the Availability Digest

www.availabilitydigest.com
@availabilitydig

Microsoft's Azure Cloud Goes Down - Again

December 2014

Public clouds are gaining increasing acceptance by both large and small companies to host their applications. Companies can avoid the upfront costs of servers, storage, networks, and cooling as well as the ongoing costs of data center space, power, and staffing. All they pay is a fee for their use of the facilities owned by the cloud provider.

Cloud computing is certainly acceptable for ordinary applications. But is it reliable enough to host a company's mission-critical applications? Research by Infosys suggests that about four in five enterprises plan to move their mission-critical workloads into the public cloud.

However, there continue to be examples of massive cloud failures that have taken down applications for hours and in some cases even days. If a company decides to host a mission-critical application in a public cloud, it must have plans as to how to continue to offer the services of the application should the cloud fail.

A case in point is the recent global failure of Microsoft's Azure cloud, a failure lasted for eleven hours. It adds to a string of Azure outages over the last few years.

The Azure Cloud Failure of November 2014

Similar to Amazon, Microsoft has built multiple Azure availability regions around the globe. Each region can comprise one or more facilities. A company can host an application in multiple regions so that if a region fails, the application can continue to run in another region. As of this date, Microsoft is operating nineteen Azure regions around the world.

Around 1 AM UTC (Coordinated Universal Time, or about 8 PM Eastern Standard Time in the U.S.), on Tuesday, November 18, 2014, 2Microsoft's Azure cloud went dark in twelve of its nineteen global regions. The regions affected included Central US, East US, East US 2, North Central US, South Central US, West US, North Europe, West Europe, East Asia, Southeast Asia, Japan East, and Japan West.

Many Azure services were affected, including Azure Storage, Virtual Machines, SQL, Visual Studio, HDInsights (Hadoop for Big Data Analytics), Active Directory, Azure Backup Services, and the Service Health Dashboard.[1] Hosted web sites went down, including some of Microsoft's own web sites. Access to Office 365 and Xbox Live, the popular multiplayer game site, was interrupted. Microsoft's MSN news and web site was inaccessible.

Of particular importance was the loss of the Service Health Dashboard. The Dashboard posts the status of all Azure services and informs customers of problems and expected times to solution. Without the

[1] See Update on Azure Storage Service Interruption, *Microsoft Blog*; November 19, 2014, for a complete list of Azure services affected.

Dashboard, Microsoft could not even tell its customers that it was having a problem. After about three hours, Microsoft resorted to Twitter in an attempt to keep customers informed.

It was not until 11:45 AM UTC, eleven hours later, that Azure was brought back online.

Microsoft warned customers who had been affected that they "will see a data gap during the impacted window." "Data gap" was not defined, but probably meant that updates that were attempted during the outage may have to be repeated. Data integrity of customer databases may also have been an issue.

In a post mortem, Microsoft identified a failed upgrade to its Azure Storage system as the culprit.

Azure Storage Services

Azure Storage services provide scalable, durable, and highly available storage for applications and for virtual machines running on Azure. Azure Storage offers support for blobs, tables, queues, and files:

- **Blob storage** stores any type of text or binary data, such as a document, media file, or videos.
- **Table storage** stores structured datasets. It allows rapid development and fast access to large quantities of data.
- **Queue storage** provides reliable messaging between application components for workflow processing.
- **File storage** offers shared storage for legacy applications. Azure virtual machines and cloud services can share file data across application components.

An Azure customer storage account can contain up to 500 terabytes of data. Data in a storage account is replicated to ensure durability and highly availability. There are several options for replicating storage account data:

- *Locally redundant storage* maintains three copies of the data within a single facility in a single region.
- *Zone-redundant storage* maintains three copies of the data replicated across two to three facilities, either within a single region or across two regions.
- *Geo-redundant storage* (the default) maintains six copies of the data replicated three times within the primary region and three times in a secondary region hundreds of miles away from the primary region.
- *Read-access geo-redundant storage* allows read access to data at the secondary region in the event that the primary region becomes unavailable.

What Caused the Failure?

The Azure outage was caused by two factors - a faulty upgrade and an improper deployment.

The upgrade was being made by Microsoft to its Azure Storage service. The upgrade was intended to improve the performance of the service. It had been extensively tested for several weeks in a subset of customer-facing storage services for Azure tables. Microsoft calls this procedure "inflighting." It is intended to identify issues before broad deployment. The inflighting demonstrated a notable improvement in storage performance. With this result in hand, Microsoft proceeded to roll out the upgrade to all of its regions.

Unfortunately, a decision was made to roll out the configuration change to all regions simultaneously. Microsoft's standard protocol was not followed, which specified that changes should be rolled out in incremental batches. Thus, all regions received the update within a short period of time.

As soon as some of the regions came online with the Azure Storage performance upgrade, problems arose. The Storage blob front-ends went into an infinite loop and could take on no further traffic. This issue had gone undetected during the testing of the update. Other services built on top of blob storage began to experience outages.

By the time Microsoft staff had identified the problem, twelve of its nineteen regions were affected. The Azure cloud went down around most of the globe.

Microsoft immediately started rolling back the Storage upgrade. However, this required restarting the Azure Storage front-ends. Restarting the front-ends significantly delayed the recovery.

It was not until eleven hours later that the failed upgrade was fully rolled back and the Azure cloud was returned to full functionality.

Microsoft's Mitigation Steps

Following its post-mortem, which in the spirit of full transparency was published in detail, Microsoft initiated the following steps to improve the reliability of Azure:

- It fixed the infinite-loop bug in the CPU improvement logic for blob front-ends before the upgrade was redeployed.
- It improved its recovery methods to minimize recovery time.
- It ensured that its deployment tools enforce the standard protocol of applying production changes in incremental batches.
- It improved its Service Health Dashboard infrastructure to make it more resilient.

Azure's History of Outages

This was the second large-scale Azure outage in less than three months. The earlier outage caused significant Azure downtime in the U.S. Japan, and Brazil. In some cases, the outage lasted nearly a week.

On March 13, 2009, a network issue caused applications to crash in a beta version of Azure; and the recovery service became overwhelmed and crashed. Azure was down for almost a day.[2]

On February 29, 2012, the entire Azure cloud was struck with a leap-day programming error. Azure went down for 32 hours.[3]

On October 30, 2013, a failed upgrade that was prematurely deployed across all regions took out a single point of failure and prevented any new applications from being deployed. It took a day and a half to correct the problem.[4]

[2] More Never Agains III, *Availability Digest*; July 2009.
http://www.availabilitydigest.com/public_articles/0407/more_never_agains_3.pdf
[3] Windows Azure Cloud Succumbs to Leap Year, *Availability Digest*; March 2012.
http://www.availabilitydigest.com/public_articles/0703/azure.pdf
[4] Windows Azure Downed by Single Point of Failure, *Availability Digest*; November 2013.
http://www.availabilitydigest.com/public_articles/0811/azure.pdf

Microsoft specifies an availability of 99.9% in its SLA for all of its services.. This represents about eight hours of downtime per year. In the last three years, Azure outages have blown through its SLA.

Summary

According to one blogger, Microsoft's compensation for violating its SLA is likely to be service credits. He notes that this is tantamount to saying "We're sorry for the crummy service – here's some more crummy service as compensation."

Azure's availability history certainly raises serious questions about the use of public clouds for mission-critical applications. Microsoft is competing with cloud platforms offered by other major companies such as Amazon, Google, and IBM. Lydia Leong, a VP and distinguished analyst at Gartner, notes that "Microsoft's disastrous inability to keep Azure outages confined to a single region is a major red flag for enterprises considering Azure."

This incident indicates that Microsoft's inflighting approach may not be totally effective. Why was the several-week inflight test conducted only on Azure Storage tables? What about blobs, queues, and files? Maybe it was lucky that only blob storage was affected.

In any event, it is a good bet that Microsoft is tuning up its inflight procedures and will ensure that upgrades are rolled out in a controlled fashion in the future. In fact, a problem with an upgrade to the Azure database the following week on November 24th was contained by doing just that.

Acknowledgements

Material for this article was taken from the following sources:

Azure outage raises questions about public cloud for mission-critical apps, *Business Cloud News*; November 19, 2014.
Microsoft Explains What Went Wrong in Latest Global Azure Outage, *CRN*; November 19, 2014.
Update on Azure Storage Service Interruption, *Microsoft Blog*; November 19, 2014.
What Happened? Microsoft's Fritz Discusses Azure Outage and Building Azure Sites, *Visual Studio Magazine*; November 19, 2014.
Microsoft Delivers A Post Mortem – The Reasons Behind the Global Azure-Alypse, *Forbes*; November 20, 2014.
Microsoft Says 11-Hour Azure Outage Was Caused by System Update, *Entrepreneur*, November 20, 2014.
Azure outage Tuesday produced disruptions to MSN website, Office 365, Xbox Live, and third-party services, as well as possible data integrity problems, *Information Week*; undated.
Introduction to Microsoft Azure Storage, *Microsoft Documentation*.