurred. The corruption had been replicated to all of the systems.

Unfortunately, the bank had not made backups. It had no way to restore the database. It was able to get some data from unrelated systems and from some of its partners. Partial operations took over three weeks to restore. However, much of the database was never recovered. It took the bank months to resolve all of the disputes.

⁵ *Why Back Up?*, *Availability Digest*; April 2009.

http://www.availabilitydigest.com/public_articles/0404/journalspace.pdf.

⁶ *Innocuous Fault Leads to Weeks of Recovery*, *Availability Digest*; December 2008.

http://www.availabilitydigest.com/public_articles/0312/simple_fault.pdf.

The Backup/Restore Plan That Didn't Work

A major collections-services company maintains a large RAID-based credit database to provide credit information for its subscribers. It backs up its database every night to an offsite service and maintains a spare server to take over should the primary server fail. The restore procedure is periodically tested to ensure that it works properly.

In November, 2006, one of the RAID disks failed.⁷ However, the system continued to operate properly; and management decided not to replace the failed disk immediately in order to save money. Bad decision! The primary server failed the next month, and the credit-reporting system went down.

The company attempted to rebuild the database on the backup server from the offsite backup copy. Only then did it realize that recovery testing had never been done using the backup server – only the primary server. The backup server was configured with a single disk rather than a RAID system, and the configuration difference prevented the restore from being successful.

After a long weekend, the restore problems were finally resolved only to find that the backup server did not have the capacity to handle the Christmas load. It was 21 days before the primary server was repaired and normal service could be restored.

The \$38 Billion Keystroke

The Alaska Permanent Fund receives and invests proceeds from the sale of Alaskan oil and minerals. The fund pays a yearly dividend to all Alaskan residents. The fund balance as of 2006 was about \$38 billion.

On a fateful day in July, 2006, a computer technician mistakenly deleted the oil fund database.⁸ This was not a big problem because the data also existed on a redundant backup disk. However, under the pressure of the moment, the technician also managed to reformat the backup disk.

The backup tapes were retrieved from storage. Only then did the magnitude of the disaster become apparent. The tapes were unreadable. The triple redundancy that was built into the system was not enough. 800,000 scanned images representing transactions over the last nine months were lost.

Over the next several days, staff tried vainly to salvage the data but were unsuccessful. Fortunately, there was a fourth level of backup – the paper documents themselves, stored in over 300 cardboard boxes. Each of the 800,000 documents had to be rescanned and sent through quality control. It took 70 people working nights and weekends for almost two months to complete the recovery, but complete it they did.

Smartphone Sidekick Service Loses All Its Subscribers' Data

Sidekick, a popular smart phone provided by Microsoft and marketed by T-Mobile, suffered an outage in early October, 2009, that wiped out all of the data of its one million worldwide subscribers.

The Sidekick service stores each subscriber's data in its central data center. Of note is that in contrast to other smartphone services, Sidekick does not provide a means for subscribers to back up their data locally. Subscribers are totally dependent upon being able to retrieve their data from the central servers.

In October, 2009, an upgrade to the Sidekick storage area network was undertaken without a proper backup.⁹ The upgrade went wrong and wiped out the online primary and backup databases. All of the

⁷ Don't Wait for the Other Shoe to Drop, *Availability Digest*, February 2007.

http://www.availabilitydigest.com/private/0202/other_shoe.pdf.

⁸ The Alaska Permanent Fund and the \$38 Billion Keystroke, *Availability Digest*, April 2007.

http://www.availabilitydigest.com/private/0204/alaska_oil_fund.pdf.

subscribers' address books, calendars, photos, and other data were gone. Incredibly, there was no backup copy of the database.

T-Mobile suspended sales of Sidekick and offered to allow customers to withdraw from their contracts. This wasn't enough to ward off the inevitable. A class action lawsuit was launched against Microsoft and T-Mobile. It alleged:

"T-Mobile and Microsoft promised to safeguard the most important data their customers possess and then apparently failed to follow even the most basic data protection principles. What they did is unthinkable in this day and age."

JPMorgan Chase Taken Down by Replicated Corruption for Two Days

JPMorgan Chase is one of the big four banks in the United States. If the bank's operations are compromised, it is felt by millions of customers.

In September, 2010, exactly this happened when the bank lost its online banking services, used by over 16 million online customers.¹⁰ The problem occurred in a large Oracle database managed by an Oracle cluster. EMC SANs provided the data storage. The database held authentication data and user profiles for its online customers.

It appears that an Oracle bug corrupted key files in the authentication database. This corruption was dutifully replicated by the EMC SANs so that both the active and mirrored SANs were corrupted. With no authentication, the online applications became inaccessible.

In addition to online banking services, connection to the ACH (Automated Clearing House) for scheduled payments was lost as well as access to private-client trading portfolios. Online loan applications died. Web and mobile applications were down.

Because both the active and standby data-storage units were corrupted, the bank's only option was to rebuild the database. Services were not restored until two days later.

Even then, the applications suffered hours of poor performance as users logged on and tried to catch up. Online bill payments were delayed up to three days. \$132 million in ACH scheduled payments were held up. Also held up were hundreds of auto-loan and student-loan applications.

Summary

Data centers are taken down by storage-subsystem failures, even when triple redundancy comprising remote online copies and offline backups are used. There are common threads in these failures. In our nine horror stories above, four incidents were caused by system and maintenance staff. Three failures were caused by replicated corruption, and two involved failed backup procedures that had not been thoroughly tested.

As we concluded in the first article of this series, we cannot possibly imagine all of the disasters that can impact our data centers. In our Business Continuity Planning efforts, we should not focus on the specific events that can impact IT services. Rather, we should focus on what we will do when we lose critical portions of the infrastructure that provides our IT services. We know that we will lose infrastructure at some point. We just cannot predict why.

⁹ Sidekick: Your Data is in "Danger", *Availability Digest*, November 2009.

http://www.availabilitydigest.com/public_articles/0411/sidekick.pdf.

¹⁰ JPMC Downed by Replicated Corruption, *Availability Digest*, November 2010.

http://www.availabilitydigest.com/public_articles/0511/jpmc.pdf.

In upcoming articles, we will look at other classes of data-center failures, including network faults and upgrades gone wrong.