the Availability Digest

www.availabilitydigest.com
@availabilitydig

Windows Azure Downed by Single Point of Failure

November 2013

Clouds are expected to be highly redundant and resilient to any single failure. There is always another component that can take over in the event of a failure. Right?

Wrong! As it turns out, this is not always true. The Microsoft Azure cloud has a single point of failure, and this component failed at the end of October, 2013. The failure caused a worldwide partial compute outage. While the glitch did not prevent cloud applications from running, it took down certain cloud management functions for a day and a half. Specifically, new applications could not be put into service.

The Failure

The Azure cloud is distributed over eight worldwide regions – East, North Central, South Central, and West U.S.; North and West Europe; and East and Southeast Asia. Each of these regions is fault-isolated from the other regions so that a problem that impacts one region presumably will not affect other regions.

However, at 2:35 AM UTC on Wednesday, October 30, 2013, Azure users were greeted with the following ominous message:

> "We are experiencing an issue with Compute in North Central US, South Central US, North Europe, Southeast Asia, West Europe, East Asia, East US and West US. We are actively investigating this issue and assessing its impact to our customers. Further updates will be published to keep you apprised of the impact. We apologize for any inconvenience this causes our customers."

At 10:30 AM later that day, the Microsoft team reported that the problem had been addressed and that the company was in the process of correcting the problem. The fault lay in a management service called Swap Deployment. Swap Deployment allows developers to move cloud applications under development to production by swapping a virtual IP address. Microsoft noted that Swap Deployment operations could cause errors and suggested that service management functions be delayed until the problem was fixed. Though the problem did not affect any applications currently in production, new applications could not be deployed.

The frustration being felt by Azure developers is reflected in the following post to Microsoft's TechNet website:

> "I am getting the following message when I try to do anything with my cloud service:
>
> *"Windows Azure is currently performing an operation with*

> *x-ms-requestid9eaaa9d657ad2e838b4c55dec6fdc4b9 on this deployment that requires exclusive access.*
>
> "This happens when I try to delete the production deployment or delete the cloud service itself. When I try to deploy a new staging deployment, it says my staging slot is filled, but it shows nothing is deployed to staging.
>
> "Is this a common occurrence? The client I'm working with is getting nervous over the technical issues. I'm the one who convinced him to go with Azure for hosting when we are going into production next year and I'd like to be able to assure him that his is just a temporary hiccup in deployment. It isn't affecting the test site so that helps but not being able to deploy a new version is an issue."

Microsoft's answer to this plea was less than helpful:

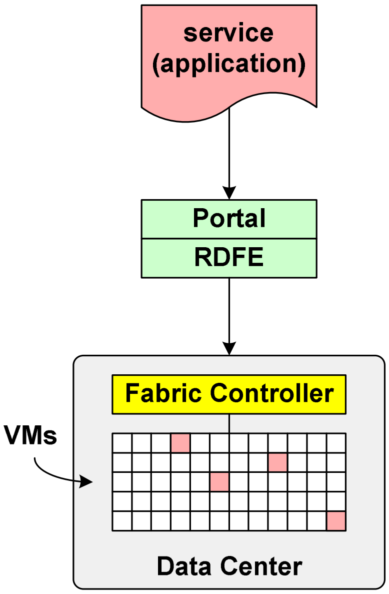
> "The Azure Management API (AMAPI) can be temperamental. There's not much you can do about this one other than let the operation internally timeout. I have seen hosted service deployments be stuck in a transitioning state for extended periods of time (4-8 hours) although this is uncommon.
>
> "You should open a support incident with Microsoft. Even if it resolves itself faster than they get to you, it is good to let them know that this is happening."

The user responded that he had opened up a support ticket but was horrified to find that this counted as one of his two allowed tickets per year under his MSDN subscription, even though the fault was Microsoft's.

Finally, at 10:45 AM the next day, Thursday, October 31st, a day and a half later, Microsoft posted the following notice:

> ""As of 10:45 AM PST, the partial interruption affecting Windows Azure Compute has been resolved. Running applications and compute functionality was unaffected throughout the interruption. Only the Swap Deployment operations were impacted for a small number of customers. As a precaution, we advised customers to delay Swap Deployment operations until the issue was resolved. All services to impacted accounts have been restored."

The Cause



The Azure cloud provides both a staging environment and a production environment for an application. The staging environment lets users test their systems before putting them into production. The production environment provides all of the facilities needed by the application to interact properly with the outside world.

Moving an application from a staging environment to a production environment is done by the management utility Swap Deployment. Swap Deployment initiates a virtual IP address swap between the staging and production environments for applications.

The culprit in the Swap Development utility was a module called Red Dog Front End (RDFE). RDFE provides the publicly exposed management portal and the service management API. User requests are fed through RDFE to the fabric front end, which disperses requests through aggregators and load balancers to Fabric Controllers. The Fabric Controllers in each data center direct the cloud's virtual machines and other resources.

When an application (a *service* in Azure terms) is to be deployed, it is passed to the Azure Portal Service. The Portal Service invokes RDFE to properly format the service, and the service is then sent to the Fabric Controller in an appropriate data center based on where the user has requested that his application run. Each data center comprises myriad servers organized into independent fault domain clusters. Each cluster contains about 1,000 servers. The Fabric Controller will instantiate versions of the application running as virtual machines in at least two clusters to maintain high availability (the Azure SLA calls for an availability of 99.95%).

The problem arose when Microsoft made an update to RDFE. The change was tested on a small number of nodes within a single cluster and worked perfectly. The Azure developers then pushed the change out to all of the data centers worldwide, and that is when the problem exposed itself. Though existing applications continued to work, the Swap Deployment management facility was broken; and new applications could not be deployed in any of the data centers worldwide.

Due to the way that Azure is built, there can only be one RDFE in the entire cloud. Therefore, the RDFE is a single point of failure. The developers could not run one instance of an RDFE and then deploy it to other instances only after it had proven itself in production.

Summary

Though this outage did not affect existing production applications, it certainly was irritating to heavy users. Imagine having a tight deadline to get an application into service, only to be blocked by an outage such as this. Regardless of whom it affected, a worldwide outage may certainly damage confidence in Microsoft's ability to manage a large distributed network.

It was only a little over a year and a half ago that the entire Azure cloud went down for over thirty hours, compute capacity and all. This problem was due to a software bug in the way that Microsoft developers calculated Leap Day.[1]

These two outages lead to an interesting observation. No matter the system, there is one single point of failure, and that is software. A software bug that is allowed to go into production can infect every system in the cloud.

Acknowledgements

Material for this article was taken from the following sources:

Windows Azure Compute cloud goes TITSUP PLANET-WIDE, *The Register*; October 30, 2013.
Whoopsie: Windows Azure stumbles again, *Gigaom*; October 31, 2013.
Microsoft's Windows Azure cloud hit by worldwide management interruption, *PC World*; October 31, 2013.
Microsoft Azure Swap Deployment Feature Restored After Global Outage, *The Whir*; October 31, 2013.
Microsoft's Windows Azure Hit With Global Compute Performance Glitch, *CRN*; October 31, 2013.
The TRUTH behind Microsoft's Azure global cloud mega-cock-up, *The Register*; November 8, 2013.
Inside Windows Azure: The Cloud Operating System, *Presentation, 2011 Microsoft Tech-Ed – COS 301*; May 16, 2011.

[1] Windows Azure Cloud Succumbs to Leap Day, *Availability Digest*; March 2012.
http://www.availabilitydigest.com/public_articles/0703/azure.pdf