the Availability Digest

www.availabilitydigest.com
@availabilitydig

Salesforce Takes a Dive

July 2016

In May, 2016, Salesforce suffered an outage of a portion of its SaaS (software as a service) cloud that lasted almost six days. This failure probably ranks as one of the worst cloud failures ever, beating the recent cloud failures of such giants as Amazon and Microsoft.

The root cause of the outage was multi-faceted. The failure of a circuit breaker in its Washington data center was compounded by a latent, unknown bug in its storage arrays in its Chicago data center.

Salesforce's recovery was hampered by an inability to fail over from it Washington data center to its backup data center in Chicago. We have noted many times in the *Digest* that failover to another data center is an iffy process if it has not been thoroughly tested. It appears that Salesforce did not give failover testing the attention it deserved.

Salesforce

Salesforce was founded in 1999. It is a cloud computing company headquartered in San Francisco, California, U.S.A. It provides software as a service to its customers via its CRM (customer relationship management) product.

CRM provides the facilities that companies need to analyze and manage customer interactions and data throughout the customer lifecycle. Its goal is to improve business relationships with customers, assisting in customer retention, and driving sales growth. CRM systems can provide customer-facing staff with detailed corporate information, personal information, purchase history, buying preferences, and concerns.

Salesforce Instances

Salesforce groups customer systems together in *instances* that link data centers and databases together in organized groups. There are about three dozen instances serving North America (NA instances), eight in Europe (EU instances), five in the Asia-Pacific region (AP instances), and 50 other "sandbox" (CS) instances.

The NA14 Instance Outage

On Monday, May 9, 2016, at 5:46 pm PDT time, the Salesforce technology team noticed a service disruption of its NA14 instance that primarily served customers on the U.S. west coast. Customers on the NA14 instance were unable to access Salesforce services.

The First Root Cause – A Faulty Circuit Breaker

The problem was traced to a power failure caused by a faulty circuit breaker in the Washington, D.C. (WAS) data center.

Salesforce uses redundant intelligent circuit breakers in its data centers to segment power from the data center universal power supply to different rooms. One of these circuit breakers failed. However, the failure created an uncertain power condition. The backup circuit breaker could not confirm the state of the problem breaker, and this led to the redundant breaker not closing to activate the backup feed.

The Second Root Cause – An Inoperable Backup System

In order to restore service as quickly as possible, the Salesforce technical team decided to perform a site switch, moving the active NA14 instance from the WAS data center to its Chicago (CHI) data center. The switchover was completed at 7:39 pm, and service for the NA14 customers was restored. The technical team then started to recover the WAS NA14 instance via a local backup.

The NA14 instance performed properly through most of the night. But at 5:41 am on May 10th, the team noticed a degradation in performance of the NA14 instance. At 6:31 am, the degradation escalated to a service disruption as the result of a database cluster failure on the NA14 instance in the CHI data center.

The database failure resulted in file discrepancies in the NA14 database in the CHI data center. These discrepancies were replicated to the WAS database, thus corrupting it. All attempts to repair the discrepancies in the CHI data center failed. Salesforce was now in a position that it could not use the NA14 instance in either data center, and it could not update the WAS data center from the CHI data center.

The Salesforce team was able to restore the CHI database. However, the NA14 instance continued to run in the CHI data center in a degraded state through Wednesday, May 11. The team halted several internal jobs and staggered initial customer activity coming into the instance to try to smooth the workload.

With help from the storage array vendor, an unknown firmware bug was found in the storage array. It was exposed by the increased traffic volume coming into the array from the backlog of customer traffic that had built up during the time of the initial disruption. The bug increased the time to write to the array. As a consequence, the database experienced timeout conditions, and database writes were unable to complete successfully.

This ultimately caused a file discrepancy in the database, and the database cluster failed and could not be restarted. It was these file discrepancies that been replicated to the WAS database before the CHI database failed, making the WAS database unusable. At this point, neither NA14 instance could be used.

The Salesforce team decided to restore the WAS data center from a local backup of the NA14 instance. However, all data that had been entered by customers since the switchover to the CHI data center was lost.

The team used redo logs from the CHI data center to replay the lost data at WAS. Unfortunately, this action could not be completed before the start of peak customer activity on Wednesday, May 11th. Rather than impacting another day of customer activity, the team decided to halt the replay of the redo log. All data from 2:53 am to 6:29 am on May 10th was not applied to the recovered NA14 instance.

On Thursday morning, May 12th, Salesforce posted a message to its status web page saying:

> "The NA14 instance continues to operate in a degraded state. Customers can access the Salesforce service, but we have temporarily suspended some functionality such as weekly exports and sandbox copy functionality."
>
> "The service disruption was caused by a database failure on the NA14 instance, which introduced a file integrity issue in the NA14 database. The issue was resolved by restoring NA14 from a prior backup, which was not impacted by the file integrity issues. We have determined that data written to the NA14 instance between 9:53 UTC and 14:53 on May 10, 2016 could not be restored."

Salesforce CEO Marc Benioff apologized for the outage on Twitter, providing customers with his email address.

All activity was finally restored on Sunday, May 15th, almost a week after the initial problem was discovered.

Lessons Learned

The initial cause of this outage was the failure of a redundant intelligent circuit breaker in the WAS data center. However, the service should have been restored by switching over to the CHI data center. The outage should have lasted no more than the two hours required to bring up the CHI data center.

However, a firmware bug in the CHI data center corrupted the CHI NA14 database and then crashed it. This corruption was replicated to the WAS database. Now neither NA14 instances could be used. It took days to return the NA14 instance to full service.

As we have emphasized in several *Digest* articles, you can't count on the success of a failover unless you have thoroughly tested it. This includes testing the backup system under full load. This clearly was not done by Salesforce; otherwise, they would have found the latent firmware bug. Failover testing is a risky and expensive activity. However, it is certainly better than taking a multi-day outage. Salesforce's experience proves this point.

Postscript

Salesforce later notified its customers that it had been able to restore the lost transaction data from the time between 2:53 am and 6:29 am on May 10 from a separate copy of the NA14 instance. This data was available to customers upon request so that they could manually re-enter the data.

Acknowledgements

Information for this article was taken from the following sources:

Salesforce outage persists across US, CEO wades in, *ZDNet*; May 11, 2016.
Salesforce experienced an outage and service disruption to the NA14 instance, sending customers to Twitter to complain and organizations to evaluate the best way to work with cloud software providers, *Information Week*; May 12, 2016.
RCM for NA14 Disruption of Service, *Salesforce document*; May 16, 2016.
Circuit breaker failure was initial cause of Salesforce service outage, *ZDNet*; May 18, 2016.